Abstracts of papers presented
at the 2010 meeting on

THE BIOLOGY OF GENOMES

May 11–May 15, 2010

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Abstracts of papers presented
at the 2010 meeting on

THE BIOLOGY OF GENOMES

May 11–May 15, 2010

Arranged by

Susan Celniker, *Lawrence Berkeley National Laboratory*
Andy Clark , *Cornell University*
Chris Ponting, *University of Oxford, UK*
George Weinstock, *Washington University School of Medicine*

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Major sponsorship for this meeting was provided by **Roche.**

Additional funding provided by **Illumina, Inc.,** and the **National Human Genome Research Institute,** a branch of the **National Institutes of Health.**

Contributions from the following companies provide core support for the Cold Spring Harbor meetings program.

Corporate Sponsors

Agilent Technologies
AstraZeneca
BioVentures, Inc.
Bristol-Myers Squibb Company
Genentech, Inc.
GlaxoSmithKline
Hoffmann-La Roche Inc.
Life Technologies (Invitrogen & Applied Biosystems)
Merck (Schering-Plough) Research Laboratories
New England BioLabs, Inc.
OSI Pharmaceuticals, Inc.
Sanofi-Aventis

Plant Corporate Associates

Monsanto Company
Pioneer Hi-Bred International, Inc.

Foundations

Hudson-Alpha Institute for Biotechnology

Cover: Illustration by Marty Macaluso, www.quicksketch.com.

THE BIOLOGY OF GENOMES

Tuesday, May 11 – Saturday, May 15, 2010

Tuesday	7:30 pm	**1** Functional and Cancer Genomics
Wednesday	9:00 am	**2** Genetics of Complex Traits
Wednesday	2:00 pm	**3** Poster Session I
Wednesday	4:30 pm	Wine and Cheese Party *
Wednesday	7:30 pm	**4** High-Throughput Genomics and Genetics
Thursday	9:00 am	**5** Computational Genomics
Thursday	2:00 pm	**6** Poster Session II
Thursday	4:30 pm	**7** ELSI Panel Discussion
Thursday	7:30 pm	**8** Evolutionary Genomics
Friday	9:00 am	**9** Population Genomic Variation
Friday	2:00 pm	**10** Poster Session III
Friday	4:30 pm	GUEST SPEAKERS
Friday	6:00 pm	Banquet
Saturday	9:00 am	**11** Genetics and Genomics of Non-Human Species

---

Poster sessions are located in *Bush Lecture Hall*

* *Airslie Lawn*, weather permitting

Mealtimes at Blackford Hall are as follows:

Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm

Bar is open from 5:00 pm until late

Printed on 100% recycled paper.

PROGRAM

TUESDAY, May 11—7:30 PM

SESSION 1 FUNCTIONAL AND CANCER GENOMICS

Chairperson: **L. Ding,** Washington University School of Medicine, St. Louis, Missouri
P. Farnham, University of California, Davis

WEDNESDAY, May 12—9:00 AM

SESSION 2 GENETICS OF COMPLEX TRAITS

Chairperson: **D. MacArthur,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
C. Ober, University of Chicago, Illinois

WEDNESDAY, May 12—2:00 PM

SESSION 3 POSTER SESSION I

WEDNESDAY, May 12—4:30 PM

Wine and Cheese Party

SESSION 4 HIGH THROUGHPUT GENOMICS AND GENETICS

Chairperson: **B. Graveley,** University of Connecticut Health Center, Farmington
J. Shendure, University of Washington, Seattle

DNA methylome map reveals conserved role of gene body methylation in regulating alternative promoters
Ting Wang, Alika Maunakea, Raman Nagarajan, Steve Jones, Tracy Ballinger, David Haussler, Marco Marra, Martin Hirst, Shaun Fouse, Brett Johnson, Chibo Hong, Joseph Costello.
Presenter affiliation: Washington University, St. Louis, Missouri. 124

Genome-wide maps of nucleosome organization in three primary human cell types identify specific mechanisms that govern nucleosome organization
Anton Valouev, Steven.Johnson, Scott Boyd, Cheryl Smith, Andrew Fire, Arend Sidow.
Presenter affiliation: Stanford University School of Medicine, Palo Alto, California. 125

THURSDAY, May 13—9:00 AM

SESSION 5 COMPUTATIONAL GENOMICS

Chairperson: **L. Duret,** CNRS, Université Lyon 1, Villeurbanne, France
J. Wortman, University of Maryland School of Medicine, Baltimore

Biased gene conversion and the evolution of human genomic landscapes
Laurent Duret.
Presenter affiliation: CNRS, Université Lyon 1, Villeurbanne, France. 126

Analysis of 1000 genomes exon capture pilot data
Amit R. Indap, Wen Fung Leong, Christopher L. Hartl, Kiran V. Garimella, Fuli Yu, Richard A. Gibbs, Gabor T. Marth, 1000 Genomes Project Exon Sequencing Group.
Presenter affiliation: Boston College, Chestnut Hill, Massachusetts. 127

***De novo* assembly of RNAseq for transcriptome reconstruction and characterization from yeasts to human**
Moran Yassour, Manfred Grabherr, Joshua Z. Levin, Mike Berger, Pamela Russell, Jessica Alfoldi, Andi Gnirke, Federica Di Palma, Kerstin Lindblad-Toh, Nir Friedman, Aviv Regev.
Presenter affiliation: The Broad Institute of MIT and Harvard, Cambridge, Massachusetts; The Hebrew University, Jerusalem, Israel. 128

THURSDAY, May 13—2:00 PM

SESSION 6 POSTER SESSION II

SESSION 7 ELSI PANEL and DISCUSSION

RETURNING RESEARCH RESULTS TO PARTICIPANTS IN LARGE-SCALE GENOMICS STUDIES

Moderator: **Susan M. Wolf, J.D.,** University of Minnesota

Panelists: **Yann Joly, DCL,** McGill University
Kazuto Kato, Ph.D., Kyoto University
Jane Kaye, Ph.D., Oxford University
Isaac Kohane, M.D., Ph.D., Harvard Medical School

As biobanks and large-scale genomics research projects proliferate around the globe, debate is growing about whether, when, and how to return individual research results and incidental findings to the people who donated their samples for study. The topic raises difficult questions about the extent of the ethical and legal duties that may, or may not, be owed to research participants. It also has important operational implications.

The traditional approach—at least in most large-scale studies involving researchers who are not clinicians—has been not to return any research results. In fact, individual identifiers in such studies are frequently removed, making the whole question of whether to return results essentially moot as it is impossible to associate the data to a named person when no identifiers are available. At the opposite extreme, some have called for the return of all individual research results, or at least giving participants the option to choose what information they will receive; proponents of this approach argue that any withholding of data is unduly paternalistic. A middle ground advocated by some is to strike a balance, returning results only in cases where the results are scientifically validated and clinically actionable. The question of when, if ever, to return incidental research findings (findings beyond the aims of a study but which have potential clinical or reproductive significance) is an important part of the debate. Whatever position is taken on these issues will have major implications for the operation of the study and for the burdens on both researchers and participants. This panel will provide a range of perspectives on these questions from four parts of the world: Canada, Japan, the United States, and the U.K

SESSION 8 EVOLUTIONARY GENOMICS

Chairperson: **L. Kruglyak,** Princeton University, New Jersey
D. Petrov, Stanford University, California

Human-specific loss of regulatory DNA and the evolution of human-specific traits
Cory Y. McLean, Philip L. Reno, Alex A. Pollen, Abraham I. Bassan, Terence D. Capellini, Catherine Guenther, Vahan B. Indjeian, Bruce T. Schaar, Douglas B. Menke, Aaron M. Wenger, Gill Bejerano, David M. Kingsley.
Presenter affiliation: Stanford University, Stanford, California. 244

Emerging speciation in ocean microbes is driven by a few ecologically-relevant genomic loci
B. Jesse Shapiro, Jonathan Friedman, Otto X. Cordero, Sarah Preheim, Eric Alm, Martin Polz.
Presenter affiliation: Massachusetts Institute of Technology, Cambridge, Massachusetts. 245

FRIDAY, May 14—9:00 AM

SESSION 9 POPULATION GENOMIC VARIATION

Chairperson: **D. Adams,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
R. Nielsen, University of California, Berkeley

The sequence and analysis of 17 laboratory and wild-derived mouse genomes, and high resolution QTL analysis in a heterogeneous stock cross
Jonathan Flint, Thomas M. Keane, Binnaz Yalcin, Jim Stalker, Kim Wong, Xiangchao Gan, Petr Danecek, Avigail Agam, Martin Goodson, Guy Slater, Ian Jackson, Laura Reinholdt, Leah Rae Donahue, Steve Brown, Ewan Birney, Allan Bradley, Chris Ponting, Richard Mott, Richard Durbin, David J. Adams.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 246

Results and lessons from the 1000 Genomes Project pilot, and moving on to much more
Richard Durbin, on behalf of the 1000 Genomes Project.
Presenter affiliation: Wellcome Trust Sanger Institute, Cambridge, United Kingdom. 247

SESSION 10 POSTER SESSION III

Single base resolution of medaka fish DNA methylomes
Wei Qu, Shin-ichi Hashimoto, Atsuko Shimada, Yutaka Suzuki, Sumio Sugano, Hiroyuki Takeda, Shinichi Morishita.
Presenter affiliation: The University of Tokyo, Chiba, Japan. 254

A deletion in chromosome 5 generating a chimaeric gene is a protective factor against ischaemic stroke
R. Rabionet, S.Villatoro, J. Aigner, J. Jimenez-Conde, R. Elosua, L. Armengol, I. Fernandez-Cadenas, J. Montaner, E. Marti, J. Roquer, X. Estivill.
Presenter affiliation: Centre for Genomic Regulation, Barcelona, Spain. 255

Construction of the first SSR-based linkage map of flax (*L. usitatissimum* L.) and localization of QTLs underlying fatty acid composition
Raja Ragupathy, Scott Duguid, Sylvie Cloutier.
Presenter affiliation: Cereal Research Centre, Winnipeg, Canada. 256

Online quantitative transcriptome analysis
Regina Bohnert, Jonas Behr, Andre Kahles, Geraldine Jean, Gunnar Raetsch.
Presenter affiliation: Max Planck Society, Tuebingen, Germany. 257

Exon capture and re-sequencing in rhesus macaques for identification of SNPs in genes expressed in the brain
M. Raveendran, G.L. Fawcett, M. Bainbridge, F. Yu, J. Yu, D. Muzny, R.A. Harris, A. Milosavljevic, R.A. Gibbs, J. Rogers.
Presenter affiliation: Baylor College of Medicine Human Genome Sequencing Center Houston, Texas. 258

Transposable element landscape characterization in five bat genomes using 454 sequence data
David A. Ray, Heidi Pagan.
Presenter affiliation: Mississippi State University, Mississippi State, Mississippi. 259

Enhancing the annotation of genomes in Ensembl
Stephen Searle, Bronwen Aken, Julio Banet, Susan Fairley, Felix Kokocinski, Magali Ruffier, Amy Tang, Jan Vogel, Simon White.
Presenter affiliation: Wellcome Trust Sanger Institute, Cambridge, United Kingdom. 281

High resolution analysis of CNV breakpoints reveals potentially predisposing sequence motifs and variable mechanisms of genomic rearrangement.
Hung-Chun Yu, Chad Haldeman-Englert, Elizabeth A. Geiger, Hongbo M. Xie, Juan C. Perin, Xiaowu Gai, Tamim H. Shaikh.
Presenter affiliation: University of Colorado, Aurora, Colorado. 282

Tandem repeat sequences as causative cis eQTLs for protein-coding gene expression variation—The case of *CSTB*
Andrew J. Sharp, Christelle B. Borel, Eugenia Migliavacca, Emmanouil T. Dermitzakis, Maryline Gagnebin, Stylianos E. Antonarakis.
Presenter affiliation: University of Geneva Medical School, Geneva, Switzerland. 283

Comparing the transcriptional circuits controlling human and mouse hematopoiesis
Tal Shay, Vladimir Jojic, Noa Novershtern, The Immunological Genome Project Consortium, Benjamin L. Ebert, John L. Rinn, Daphne Koller, Aviv Regev.
Presenter affiliation: Broad Institute, Cambridge, Massachusetts. 284

The design of whole genome resequencing for association studies
Yufeng Shen, Itsik Pe'er.
Presenter affiliation: Columbia University, New York, New York. 285

Methylation detection in a MCF-7 cell line using ultra high-throughput bisulfite-sequencing with the SOLiD™ System
Vrunda Sheth, Stephen F. McLaughlin, Zheng Zhang, Christina Chung, Melissa A. Barker, Victoria L. Boyd, Heather E. Peckham.
Presenter affiliation: Life Technologies, Beverly, Massachusetts. 286

In-depth metabolic characterization of genetic loci underlying serum-lipids
S.-Y. Shin, A.-K. Petersen, W. Römisch-Margl, G. Zhai, K. Small, R. Wang-Sattler, E. Grundberg, J. Ried, A. Döring, H.-E. Wichmann, M. Hrabé de Angelis, H.-W. Mewes, T. Illig, T.D. Spector, J. Adamski, K. Suhre, C. Gieger, N. Soranzo.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 287

Polymorphic LTR retrotransposons can terminate transcripts at a distance, causing mouse lineage variation
Jingfeng Li, Keiko Akagi, Yongjun Hu, Natalia Volfovsky, Robert M. Stephens, David E. Smith, David E. Symer.
Presenter affiliation: Ohio State University, Columbus, Ohio. 302

A next-generation of methods for characterizing complex balanced rearrangements contributing to developmental disorders
Michael E. Talkowski, Bhavana Muddukrishna, Carl Ernst, Andrew Kirby, Toshiro Ohsumi, Mark Borowsky, Mark J. Daly, Cynthia C. Morton, James F. Gusella.
Presenter affiliation: Massachusetts General Hospital/Brigham and Women's Hospital and Harvard Medical School Boston, Massachusetts; Broad Institute, Cambridge, Massachusetts. 303

HIF-1α ChIP-Seq analysis of cancer cell line DLD-1
Kousuke Tanimoto, Katsuya Tsuchihara, Yutaka Suzuki, Sumio Sugano.
Presenter affiliation: the University of Tokyo, Kashiwa, Japan. 304

Evaluating the efficacy of cross-species microarray-based genomic capture and its application to targeted sequencing in a nonhuman primate model for HIV/AIDS research
K. Mondal, J.K. Davis, V.C. Patel, A.C. Shetty, Z.P. Johnson, G.Silvestri, M.E. Zwick, James W. Thomas.
Presenter affiliation: Emory University School of Medicine, Atlanta, Georgia. 305

Global analysis of RNA and protein changes in response to osmotic stress
Scott E. Topper, M. Violet Lee, Joshua J. Coon, Audrey P. Gasch.
Presenter affiliation: University of Wisconsin-Madison, Madison, Wisconsin. 306

Cufflinks—Transcript assembly, abundance estimation, and differential expression with RNA-Seq
Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, Lior Pachter.
Presenter affiliation: University of Maryland, College Park, College Park, Maryland; University of California, Berkeley, Berkeley, California. 307

Next generation whole exome sequencing in familial cancer
Lisa R. Trevino, David A. Wheeler, Kyle Chang, Donna M. Muzny, Jeffrey G. Reid, Richard A. Gibbs, Sharon E. Plon.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 308

Comparing genomic sequence of select large stretches of inbred rat strains using three different sequencing platforms in tandem
Michael Tschannen, Elizabeth Worthey, Kathrin Saar, Marek Tutaj, Oliver Hummel, Giannino Patone, Wei Chen, Howard Jacob, Norbert Hubner.
Presenter affiliation: Medical College of Wisconsin, Milwaukee, Wisconsin. 309

The transcriptomes of two heritable cell types help illuminate the circuit governing their differentiation
Brian B. Tuch, Quinn M. Mitrovich, Oliver R. Homann, Aaron D. Hernday, Francisco M. De La Vega, Alexander D. Johnson.
Presenter affiliation: University of California, San Francisco, San Francisco, California; Life Technologies, Foster City, California. 310

Construction of a real-time disease weather map
Stephen W. Turner, Eric E. Schadt.
Presenter affiliation: Pacific Biosciences, Menlo Park, California. 311

Origins and evolution of sulfadoxine resistance in human malaria parasite, *P. falciparum*
Sumiti Vinayak, Md Tauqeer Alam, Kanungnit Congpuong, Chansuda Wongsrichanalai, Laurence Slutsker, Ananias A. Escalante, John W. Barnwell, Venkatachalam Udhayakumar.
Presenter affiliation: Centers for Disease Control and Prevention, Atlanta, Georgia; Atlanta Research and Education Foundation, Decatur, Georgia. 312

Strand-specific RNA sequencing of HepG2 cells identifies genes that are differentially expressed, alternatively spliced and allelically imbalanced in response to TGF-beta
Stefan Enroth, Ola Wallerman, Brian Tuch, Catalin Barbacioru, Madhu Bysani, Robin Andersson, Stefan Thermén, Aristidis Moustakas, Carl-Henrik Heldin, Niclas Eriksson, Sarah Stanley, Jian Gu, Scott Kuersten, Melissa Barker, Jan Komorowski, Kevin McKernan, Francisco M. De La Vega, Claes Wadelius.
Presenter affiliation: Uppsala University, Uppsala, Sweden. 313

1000 Genomes Project—Data flow and quality assurance
Xiangqun Zheng-Bradley, Laura Clarke, Richard Smith, Chunlin Xiao, Martin Shumway, Steve Sherry, Paul Flicek, 1000 Genomes Project DCC.
Presenter affiliation: European Bioinformatics Institute, Hinxton, United Kingdom. 341

Mosaic recombination in gene families—Genome structure change and host-parasite coevolution
Martine M. Zilversmit, Ella K. Chase, Natalia Tichshenko, Diego Czul, Karen P. Day, Gil McVean, Philip Awadalla.
Presenter affiliation: University of Montreal, Montreal, California; University of Oxford, Oxford, United Kingdom; NYU Langone Medical Center, New York, New York. 342

FRIDAY, May 14—4:30 PM

GUEST SPEAKERS

Cori Bargmann
Rockefeller University

"Genetic variation in animal behavior—Genes, neurons, and maybe some principles" 343

Martin Blaser
New York University

FRIDAY, May 14

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SESSION 11 GENETICS AND GENOMICS OF NON-HUMAN SPECIES

Chairperson: **K. Pollard,** Gladstone Institutes, San Francisco, California
M. Purugganan, New York University, New York

Genomes, transcriptomes, methylomes and smRNAomes of *Arabidopsis* accessions
Ronan O'Malley, Ryan Lister, Robert Schmitz, Jarrod Chapman, Issac Ho, Jason Affourtit, Zhoutao Chen, Brian Desany, Srinivasan Maithreyan, James Knight, Daniel Rokshar, Michael Egholm, Tim Harkins, Joseph Ecker.
Presenter affiliation: The Salk Institute for Biological Studies, La Jolla, California. 349

A fine-scale genetic map of the chimpanzee genome from sequence variation data
Oliver Venn, Adi Fledel-Alon, Adam Auton, Cord Melton, Susanne Pfeifer, Ryan Hernandez, Rory Bowden, Zamin Iqbal, Simon Myers, Peter Donnelly, Molly Przeworski, Gilean McVean.
Presenter affiliation: University of Oxford Oxford, United Kingdom. 350

De novo assembly and evolutionary analyses of liver-expressed genes in 16 mammal species
John C. Marioni, George H. Perry, Pall Melsted, Ying Wang, Katelyn Michelini, Matthew Stephens, Jonathan K. Pritchard, Yoav Gilad.
Presenter affiliation: University of Chicago, Chicago, Illinois. 351

ZNF274 RECRUITS THE HISTONE METHYLTRANSFERASE SETDB1 TO THE HUMAN GENOME

Peggy J Farnham
University of California-Davis, Genome Center, Davis, CA, 95616

Only a small percentage of human transcription factors (e.g. those associated with a specific differentiation program) are expressed in a given cell type. Thus, cell fate is mainly determined by cell type-specific silencing of transcription factors that drive different cellular lineages. Several histone modifications have been associated with gene silencing, including H3K27me3 and H3K9me3. We have previously shown that the two largest classes of mammalian transcription factors are marked by distinct histone modifications; homeobox genes are marked by H3K27me3 and zinc finger genes are marked by H3K9me3. Several histone methyltransferases (e.g. G9a and SETDB1) may be involved in mediating the H3K9me3 silencing mark. We have used ChIP-chip and ChIP-seq to demonstrate that SETDB1, but not G9a, overlaps with H3me3K9. A current model is that SETDB1 is recruited to specific genomic locations via interaction with the corepressor TRIM28 (KAP1), which is in turn recruited to the genome via interaction with zinc finger transcription factors that contain a Kruppel-associated box (KRAB) domain. However, specific KRAB-ZNFs that recruit TRIM28 (KAP1) and SETDB1 to the genome have not been identified. We have shown that ZNF274 (a KRAB-ZNF that contains 5 finger domains), can interact with KAP1 in vitro and, using genome-wide ChIP-seq, we show that ZNF274 binding sites co-localize with SETDB1, KAP1, and H3K9me3. Knockdown of ZNF274 with siRNAs reduces the levels of KAP1 and SETDB1 at target sites. Thus our studies provide in vivo support for the model that KRAB domain-containing ZNFs can recruit histone methyltransferases to specific sites in the human genome.

TRANSCRIPTION BINDING VARIATION IN EUCARYOTES

M Snyder, M Kasowski, W Zheng, F Grubert, C Heffelfinger, M Hariharan, A Asabere, S Waszak, L Habegger, J Rozowsky, M Shi, A Urban, K Karczewski, H Zhao, E Mancera, L Steinmetz, M Gerstein, J Korbel

Stanford University, Genetics, Stanford, CA 94305

Variation in transcriptional regulation is thought to be a major cause of phenotypic diversity. Widespread differences in gene expression among individuals of a species have been observed, yet few studies have examined the variability of transcription factor (TF) binding, and thus the extent and underlying genetic basis of TF binding diversity is largely unknown. We mapped differences in transcription binding among individuals and elucidated the genetic basis of such variation on a genome-wide scale for both yeast and humans. For humans we mapped the binding sites of RNA Polymerase II (PolII) and a key regulator of immune responses, NFκB (p65), in ten lymphoblastoid cell lines and found that 25% and 7.5% of the respective binding regions differed between individuals. Binding differences were frequently associated with SNPs and genomic structural variants.
To further understand the genetic basis of transcription factor binding variation, we mapped the binding sites of Ste12 in pheromone-treated cells of 43 segregants of a cross between 2 highly diverged yeast strains and their parental lines. We found that the majority of TF binding variation is cis-linked and that many variations are associated with polymorphisms in the binding motifs of Ste12 as well as those of several proposed Ste12 cofactors. We also identified trans factors that modulate Ste12 binding to specific promoters. Yeast and human transcription factor binding strongly correlates with gene expression showing that binding variation is functional. Overall these different studies identified genetic regulators of molecular diversity among individuals and provide novel insights into variation in eucaryotes and mechanisms of gene regulation.

RELAPSE-SPECIFIC MUTATIONS IN AN AML GENOME DISCOVERED BY WHOLE GENOME SEQUENCING

Elaine Mardis [1], Li Ding[1], John Welch[2], David Larson[1], Ken Chen[1], Michael McLellan[1], Heather Schmidt[1], Ling Lin[1], Vince Magrini[1], Tammi Vickery[1], Richard Wilson[1], Timothy Ley[2]
[1]Washington University School of Medicine, The Genome Center, 4444 Forest Park Blvd Campus Box 8501, St. Louis, MO, 63108, [2]Washington University School of Medicine, Division of Oncology, 500 South Kingshighway SW Tower, St. Louis, MO, 63110

Acute myeloid leukemia (AML) is a highly malignant hematopoietic disease that affects about 13,000 adults in the United States annually, causing ~9,000 deaths. Whole-genome sequencing of two AML genomes have identified novel mutations that may be relevant for disease initiation, and (in the case of IDH1 mutations) predictive of poor outcome. To better understand the mutations associated with AML progression and relapse, we sequenced the relapse genome of an AML patient obtained 11 months after initial diagnosis. The patient initially presented with FAB M1 AML with normal cytogenetics and 100% blasts in the marrow. She was treated with standard induction chemotherapy, and consolidated with high dose AraC and an autologous stem cell transplant. She relapsed at 11 months with M1 AML, 78% blasts, and a novel t(10;12) in 100% of her metaphases. The ten somatic mutations present in this AML genome at initial presentation were previously reported (Ley et al Nature 2008). By sequencing the relapse tumor genome, we discovered two genic mutations that were missed in the de novo genome due to coverage issues (DNMT3A L723FS; LOC728896 F46L). Most importantly, we discovered four novel relapse-specific genic mutations, including a missense mutation in ETV6/TEL (R105P), an 11bp insertion in STK4 (at A461, predicted to cause loss of function due to a frameshift), and a missense mutation in MYO18B (A2317T). The ETV6/TEL gene is frequently involved in translocations in acute leukemias, and both STK4 and MYO18B are putative tumor suppressors. Although the mutations in STK4 and MYO18B were not detected in 21 additional AML relapse genomes, a mutation at the same position in ETV6 was detected in an additional de novo AML genome (R105Q). Paired-end data revealed a novel WNK1-WAC fusion at base pair resolution, resolving the relapse-specific clonal cytogenetic abnormality, t(10;12)(p12;p13). The resulting fusion retains the full catalytic domain of WNK1 but truncates its C-terminus; the fusion cDNA is expressed in the tumor. Functional analysis is in progress. Ultra deep read count analysis revealed that none of the relapse-specific mutations could be detected in the de novo AML sample, or in subsequent bone marrow specimens obtained at 43, 121, or 170 days after presentation (relapse occurred at day 338); the sensitivity of the assay allowed us to accurately detect 1 cell in 1,000 carrying a mutation. Our study establishes whole-genome sequencing as an unbiased approach for discovering AML relapse-specific mutations. Comparison of the de novo and relapse samples revealed additional somatic mutations at relapse that were almost certainly relevant for the progression

TUMOR PROGRESSION REVEALED BY SEQUENCING 100 SINGLE CELLS IN A HETEROGENEOUS BREAST CARCINOMA

Nicholas E Navin[1,2], Jude Kendall[1], Kerry Cook[1], Jennifer Troge[1], James Hicks[1], Michael Wigler[1]

[1]Cold Spring Harbor Laboratory, Wigler Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY, 11724, [2]Stony Brook University, Molecular Genetics & Microbiology, 100 Nicolls Road, Stony Brook, NY, 11733

Genomic analysis by microarray, and more recently DNA sequencing, has provided important insights into the role of copy number variation in human cancer. However, these methods can yield only approximate results when applied to mixed populations of rapidly evolving cells. In such cases our understanding would be improved by dissecting genetic events at the single cell level. We have therefore developed a method of single nucleus sequencing (SNS) to quantify the genome copy number of individual nuclei. Using the Illumina GA2 platform we have shown that a single lane of sequence reads are sufficiently distributed across the genome to measure copy number at a resolution of about 50kb. We validated our method in a normal fibroblast cell line (SKN1) that has been deep-sequenced along with a genetically complex breast cancer cell line (SK-BR-3). We then used SNS to analyze 100 single cells isolated from a heterogeneous Basal-like breast carcinoma. From this data, we constructed a detailed phylogenetic lineage, showing that the majority of cells belong to one of five major clonal subpopulations. We compared these subpopulations to infer a step-wise progression in which much of the genome is deleted, followed by endoreduplication to generate a highly aneuploid genome that acquires many focal amplifications and deletions of cancer genes including KRAS, EFNA5 and COL4A5. Additionally, we observed a rare subpopulation of diploid cells that contain single random amplifications and deletions that are not present in the major aneuploid subpopulations and perhaps represent an unstable precursor. Our data strongly support the polyclonal evolution model for tumor progression, in which the majority of tumor cells continue proliferate and undergo clonal expansions to form the mass of the tumor.

APPROACHING A COMPREHENSIVE VIEW OF CANCER GENOMES

Li Ding, Ken Chen, Michael D. McLellan, David E. Larson, Christopher C. Harris, Daniel Koboldt, Nathan Dees, Dong Shen, David Dooling, John Wallis, Sean McGrath, Todd Wylie, Kim Delehaunty, Lisa Cook, Vincent Magrini, Rachel M. Abbott, Lucinda L. Fulton, Robert S. Fulton, George Weinstock, Matthew J. Ellis, Timothy J. Ley, Elaine R. Mardis, and Richard K. Wilson

The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd., St. Louis MO 63108.

Massively parallel DNA sequencing technologies provide an unprecedented ability to characterize entire genomes, in an unbiased manner, for genetic changes associated with tumor initiation, growth, and metastasis. Building upon our large collection of data from acute myeloid leukemia, glioblastoma multiforme, breast cancer, lung adenocarcinoma, and serous cystadenocarcinoma, we have developed a set of high throughput software solutions for discovering and analyzing a wide range of variation types, both in individual and in a population of cancer genomes. Algorithmic approaches we developed to perform structural variation analysis of paired end reads, coupled with a targeted capture-based validation strategy, have identified a multitude of such sites in these tumor genomes. This large data set also allowed us to perform a comparative analysis of different tumor types aimed at identifying cancer type-specific mutational profiles, rates, and spectra. Furthermore, the comparison between primary tumor and metastasis in breast cancer, as well as between de novo tumor and relapse in acute myeloid leukemia, identified genetic changes associated with tumor growth, relapse, and metastasis. The differential mutation frequencies and structural variation patterns between primary and metastatic breast tumors suggest that metastatic tumors may arise from minor subpopulations of cells within the primary. On the other hand, the comparison between de novo tumor and relapse revealed clonal characteristics of acute myeloid leukemia.

EVOLUTION OF THE POPULATION OF CANCEROUS CELLS IN A HEPATOCELLULAR CARCINOMA PATIENT

Xuemei Lu*[1], Weiwei Zhai*[1], Jue Ruan*[1], Yong Tao[1], Yu Wang[1], Jun Cai[1], Shaoping Ling[1], Shiou-Hwei Yeh[2], Pei-Jer Chen[2], Chung-I Wu[1]

[1]Beijing Institute of Genomics, Chinese Academy of Sciences, 7 Beitucheng Xilu, Beijing, 100029, China, [2]National Taiwan University Medical College, Graduate Institute, 1 Jen-Ai Rd., Taipei, 10617, Taiwan

"... natural selection as a force for the good maintains the fittest in a species. But when we turn to the competition between the individual cells within a single animal, … natural selection has now become a liability. " All cancers are caused by somatic mutations in the population of cells. The cancer-causing mutations (the drivers) can be inferred by functionality or by the signature of natural selection. Here we adopted the approach of population genetics in searching for advantageous mutations in natural populations to identify the mutations that drive tumor cells to proliferate. We collected primary hepatocellular tumor (T0) and its six adjacent normal liver tissue (N1…N6), two recurrent tumors (RC1 and RC2) which emerged 15 months after the operation to the T0, and the RC1's adjacent normal tissue (N0) from a hepatocellular carcinoma (HCC) patient . We have sequenced the genomes of N0 and RC1 with AB SOLiD and Illumina GAII to identify SNP, CNV and SV, providing the first whole genome analysis of somatic mutations, for HCC. Tumor and normal specific SNPs called from whole genomic short-read sequencing were genotyped for all those samples mentioned above with Sequenom and PCR sequencing. The patterns of somatic mutations indicate that tumors are highly clonal even with respect to aneuploidy generation, whereas normal tissues may generally lack clonality. RC2 is somewhat distinct. The two mutations and Δ5q occurred in T0 and RC1 after their separation from RC2, associating with the "faster" proliferation. It also suggests that metastasis happened during different phases of tumor evolution. Several cancer suppresser genes are in the region of Δ5q which is the more likely candidate for "driver mutation". Most of SNVs exist in the region with partial ploidy. Partial aneuploidy contributes to genetic diversity, enabling natural selection.

CHIP-SEQ REVEALS EVOLUTIONARILY HIDDEN HEART ENHANCERS

Axel Visel[1,2], Matthew J Blow[1,2], Bing Ren[3], Brian L Black[4], Edward M Rubin[1,2], Len A Pennacchio[1,2]

[1]Lawrence Berkeley National Laboratory, Genomics Division, 1 Cyclotron Road, Berkeley, CA, 94720, [2]U.S. Department of Energy Joint Genome Institute, Genetic Analysis, 2800 Mitchell Drive, Walnut Creek, CA, 94598, [3]University of California San Diego School of Medicine, Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, CA, 92093, [4] University of California San Francisco, Cardiovascular Research Institute and Department of Biochemistry and Biophysics, 600 16th Street, San Francisco, CA, 94158

The success of enhancer prediction by extreme non-coding sequence conservation across vertebrate species varies depending upon the tissue being studied. Conservation-based screens of the human genome have identified a 20-fold larger number of developmental forebrain enhancers than developmental heart enhancers. This puzzling discrepancy suggests either a relative paucity of heart enhancers compared to other tissues, or the failure of evolutionary constraint to identify them. To explore this question, we used ChIP-seq with the enhancer-associated protein p300 from mouse embryonic heart and forebrain tissues. This conservation-independent strategy for discovery of *in vivo* enhancers identified several thousand candidate sequences for both heart and forebrain. In striking contrast to forebrain, candidate heart enhancers were under substantially relaxed evolutionary sequence constraint. Only 6% of predicted heart enhancers were highly constrained (phastCons scores >600) compared to 44% in forebrain. Conversely, a substantial proportion (24%) of predicted heart enhancers was under no detectable sequence constraint, compared to a seven-fold smaller fraction (3.5%) in forebrain. To test if these ChIP-seq predictions represent *bona fide* enhancers despite the absence of sequence constraint, we tested more than 130 heart candidate regions with high, moderate, weak or no sequence constraint in a transgenic mouse assay. More than 60% of candidate sequences were reproducible *in vivo* enhancers active in the heart and, importantly, no significant difference in success rate was observed between the four constraint bins. The large, previously concealed population of poorly conserved heart enhancers identified in this study highlights the strengths of epigenomic strategies for discovery of *in vivo* enhancers and provides evidence for marked global disparities in evolutionary constraint between enhancers involved in different biological processes.

CHARACTERIZATION OF 1000 BREAST CANCER GENOMES AND TRANSCRIPTOMES

Christina Curtis*[1,2], Suet-Feung Chin*[1,2], Sohrab Shah*[3], METABRIC Consortium[1,2,3,4], Simon Tavaré[1,2], Samuel Aparicio[3], Carlos Caldas[1,2]

[1]University of Cambridge, Oncology, Hills Road, Cambridge, CB2 2XZ, United Kingdom, [2]Cancer Research UK, Cambridge Research Institute, Robinson Way, Cambridge, CB2 0RE, United Kingdom, [3]British Columbia Cancer Research Agency, Molecular Oncology, 675 W10th Avenue, Vancouver, V5Z 1L3, Canada, [4] University of Nottingham, Histopathology, Hucknall Road, Nottingham, NG5 1PB, United Kingdom

Breast cancer is driven by the acquisition of key genetic aberrations that confer clonal growth advantages. The identification of recurrent genomic alterations that impact gene expression can facilitate the elucidation of such 'driver' events. Given the substantial inter-individual variability of clinicopathological characteristics amongst breast cancer patients, genome-wide measurements of multiple data types combined with clinical variables provide an invaluable resource to dissect the complexity of this disease. Here we describe an integrated analysis of copy-number, allelic-ratios, and gene-expression for 1000 primary tumors aimed at further characterizing the genomic and transcriptional landscape of breast cancer.

High-density Affymetrix SNP 6.0 arrays were employed to assay allele-specific and total copy number on 1000 fresh frozen tumors with a minimum of 5 years clinical history. Matched RNA from 824 samples was hybridized to Illumina HT-12 arrays for gene-expression analysis. Additional orthogonal data includes deep sequencing of a subset of cases to survey the mutational spectrum of critical cancer loci.

Through the joint analysis of diverse data types, which reflect alterations at both the DNA and mRNA level, we identified novel breast cancer subtypes with distinct clinical outcomes. We further characterized the genomic and transcriptional landscape of breast cancer in terms of focal alterations, preferential allelic amplification, and ploidy. By identifying both *cis* and *trans*-acting copy number alterations, we have generated a systematic overview of pathway disruption amongst subtypes suggesting novel targets for therapeutic agents in specific patient sub-populations.

LOSS-OF-FUNCTION MUTATIONS IN HEALTHY HUMAN GENOMES: IMPLICATIONS FOR CLINICAL GENOME SEQUENCING

Daniel G MacArthur[1], Suganthi Balasubramanian[2], Ni Huang[1], Adam Frankish[1], Zhang Zhengdong[2], Lukas Habegger[2], Xinmeng Mu[2], Matthew Bainbridge[3], Bryndis Yngvadottir[1], 1000 Genomes Consortium[1], Jennifer Harrow[1], Richard A Gibbs[3], Matthew E Hurles[1], Mark B Gerstein[2], Chris Tyler-Smith[1]

[1]Wellcome Trust Sanger Institute, various departments, Genome Campus, Hinxton, CB10 1SA, United Kingdom, [2]Yale University, Gerstein Lab, 266 Whitney Ave, New Haven, CT, 06520, [3]Baylor College of Medicine, Molecular and Human Genetics, Baylor Plaza, Houston, TX, 77030

For the clinical benefits of high-throughput sequencing technologies to be fully realised it must be possible to distinguish medically relevant genetic variants from benign polymorphisms in individual human genomes.

Here we present a comprehensive analysis of common variants predicted to severely affect gene function, using data from all three pilots of the 1000 Genomes Project (low-coverage sequencing of ~180 individuals, high-coverage sequencing of six individuals and sequencing of ~1,000 genes in over 600 individuals), including all nonsense and splice site-disrupting SNPs, frame-shift-inducing indels and structural variants predicted to disrupt one or more protein-coding exons.

We show that the number of predicted loss-of-function (LOF) variants in putatively healthy individuals is surprisingly high. These variants are highly enriched for artefacts - thus providing sensitive indicators of sequencing and annotation quality - but display a frequency spectrum suggesting many are mildly deleterious to human health.

We present catalogues of LOF variants and LOF-tolerant genes for use in distinguishing between benign and pathogenic variants in clinical genome sequencing. We also demonstrate profound differences in functional, evolutionary, expression and interaction properties of LOF-containing genes and genes implicated in severe Mendelian disease, allowing prioritisation of novel variants for functional follow-up.

HIGH THROUGHPUT RNA SEQUENCING REVEALS GENETIC DETERMINANTS AND MECHANISMS REGULATING HUMAN EXPRESSION QUANTITATIVE TRAITS LOCI

Jacek A Majewski[1,2], Zibo Wang[1,2], Amandine Bemmo[1,2], Kevin Ha[1,2], Emilie Lalonde[1,2], Tony Kwan[2], Tomi M Pastinen[1,2]

[1]McGill University, Department of Human Genetics, 740 Dr. Penfield, Montreal, H3A1A4, Canada, [2]Centre d'innovation, Génome Québec et Université McGill, 740 Dr. Penfield, Montreal, H3A1A4, Canada

Expression levels of many human genes are under genetic control. Such genes, known us expression quantitative trait loci (eQTLs), have been identified by their association to cis-acting variants – in most cases single nucleotide polymorphisms. It is presumed that the associated genetic variants, or variants in linkage disequilibrium with the marker SNP, exert control over the expression levels of the gene. However, to date few actual causative variants have been identified. Two competing hypothesis postulate either control at the level of transcription, or post-transcriptionally at the level of mRNA stability. Here we use ultra-deep mRNA sequencing to investigate the mechanisms underlying eQTLs by focusing on deep coverage – over 200 million sequence tags - of two fully sequenced HapMap samples. Our RNA preparation method allows us to extract information from both exonic (mRNA) and intronic (pre-mRNA) sequences. For a set of highly confident eQTLs, we find that there is a very high correlation between the levels of mature and unprocessed RNA, implying that 80% of eQTL variance can be explained by genetic control at the level of transcription. To explain the remaining 20% of the variance, we focus on the outliers which exhibit discordant mRNA and pre-mRNA expression patterns. We find dozens of SNPs affecting splicing patterns within genes, many of which in turn affect the overall RNA levels. We observe three predominant mechanisms of action: 1) SNPs that affect the usage of existing splice sites leading to alternative splice site choices, likely resulting in non-sense mediated decay, 2) SNPs activating the usage of pseudoexons and resulting in premature stop codons, 3) SNPs altering splicing within the 5'UTRs, likely affecting the stability of the resulting mRNA. We focus in detail on four examples, OAS1, MRPL43, MMAB, and USMG5 to illustrate and dissect the exact downstream effects of the causative genetic variants.

SYNTHETIC ASSOCIATIONS ARE UNLIKELY TO ACCOUNT FOR MOST COMMON DISEASE GENOME-WIDE ASSOCIATION SIGNALS

Jeffrey C Barrett, Carl A Anderson, Nicole Soranzo, Eleftheria Zeggini

Wellcome Trust Sanger Institute, Human Genetics, Wellcome Trust Genome Campus, Hinxton, CB10 1HH, United Kingdom

Genome wide association studies have revealed hundreds of bona fide, replicable regions associated with dozens of complex diseases and traits. These studies are explicitly designed to discover associations to common polymorphisms, almost always of weak effect. Thus far they explain only a small fraction of the total heritability of complex disease. A great deal of recent attention has focused on explaining this "missing heritability", including a recent suggestion that "synthetic associations" generated by many rare variants of high penetrance within a few megabases of common alleles are actually responsible for the signals observed in GWAS. While this scenario is plausible, we believe it is likely to be uncommon for several reasons:

a) Rare variants of large effect – those most likely to yield synthetic associations - can be readily detected by well-powered linkage scans. We substantiate this argument with power calculations and highlight an example from the literature: the *NOD2* susceptibility locus for Crohn's disease. This locus exemplifies the synthetic association hypothesis, at which three rare, highly penetrant mutations fully explain the observed association to nearby common variants; as our power calculations predict, it is one of the few complex disease genes mapped by linkage.

b) We describe why the empirical evidence (sickle cell disease and syndromic hearing loss) provided in support of synthetic associations is not appropriate for common complex traits and highlight reasons why it is incorrect to extrapolate to traits with an underlying polygenic genetic architecture. Briefly, both these loci contain rare variants that explain a significant proportion of the phenotypic variance (100% and 50% respectively) and such loci do not exist in common complex traits (otherwise they would have been detected using linkage analysis, as indeed *HBB* and *GJB2/GJB6* were).

c) Finally, we discuss published resequencing efforts, in which very little evidence has been unearthed pointing toward rare variants explaining signals from GWAS.

We believe the consideration of these wider lines of evidence support the conclusion that synthetic associations explain very few GWAS signals.

A SCALABLE CLASS OF MULTIPLE LOCUS METHODS FOR GENOME-WIDE ASSOCIATION STUDIES

Gabriel Hoffman[1], Benjamin Logsdon[1], Chuan Gao[1], Abra Brisbin[1], Jason Mezey[1,2]

[1]Cornell University, Biological Statistics and Computational Biology, Biotechnology Building, Ithaca, NY, 14853, [2]Weill Cornell Medical College, Genetic Medicine, 1305 York Ave, New York, NY, 10065

All corroborated disease loci that have been discovered in genome-wide association studies (GWAS) were identified by independently analyzing each genetic marker in a study. While the success of individual marker analysis is unequivocal, it is well appreciated that multiple locus algorithms can have better performance, particularly in the identification of weaker associations due to small locus effects, weak linkage, or low minor allele frequencies. Yet, computational limitations and concerns about performance on large datasets have prevented widespread application of these methods by practitioners of GWAS. To address these concerns, we have developed a highly scalable class of multiple locus algorithms for simultaneous analysis of all markers in a GWAS, and we have undertaken an extensive simulation and data analysis assessment of algorithm performance. The foundation of our family of methods is a penalized generalized linear model (GLM) that provides great versatility for analyzing case/control and continuous phenotypes, and allows the incorporation of covariates. We have implemented a range of theoretically well-founded penalties, including standard (i.e. lasso, ridge, SCAD) and non-standard (i.e. mixture) approaches. We have also incorporated random covariate and related techniques to increase mapping power in the presence of missing latent covariates. In order to scale these methods to massive GWAS datasets we have implemented coordinate-wise gradient descent algorithms for likelihood analysis and variational Bayes algorithms for approximate Bayesian analysis. Our algorithms scale extremely well and are typically able to complete an analysis of a GWAS that included a thousand samples and one-million markers in less than 24 hours on a standard desktop (with large memory capacity). Through our simulations, we have identified broad conditions under which different penalties perform better. Through our analysis of available GWAS data for a number of diseases, we demonstrate that our algorithms identify additional high-confidence “hits” when compared to individual marker analysis.

RARE VARIANTS CONTRIBUTE TO ASTHMA SUSCEPTIBILITY

Carole Ober[1], Dara G Torgerson[1], Daniel Capurso[1], Scott R Weiss[2], Deborah A Meyers[3], Kathleen C Barnes[4], Eugene R Bleecker[3], Benjamin A Raby[2], Rasika A Mathias[4], Penelope E Graves[5], Fernando D Martinez[5], Dan L Nicolae[1]

[1]University of Chicago, Human Genetics, Chicago, IL, 60637, [2]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, [3]Wake Forest University, Genomics Center, Winston-Salem, NC, 27157, [4] Johns Hopkins University, Ashtma and Clinical Immunology, Baltimore, MD, 21244, [5]University of Arizona, Respiratory Science Center, Tucson, AZ, 85724

Genome-wide association studies of asthma have identified only a small proportion of the heritability, similar to other common diseases. In this study we sequenced the coding exons and flanking sequences of 9 asthma candidate genes to assess the role of rare variants in asthma risk. We selected genes that were previously associated with asthma and showed signatures of purifying selection for resequencing studies in 513 asthma cases and 515 controls. In European Americans, individual rare (<5%) variants in 4 genes were enriched in asthma cases compared to controls, whereas no individual variants were enriched in African American cases. The largest number of associated variants was in the *CFTR* gene, in which 7 alleles were significantly enriched in the cases compared to controls: 3 had frequencies of 3-6% in cases and were absent in controls ($p= 0.0003 - 1.5 \times 10^{-7}$) and 4 had frequencies 4-5% in cases and <1% in controls ($p= 0.0008 - 1.9 \times 10^{-5}$). One associated variant in *CFTR* as nonsynonymous and 6 were noncoding (intronic); odds ratios ranged from 8.8 to >12.7. None of the associated variants are known cystic fibrosis mutations, suggesting that highly penetrant, noncoding rare variation in the *CFTR* gene specifically affects asthma risk. Our study further supports the hypothesis that a significant proportion of the heritability of asthma is due to rare variation.

A GENE-BASED APPROACH TO JOINT ANALYSIS OF MULTIPLE RELATED PHENOTYPES

Katherine I Morley[1], David G Clayton[2], Jeffrey C Barrett[1]

[1]Wellcome Trust Sanger Institute, Human Genetics, Wellcome Trust Genome Campus, Cambridge, CB1 3BP, United Kingdom, [2]Cambridge Institute for Medical Research, Medical Genetics, Wellcome Trust/MRC Building, Addenbrooke's Hospital, Cambridge, CB2 0XY, United Kingdom

Genome-wide association studies have been very successful for autoimmune (AI) diseases such as type 1 diabetes (T1D) and Crohn's disease (CD), identifying over 80 associated loci. Some of these findings implicate novel biological pathways, generating hypotheses about how these may be shared by AI disorders.

Joint association analysis of AI diseases using raw data may improve power to detect loci shared between diseases. However, any method must allow for heterogeneity in disease-variant associations at both the phenotypic and genotypic levels. Although a number of shared loci have been discovered, allelic heterogeneity exists. For example, the tryptophan variant of R602W in *PTPN22* is associated with increased risk of T1D and rheumatoid arthritis (RA), but reduced risk of CD. Additionally a single locus may be associated with multiple AI phenotypes but the causal SNP may differ. For example, different SNPs at the *IL2RA* locus are associated with RA, multiple sclerosis and T1D.

No method deals with all these issues. We address the problem using an elastic net within a multinomial logistic regression framework to conduct gene-based tests of association. This method selects SNPs within a region that jointly provide the strongest evidence for association to a set of diseases, but explicitly allows for heterogeneity on both sides of the disease-locus association.

We compare our method to other strategies for gene-based association using binary or multinomial logistic regression models. In scenarios where a single causal SNP is associated with two diseases the power of this method is very close to that of multinomial logistic regression analyses of each SNP, adjusting for multiple testing (generally the most powerful strategy). With more than one causal SNP, the multinomial elastic net is more powerful than any other analytical strategy, even if each causal SNP is only associated with a single disease. This new method thus utilises the increase in power provided by combining disease samples without compromising the ability to identify variants that are only associated with some traits, or which show opposite directions of association with different diseases.

THE *DROSOPHILA* GENETIC REFERENCE PANEL: WHOLE GENOME ASSOCIATION MAPPING OF QUANTITATIVE TRAITS, AND A NEW TOOL FOR *DROSOPHILA* GENETICS.

Stephen Richards[1], Dianhui Zhu[1], Yi Han[1], Julien Ayroles[2], Mary Anna Carbone[2], Trudy Mackay[2], Eric Stone[2], Richard A Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, 1 Baylor Plaza, Houston, TX, 77030, [2]N.C. State University, Department of Genetics, Box 7614, Raleigh, NC, 27695

The *Drosophila* Genetic Reference Panel is a community resource of 192 sequenced inbred *Drosophila melanogaster* lines with measured quantitative traits. Forty lines have been sequenced to a minimum of 12X coverage using both 454 and Illumina sequencing platforms. The remaining 152 lines are being sequenced using the Illumina platform. Alignments have been generated using BWA, and polymorphisms identified using Atlas SNP2. We find approximately 500,000 SNPs, and 50,000 indels per inbred line, relative to the reference sequence. Extensive quality control genotyping has ensured both sequence and strain sample integrity, matching and true homozygosity for all of the lines. Lines with excess heterozygosity have been replaced to optimize the power of association studies.

Initial whole genome association analyses have identified between 20-100 polymorphisms associated with complex traits with p-values of 10^{-5} or less. Interestingly, most of these genes are novel, and many have pleiotropic effects on multiple traits. Thus, we believe this tool will provide a new understanding of *Drosophila* genetics, complementary to the large body of work based on induced mutational screens. We also present multiple methods for the validation of associations.

As a community resource we are working to facilitate the use of the reference panel in as many settings as possible. The fly stocks are currently available from the Bloomington *Drosophila* Stock Center. We are also preparing a website for genome wide associations allowing input of phenotypes measured on the lines, and outputting associated polymorphisms and p-values. We hope this web tool will allow any *Drosophila* geneticist with the ability to work on flies to utilize the reference panel – even high school students counting bristles.

TRANSCRIPTION FACTOR POLYMORPHISMS AND COMPLEX TRAITS: A THERMODYNAMIC MODEL OF GENETIC INTERACTIONS.

Justin P Gerke[2,1], Jason Gertz[3,1], Barak A Cohen[1]

[1]Washington University School of Medicine, Genetics, 4444 Forest Park Parkway, St. Louis, MO, 63108, [2]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Washington Road, Princeton, NJ, 08544, [3]Hudson Alpha Institute, Genomics, 601 Genome Way, Huntsville, AL, 35806

Understanding the molecular basis of complex traits is a major challenge facing modern geneticists. By mapping Quantitative Trait Loci (QTL) much progress has been made in understanding the genetic basis of complex traits. In contrast, our understanding of the molecular basis of these traits lags far behind. Advancing this field requires understanding the molecular mechanisms through which causal polymorphisms exert their effects on phenotype. Using the yeast *S. cerevisiae* as a model system, we showed that variation in sporulation efficiency between two natural isolates is due, almost entirely, to four polymorphisms that reside in three transcription factors (TFs), *IME1*, *RME1*, and *RSF1*. The causative variants are both coding (in IME1 and RSF1) and non-coding (in the promoters of *IME1* and *RME1*). Genetic interactions (epistasis) between these TF loci explain 87% of the phenotypic variation in this trait. To better understand the molecular mechanisms that underlie these genetic interactions we constructed a thermodynamic model of natural variation in sporulation efficiency. The model, which incorporates the non-linear biochemical reactions that underlie genetic interactions, suggests a general molecular mechanism that generates epistasis between polymorphisms. The model produces testable hypotheses about the concentrations and affinities of specific molecular species in different genetic backgrounds. Our first experimental test of this model suggests that we are indeed capturing many of the salient features that underlie genetic interactions. We anticipate that this biophysical framework for modeling genetic variation will be a useful tool for understanding the molecular basis of complex traits. Our results also point to variation in transcription factors as a major source of phenotypic diversity within species.

DIRECT EFFECTS OF ENVIRONMENTAL PERTURBATION ON CIS-REGULATION ASSESSED BY ALLELIC EXPRESSION

V. Adoue[1], E. Grundberg[1], B. Ge[1], T. Kwan[1], K. L Lam[1], V. Koka[1], O. Nilsson[2], Q. L. Duan[3], S. T Weiss[3], B. Raby[3], K. G Tantisira[3], T. Pastinen[1]

[1]McGill Univ., Hum Genet, 740 Dr. Penfield, Montreal, H3A 1A4, Canada, [2]Uppsala Univ., Surgery, Akademiska Sj, Uppsala, 751 85, Sweden, [3]HMS, Channing Labs, 181 Longwood Av., Boston, MA, 02115

The effects of environmental perturbation on cis-eQTLs are unknown. Using 100 primary human osteoblasts from Swedish donors and following culture in either standard conditions or after induction with BMP-2, dexamethasone (DEX), and PGE2 we collected expression data (Illumina Ref-8) and correlated expression levels to SNP data. Only ~1.4% of the cis-eQTLs showed treatment-specificity at high confidence. However, in follow-up analysis by direct measurement of cis-variation using genome-wide allelic expression (AE) tests (NG 2009;41:1216) most treatment-specific effects failed to validate. DEX-specific cis-variants in the MYO6 and TNC loci could be fine-mapped to putative distal regulatory elements. Strong interaction between treatment and cis-variation appear rare and detection of such requires allelic expression studies.
We are now applying the genome-wide AE test for specific discovery of functional variants underlying differences in drug response using resting and stimulated Lymphoblastoid Cell Lines (LCL) from asthmatic children included in the Childhood Asthma Management Program (CAMP). CAMP includes children who demonstrate variability in clinical response to corticosteroids. We analyzed LCLs from patients representing extremes of response and observed allelic variation at different doses and timepoints upon DEX stimulation. Preliminary data analysis indicates up to 30 transcripts with DEX-specific allelic cis-regulation. Most effects are detected at 4h timepoint and some show dose response, suggesting direct treatment-specific interactions. We focus now on comparing differences in DEX-specific cis-regulation among cell lines derived from patients with varying clinical response and mapping cis-regulatory variants underlying DEX-specificity. Our approach can be generically applied for isolation of pharmacogenetic variants.

ACCURACY OF ILLUMINA GENOME ANALYZER AND HISEQ 2000: WHAT DEPTH OF COVERAGE DO YOU REALLY NEED?

Subramanian S Ajay[1], Stephen C Parker[1], Hatice Ozel Abaan[1], Jamie K Teer[1], Praveen F Cherukuri[1], Nancy F Hansen[1], Pedro Cruz[1], William A Gahl[2], James C Mullikin[1], Elliott H Margulies[1]

[1]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, [2]Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892

The development of high-throughput sequencing technologies has made it feasible to sequence whole human genomes at a fraction of the cost and time than was previously possible. This ability can be leveraged to routinely use whole-genome sequencing as a clinical diagnosis tool. With this objective, we present here results from analysis of a clinical sample that was sequenced using both the Illumina GAIIx and HiSeq 2000 instruments.

With 100bp paired-end reads aligned to the reference genome, we were able achieve more than 75x coverage from two flowcells on the HiSeq 2000 platform and 35x coverage from two flowcells on the GAIIx platform, giving us a total coverage of more than 100x. Using bioinformatics algorithms, genotype calls were made to detect single nucleotide variants (SNVs) across the whole genome. We compared calls made from the two different datasets to genotype calls made on an array-based technology. In addition, we used these datasets to address what coverage is required to attain a certain level of comprehensiveness and accuracy, allowing us to make an informative decision about the depth of coverage needed for future whole-genome sequencing endeavors.

To tackle biological questions, SNVs were subjected to further analysis, including identification of synonymous and non-synonymous base changes. In addition, positions identified as variants were evaluated for their potential to dramatically change DNA structure and for their overlap with evolutionarily constrained regions. This helps us prioritize non-coding variants for their potential to be biologically important. Finally, identification of copy number variations and gross rearrangements of the genome are also providing insights into the nature of disease.

These results will help better develop methodologies to provide a confident approach in solving clinical cases where the underlying disease mechanism remains unknown.

GENOME-WIDE IDENTIFICATION OF SMALL INSERTIONS AND DELETIONS IN THE 1000 GENOMES PILOT PROJECT

Cornelis A Albers[1,2], Gerton A Lunter[3], Quang S Le[2], Daniel MacArthur[2], Willem H Ouwehand[1,2], Richard Durbin[2], 1000 Genomes Consortium [1,2,3]

[1]Univ of Cambridge, Haematology & NHSBT, Cambridge, United Kingdom, [2]Sanger Institute, United Kingdom, [3]Univ of Oxford, WTCHG, United Kingdom

We compiled a genome-wide set of over a million insertions and deletions ranging from 1 to 50 nucleotides in the 1000 Genomes Project pilot 1 data set using the novel Bayesian method Dindel. The main idea is to consider candidate indels and other sequence variants obtained from different approaches such as read-mappers and assembly-based methods, and then evaluate support for these candidates by performing a Bayesian gapped realignment of reads to a set of haplotypes potentially segregating in the population. We use a Bayesian EM caller to estimate allele frequencies and posteriors.

We observed that the rate of indels due to sequencing errors is as high as 1% in long homonucleotide runs, and that ~10 % of the called indels occurred in these. We explicitly account for these context-specific error rates in the realignment.

The method successfully identified all capillary-validated indels in a 3.8 kb region from a candidate gene study in 96 individuals without false-positives on artificially downsampled coverage ranging from 10X to 100X.

We identified in total 1.03 M indels in 170 individuals using candidate indels generated by various methods. Deletions were more common than insertions, noting that there is better power to detect deletions. The length distribution of indels did not differ between populations. YRI had ~3 times more private indels than CEU and JPT/CHB. We called ~1400 indels in CCDS coding regions; 60% resulted in a frame shift, although 3n-indels were strongly overrepresented relative to non-coding regions. In each population the non-reference allele frequency distribution of frame-shift indels was shifted towards zero compared to that of in-frame indels.

Genotypes appear to be highly consistent with the local genealogy independently estimated from HapMap3 haplotypes. This suggests that most indels are amenable to imputation for disease association testing. Validation of a subset of the predicted indels is currently in progress.

STUDYING ANIMAL DOMESTICATION BY BRAIN TRANSCRIPTOME SEQUENCING

Frank W Albert[1], Michel Halbwax[1], Jose A Blanco Aguiar[2,3], Miguel Carneiro[2], Sylvia Kaiser[4], Irina Plyusnina[5], Lyudmila Trut[5], Rafael Villafuerte[3], Nuno Ferrand[2], Per Jensen[6], Svante Pääbo[1]

[1]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Deutscher Platz 6, Leipzig, 04103, Germany, [2]Universidade do Porto, CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, Vairao, 4485-661, Portugal, [3]Instituto de Investigacion en Recursos Cinegéticos IREC, CSIC-UCLM-JCCM, Ronda de Toledo, s/n, Ciudad Real, 13071, Spain, [4] University of Münster, Department of Behavioural Biology, Badestr. 13, Münster, 48149, Germany, [5]Siberian Branch of the Russian Academy of Sciences, Institute of Cytology and Genetics, Prospekt Lavrentyeva 10, Novosibirsk, 630090, Russia, [6]Linköping University, Division of Zoology, IFM Biology, Linköping, 581 83, Sweden

The genetic basis of animal domestication is currently not well understood. Which genes were important when humans converted wild into domestic animals, and which genes differ today in their structure or expression between domestic animals and their wild relatives? A promising approach is to study gene expression differences between domestic and wild animals. Current next-generation sequencing technologies allow transcriptome analyses in non-model species that are difficult to study with microarray-based gene expression measurements.

To learn more about functional genetic differences between domestic and wild animals, we have used Illumina sequencing to analyze mRNA from the brains of several pairs of domestic and wild species: pigs, rabbits, Guinea pigs and dogs/wolves. We compare these data to those from a long-running selection experiment for tameness and aggression in rats. Analyses are underway to study the impact of domestication and artificial selection on gene expression levels, transcript structure and sequence polymorphism in expressed transcripts.

UNBIASED RECONSTRUCTION OF A MAMMALIAN TRANSCRIPTIONAL NETWORK MEDIATING THE DIFFERENTIAL RESPONSE TO PATHOGENS

Ido Amit[1,2], Manuel Garber[1], Nicolas Chevrier[1], Ana Paula Leite[1,2], Thomas Eisenhaure[1], Mitchell Guttman[1,2], Jen Grenier[1], Or Zuk[1], Alex Meissner[1], David E Root[1], Nir Hacohen [1], Aviv Regev[1,2]

[1]Broad Institute of MIT and Harvard, Broad, 7 Cambridge Center, Cambridge, MA, 02142, [2]MIT, Department of Biology, MIT, Cambridge, MA, 02142, [3]Harvard, Medical School, 333 Longwood Avenue, Boston, MA, 02115, [4] MGH, Center for Immunology and Inflammatory Diseases, 149 13th St, Charlestown, MA, 02129

Models of mammalian regulatory networks controlling gene expression have been inferred from genomic data, yet have largely not been validated. We present an unbiased strategy to systematically perturb candidate regulators and monitor cellular transcriptional responses. We apply this approach to derive regulatory networks that control the transcriptional response of mouse primary dendritic cells (DCs) to pathogens. Our approach revealed the regulatory functions of 125 transcription factors, chromatin modifiers, and RNA binding proteins and constructed a network model consisting of two dozen core regulators and 76 fine-tuners that help explain how pathogen-sensing pathways achieve specificity. This study establishes a broadly-applicable, comprehensive and unbiased approach to reveal the wiring and functions of a regulatory network controlling a major transcriptional response in primary mammalian cells.

MULTIVARIATE ANALYSIS OF RATE CO-VARIATION OF DIFFERENT TYPES OF MUTATIONS IN THEIR GENOMIC CONTEXT

Guruprasad Ananda[4,5], Anton Nekrutenko[1,4,5], Francesca Chiaromonte[2,4,5], Kateryna Makova[3,4,5]

[1]Penn State University, Department of Biochemistry and Molecular Biology, 108 Althouse Lab, University Park, PA, 16802, [2]Penn State University, Department of Statistics, 326 Thomas Building, University Park, PA, 16802, [3]Penn State University, Department of Biology, 208 Mueller Lab, University Park, PA, 16802, [4] Penn State University, Center for Comparative Genomics and Bioinformatics, 501 Wartik Lab, University Park, PA, 16802, [5]Penn State University, Integrative Biosciences Program, Life Sciences Building, University Park, PA, 16802

While the abundance of completely sequenced genomes has greatly facilitated our understanding of regional heterogeneity in rates of individual mutation types, the co-variation in rates of multiple mutation types has remained largely unexplored, hindering a deeper understanding of mutagenesis. In this study, we used linear and non-linear multivariate analysis tools to explore rate co-variation among four mutation types, and associate it to multiple genomic features simultaneously. We observed a concordant and largely linear co-variation among rates of nucleotide substitutions, small insertions and small deletions. In contrast, microsatellite mutability did not display co-variation with any of the other three rates studied. GC content, distance to telomere, and local recombination rates were found to be significant predictors of mutation rate co-variation, corroborating the role of these features as predictors of mutagenesis. Our analysis also uncovered the significance of novel genomic predictors of mutation rate co-variation; namely, nuclear lamina binding regions and methylated non-CpG sites. Thus, co-variation in the rates of different mutation rates might be explained by shared local genomic landscapes. Interestingly, we observed strong non-linearities among the genomic predictors explaining co-variation in mutation rates. The genomic loci driving these non-linear behaviors are located either on chromosome X or at a certain distance to telomeres, suggesting unique environments in these portions of the genome. Based on the role of various genomic predictors, we speculate about the importance of different molecular mechanisms (e.g., replication and recombination) in generating mutations. Importantly, our multivariate analysis approach can provide improved background corrections for computational methods that identify potentially functional regions of a genome – these corrections would employ composite scores, encompassing rates of multiple mutation types simultaneously.

VIRUS METAGENOMICS; VIRUS ENRICHMENT AND DEEP SEQUENCING REVEALS NEW HUMAN VIRUSES AND STRAINS

Fredrik Lysholm[2], Tobias Allander[3], Bengt Persson[1,2], Björn Andersson[1]

[1]Karolinska Institutet, Dept. of Cell and Molecular Biology, Berzelius vag 35, Stockholm, 171 77, Sweden, [2]Linköping University, IFM Bioinformatics, Linköping, 581 83, Sweden, [3]Karolinska Institutet, Dept. of Microbiology Tumour and Cell Biology, Stockholm, 171 77, Sweden

Virus infections cause many of the largest health problems in the world. It is likely that there is a multitude of unknown viruses that infect humans and it is has been suggested that viruses are involved in causing many common diseases, such as diabetes and MS. The discovery rate of new viruses has until now been slow. Deep sequencing methods have made it possible to more efficiently characterize human viruses. Our strategy includes enrichment of virus particles, 454 shotgun sequencing and bioinformatics analyses. We have sequenced pooled samples from the respiratory tract, serum, feces and other body fluids, from different patient groups, including people with common infection symptoms, different autoimmune disorders, and other diseases. In addition, several tumor samples have been sequenced. The results have revealed the presence of a multitude of known viruses, new virus strains and new virus species. Two new human viruses, Human Bocavirus and KI Polyomavirus have been published.
I will present the analysis of the sequence data accumulated thus far (10 libraries sequenced by 454)and a broader description of the known and novel viruses, bacterial, phage and human sequences found. The results of deeper analyses of variable virus families, including TTV and picornaviruses, where a multitude of new variants have been found, will be shown, as well as descriptions of which viruses are present in specific patient groups.

WHOLE GENOME RESEQUENCING REVEALS LOCI UNDER SELECTION DURING CHICKEN DOMESTICATION

Carl-Johan Rubin[1], Michael C Zody[1,2], Jonas Eriksson[1], Jennifer R Meadows[1], Ellen Sherwood[3], Matthew T Webster[1], Ted Sharpe[2], Francois Besnier[4], Örjan Carlborg[4], Paul Siegel[5], Kerstin Lindblad-Toh[1,2], Leif Andersson[1,4]

[1]Uppsala University, Medical Biochemistry and Microbiology, Box 582, Uppsala, 75123, Sweden, [2]Broad Institute of Harvard and MIT, Genomics, 7 Cambridge Center, Cambridge, MA, 02142, [3]Karolinska Institutet, Cell and Microbiology, Fogdevreten, Stockholm, 17177, Sweden, [4] Swedish University of Agricultural Sciences, Animal Breeding and Genetics, Box 7023, Uppsala, 75007, Sweden, [5]Virginia Polytechnic Institute and State University, Animal and Poultry Science, Blacksburg, VA, 24061-0306

Domestic animals are excellent models for genetic studies of phenotypic evolution. They have evolved genetic adaptations to a new environment, the farm, and have been subjected to strong human-driven selection leading to remarkable phenotypic changes in morphology, physiology and behaviour. Here we have used massively parallel sequencing to identify selective sweeps of favourable alleles as well as candidate mutations that have played a prominent role during chicken domestication and subsequent specialization into broiler (meat-producing) and layer (egg-producing) chickens. We have generated 44.5x coverage of the chicken genome using pools of genomic DNA representing eight different populations of domestic chickens as well as red junglefowl, the major wild ancestor. We report more than seven million SNPs, about 1,300 deletions and a number of putative selective sweeps. One of the most striking selective sweeps found in all domestic chickens occurred at the locus for thyroid stimulating hormone receptor (TSHR), which has a pivotal role for metabolic regulation and photoperiod control of reproduction in vertebrates. Several of the selective sweeps detected in broilers overlapped genes associated with growth, appetite and metabolic regulation. We found little evidence that selection for loss-of-function mutations played a prominent role during chicken domestication, but we detected two deletions in coding sequences that we suggest are functionally important. This study has direct application to animal breeding and enhances the importance of the domestic chicken as a model organism for biomedical research.

POPULATION GENETIC INFERENCE USING LOW COVERAGE SEQUENCING DATA

Adam Auton[1], Ryan Hernandez[2], Gil McVean[1,3], The 1000 Genomes Project[1]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Dr, Oxford, OX3 7BN, United Kingdom, [2]University of Chicago, Department of Human Genetics, University of Chicago, Chicago, IL, 60637, [3]University of Oxford, Department of Statistics, South Parks Road, Oxford, OX1 3TG, United Kingdom

In population-scale genome sequencing a common experimental design is to sequence a large number of individuals at low coverage (4-6X). Such a design is powerful for discovery of novel genetic variants in the population, but can be problematic for population genetic studies as accurate genotypes cannot be obtained without the use of LD and imputation, which has the potential to bias results and which typically requires populations to be analyzed independently.

For these reasons, we have developed methodology for estimating population genetic parameters, including the level of diversity, the site-frequency spectrum and the degree of differentiation between populations that analyses multiple populations simultaneously, incorporates uncertainty arising from sequence data and does not rely on imputation. Specifically, the method integrates over the uncertainty in the data at each site in the sample, combining population genetics modeling with machine-generated estimates of base quality to calculate the likelihood of key parameters. The method is validated through analysis of sequence data from the 1000 Genomes Project where independent genotypes are available from the HapMap project. We compare the method to estimates from genotypes inferred through imputation and describe the results of applying the methodology to population-scale genome-sequencing data sets in humans and other species.

CAPTURING THE RATE OF GERM-LINE AND SOMATIC MUTATIONS IN GENOME RESEQUENCING OF CHILDHOOD LEUKEMIA FAMILIES USING A "QUARTET" DESIGN

Philip Awadalla, Jon M Keebler, Julie Hussin, Mathieu Lariviere, Diego Czul, Daniel Sinnett

Universite de Montreal, Pediatrics, 3175 Cote Ste Catherine, Montreal, H3T 1C5, Canada

Although acute lymphoblastic leukemia (ALL) is the most common childhood cancer, factors governing susceptibility to this disease have not yet been identified. Little is known about the molecular basis of childhood ALL, although the clinical, pathological, and immunophenotypic features have been well documented. Here we sequenced the entire exomes of tumor (malignant) tissue, healthy tissue from the same individual, and from parents (quartet) from sets of leukemia patients/families (n=30) which either 1) carry a translocation, 2) the disease phenotype is associated with hyperdiploidy or 3) that do not have known genomic alterations. We also sequenced the full genome of tumor tissue and healthy tissue of an ALL family with two siblings having ALL with different ages of onset. A set of controls includes sequencing a set of twin families to obtain a background rate of de novo somatic and germ-line mutations where we demonstrate a clear excess of somatic mutations in our ALL families. Our approach involves making de novo mutation calls with a probabilistic framework that makes use of the relatedness between the individuals, including different tissue types, and produces a joint probability for the entire pedigree for each site. By dependently rather than independently determining the family members' genotypes, we take full advantage of the information contained within the pedigree. Our framework includes the implementation of an Expectation-Maximization algorithm to simultaneously produce direct estimates for the de novo point mutation rate, the population mutation rate (μ), and the sequencing error rate. We obtained a full catalogue of both germ-line and somatic de novo mutations many of which occur in coding regions of known genes associated with various cancers, indicating potential driver mutations of cancer. We have also catalogued a number of driver and passenger mutations with high predictive phenotypic impact.

A POPULATION GENOMICS APPROACH TO EXPLORE THE MOLECULAR BASIS FOR ATHLETIC PERFORMANCE TRAITS IN NORTH SWEDISH TROTTERS

Jeanette Axelsson[1], Jennifer Meadows[2], Lisa Andersson[1], Hanna Smedstad[1], Aneta Ringholm[2], Knut Roed[3], Leif Andersson[2], Sofia Mikko[1], Gabriella Lindgren[1]

[1]Molecular Breeding and Genetics, Dept. of Breeding and Genetics, P.O.Box 597, Uppsala, 75124, Sweden, [2]Dept. of Medical Biochemistry and Microbiology, Uppsala University, P.O.Box 582, Uppsala, 75123, Sweden, [3]Dept. of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, P.O.Box 8146, Oslo, N0033, Norway

We have compared the genetic makeup of North Swedish trotters (NST), North Swedish horses (NS) and Standardbreds (S) by performing a genome scan using 144 microsatellite markers (10 individuals/breed) and the Illumina EquineSNP50 BeadChip (12 individuals/breed). The total genotyping rate in individuals analyzed on the BeadChip was 0.982. 17 887 SNPs had a MAF < 0.1 and 1331 SNPs failed the missingness test (>0.1). In total 36 415 SNPs remained after frequency and genotype pruning. The three breeds separated into three distinct clusters in the identity by state (IBS) cluster analysis. All analyses above were preformed in the software PLINK.

In the microsatellite data set, the average number of alleles per locus was highest for the NS, 4.75, which also showed the highest expected heterozygosity, He, 0.666. The S showed the lowest He, 0.627, and also the lowest average number of alleles per locus, 4.40. The NST had a He of 0.649 and 4.53 as average number of alleles per locus. Following this the NS also showed the highest observed heterozygosity, Ho, 0.709. Contrary to He, however, the S showed the second highest Ho, 0.699, and the NST showed the lowest Ho, 0.692. The rarefied allelic richness average over loci was highest for the NS (3.23), and NST (3.17) and lowest for the S (3.02). The private allelic richness was highest for the S (0.96) and was 0.89 and 0.86 for the NS and the NST respectively.

The genetic distance between the NST and the S (0.503) is smaller than between the NS and the S (0.524). The genetic distance between the NST and the NS was estimated at 0.367. These results support that crossbreeding has occurred between the Standardbred and the North Swedish trotter.

GAMBIT: A CROSS-PLATFORM, LOW-MEMORY VISUALIZATION & ANALYSIS TOOLKIT FOR NEXT-GENERATION SEQUENCING

Derek Barnett, Gabor Marth

Boston College, Biology, 140 Commonwealth Ave., Chestnut Hill, MA, 02467

Next-generation sequencing is now a mainstay for polymorphism/mutation discovery, detecting structural or copy number variations, novel transcript discovery and counting-based expression analysis, etc. Although pipelines for automated analysis exist, manual data review remains essential for sequencing software developers (to check program performance), data analysts (for spot-checking), and scientists (for hypothesis generation). Visualization requires considerable flexibility from viewer software. A fully-featured viewer application must support the visualization of organismal reference sequence, multiple resequenced samples, and deep-coverage sequence alignments. Sequences are best evaluated in the context of genome annotations such as gene models, regulatory units, repeat elements, and genetic variation features. Such annotations must therefore be integrated into the viewer.

Analysis of a growing amount of sequence data currently requires specialized expertise and is rapidly becoming a nontrivial bottleneck. The needs of smaller biology laboratories can be served if the visualization software integrates essential analytical functionality such as PCR primer design (e.g. for SNP validation experiments), importing custom annotations, and exporting data from specific regions for detailed and focused analysis. Gambit is a cross-platform, low-memory, graphical program for sequence visualization and analysis. A researcher is able to upload an alignment file showing the sequencing data from samples of interest. An overall view of the alignment depth along chromosomes is provided. Essential properties of sequencing reads (e.g. sample identity, alignment orientation, mismatches & indels) are displayed as graphical cues. Views are customizable to hide/highlight certain alignment types.

Gambit takes advantage of the indexing features of the BAM sequence alignment format that allows rapid panning and zooming across chromosomes, with only minimal startup time, and re-rendering delay. Gambit is multi-threaded, handling time-intensive steps “behind the scenes”, which keeps the interface responsive. The application is also plugin-aware, so that anyone with programming skills can readily include their own custom features: support for new file formats, additional metrics, or even custom views. A beta version of Gambit is available for download at http://code.google.com/p/gambit-viewer .

ORANGUTAN *(PONGO PYGMAEUS)* MOBILE ELEMENTS: THE EXTINCTION OF ALU

Miriam K Konkel[1], Jerilyn A Walker[1], Brygg Ullmer[2], Leona G Chemnick[3], Oliver A Ryder[3], Robert Hubley[4], Arian F A Smit[4], Mark A Batzer[1], for the Orangutan Genome Sequencing and Analysis Consortium[5]

[1]Louisiana State University, Department of Biological Sciences, Biological Computation and Visualization Center, 202 Life Sciences Building, Baton Rouge, LA, 70803, [2]Louisiana State University, Department of Computer Sciences, Center for Computation and Technology (CCT), 216 Johnston Hall, Baton Rouge, LA, 70803, [3]Beckman Center for Conservation Research (CRES), Zoological Society of San Diego, San Diego Zoo, San Diego, CA, 92112, [4] Institute for Systems Biology, Computational Biology, 1441 North 34th Street, Seattle, WA, 98103, [5]Washington University School of Medicine, Genome Sequencing Center, 4444 ForestPark Ave, St. Louis, MO, 63108

Orangutans *(Pongo pygmaeus)* are the only living Asian ape and are highly endangered. We investigated the mobile DNA composition (mobilome) of the orangutan draft genome sequence derived from a female of Sumatran origin *(Pongo pygmaeus abelii)*. Similar to other primate genomes, about half of the orangutan draft genome sequence is comprised of repetitive sequences. As expected, no DNA transposon activity was detected in the orangutan lineage. L1 (long interspersed element 1, LINE1) is the only active autonomous non-LTR retrotransposon in the orangutan lineage and shows a mostly linear evolution. The orangutan-specific L1 lineage appears to be derived from L1PA3. SVA elements have been active throughout the evolution of orangutans and appear to be currently undergoing retrotransposition. Similar to L1, the orangutan-specific SVA subfamilies show a mostly linear evolution. We found evidence of expansion of SVA and L1, with ~1800 and ~4700 orangutan lineage-specific insertions, respectively. This translates to a retrotransposition rate comparable to other sequenced primates. In contrast, *Alu* elements appear to have propagated at a very low rate in orangutans. On the basis of computational and wet bench analyses, we estimate that the draft genome sequence contains only ~250 lineage-specific insertions. The identification of polymorphic and population-specific *Alu* insertions indicates that *Alu* retrotransposition is ongoing albeit at an extremely low rate. The quiescence of *Alu* retrotransposition in the orangutan-lineage is particularly surprising, as all other primate species studied to date show evidence of an appreciable *Alu* retrotransposition rate.

GENOME SEQUENCING AND MicroRNA DISCOVERY IN THE BASAL FLATWORM *MACROSTOMUM LIGNANO*

Daniil Simanov[1], Patrick van Zon[1], Ewart de Bruijn[1], Sam Linsen[1], Katrien de Mulder[1], Edwin Cuppen[1], Andres Canela[2], Gregory J Hannon[2], Dita B Vizoso[3], Lukas Scharer[3], Peter Ladurner[4], Eugene Berezikov[1]

[1]Hubrecht Institute, University Medical Center Utrecht, Uppsalalaan 8, Utrecht, 3584CT, Netherlands, [2]Cold Spring Harbor Laboratory, Watson School of Biological Sciences and Howard Hughes Medical Institute, One Bungtown Road, Cold Spring Harbor, NC, 11724, [3]University of Basel, Zoological Institute, Vesalgasse 1, Basel, 4051, Switzerland, [4] University of Innsbruck, Institute of Zoology, Technikerstrasse 25, Innsbruck, A-6020, Austria

Macrostomum lignano is a free-living flatworm with high regeneration capacity facilitated by neoblasts, the stem cell system of the worm. Due to its small size, short generation time, amenability to genetic manipulation and easy maintenance in laboratory conditions, *M. lignano* is a promising invertebrate experimental model for stem cell research. We have initiated *de novo* genome sequencing of this organism using hybrid 454 and Solexa/Illumina approach. The estimated genome size of *M. lignano* is 600 Mb. Initial 5x coverage of the genome produced ~300Mb of assembled contigs with N50 size of 1.7 kb. While this low-coverage assembly is not productive for annotation of protein-coding genes, it is already efficient for discovery of microRNAs – small RNAs that regulate gene expression at the posttranscriptional level and are involved in various cellular processes, including maintenance and differentiation of stem cells. More than 40% of the 5.4 million small RNA reads generated by SOLiD technology from irradiated (=neoblast-depleted) and non-irradiated worms mapped to the assembly and allowed annotation of more than 50 miRNA genes. Levels of several miRNAs are significantly decreased in the irradiated worms, suggesting neoblast-specific expression of these miRNAs. The progress on the genome sequencing and small RNA analysis in *M. lignano* will be presented.

ASSAYING THE DISTRIBUTION OF SEQUENCE VARIANTS WITHIN A VIRAL PATIENT SAMPLE

Henry R Bigelow[1], Michael G Ross[1], Filipe J Ribeiro[1], Bruce D Walker[2], Michael C Zody[1], Todd M Allen[2], Matthew R Henn[1], David B Jaffe[1]

[1]Broad Institute, Sequencing, 320 Charles Street, Cambridge, MA, 02141, [2]Ragon Institute of MGH, MIT, and Harvard, Virology, 149 13th Street, Charlestown, MA, 02129

Over the course of infection, single stranded RNA viruses such as HIV rapidly replicate, mutate, and are selected against by the patient's immune system. This process leads to a highly diverse and dynamic viral population, and for both clinical and theoretical reasons it is important to be able to assay this population accurately. Such an assay would need to be sensitive to rare variants that could evade the patient's immune system and in a short period of time become common.

Deep sequencing of patient samples could provide such an assay. Here we address a key technical challenge: *bona fide* rare variation tends to be swamped by sequencing error. To maximize the potential to distinguish between these two, we developed the appropriate computational model, which takes as input base calls along with quality scores, and has no prior on the distribution of variants at a given locus. While this model is computationally intractable to direct analysis, it can be interrogated using sampling, which we do using a Markov Chain Monte Carlo method.

In this way we estimate the abundance of each variant within a given sample and provide a confidence interval for each such estimate. We tested our method using Illumina sequence from 24 HIV patient samples, and assessed the methods using clonal control samples and also by sequencing the same samples using a different sequencing technology (454).

METAPHASE SPINDLE PROTEOME REVEALS POTENTIAL FURROW INITIATION FACTORS

Mary Kate Bonner[1], Ali Sarkeshik[2], Dan S Poole[1], John Yates III[2], Ahna R Skop[1]

[1]UW-Madison, Genetics Department, 425-G Henry Mall, Madison, WI, 53706, [2]Scripps Research Institute, Department of Chemical Physiology, 10550 North Torrey Pines Rd., La Jolla, CA, 92037

Cytokinesis is an important and fundamental process in the development of all organisms. The factors that establish the cleavage furrow have remained mysterious and have eluded many for over 130 years. In order to identify factors required for early steps in cytokinesis, mitotic spindles from synchronized Chinese Hamster Ovary (CHO) cells were isolated. Proteins enriched from isolated metaphase-enriched spindles were identified by multidimensional protein identification technology (MudPIT) in collaboration with the Yates Lab at Scripps. A comparative genomics analysis between the spindle and the midbody proteome (Skop *et al.*, 2004), identified potential candidates.

Results from multiple MudPIT data sets identified ~1500 proteins with two or more peptide hits. We compiled the spindle proteome by averaging four MudPIT data sets. Comparison to the midbody proteome has yielded a list of proteins unique to the metaphase spindle. We prioritized our list of candidates by sorting metaphase specific proteins, using Babelomics (http://babelomics.bioinfo.cipf.es/) and Microsoft Access. We are particularly interested in membrane-cytoskeleton remodeling proteins, as these factors are likely involved in establishing and regulating the acto-myosin contractile ring. We are currently screening several homologs of the identified mammalian candidates in *C. elegans* using RNAi. Potential candidates include factors that function in furrow formation. We are further characterizing candidates using in vivo microscopy, genetics and cell biological techniques.

MKB is supported by an NHGRI training grant to the Genomic Sciences Training Program (5T32HG002760). The work from the Yates lab is supported by the National Center for Research Resources of the National Institutes of Health by a grant to T.N.D. entitled 'Comprehensive Biology: Exploiting the Yeast Genome,' P41RR11823. A.R.S. is funded by an NSF CAREER Award (MCB-0546398).

EXPLORING SYNTHETIC GENETIC INTERACTION NETWORKS BY HIGH-THROUGHPUT RNAI

Thomas Horn[1], Thomas Sandmann[1], Bernd Fischer[2], Wolfgang Huber[2], Michael Boutros[1]

[1]German Cancer Research Center and Univ. Heidelberg, Signaling and Functional Genomics, Im Neuenheimer Feld 580, Heidelberg, 69120, Germany, [2]EMBL, Genome Biology Program, Meyerhofstr. 1, Heidelberg, 69117, Germany

Synthetic genetic interaction analysis has provided key insights by highlighting functional relationships between genes on a genome-wide scale. Genetic interactions can identify components that are otherwise masked due to buffering, dissect complex phenotypes, and circumvent early lethality in metazoans. The profile of genetic interactions of a gene with all other genes is a sensitive assay for a comprehensive view on its molecular functions. We employ high-throughput RNAi approaches to systematically identify genetic interactions in cell-based assays using Drosophila as a model system.

We developed methods to quantitatively perform, validate and analyze double RNAi experiments in a high-throughput format. Phenotypes were assessed using automated imaging to simultaneously record cell number and evaluate multi-parametric changes. Novel computational approaches, including methods for the exclusion of off-target effects, experimental fluctuations and robust error estimation were developed in order to build multi-dimensional and quantitative maps of genetic interactions based on RNAi phenotypes.

We have applied these technologies to construct a genetic interaction map of several signal transduction pathways using a matrix of combinatorial RNAi experiments in cultured cells. Mathematical modelling of the combinatorial perturbations allowed us to reconstruct known molecular pathways, to predict novel molecular interactions, and to predict new functions of genes that were confirmed by independent biochemical experiments.

GENOMIC ANALYSIS OF GLOBAL VILLAGE DOG POPULATIONS REVEALS COMPLEX DOMESTICATION HISTORY OF DOMESTIC DOGS FROM GRAY WOLVES

Adam R Boyko[1], Ryan H Boyko[2], Corin M Boyko[2], Elaine A Ostrander[3], Robert K Wayne[4], Carlos D Bustamante[1]

[1]Stanford, Genetics, 300 Pasteur Dr, Stanford, CA, 94305, [2]UC-Davis, Anthropology, 1 Shield Ave, Davis, CA, 95616, [3]NIH, NHGRI, 50 South Dr, Bethesda, MD, 20892, [4]UCLA, EEB, Chas Young Dr, Los Angeles, CA, 91302

Written into the genome of modern domestic dogs are the genetic footprints of the demographic and selective forces underlying their transition from gray wolves. However, mtDNA sequencing and SNP genotyping of wolves and dog breeds have thus far yielded conflicting accounts of dog origins. Village dogs are potentially highly informative for domestication history because of their high genetic diversity, their geographic population structure, and the small amount of genetic drift since domestication relative to breed dogs. We present a novel dataset comprised of 58,000 SNPs in 260 village dogs, 150 gray wolves, and hundreds of modern and ancient breed dogs that reveals strong evidence for a complex domestication history in dogs that differs substantially from current theories of a single, recent East Asian origin of domestic dogs.

Principal component analysis uncovers at least four radiations of dogs from a central Middle Eastern cluster, including a radiation into sub-Saharan Africa, northern Africa, East Asia and the Pacific, and Europe. PCA and haplotype sharing patterns between village dogs and wolves show both a primary domestication origin in the Middle East and regional introgression with sympatric wolf populations. Wolves have significantly less linkage disequilibrium than village dogs, consistent with a domestication bottleneck shared by breed and village dogs. In contrast, wolves have more IBD tracts than village dogs, implying village dogs have subsequently maintained higher effective population sizes than wolves.

Several traits associated with domestication---including body size, fur type and floppy ears---show significant variation within or between village dog populations. Analysis of known QTLs for these traits in breed dogs reveals they account for much of the phenotypic variation of village dogs as well. Tracing allele and haplotype sharing patterns between village dog and wolf populations around these loci reveals a rich picture of the genetic and phenotypic evolution of domestic dogs.

EXPRESSED, RARE, DELETERIOUS GENETIC VARIANTS DISTINGUISH BREAST CANCER SUB TYPES

Christopher D Brown[1], Thomas Stricker[1,2], Megan E McNerney[1,2], Ralf Kittler[1], Subhradip Karmakar[1], Kevin P White[1]

[1]University of Chicago, Department of Human Genetics and Institute for Genomics and Systems Biology, 900 E. 57th St., Chicago, IL, 60637, [2]University of Chicago, Department of Pathology, 900 E. 57th St., Chicago, IL, 60637

Breast cancer is a heterogenous disease, representing a mix of tumors with different histologies, expression of biomarkers and gene expression profiles. Importantly, expression of biomarkers such as Estrogen Receptor (ER), the Progesterone Receptor (PR), or the Human Epidermal Growth Factor Receptor 2 (HER2) can determine choices in therapy. Indeed, one of the most aggressive classes of breast cancer is defined by the absence of these biomarkers; so-called triple negative breast cancer (TNBC) cannot be treated by inhibiting hormone receptors or by targeting HER2. To examine the genetic basis underlying the different subtypes of breast cancer, we sequenced the transcriptomes of eleven ER positive (ER+) breast cancers and fifteen TNBCs. In total, we identified over 70,000 allelic variants, many of which are presumed to be of germline origin as they commonly occur in the human population. However, a smaller fraction of allelic variants we identified are not common alleles and are predicted to be deleterious to protein function. Genotyping of adjacent normal tissue demonstrate that a small fraction of these variants have a somatic origin. These rare, deleterious alleles deifferentiate less aggressive ER+ tumors from more aggressive triple negative tumors. In contrast, common variants fail to segregate the tumor types. Thus, the previously defined clinicopathologic subtypes of breast cancer have a common genetic basis that is unique from the other subtypes. Dozens of rare, deleterious alleles contribute to each genetic subtype we observed, indicating that these subtypes are may be by many loci acting in combination. Furthermore, these combinations of alleles often occur in multiple components of the same molecular pathways.

GENOME-WIDE PATTERNS OF POPULATION STRUCTURE AND ADMIXTURE AMONG HISPANIC/LATINO POPULATIONS

Katarzyna Bryc*[1], Christopher Velez*[2], Tatiana Karafet[3], Andres Moreno-Estrada[1,4], Andy Reynolds[1], Adam Auton[1,5], Michael Hammer[3], Carlos D Bustamante*[1,4], Harry Ostrer*[2]

[1]Cornell University, Dept Biol Stat and Comp Biology, 1187 Comstock Hall, Ithaca, NY, 14850, [2]NYU School of Medicine, Dept of Pediatrics Hum Gen Program, 550 First Avenue, MSB 136, New York, NY, 10016, [3]University of Arizona, ARL Div of Biotech and Dept of Ecology and Evol Biol, 1041 E. Lowell St., Tucson, AZ, 85721, [4] Stanford University, Dept of Genetics, Mail Stop-5120, Stanford, CA, 94305, [5]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom

Hispanic/Latino populations possess a complex genetic structure that reflects recent admixture among and potentially ancient substructure within Native American, European, and West African source populations. We quantify genome-wide patterns of SNP and haplotype variation among 100 individuals with ancestry from Ecuador, Colombia, Puerto Rico, and the Dominican Republic genotyped on the Illumina 650K SNP panels as well as 112 Mexicans genotyped on the Affymetrix 500K platform. Intersecting these data with previously collected high-density SNP data from 4,305 individuals, we use principal component analysis and clustering methods to investigate genome-wide patterns of African, European, and Native American population structure within and among Hispanic/Latino populations. Comparing autosomal, X and Y chromosome, and mtDNA variation, we find evidence of a significant sex bias in admixture proportions consistent with disproportionate contribution of European male and Native American female ancestry to present day populations. We also find that patterns of linkage-disequilibria in admixed Hispanic/Latino populations are largely impacted by the admixture dynamics of the populations with faster decay of LD in populations of higher African ancestry. Finally, we reconstruct fine-scale chromosomal patterns of admixture and estimate time since admixture from the lengths of ancestry tracts for each of the admixed populations.

* These authors contributed equally.

TISSUE-SPECIFIC REWIRING OF SIGNALING PATHWAYS THROUGH ALTERNATIVELY SPLICED DISORDERED SEGMENTS

Marija Buljan[1], Alex Bateman[1], Madan Babu[2]

[1]Wellcome Trust Sanger Institute, Informatics Department, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom, [2]MRC Laboratory of Molecular Biology, Structural Studies, Hills Road, Cambridge, CB2 0QH, United Kingdom

It is a frequent phenomenon that the same gene takes part in signaling pathways that have different, sometimes even opposing, outcomes. Intriguingly, mechanisms that ensure the specificity of the transmitted signals are still unclear. In this study we present the argument for the importance of finely regulated alternative splicing of disordered protein segments in attaining this specificity. We observe that tissue-specific exons encode protein segments enriched in intrinsically disordered regions with overrepresented binding peptides and post-translationally modified sites. Functional relevance of the observed phenomenon is further indicated with significant evolutionary conservation of the tissue-specific disordered regions and predicted binding peptides. By alternatively splicing functional disordered segments, an individual gene can achieve functional versatility without compromising the structural stability of its protein products. Since the mechanisms for regulation of signaling specificity are frequently disrupted in cancer and other diseases, it is important to understand the role of proteins' functional disordered residues in the regulation of signaling cascades.

ENRICHING FOR INDELS IN COMPLEX DISEASE BY DEEP SEQUENCING.

Lara M Bull-Otterson*[1], Alex Coventy*[2], Andrew G Clark[2], Alan R Templeton[3], Thomas J Rea[4], Charles F Sing[4], Jacy Crosby[4], Xiaoming Liu[5], Taylor Maxwell[5], Eric A Boerwinkle[5], Richard A Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Molecular & Human Genetics, 1BaylorPlaza, Houston, TX, 77030, [2]Cornell University, Molecular Biology and Genetics, TowerRd, Ithaca, NY, 14853, [3]Washington University, Biology and Genetics, Monsanto419, St.Louis, MO, 63130, [4] University of Michigan, Human Genetics, 1241ECatherine, AnnArbor, MI, 48109, [5]University of Texas HSC-H, Div of Epidemiology, 1200HermanPressler, Houston, TX, 77030

* Authors contributed equally

Genetic association studies of complex disease commonly focus on variations caused by single base-pair substitutions while few studies consider the large assortment of insertions and deletions (INDELs) that also make up natural genetic variation in the face of disease. In a large population-based cohort of 13,715 individuals from the Atherosclerosis Risk in Communities (ARIC) study, we examined the rare and common INDELs in HHEX and KCNJ11, genes historically associated with diabetes. Our analysis focuses on these genes in relation to the population genetics of INDELs and the impact of INDEL variations on metabolic risk factors for diabetes and coronary heart disease. Using Sanger sequencing of the full gene regions, we detected significantly associated INDELs for multiple metabolic phenotypes that were internally replicable. From the initial 76 INDELs, 44 were validated by 454FLX sequencing and ranged in size from 1-9bp. We examined the INDELs in context with the surrounding haplotype background and the linkage disequilibrium between putative functional mutations and the INDELs. We then examined the correlation between phenotypic variation and regions of conservation in orthologous sequences. Finally, we compared the INDELs in the extreme phenotypic deciles to those found in the full distribution to examine the consequence of a common sampling strategy in wide use today. These data provide the first glimpse at the population genetic dynamics of INDEL variations in a population-based sample.

FITNESS AND STRUCTURE LANDSCAPES FOR PRE-MIRNA PROCESSING

Ralf A Bundschuh[1], Juliette de Meaux[2], Michael Lässig[3]

[1]Ohio State University, Department of Physics, 191 West Woodruff Av, Columbus, OH, 43210, [2]Max Planck Institute for Plant Breeding Research, Department of Plant Breeding and Genetics, Carl-von-Linné-Weg 10, Köln, 50829, Germany, [3]University of Cologne, Institute for theoretical Physics, Zülpicher Str. 77, Köln, 50937, Germany

The processing from pre-miRNA to mature miRNA in plants involves a mechanism, which depends on an extended stem in the secondary structure of the pre-miRNA. Here, we show how natural selection acts on this secondary structure to produce *evolutionary conservation* of the processing mechanism together with *modularity* of the pre-miRNA molecules, making this molecular function independent of others. Our main results are:

1. Selection on miRNA processing can be described by a fitness landscape which depends on the free energy ΔG of the processing stem as quantitative molecular phenotype. Stem structures with $\Delta G > 17$kcal/mol have a constant fitness value; less stable stems incur a relative fitness cost. We infer this fitness landscape by a genome-wide analysis of *Arabidopsis thaliana*, comparing the stem free energy distribution of 130 functional pre-miRNAs with a background ensemble of random RNA molecules.

2. This fitness landscape (together with the overall sequence divergence) predicts the divergence of the free energy phenotype, as well as the selection coefficients of pre-miRNA sequence changes. These predictions are in excellent quantitative agreement with the results of a genome-wide cross-species comparison of orthologous pre-miRNAs between *A. thaliana* and *A. lyrata*. We conclude that there is a strong evolutionary constraint on processing phenotype and function, although individual sequence changes are only under weak selection. Our analysis shows how natural selection interacts with the complex genotype-phenotype mapping induced by RNA folding.

3. Actual pre-miRNA structures are modular: their stem free energy is significantly less affected by deleterious mutations in the remainder of the molecule than for random RNA molecules. This modularity is driven by mutational load favoring structures with more stable stems. This suggests a general evolutionary mechanism by which selection, mutations, and genetic drift generate modularity, a feature that is important for the independence of molecular functions in RNA, proteins, and extended molecular networks.

PUNCTUATED OR GRADUAL? TIMING THE ACCELERATION OF HUMAN ACCELERATED REGIONS WITH NEANDERTAL DNA SEQUENCES

Hernán A Burbano[1], Richard E Green[1,2], Tomislav Maricic[1], Marco de la Rasilla[3], Antonio Rosas[4], Michael Lachmann[1], Svante Pääbo[1]

[1]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Deutscher Platz 6, Leipzig, 04103, Germany, [2]University of California, Biomolecular Engineering, 1156 High St, Santa Cruz, CA, 95064, [3]Universidad de Oviedo, Departamento de Historia, Calle del Teniente Alfonso Martinez, Oviedo, 33011, Spain, [4] Museo Nacional de Ciencias Naturales, Departamento de Paleobiología, Calle de José Gutierrez Abascal 2, Madrid, 28006, Spain

Recent comparative genomics analyses have identified regions that are both highly conserved in vertebrate evolution yet fast-evolving on the human lineage. One intriguing possibility is that these human accelerated regions (HARs) are spots of adaptive evolution in human ancestors. Alternatively, however, the biased substitution spectrum of these elements, and their proximity to recombination hotspots, suggest that biased gene conversion could have caused the observed acceleration.
One limitation in investigating the forces shaping HARs is that it is unknown exactly when, in the last 5-7 million years since the last human-chimpanzee common ancestor, the human substitutions occurred. To get temporal resolution on the evolution of HARs, we are using DNA sequences from Neandertals, the closest extinct relative of current human populations, to determine which substitutions in HARs pre- and postdate the divergence to Neandertals.
Using targeted sequence capture on glass microarrays, we captured Neandertal DNA sequences of previously identified HARs from a ~43,000-year-old Neandertal specimen from El Sidrón, Spain, attaining multifold coverage. To generate a picture of which substitutions in HARs are fixed versus polymorphic in present-day humans, we also sequenced the same regions in 50 individuals from the Human Diversity Panel. We will present the substitutions in HARs and compare them with genome-wide average rates of substitutions.

A GENERAL METHOD FOR ASSEMBLING GENOMES FROM ILLUMINA DATA

Joshua N Burton, Iain A MacCallum, Sante Gnerre, Dariusz Przybylski, Filipe Ribeiro, Bruce Walker, Ted Sharpe, Giles Hall, Carsten Russ, Chad Nusbaum, David B Jaffe

Broad Institute, Sequencing, 320 Charles St, Cambridge, MA, 02141

Over the past decade, DNA sequencing costs have dropped about 10,000-fold. The lowest-cost reads from current technologies are short, have a high error rate, and land unevenly on the genome. They are ideally suited to 'resequencing' applications in which a preexisting reference sequence is available; but for *de novo* genome assembly, they are challenging to work with, and results have generally been inferior to those obtained from the old (Sanger method) technology.

Here we demonstrate a practical and general laboratory/computational method for generating high-quality *de novo* assemblies of genomes at the lowest possible cost. Our method starts with 100-base Illumina paired reads from two libraries: one from fragments of size 180 bp (slightly less than twice the read length), and one from fragments of size 3000 bp, via a 'jumping' construction. These two libraries use off-the-shelf methods and provide power that could not be obtained from a single library. We also demonstrate experimental methods for jumping longer fragments to yield a third library, providing even greater potential for long-range connectivity.

We assembled these data using our new version of the ALLPATHS algorithm. This algorithm has been scaled up to work on large genomes and made robust to idiosyncrasies in library construction, variation in coverage, and run-to-run variability in sequence quality, all of which have been critical problems for genome assembly.

We tested our method using a suite of 16 genomes, including 9 for which a reference sequence was available and 7 from completely new samples. These genomes range in size from 2 to 2600 Mb, and in GC content from 19% to 71%. Using the preexisting reference sequences, we assess the completeness, continuity, and accuracy of these assemblies, finding that in most cases their quality exceeds the general quality level of draft assemblies that had been achieved using Sanger sequencing.

COMPARATIVE ANALYSIS OF TRANSCRIPTION IN FOUR YEAST SPECIES USING RNA-SEQ

Michele Busby[1], Jesse Gray[2], Michael Springer[2], Chip Stewart[1], Jeffrey Chuang[1], Michael Greenberg[2], Gabor Marth[1]

[1]Boston College, Biology, 140 Commonwealth, Newton, MA, 02467, [2]Harvard Medical School, Neurobiology, 220 Longwood Ave, Boston, MA, 02115, [3]Harvard Medical School, Systems Biology, 220 Longwood Ave, Boston, MA, 02115

Comparative gene expression studies can be used to see which genes are diverged across species, not just in terms of sequence, but also in terms of function. RNA-Seq provides a method of measuring gene expression that is independent of probe design and therefore ideally suited for cross-species comparisons. We used this approach to compare the transcription level of orthologous genes in four Saccharomyces species.

We found that the transcription level of single-copy orthologous genes was well-correlated between species, and genes which are known to have diverged in expression during evolution were easily identified. For example, genes involved in the nitrogen catabolite repression pathway were clearly expressed differently in cerevisiae versus the other three yeast species.

We also looked at gene transcription in evolutionary lineages after a gene duplication event. We found several examples where the syntenic gene maintained a transcription level consistent with the orthologous genes in the other species, while the paralog, located in a different genomic context, had a markedly decreased level of transcription.

In the course of this study, we developed a simple and statistically robust method for measuring differential gene expression in RNA-Seq (shotgun) reads. To develop this method, we relied on biological replicates as a null model. We found that our best results were achieved when we used the raw count of reads which uniquely mapped to the region of the genome where the gene is located as our primary measurement of gene expression, and we did not normalize by transcript length. We observed that variations in the measurement of gene expression across replicates was greater than is predicted by Poisson variation. We quantified the magnitude of this variation based on the replicates and modified the conventional X2 test to allow for calibration based on these fluctuations. We then use this test to identify statistically significant, and therefore likely biologically meaningful, differences in gene expression.

IDENTIFYING THE GENETIC DETERMINANTS OF TRANSCRIPTION FACTOR ACTIVITY

Eunjee Lee[1], Harmen J Bussemaker[1,2]

[1]Columbia University, Biological Sciences, 1212 Amsterdam Ave, New York, NY, 10027, [2]Columbia University, Center for Computational Biology and Bioinformatics, 1130 St Nicholas Ave, New York, NY, 10032

Understanding how phenotype relates to genotype is one of the central goals of biology. Analysis of parallel genotyping and expression profiling data has shown that mRNA expression level is highly heritable. While the molecular mechanisms underlying the heritability of gene expression levels are poorly understood, they are expected to often involve mediation by transcription factors (TFs). We here present a transcription-factor-centric and sequence-based method for dissecting the transcriptional response to genetic perturbations. In our approach, we first predict the affinity with which each TF binds to the promoter region of each gene using quantitative prior information about the DNA binding specificity. Next, we perform genomewide linear regression of differential mRNA expression on predicted promoter affinity to estimate segregant-specific TF activity as a quantitative phenotype. Genetic mapping of the TF activity trait allows us to identify the activity quantitative trait loci ("aQTLs") whose inheritance modulates the regulatory activity of each specific TF. Our method has a greatly improved statistical power to detect regulatory mechanisms underlying the heritability of genomewide mRNA expression. Specifically, it identifies seven times as many locus-TF associations and more than twice as many trans-acting loci from a genetic cross between two haploid yeast strains as all existing methods combined. We validated our ability to predict locus-TF associations in yeast using gene expression profiles for allele replacement strains. Furthermore, application to mouse data from an F2 intercross identified an aQTL on chromosome VII modulating the activity of Zscan4 in liver cells, demonstrating that our method also works in higher eukaryotes.

IMPROVING THE HIGH-QUALITY DRAFT SWINE GENOME REFERENCE

Mario Caccamo

The Genome Analysis Centre, Bioinformatics, Norwich Research Park, Norwich, NR4 7UH, United Kingdom

On behalf of the International Swine Genome Sequencing Consortium

The availability of a high-quality genome reference sequence is a key resource to drive scientific discoveries in molecular biology. Whilst the recent developments of high-throughput sequencing technology have accelerated and reduced the cost of generating genomic data it remains a challenge to assemble high-quality contiguous sequences that can reliably represent the genome of the organism under study. Here we present the quality improvement work being undertaken by the Swine Genome Sequencing Consortium to generate a high-quality draft reference genome sequence for the pig. The development of the pig sequence exploits the availability of a physical map that integrates fingerprinted clones from 4 different BAC libraries as well as a dense panel of radiation-hybrid markers. A minimal tile path of BAC clones representing 98.3% of the physical map was subjected to hierarchical shotgun Sanger sequencing, with preference given to clones from the CHORI-242 library, which was constructed from a single Duroc sow (TJ Tabasco). Sequence coverage for each BAC clone was supplemented using automated primer walking from the ends of assembled sequence contigs to obtain an equivalent to 6x shotgun coverage of each BAC clone. A number of genomic regions of specific interest to the pig research community and the ENCODE regions have been sequenced to finished quality. In the latest release of the pig genome (Sscrofa9), 114Mb of the 2.97 Gbp represented in sequence contigs have been finished. To further improve the quality of the genome, we are now integrating whole genome shotgun (WGS) Illumina sequences from the same Duroc sow generated by the Beijing Genomics Institute (China) and the Wellcome Trust Sanger Institute (UK) into the assembly. These reads were assembled using a stringent algorithm to ensure that the resulting contigs were free of misassemblies. Although short, the these contigs provided an invaluable resource to reliably close or reduced the size of 20% of the gaps. This approach is complemented by the alignment of WGS Sanger reads generated by the National Livestock Research Institute (Korea) and full-length cDNAs to correct the orientation and order of contigs. This strategy preserves the underlying sequence architecture defined by the physical map and minimizes the confounding effect of repeats. This information will be integrated in future releases of the pig genome reference sequence.

PATTERNS OF GENETIC CHANGE IN DROSOPHILA CODING SEQUENCE REVEAL THE GENERIC MODULARITY OF PROTEINS, AND THE UBIQUITY OF EPISTASIS.

Benjamin J Callahan[1], Richard A Neher[2], Peter Andolfatto[3], Boris I Shraiman[2]

[1]Stanford University, Applied Physics, 476 Lomita Mall, Stanford, CA, 94305, [2]University of California, Santa Barbara, Kavli Institute for Theoretical Physics, Santa Barbara, CA, 93106-4030, [3]Princeton University, Department of Ecology and Evolutionary Biology, Princeton, NJ, 08544

Protein structure is generically modular, a property which extends to protein sequence and results in characteristic patterns in the distribution of changes along protein coding sequence. We consider this effect in the Drosophila genome by use of correlation functions between divergences and polymorphisms as a function of sequence separation. Amino acid evolution is significantly correlated on a length scale of O(20) amino acids, the characteristic length scale of protein sequence modules. We call this phenomenon *local clustering* and its cause is two-fold: the character of selection (e.g. level of constraint) is correlated within a module, and epistasis is more likely within a module. Epistasis is revealed in the increased probability that nearby divergences occur on the same lineage, and by a significant tendency for nearby same-lineage divergences to conserve the total local charge. Overrepresentation of the doubly mutant haplotype at nearby polymorphic sites also suggests epistasis. While heterogeneity in constraint contributes greatly to clustering, we conclude that epistasis is essential to at least 10% of amino acid changes.

RAPID INTEGRATION OF NOVEL GENES INTO CELLULAR NETWORKS

John A Capra[1,2], Katherine S Pollard[1], Mona Singh[2]
[1]University of California, San Francisco, Gladstone Institutes, 1650 Owens St, San Francisco, CA, 94158, [2]Princeton University, Dept. of Computer Science and the Lewis-Sigler Institute for Integrative Genomics, 35 Olden St, Princeton, NJ, 08540

Gene duplication has long been appreciated as a source of raw material for genetic and functional innovation. More recently, attention has focused on other mechanisms that generate new genes. For example, *de novo* gene creation from previously non-coding sequence has been found in fungi, flies, and mammals---with estimates that as many as 12% of new genes in fly and 6% in human were created via this mechanism. Processes such as exon shuffling, incorporation of mobile elements, and gene fission and fusion can also create new genes with sequence and structure combinations that are novel to a genome. Despite the prevalence of new gene creation, very little is known about the functions of recently evolved genes and how these functions change over time.

To investigate these processes, we classified all genes in *Saccharomyces cerevisiae* according to their age and mechanism of creation. This enabled us to analyze the relationship between how and when a gene was created and the roles it plays in the cell. We observed significant differences in genes' annotation, essentiality, interactions, and specific functions based on their evolutionary histories.

Our primary finding is that new genes become more functionally integrated into the cell with time, but that the dynamics of this process differ significantly between duplicated and totally novel genes. For example, new proteins are generally less connected in physical interaction networks than older proteins, but duplicate proteins are more integrated than novel proteins of similar age. Similarly, older proteins are more likely to be essential and to play roles in well-characterized cellular processes than new proteins, especially those created by mechanisms other than duplication. Furthermore, recently created duplicated genes show evidence of being involved in adapting existing functions to environmental changes, while no significant functional enrichment was found among new genes created by other mechanisms. Thus, genes with novel sequences are initially less integrated into cellular networks than duplicated genes, but they are more likely to experience a dramatic gain in function after creation.

This study shed light on several other aspects of gene evolution. First, we found a significant preference among all proteins to interact with other proteins that share both the same age *and* mechanism of creation. Second, the regulation of genes differs based on their age and origin, with novel genes being significantly less regulated than other groups. Finally, analysis of new proteins in the context of the interaction network identified areas of potential recent adaptation.

NEXT-GENERATION TARGETED RESEQUENCING TO INVESTIGATE POPULATION HISTORY OF GIBBON SPECIES

Lucia Carbone[1], Sung Kim[2], Alan R Mootnick[3], David Li[1], Pieter J deJong[1], Jeffrey D Wall[2]

[1]Children's Hospital Oakland Research Institute, CHORI, 5700 Martin Luther King Jr, Oakland, CA, 94609, [2]University of California San Francisco, Institute for Human Genetics, 513 Parnassus Avenue, S965, San Francisco, CA, 94143, [3]Gibbon Conservation Center, Gibbon Conservation Center, PO Box 800249, Santa Clarita, CA, 91380

Gibbons are small apes that inhabit South East Asia and shared a common ancestor with the other hominoids about 18 million years ago. Gibbon species display an extremely high rate of chromosomal rearrangements, 20 fold higher than most of the other mammals. Moreover, with 15 recognized species they show a taxonomic diversity greater than the other hominoids. We have already shown that the increase frequency of chromosomal rearrangements in gibbons can be explained by lower epigenetic repression of transposable elements (Carbone et al. 2009). We are now investigating factors that might explain the high fixation rate of chromosomal rearrangements in this species group. In particular we are speculating that population dynamics and other aspects of gibbon population history might have played a role.
To address this problem we have been gathering sequence data using targeted next-generation sequencing (Illumina GAII) in non-exonic sequences. For this study we used our collection of genomic DNA of more than 20 unrelated individuals from 9 different gibbon species. We are currently examining this unique dataset and we will discuss the results of this analysis, including the possible implications with the rapid karyotype evolution observed in gibbon species.

Carbone, L. et al. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. PLoS Genet 5, e1000538 (2009).

DECIPHERING MAMMALIAN TRANSCRIPTOME COMPLEXITY BY DEEP-CAGE

Piero Carninci[1], Timo Lassmann[1], Hazuki Takahashi[1], Charles Plessy[1], Nicolas Bertin[1], Geoffrey Faulkner[2], Nadine Hornig[3], Carrie Davis[4], Valerio Orlando[3], Thomas Gingeras[4], Yoshihide Hayashizaki[1]

[1]RIKEN, Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan, [2]The Roslin Institute and R(D)SVS University of Edinburgh, Roslin Institute, Midlothian EH25 9PS, Edinburgh, EH25 9PS, United Kingdom, [3]Dulbecco Telethon, Epigenetics and Genome Reprogramming lab, Via del Fosso di Fiorano 64, Rome, 00143, Italy, [4] CSHL, Functional Genomics, Cold Spring Harbor, NY, NY, 11724

The next generation sequencing technologies are profoundly influencing our way to study biology. We have previously developed cap-analysis gene expression (CAGE) to simultaneously mRNA/noncoding RNA starting sites and simultaneously detect their expression and transcriptional networks. We also adapted CAGE to few nanogram of RNA (nanoCAGE). CAGE technology was coupled to deep sequencing (DeepCAGE) to achieve comprehensive coverage of transcription starting site, and more recently also miniaturized for use with small amount of material (nanoCAGE) and to assign newly discovered promoters/5' ends to the remaining part of the RNA sequence (CAGEscan). We have adapted several of these technologies for multiple sequencers, including the 454 Life Sciences, Illumina, SOLiD and Heliscope.

DeepCAGE is being also used on subcellular fractionation as a part of the ENCODE project. We have detected clear patters of expression of retrotransposon elements (RE) in a panel of human and mouse tissues, which have a regulatory role. Now, we have clearly identified specific patterns of RE-derived RNA in different cell compartments. For instance, Line derived RNAs are preferentially bound to chromatin, while other specific patterns have been identified also for nucleoplasm, nucleolus, cytoplasmic, nuclear and polysomal RNA fractions. Altogether, these data are pointing at a potential function of these elements. In the C2C12 muscle cells, perturbation of the expression of Line elements can either accelerate or repress differentiation of myoblasts into myotubes.

A POPULATION GENETIC APPROACH TO MAPPING NEUROLOGICAL DISORDER GENES USING DEEP RESEQUENCING

Ferran Casals*[1], Rachel A Myers*[1,2], Julie Gauthier[3], Jonathan E Keebler[1,2], Adam R Boyko[4], Carlos D Bustamante[4], Amelie M Piton[3], Dan Spiegelman[3], Edouard Henrion[3], Martine M Zilversmit[1], Julie Hussin[1], Jacki Quinlan[1], Yan Yang[3], Ron Lafrenière[3], Alexander Griffing[2], Eric A Stone[2], Guy A Rouleau[1,3], Philip Awadalla[1,2,3]

[1]Université de Montréal, Centre de Recherche CHU Sainte-Justine, 3175 Côte-Ste-Catherine, Montréal, H3T 1C5, Canada, [2]North Carolina State University, Bioinformatics Research Centre, Box 7566, Raleigh, NC, 27659, [3]Centre d'excellence en neuromique, Université de Montréal, 2099 Alexandre-de-Sève, Montréal, H2L 2W5, Canada, [4] Stanford University School of Medicine, Dpt Genetics, 300 Pasteur Drive, Stanford, CA, 94305

Deep re-sequencing of coding regions in human genomes coupled with parametric and non-parametric statistical approaches are necessary to identify potentially causal rare variants for complex disorders. We present the results of a study to identify rare variants associated with ASD and Schizophrenia, the first to use population genetic methods. Nonparametric approaches identified three genes (MAP1A, GRIN2B and CACNA1F) with significant excesses of rare missense mutations in either one or both disease cohorts. In a broader context, we also find the variation in cases is best explained by population models including selection and complex demography over either neutral models or models including only complex demography. Mutations in disease associated genes explain much of the difference in the overall site frequency spectrum among the cases versus controls. This study demonstrates that genes associated with complex disorders can be mapped using resequencing and analytical methods, using sample sizes far smaller than those required by GWA studies. Additionally, our findings support the hypothesis that rare mutations are required to explain the etiology of these complex disorders.
*equal contribution

SAMPLE SEQUENCING THE DYNAMIC REPEAT STRUCTURE OF SNAKE GENOMES

Todd A Castoe[1], Kathryn Hall[1], Marcel Guibotsy Mboulas[2], A. P. Jason de Koning[2], Cedric Feschotte[1], David D Pollock[2]

[1]University of Colorado School of Medicine, Biochemistry and Molecular Genetics, 12801 17th Ave, Aurora, CO, 80045, [2]The University of Texas at Arlington, Biology, 501 S. Nedderman Drive, Arlington, TX, 76019

Genomic sequencing efforts have mostly focused on creating nearly complete and highly annotated genome sequences, but target species are chosen based on limited prior genome information. This strategy has provided only a limited diversity of complete genomes among vertebrates and great gaps in our understanding of genome diversity and dynamics across vertebrates. Modest sequencing of random samples from vertebrate species using next-generation sequencing methods is a promising cost-effective alternative for amassing comparative data for understanding vertebrate genome evolution and diversity. To test this approach, we sampled moderate fractions (2-4.5%) of the genome of two snake species, the copperhead (Agkistrodon contortrix) and the Burmese Python (Python molurus). This modest sampling allowed a fairly comprehensive perspective on the transposable element and simple sequence repeat structure of these two snake species, and provides novel insight into the repeat structure of snake genomes. Our results demonstrate the utility of such sample sequencing across a diversity of vertebrates for great resolution of vertebrate genome evolution, and the great genome diversity present within snakes.

TRANSCRIPTOME CHARACTERIZATION OF PLANARIAN SCHMIDTEA MEDITERRANEA BY MASSIVE PARALLEL SEQUENCING

Catherine Adamidi, Yongbo Wang, Xintian You, Christoph Dieterich, Nikolaus Rajewsky, Wei Chen

Max-Delbrueck-Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, Robert-Rössle-Straße 10, Berlin, 13125, Germany

Planarian Schmidtea mediterranea has extraordinary regenerative capacities and is becoming a model organism for the study of regeneration, tissue homeostasis and stem cell biology. Although the genome annotation endeavors are rapidly ongoing, until now its transcriptome characterization has not been completed and is still largely dependent on computational prediction . Here, by integrating 454 and Solexa sequencing technology, we sequenced the poly A+ transcripts from the whole worm of S. mediterranea at an unprecedented depth and do novo assembled the transcriptome. In total, 26,669 candidate transcripts with mean length 1,300 bp were obtained, out of which over 99% can be aligned to the current genome reference sequence. The transcribed regions and the exons covered 35.3% and 7.4% of the genome, respectively. Our results provide the most comprehensive transcriptome reference resources for further functional research. In addition, the robust transcriptome characterization pipeline established in this study could be widely applied to the study in other non-model organisms whose genome sequence has not been determined yet.

INTEGRATED GENOME ANALYSIS OF GENETIC NETWORKS REGULATED BY EYELESS DURING RETINAL DEVELOPMENT IN DROSOPHILA

Yiyun Chen[1], Yumei Li[1], Keqing Wang[1], Rui Chen[1,2,3]

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, HGSC, One Baylor Plaza, Houston, TX, 77030, [3]Baylor College of Medicine, Developmental Biology Program, One Baylor Plaza, Houston, TX, 77030

As a key regulator of retinal development in metazoan, *eyeless* (*ey*) functions at the top of the genetic hierarchy and is sufficient to initiate the entire genetic cascade controlling retinal cell specification, determination, and differentiation. To gain comprehensive understanding of the genetic network controlled by ey during retinal development, we have first performed comprehensive gene expression profiling using microarray. Gene expression profiles have been generated for 19 tissue and genotype combinations, which allow us to identify and group of genes regulated by *ey* during retinal development. In parallel, genome wide Ey occupancy has been profiled by ChIP-Seq in developing eye imaginal discs. A total of more than 20 million uniquely mapped reads have been obtained from Ey ChIP and more than 3000 potential Ey binding regions have been identified. By comparing with the profile of PolII and Histon H3K4Me3 for the same tissue, we are able to classify these 3000 regions into promoters and enhancers. Combining with gene expression and Ey binding profile, downstream targets and pathways that are directly regulated by Ey in the *Drosophila* genome have been identified. To our surprise, ey appears to directly regulate genes from virtually all major known genetic pathways involved in retinal development, including the *hh*, *dpp*, *wg*, *Notch*, *EGFR*, and cell cycle control pathway. Furthermore, ey frequently regulates multiple members in the same pathway, presumably allowing more complex and fine-tuned control. In addition, we have found evidence that *ey* can function as both an activator and a repressor in terms of regulating downstream gene expression. Finally, in addition to known genes and pathways, a large set of novel ey targets have been identified. Functional analysis of these candidate genes is currently underway and several novel genes have been found to cause retinal developmental defects when mutated. Further studies of these genes will provide important insights of Ey function.

AN INTEGRATIVE COMPUTATIONAL PIPELINE TOWARDS LARGE-SCALE ACCURATE DISCOVERY OF GENOMIC STRUCTURAL VARIATION IN CANCER

Ken Chen, Lei Chen, John Wallis, Li Ding, George Weinstock, Elaine R. Mardis, and Richard K. Wilson

The Genome Center, Washington University School of Medicine, 4444 Forest Park Blvd., St. Louis MO 63108

Recent progress on next generation sequencing of whole genomes has greatly expanded the scope and the resolution of structural variant (SV) analysis. To facilitate the systematic discovery of somatic SVs in hundreds of cancer genomes with a reasonable validation rate, we developed an integrative SV discovery pipeline that includes a suite of algorithms for variant detection, read re-mapping, and localized assembly. Among the components in this pipeline, the BreakDancer algorithm and the iterative graph routing assembler (TIGRA) are our key innovations. Our pipeline was able to ascertain twice as many nucleotide resolution SVs in a personal genome (NA18507) than a whole genome assembly approach reported recently. Meanwhile, it can accurately pinpoint somatic SVs in cancer genomes, as demonstrated by numerous validated novel somatic SVs discovered in various cancer genomes. The breakpoint sequences produced by targeted assembly have allowed us to further examine the genotypes and to predict the putative formation mechanisms of each SV in a large number of cancer genomes. We will present these results as well as recent improvements to our procedures that lower the false positive rate in this most challenging of discovery processes.

MAPSPLICE: MAPPING RNA-SEQ READS FOR SPLICE DISCOVERY

Kai Wang[1], Darshan Singh[3], Zheng Zeng[1], Stephen J Coleman[2], Xiaping He[4], Piotr Mieczkowski[4], Charles M Perou[4], James N MacLeod[2], Derek Y Chiang[1,4], Jan F Prins[3], Jinze Liu[1]

[1]University of Kentucky, Department of Computer Science, Lexington, KY, 40506, [2]University of Kentucky, Maxwell H. Gluck Equine Research Center, Department of Veterinary Science, Lexington, KY, 40506, [3]University of North Carolina, Department of Computer Science, Chapel Hill, NC, 27599, [4] University of North Carolina, Department of Genetics and UNC Lineberger Comprehensive Cancer Center, Chapel Hill, NC, 27599

The vast majority of genes can undergo alternative splicing among multiple transcript isoforms. Deep sequencing of cDNA libraries provides quantitative snapshots on the diversity of alternative splicing. In particular, the counts of sequence reads spanning multiple exons can be used to infer the relative proportion of transcript isoforms. Thus, algorithms that reliably map sequence reads to splice junctions are critical for bioinformatics analyses of alternative splicing. While many computational approaches align short sequence reads to splice junction databases, these databases are incomplete and must be custom-built for different read lengths. We present a new algorithm, called MapSplice, to discover splice junctions from the alignment of mRNA-seq reads to any reference genome. MapSplice combines global and local search to find optimal spliced alignments in a memory-efficient implementation. Our method tolerates read errors relative to the reference genome (due to SNP variation or incorrect base calls) and utilizes base-call quality scores to maximize the probability of an alignment. A classification strategy predicts canonical as well as non-canonical splice junctions, based on the quality of read alignments and the distribution of aligned reads that include the given junction. We will present MapSplice analyses of 75bp mRNA-seq datasets to identify alternative splicing differences between basal and luminal classes of breast cancers.

REWIRED SIGNAL TRANSDUCTION PATHWAYS AMONG *SACCHAROMYCES CEREVISIAE* STRAINS.

Brian L Chin, Gerald R Fink

Whitehead Institute for Biomedical Research, Biology, 9 Cambridge Center, Cambridge, MA, 02142

What is the phenotypic consequence of individual genetic variation? The *Saccharomyces cerevisiae* S288C deletion collection provides phenotypes for every ORF in this reference strain, but it does not address the question of individual strain variability. Here, we use identical deletion libraries made in two closely related *S. cerevisiae* strains, S288C and Σ1278b, to examine control of agar adhesion. Both strains have about 300 genes that affect agar adhesion, but few of these genes affect adhesion in both strains. In other words, a gene required for adhesion in S288c will likely not be required for adhesion in Σ1278b, and vice versa. The MAP kinase pathways are a striking example of this. They have been rewired such that the filamentation MAPK, needed for adhesion in Σ1278b, is dispensable for adhesion in S288C. Instead, we find that S288C utilizes signals from other pathways thereby bypassing the filamenation MAPK pathway. The genes that bypass the requirement for the filamentation MAPK for adhesion in S288C are polymorphic in Σ1278b. This shows that the genetic variation between these strains can significantly impact what mutations will have a phenotype, to the extent that a well characterized signaling pathway can have no effect in one strain but a significant effect in another.

IDENTIFICATION AND ANALYSIS OF NOVEL, FUNCTIONAL VARIANTS BY POPULATION-BASED DEEP RE-SEQUENCING IN EIGHT DRUG TARGET GENES

Stephanie L Chissoe[1], Matthew R Nelson[1], Kijoung Song[2], Silviu-Alin Bacanu[1], Xiangyang Kong[2], Dana Fraser[1], Jennifer Aponte[1], Li Li[1], Xin Yuan[2], John Whittaker[3], Dawn Waterworth[2], Lon Cardon[2], Vincent Mooser[2]

[1]GlaxoSmithKline, Genetics, 5 Moore Drive, Research Triangle Park, NC, 27709, [2]GlaxoSmithKline, Genetics, 709 Swedeland Rd, Upper Merion, PA, 19406, [3]GlaxoSmithKline, Genetics, Third Avenue, Harlow, CM19 5AW, United Kingdom

We conducted a re-sequencing study to explore the genetic variation within eight genes encoding drug targets and assess the phenotypic effects of those variants. This knowledge may increase our confidence in target choice and point to new indications for these drugs.

Re-sequencing was performed on 2000 subjects from the CoLaus population study, which measured a range of cardiovascular, metabolic and psychiatric traits. The sequencing strategy targeted exons and exon-intron boundaries, including 10,248 coding bases, and identified 515 variants. The vast majority, 437 (85%), were rare (<0.5% minor allele frequency [MAF]) and 292 (56%) were observed in one heterozygous individual, and only 41 (8%) had MAF>5%. The coding, UTR and intronic frequency spectra were similar with a small decrease in common variants within coding regions.

Genotype data was available for 25 of the 515 variants including 14 with MAF<5%. The overall heterozygous error rate was 6.78%, with the sequence data calling fewer heterozygotes. Further genotyping was conducted to confirm the new variants. Although the concordance was very high, genotyping very rare variants was challenging. Two methods were used to conduct genotype–phenotype analyses. Variants with MAF>0.5% were analyzed individually and those with MAF<0.5% were analyzed in aggregate within genes.

Results will be presented on the analysis and characterization of variants identified. This experiment has several implications for the design of future re-sequence-based studies. Because so many of the observed variants are private, follow up studies may need to be based on sequencing rather than genotyping.

DBVAR: NCBI'S DATABASE OF GENOMIC STRUCTURAL VARIATION

D M Church, T P Sneddon, J Lopez, J Garner, A Mardanov, C Clausen, N Bouk, J Paschall, M Feolo, S T Sherry, L Phan, D R Maglott, J Ostell

NCBI/NIH, IEB, 45 Center Dr, Bethesda, MD, 20892

Structural variants in the human genome, those segregating in the population as well as *de novo* events, have profound impact on human health. The current challenge lies in translating these discoveries into understanding of phenotypic consequences. To help address this challenge, we are providing a new resource called dbVar that is focused on collecting and adding value to large scale variation data.
dbVar currently includes data for several species. While data is available for multiple organisms, human is by far the most data rich. In addition to data surveying normal populations, such as HapMap, dbVar maintains data from studies making genotype-phenotype correlations for a range of diseases, including autism and cancer. In cases where sample level data are not consented for public release, the individual level data is first submitted to dbGaP (the database of Genotype and Phenotype) and summary level data are submitted to dbVar. The data are presented such that asserted variants, supporting data and experimental data are available for download.
One challenge of maintaining these data is that the submissions are typically tied to a particular genome assembly. However, meta-analysis across multiple datasets requires the placement of variants into a common coordinate system; dbVar performs this critical mapping process on submission and updates for subsequent assemblies. Variants with a genome location may be mapped to a new assembly via assembly-assembly alignment; however, the complex structure of many of these regions makes them difficult to align and data may be lost. Additionally, some of the structural variants have been identified based on misassembly in a previous version of a genome. We will discuss our efforts to improve mapping of these variants to a common coordinate system as well as our approach to correlating these events with curation being performed by the Genome Reference Consortium (GRC).

A NEW HIGH-RESOLUTION AND ULTRA-DENSE ZEBRAFISH MEIOTIC MAP.

Matthew D Clark[1], Carlos Torroja[1], John Postlethwait [2], Derek L Stemple[1]

[1]Wellcome Trust Sanger Institute, Vertebrate development and genetics, Genome Campus, Hinxton, CB10 1SA, United Kingdom, [2]University of Oregon, Institute of Neuroscience, Eugene, OR, 97403-1254

The Sanger Institute's zebrafish genome sequencing project is based on the Tübingen strain, the other widely used strain is AB. The current assembly (Zv8) is clone by clone sequence supplemented by whole genome shotgun. Zebrafish lacks homozygous strains, and has a 1.45Gb genome with large arrays of repeat, duplicated and AT-rich regions. By fertilizing eggs with inactivated sperm, and heat shock suppression of the first mitotic division we generated doubled haploid (DH) Tübingen and AB individuals. Using Illumina GA technology we generated 40x base coverage of a DH AB male and DH Tübingen female. We mapped these reads to Zv8, finding over 10 million single nucleotide polymorphisms (SNPs), and 700,000 small insertion/deletions between these completely homozygous fish. We crossed the sequenced fish to generate genetically identical heterozygous F1 hybrids, and F1s to generate F2s. Since F2s can only have the sequenced alleles at any location, and exhibit hydrid vigor, we are distributing them to the community under the name SAT (Sanger AB x Tübingen).
We selected 200,000 SNPs across Zv8 for a SNP array, and genotyped 459 F2 fish. After SNP calling with homozygous and heterozygous hint tables, and genetic filters, we made a de novo map of 141,675 SNPs with MSTmap (Wu, Bhat, Close and Lonardi. 2008). The new map has ~1SNP/10kb mapped at a resolution of 0.1cM (~60kb). The Illumina data also contain sequences not present in the clone based Zv8. The next assembly, Zv9, will contain merged clone, shotgun and Illumina data, and be confirmed not just by sequence and clone overlaps but also extensive genetic mapping. The new markers place previously unattached scaffolds and, in combination with read depth, help identify haplotypic and duplicated regions.
We believe the combination of next generation sequencing and a dense genetic map could be used in other genome projects to improve assembly and usefulness of the genome sequence. As the technology matures, dense meiotic or linkage disequilibrium maps are feasible to construct and resolve assembly problems that can not be solved by more sequence alone and which were previously addressed using detailed physical maps.

BROWSING 1000 GENOMES DATA USING ENSEMBL

Laura Clarke, Holly Zheng Bradley, Richard Smith, Eugene Kuleshea, William McLaren, Paul Flicek, The 1000 genomes Project

EBI, Vertebrate Genomics, The Wellcome Trust Genome Campus, Hinxton, Cambridge, cb10 1SD, United Kingdom

The 1000 Genomes Project is producing a deep catalogue of human variation, to provide a better baseline to underpin human genetics.

In order to make this data useful to the wider community the project will be submitting its variant calls to the appropriate archives such as dbSNP. An Ensembl style browser is also provided to enable users to see the variant calls in the context of other annotation at http://browser.1000genomes.org.This browser allows users to view the SNP sets both for high coverage individuals and for the low coverage population in genomic context and see the consequences of the snps on the gene structures.

Recent developments in both the Ensembl variation code and the Ensembl web code are enabling us to show a greater variety of variation data and make alignment data easier to visualise. Ensembl is now working with emerging structural variation archives to display this data as tracks within the genome browser. As structural variants are released based on 1000 Genomes data we hope to be able to display them on the genome in the same way.

The 1000 Genomes project provides a vast amount of alignment data in the SAM format. These files represent the alignment of all the sequence data for available for an individual to the reference genome. We can now view these files using the Ensembl system which should allow users to see how the genomic sequencing supports the variant calls the project is making. This large volume of data will hopefully help drive Ensembl variation visualization forward as more people need help to see the data in both the gene and genomic context.

MICROBIAL AND METAGENOMIC RESEARCH AT THE WASHINGTON UNIVERSITY GENOME CENTER

Sandra W. Clifton, Erica Sodergren, Makedonka Mitreva, Vince Magrini, Lucinda Fulton, Bob Fulton, Asif Chinwalla, Darina Cejkova, Lei Chen, Laura Courtney, Catrina Fronick, Hongyu Gao, Otis Hall, Betty Lobos, Kathie Mihindukulasuriya, Patrick Minx, Michelle O'Laughlin, Kymberlie Pepin, Michal Strouhal, Chad Tomlinson, Kristine Wiley, Tiffany Williams, Guohui Yao, Yanjiao Zhao, Elaine Mardis, Richard K. Wilson, George Weinstock.

The Genome Center at Washington University, St. Louis, MO 63108

Microbial genomics at the Genome Center at Washington University includes studies of both individual genomes as well as communities, and of prokaryotes, eukaryotes, and viruses. Sequencing of individual bacteria employs Illumina as well as the 454 platform. Bacteria are bar-coded, pooled, twelve to an Illumina lane, sequenced and deconvoluted, with over 90% of the genomes passing the criteria for a high quality draft sequence. Sanger, 454, and Illumina platforms are employed for metagenomic sequencing of communities focusing on 16S rRNA or shotgun sequencing. These methods are being used to characterize the microbiome of healthy individuals, as well as diseased tissues in acne, necrotizing enterocolitis, Crohn's disease, urethritis, HIV infection, periodontitis, as well as dietary effects measured in primates. Virus sequencing, to characterize the healthy virome or agents of febrile pediatric cases, employs either 454 or Illumina sequencing. Eukaryotic microbes are being studied both as reference sequencing projects as well as ocean metagenomics.

A FRAMEWORK FOR DETECTION AND INTERPRETATION OF STRUCTURAL VARIATION FROM MATEPAIR DATA

Cristian Coarfa[1], Oliver A Hampton[1], Petra Den Hollander[2], Martin M Matzuk[3], Adrian V Lee[2], Aleksandar Milosavljevic[1]

[1]Baylor College of Medicine, Molecular and Human Genetics, 1 Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, Lester and Sue Smith Breast Cancer Center, 1 Baylor Plaza, Houston, TX, 77030, [3]Baylor College of Medicine, Department of Pathology, 1 Baylor Plaza, Houston, TX, 77030

The availability of long insert mate pairs for next generation technology sequencing offers opportunities to analyze structural variants at an unprecedented scale. Compared to the the initially available short inserts in the 200-500 base pair range, currently available inserts in the 4000-6000 basepair range enable a more sensitive detection of structural variants. The increasing size of inserts and volume of reads pose a series of challenges: 1) it is imperative to have computational methods that scale well with the increasing volume of reads and can make use of existing parallel infrastructures; 2) Interpretation of the obtained structural variants requires integrative data analysis, both with respect to previous studies and with respect to genomic features of interest. To answer these challenges we have developed the Breakout software package which performs multiple functions. Given a matepair dataset, Breakout fist decomposes the input into balanced subsets, to enable good load balancing on a multi-node cluster or a multicore desktop. Next, it infers automatically the lower and upper bounds for the observed insert size, and clusters efficiently aberrant matepairs indicating breakpoints using a parallel grid iterative strategy. Resulting matepairs are filtered for redundant clones and annotated with p-values using a Poisson distribution. Breakout enables their interpretation against reference genomic features, such as gene promoters, user-specified targets, and against published structural variant datasets. We show applications of this method on cancer cell lines, with PCR validation. These pipelines are made available to the community via a Galaxy server. We showcase the use of the Genboree Discovery System for structural variants intuitive visualization, via genome browser extensions and specialized graphing tools such as Circos.

IDENTIFYING CAUSAL GENETIC VARIANTS WITH SINGLE-NUCLEOTIDE EVOLUTIONARY CONSTRAINT SCORES

Gregory M Cooper[1], David L Goode[2], Sarah B Ng[1], Arend Sidow[2,3], Michael J Bamshad[1,4], Jay Shendure[1], Deborah A Nickerson[1]

[1]University of Washington, Genome Sciences, 1705 NE Pacific St, S-433D, Seattle, WA, 98115, [2]Stanford University, Genetics, 300 Pasteur Dr, Stanford, CA, 94305, [3]Stanford University, Pathology, 300 Pasteur Dr, Stanford, CA, 94305, [4] University of Washington, Pediatrics, 1705 NE Pacific St, Seattle, WA, 98115

Massively parallel sequencing is becoming routine in human genetics, but identifying disease-causing genetic variants in individual genomes is a difficult problem. Noncoding regions (>98% of the genome) are particularly challenging to study, but isolation of causal variants is difficult even within the best-characterized subset of the genome that encodes proteins (i.e. the 'exome'). Analyses of nucleotide-level sequence conservation hold potential to address this challenge, on the assumption that purifying selection 'constrains' evolutionary divergence at phenotypically important nucleotides. Constraint scores are quantitative, do not require functional annotations (e.g. non-synonymous), and are applicable genome-wide, offering substantial advantages, in principle, over other approaches to identify causal variants. However, it remains unclear if constraint scores hold practical utility for the analysis of resequenced individuals. We therefore assessed the ability of a nucleotide-level evolutionary metric to identify causal variants in 16 published exomes, derived from 4 individuals with Freeman-Sheldon syndrome (FSS; OMIM #193700), a dominant disease caused by rare variants in MYH3; 4 individuals with Miller syndrome (OMIM #263750), a recessive disease caused by rare variants in DHODH; and 8 HapMap samples. We estimated constraints on each nucleotide of the human genome using Genomic Evolutionary Rate Profiling (GERP) in sequence alignments of 34 mammalian species that capture ~5.8 substitutions per neutral site. We found that single-nucleotide constraint scores enrich strongly and site-specifically for deleterious variants within these exomes. GERP scores were, in fact, more effective than functional definitions (i.e. non-synonymous) at enriching for the known causal variants for both diseases, and also more effective than other, more complex, methods like SIFT and PolyPhen that only consider non-synonymous variants. Since they are quantitative, constraint scores also facilitate relative comparisons that are unavailable with discrete annotations, and ranked the known causal genes highly even under disease models that allow for genetic heterogeneity. Assuming a monogenic mode of inheritance, GERP scores ranked the known causal gene as the top candidate for both diseases. We conclude that single-nucleotide evolutionary constraint scores offer clear utility for analyzing exomes. Given this proof of principle and considering that constraint scores are readily defined genome-wide, they hold exciting potential for the discovery of causal variation in arbitrary genomic segments (e.g. linkage peaks) and ultimately re-sequenced genomes.

THE GENETIC ARCHITECTURE OF IMMUNE-MEDIATED DISEASE

Chris Cotsapas[1,2,3], Benjamin F Voight[1,2,3], Kasper Lage[1,2,4], Elizabeth R Rossin[1,2,3], Benjamin M Neale[1,2,3], Mark J Daly[1,2,3]

[1]Massachusetts General Hospital, Center for Human Genetic Research, 185 Cambridge St, Boston, MA, 02114, [2]Harvard Medical School, Dept of Medicine, 185 Cambridge St, Boston, MA, 02114, [3]Broad Institute of MIT and Harvard, Medical and Population Genetics, 7 Cambridge Ctr, Cambridge, MA, 02138, [4] Massachusetts General Hospital, Pediatric Surgical Research Laboratories, 185 Cambridge St, Boston, MA, 02114

Recent genome-wide association (GWA) studies have identified numerous replicable genetic associations influencing risk of common autoimmune and inflammatory (immune-mediated) diseases. Moreover, several loci have been independently shown to influence risk to more than one such disease. This observation suggests commonality in the processes underlying disease progression and is echoed in epidemiological observations, which nonetheless have as yet been unable to propose common molecular mechanisms of pathogenesis. Here, we compare 140 loci associated to immune-mediated disease across GWA studies of six such diseases and find that ~50% of these regions are convincingly associated to multiple immune-mediated diseases. We further show that patterns of disease association group loci into clusters which encode interacting proteins, suggestive of distinct molecular mechanisms. Finally, we examine the genetic overlap between MS and Crohn Disease in more detail by comparing genome-wide meta-analyses of the two diseases across 629,000 SNPs and find association to components of the interleukin 23 receptor- and prostaglandin receptor-mediated signaling pathways in MS, suggesting a role for variation in the cell functions they control in both diseases.
This work demonstrates how functional information may be systematically retrieved from GWA data to make inferences about the molecular underpinnings of pathogenesis and provide medically relevant cellular targets for drug discovery and disease prediction efforts. We note that these approaches are relevant to any group of related traits and are likely useful in psychiatric, metabolic and pharmacological genetic research.

VAST EXCESS OF RARE VARIATION REVEALED BY RESEQUENCING 13,715 INDIVIDUALS

Alex Coventry*[1], Lara M Bull*[2], Xiaoming Liu[4], Andrew G Clark[1], Taylor J Maxwell[4], Jacy Crosby[4], James E Hixson[4], Thomas J Rea[3], Alan R Templeton[4], Eric Boerwinkle[4], Richard Gibbs[2], Charles F Sing[3]

[1]Cornell University, Ithaca, NY, 14853, [2]Baylor College of Medicine, Houston, TX, 77030, [3]University of Michigan, Ann Arbor, MI, 48109, [4] University of Texas Health Science Center, Houston, Texas.

* Contributed equally

Targeted resequencing in an exceptionally large sample is the only way to (1) obtain an unbiased sample of rare human genetic variants, (2) determine how human demographic shifts have shaped their abundance and distribution, and (3) begin to understand the role of rare variants in risk for complex disease. Using Sanger sequencing of genomic PCR amplicons, we resequenced the diabetes-associated genes *KCNJ11* and *HHEX* in 13,715 individuals from the Atherosclerosis Risk in Communities (ARIC) study, then validated amplicons harboring rare variants using 454 pyrosequencing. The expected number of segregating sites in these two genes is far higher than previous polymorphism surveys would have predicted: over 740 and comprised mostly of rare variants, including an enriched set of damaging and loss-of-function variants. Variants which arose over the last 8,000 years would primarily appear as singletons in our dataset, so we had a unique opportunity to estimate human demography in Europe after the widespread adoption of agriculture there. We found a clear signature of a much higher population growth rate than has been estimated in earlier studies based on higher-frequency variants. Because higher-frequency variants arose earlier in human history, we conclude that, consistent with historical and archaeological data, there was an acceleration in the growth rate over the last 8,000 years. In fact, rapid population growth explains the huge number of rare variants we identified, and suggests that there are hundreds of distinct, rare loss-of-function mutations in every human gene. We also used our demographic estimates to estimate the mutation rates in the two genes directly from the site-frequency spectra (SFS) we obtained.

HUMAN ISLETS PREPARED FOR CLINICAL TRANSPLANTATION EXHIBIT AN ALTERED GLYCOLYTIC PROFILE

Mark J Cowley*[1], Anita Weinberg*[2], James Cantley[3], Warren Kaplan[1], Stacey N Walters[2], Wayne J Hawthorne[4], Philip J O'Connell[4], Shane T Grey[2]

[1]Garvan Institute, Peter Wills Bioinformatics Centre, 384 Victoria St, Sydney, 2010, Australia, [2]Garvan Institute, Gene Therapy and Autoimmunity Group, 384 Victoria St, Sydney, 2010, Australia, [3]Garvan Institute, Diabetes and Obesity Research Program, 384 Victoria St, Sydney, 2010, Australia, [4] Westmead Hospital, The Centre for Transplant and Renal Research, Hawkesbury Rd, Sydney, 2145, Australia

Type I diabetes (T1D) is characterized by the destruction of pancreatic beta cells, leading to inadequate insulin production and poor control of blood sugar levels. Islet cell transplantation can potentially reverse the harmful effects of T1D, however there is a decline in graft success rate from ~70% (1 yr) to ~10% (after 5 yrs). We reasoned that some islet preparations were more capable of restoring euglycemia than others, controlled in part at the level of gene expression.

We used HGU133+2 microarrays to characterize 9 transplant grade human islet preparations. In the absence of patient outcome data, we used gene expression heterogeneity to identify perturbed molecular processes: 153 (0.28%) genes had MAD>1.5. We adapted GSEA to assess variability at the level of genesets; 5 genesets relating to hypoxia were enriched (FDR<0.001). Animal studies show that stabilization of hypoxia inducible factor 1α (HIF-1α) represents a molecular switch towards anaerobic glycolysis ablating glucose sensing and insulin secretion in islets.

Compared to laser captured microdissected human islets, key glycolytic enzymes PDK1, LDHA, and MCT4 were up-, and GCK down-regulated, indicating a switch to anaerobic glycolysis. These results were confirmed using mouse islets cultured in normoxic and hypoxic conditions, which was reversed using a HIF-1α inhibitor. We transplanted hypoxic mouse islets into immunodeficient RAG-/- mice, and 4 weeks after transplantation the same glycolytic profile was observed as seen in human islets.

The permanence of this molecular signature may explain poor islet function post transplantation and suggests early intervention in reversing the hypoxic changes may be important.

GENOME-WIDE DNASEI FOOTPRINTING IN A DIVERSE SET OF HUMAN CELL-TYPES

Gregory E Crawford[1], Alan P Boyle[1], Lingyun Song[1], Bum-Kyu Lee[2], Damian Keefe[3], Ewan Birney[3], Vishwanath R Iyer[2], Terrence S Furey[1]

[1]Duke University, Institute for Genome Sciences & Policy, 101 Science Drive, Durham, NC, 27708, [2]University of Texas at Austin, Cellular and Molecular Biology, 2500 Speedway, Austin, TX, 78712, [3]European Bioinformatics Institute, EMBL Outstation, Hinxton, Cambridge, CB10, United Kingdom

Regulation of gene transcription is largely determined by cis-elements where trans-acting factors bind. In diverse cell types, gene transcription levels are modulated by utilizing varied sets of cis-elements. Here we demonstrate that data from high-throughput DNaseI hypersensitivity (HS) assays (DNase-seq) can delineate base-pair resolution 'footprints' that precisely mark individual protein-DNA interaction sites within DNaseI HS sites across the genome. We find that footprints for specific transcription factors correlate well with ChIP-seq enrichment and correctly identify functional vs. non-functional sites computationally predicted using motifs. By analyzing 6 cell types, we have identified many footprints for various factors that are cell-type specific and many others that co-localize together. We also find that footprints reveal a unique evolutionary conservation pattern that differentiates footprinted bases from surrounding DNA. These footprints can be used in addition to ChIP-seq data to more comprehensively elucidate genomic regulatory systems.

A HIGH RESOLUTION MAP OF THE DROSOPHILA TRANSCRIPTOME BY PAIRED-END RNA-SEQUENCING

Bryce Daines[1,2], Liguo Wang[3], Hui Wang[2], Yumei Li[1,2], David Emmert[5], William Gelbart[5], Wei Li[3,4], Richard Gibbs[1,2], Rui Chen[1,2]

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, 77030, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030, [3]Baylor College of Medicine, Dan L. Duncan Cancer Center, Houston, TX, 77030, [4]Baylor College of Medicine, Department of Molecular and Cellular Biology, Houston, TX, 77030, [5]Harvard University, Department of Molecular and Cellular Biology, Cambridge, MA, 02138

We have applied RNA-Sequencing to generate a comprehensive map of the Drosophila melanogaster transcriptome by broad sampling of 12 distinct developmental stages. In total, 272 million 64-75bp paired-end reads were generated on the Illumina GA II yielding 690x sequencing coverage. Direct supporting evidence for more than 95% of annotated FlyBase genes and 90% of splicing junctions was observed. Furthermore, improvements were made to 34% of the current annotated gene models by extension of UTRs, inclusion of novel exons, and identification of novel splicing events. A total of 479 novel transcripts were identified, representing a 3.4% increase over the current Drosophila gene set. These data suggests that alternate splicing occurs in 36% of Drosophila genes, significantly less than mammals, but a 61% increase over previous estimations. Much of the increase is attributable to subtle alternate splicing forms such as tandem alternate splice sites (TASS). For one form of TASS, the NAGNAG alternate acceptor site, 241 novel cases were observed, a 4-fold increase over annotations. To make this massive dataset readily accessible multiple data tiers have been designed and made available through FlyBase providing a valuable resource to the Drosophila research community.

DETERMINING EVOLUTIONARY CHANGES IN GLUCOCORTICOID RECEPTOR BINDING SITES USING CHIP-SEQ.

Charles G Danko[1,2], Lee W Kraus[1], Adam Siepel[2]

[1]Cornell University, Department of Molecular Biology and Genetics, 467 Biotechnology Building, Ithaca, NY, 14853, [2]Cornell University, Department of Biological Statistics and Computational Biology, 102D Weill Hall, Ithaca, NY, 14853

Changes in gene expression play an important role in the diversification of primate species. This effect is mediated partially by changes in the binding locations of DNA sequence-specific transcription factors. The glucocorticoid receptor (GR) is a steroid hormone-dependent transcription factor that affects the expression of hundreds of genes in the presence of corticosteroid hormones such as cortisol and dexamethasone (DEX). GR plays an important role in regulating innate and acquired immunity, and is involved in mediating a response to both stress and anxiety in nearly all tissues, including the brain. Moreover, several aspects of GR physiology and ligand sensitivity are known to have changed during the course of primate evolution.

Here, we examine changes in GR binding sites during primate evolution. Our strategy is to systematically determine the location of GR binding sites in human, chimpanzee, and rhesus macaque using chromatin-immunoprecipitation coupled with massive parallel sequencing (ChIP-seq). We focus on CD4+ T-cells, a peripheral blood cell that plays an important role in adaptive immune response. CD4+ T-cells are isolated from human and non-human primate peripheral blood samples using magnetic separation. Samples are treated using the synthetic GR agonist DEX to separate true species-specific differences in binding from those caused by differences in environment and ligand sensitivity. Using this system, we are in the process of constructing a deep and robust dataset of GR binding in human, chimpanzee, and rhesus macaque. This dataset will represent a wealth of information useful in understanding how changes in GR binding play a role in primate evolution.

RAPID EVOLUTIONARY INNOVATION DURING AN ARCHEAN GENETIC EXPANSION

Lawrence A David[1], Eric J Alm[2,3]

[1]Massachusetts Institute of Technology, Computational & Systems Biology Initiative, 77 Massachusetts Avenue, Cambridge, MA, 02139, [2]Massachusetts Institute of Technology, Department of Civil & Environmental Engineering, 77 Massachusetts Avenue, Cambridge, MA, 02139, [3]Massachusetts Institute of Technology, Department of Biological Engineering, 77 Massachusetts Avenue, Cambridge, MA, 02139

A natural history of Precambrian life remains elusive because of the rarity of microbial fossils and biomarkers. We have employed an explicit model of macroevolution including gene birth, transfer, duplication and loss to map the evolutionary history of 3,968 gene families across the three domains of life. This model uses the topology of the gene family tree rather than just gene copy presence/absence across genomes and accounts for uncertainty in gene trees. We observe that gene transfer is the primary source of gene gain in prokaryotes, while duplication dominates in Eukaryotes. Inter-domain gene transfer is rare compared to intra-domain transfer with the notable exception of a statistically significant enrichment for Bacteria-Eukarya transfer events that corresponds to the endosymbioses of the mitochondria and chloroplasts. Surprisingly, we find that a brief period of genetic innovation during the Archean eon gave rise to 27% of major modern gene families. Genes born during this period are especially likely to be involved in electron transport, while later genes exhibit a gradually increasing usage of molecular oxygen. Our results demonstrate that reconstructing the complex interplay between organismal and geochemical evolution over Earth history is becoming a tractable goal.

OPTIMAL STUDY DESIGN FOR TARGETED RE-SEQUENCING IN POOLED DNA FOR DISEASE ASSOCIATION STUDIES

Aaron G Day-Williams, Kirsten McLay, Eleanor Howard, Alison J Coffey, Aarno Palotie, Eleftheria Zeggini

Wellcome Trust Sanger Institute, Human Genetics, Wellcome Trust Genome Campus, Hinxton, CB4 3QJ, United Kingdom

Genome-wide association studies have yielded an unprecedented number of associations between common genetic variants and complex traits. But the loci identified only explain a small proportion of the heritability of the analyzed traits. Rare variants have been postulated to harbor some of the missing heritability. This has generated a huge interest in the re-sequencing of large numbers of cases and controls to find rare variants associated with disease. Although next-generation sequencing technologies have drastically reduced the costs of sequencing, it is still prohibitively expensive to sequence thousands of individuals. This has lead to the development of numerous methods for targeted sequence enrichment. To further reduce cost and increase the number of genomes sampled many groups propose to perform the capture and sequencing in DNA pools. There are many questions that need investigation in order to optimally design a targeted re-sequencing experiment in pooled DNA including choice of sequence enrichment technology and number of individuals to pool.
We evaluate long-range PCR, array based pull down, and in-solution based pull down enrichment methods for 6 genomic regions composing 1.6Mb. The enriched regions are sequenced in DNA pools of 1, 2, 10, 20, and 50 individuals with an Illumina GA II. The 50 individuals assayed are composed of 31 HapMap individuals and 19 control individuals from the WTCCC. In addition to the extensive genotype data on these individuals, 25 of the individuals are sequenced in the ExoSeq project and 22 individuals are sequenced in Pilot 1 of the 1000 Genomes project (19 individual overlap). Utilizing all the available genotyping and sequencing data on our subjects we will present an analysis of the estimated false positive and negative rates of variant discovery and accuracy of allele frequency estimation in all 15 study designs as well as recommendations on the optimal design including which capture technologies are the best in terms of enrichment of target regions and sequencing coverage, which is the most cost-effective, and the optimal number of individuals to include in the pools.

HIGH-THROUGHPUT EVOLUTIONARY GENOMIC ANALYSIS ON LARGE PHYLOGENIES

A.P. Jason de Koning, Todd A Castoe, David D Pollock

University of Colorado Denver, School of Medicine, Biochemistry and Molecular Genetics, 12701 E 17th, MS8101, PO Box 6511, Aurora, CO, 80045

The informativeness of comparative data for making functional inferences can depend strongly on taxonomic sampling. In particular, densely-sampled comparative data is expected to be highly informative on patterns of site- and lineage-specific variability, including on coevolutionary interactions among sites and on functional and adaptive shifts. We are approaching an exciting time for evolutionary genomics, in that the number of sequenced metazoan genomes is expected to increase remarkably to 500-10,000 over the next 2-10 years. Unfortunately, making effective use of this incoming flood of data poses significant computational challenges, particularly because the computational burden of rigorous evolutionary analysis explodes with increasing data size and model complexity. We have therefore been working on ways to alleviate the poor scaling of likelihood-based phylogenetic approaches, to facilitate rapid, large-scale comparative genomic analysis using more realistic models. We present a variety of results showing how likelihood analysis can be rapidly performed on large comparative genomic data sets with nearly as little computational burden as required by parsimony-based analysis, but without appreciably sacrificing rigor. We show how these approaches can be applied in a Bayesian context to a number of important problems in comparative genomics and molecular evolution. In particular, we will present novel results on genome-wide inference of variation in selective constraint, variation of evolutionary pattern across sites in proteins, and coevolutionary interaction among sites. In each application, data analysis can be performed in merely minutes on a desktop computer, even for data sets including hundreds or thousands of species.

MASSIVELY-PARALLEL TRANSCRIPTOME SEQUENCING ANALYSIS BY CLOUD COMPUTING

Francisco M De La Vega[1], Jigntao Sun[1], Catalin Barbacioru[1], Adam Kraut[2], Bill Van Etten[2], Brian Tuch[1], Yongming Sun[1]

[1]Life Technologies, Genomic Systems R&D, 850 Lincolnc Centre Dr., Foster City, CA, 94404, [2]BioTeam Inc, Professional Services, 7 Derosier Drive, Middleton, ME, 01949

Due to growing throughput and shrinking cost, massively parallel sequencing is rapidly becoming an attractive alternative to microarrays for the genome-wide study of gene expression. The sequencing of transcripts (RNA-Seq) offers several advantages over microarray-based methods, including the ability to detect somatic mutations and accurately measure allele-specific expression. The Applied Biosystems SOLiD System can produce over a billion 50bp reads per run, enabling deep sequencing analysis of transcriptomes. Nevertheless, one of the challenges of the technology is that analyzing such ever growing volumes of data requires the implementation of commodity computer clusters that could be expensive in their high-end configurations, need dedicated facilities and incur associated maintenance and environmental costs. As an alternative to the computer cluster configuration, we explored implementing the SOLiD Whole Transcriptome Analysis pipeline on a Utility Computing environment and on the Amazon Elastic Compute Cloud (EC2). As a feasibility study, we analyzed a RNA-Seq dataset of over 700 million reads coming from a single slide from a SOLiD instrument. We compare both environments in terms of ease of deployment of applications, use, performance, and cost. The EC2 environment offers the maximum flexibility and two execution modes: virtual clusters and Map-Reduce jobs. The former is easier to deploy, although still with some effort, but not as scalable, and the later requires rewriting the algorithm code to accommodate the distribution system, which could be initially expensive and time consuming. On the other hand, the Utility Computing paradigm offers complete transparency and is highly efficient. We conclude that the analysis of sequencing data through cloud computing environments is efficient and cost effective, and could supplement or become the main IT infrastructure for large and small **LABS.**

MULTIPLE REGIONS OF STRONG PRIMATE-SPECIFIC NONCODING CONSTRAINTS IN THE *FOXP2* LOCUS

Ricardo C del Rosario, Shyam Prabhakar

Genome Institute of Singapore, Computational and Systems Biology, 60 Biopolis St, #02-01 Genome, Singapore, 138672, Singapore

Recent studies have provided significant examples of cis-regulatory sequence changes that drove morphological evolution in model organisms. However, regulatory elements that evolved uniquely among primates and contributed to primate-specific morphology have yet to be identified. The transcription factor FOXP2, which is associated with motor coordination, motor learning, and speech and language development, is well known for two human-specific amino acid substitutions that affect neuronal morphology and verbal phenotypes when introduced into mice. Intriguingly, the *FOXP2* locus is extremely rich in non-coding conservation, raising the possibility that it may also have contributed to the evolution of motor phenotypes through alterations in its pattern of expression. Specifically, we hypothesized that primate-specific regulatory sequences in the *FOXP2* locus might have contributed to the evolution of unique motor control phenotypes required for prehensile hands and feet and the arboreal lifestyle of primates. To test this hypothesis, we sought to identify genomic segments that evinced statistically significant evolutionary constraint among anthropoid primates, but little or no constraint among non-primate mammals. We scanned for primate conserved elements (human, orangutan, rhesus macaque, marmoset) using Gumby with parameters R-ratio=2 and 5 in the 800-kb *FOXP2* locus. To obtain conserved elements specific to primates, we contrasted these primate-conserved sequences with elements conserved in 24 mammals, ascertained by running Gumby with the same parameters on the 24-way non-primate mammal Multiz alignment. At a Pvalue threshold of 0.001, we identified 134 primate-conserved elements in the locus. Using a loose Pvalue threshold of 0.5 in order to rule out any possibility of mammalian conservation, we identified 836 mammal-conserved elements covering 16.1% of the 800-kb. Surprisingly, in spite of these stringent criteria, we identified 5 primate-specific conserved elements in the locus (all noncoding), the strongest of which had a Pvalue of 5.0e-13. This result suggests that multiple noncoding elements in the *FOXP2* locus potentially evolved primate-specific gain of function and provides the basis for the experimental exploration of gene expression changes that may have contributed to primate-specific motor phenotypes.

THE GENOME ANALYSIS TOOLKIT: A MAPREDUCE FRAMEWORK FOR ANALYZING NEXT-GENERATION DNA SEQUENCING DATA

A McKenna[1], M Hanna[1], E Banks[1], A Sivachenko[1], K Cibulskis[1], A Kernytsky[1], K Garimella[1], D Altshuler[1,2], S Gabriel[1], M A DePristo[1]

[1]The Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Five Cambridge Center, Cambridge, MA, 02142, [2]Massachusetts General Hospital, Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, MA, 02114

Next-generation DNA sequencing (NGS) projects, such as the 1000 Genomes Project, are already revolutionizing our understanding of genetic variation among individuals. However, the massive data sets generated by NGS—the 1000 Genome pilot alone includes nearly five terabases—make writing feature-rich, efficient and robust analysis tools difficult for even computationally sophisticated individuals. Indeed, many professionals are limited in the scope and the ease with which they can answer scientific questions by the complexity of accessing and manipulating the data produced by these machines. Here we discuss our Genome Analysis Toolkit (GATK), a structured programming framework designed to ease the development of efficient and robust analysis tools for next-generation DNA sequencers using the functional programming philosophy of MapReduce. The GATK provides a small but rich set of data access patterns that encompass the majority of analysis tool needs. Separating specific analysis calculations from common data management infrastructure enables us to optimize the GATK framework for correctness, stability, CPU and memory efficiency, and to enable distributed and shared memory parallelization. We highlight the capabilities of the GATK by describing the implementation and application of robust, scale-tolerant tools like coverage calculators and SNP calling. We conclude that the GATK programming framework enables developers and analysts to quickly and easily write efficient and robust NGS tools, many of which have already been incorporated into large-scale sequencing projects like the 1000 Genomes Project and The Cancer Genome Atlas.

DE NOVO SEQUENCING AND ASSEMBLY OF THE CUCUMBER GENOME USING EXCLUSIVELY NEXT-GENERATION SEQUENCING METHODS

Brian Desany[1], Jason Affourtit[1], Pascal Bouffard[1], Timothy Harkins[2], James Knight[1], Chinnappa Kodira[1], Jason Miller[3], Therese Mitros[4], Mohammed Mohiuddun[1], Daniel Rokhsar[4,5], Granger Sutton[3], Cynthia Turcotte[1], Yiqun Weng[6], Jack Staub[6]

[1]454 Life Sciences, R&D, 20 Commercial St., Branford, CT, 06405, [2]Roche Applied Science, Marketing, 9115 Hague Rd., Indianapolis, IN, 46250, [3]J. Craig Venter Institute, Informatics, 9704 Medical Center Dr., Rockville, MD, 20850, [4] University of California, Center for Integrative Genomics, 142 LSA #3200, Berkeley, CA, 94720, [5]Department of Energy, Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA, 04598, [6]USDA, ARS, University of Wisconsin, Vegetable Crops Research Unit, 1575 Linden Dr., Madison, WI, 53706

De novo reconstruction of complex genomes by random shotgun methods is made difficult by the presence of genomic repeats. However, if individual sequence reads are long enough, they can span isolated repetitive sequences and reduce the magnitude of the problem. For this reason, sequencing of complex genomes longer than 100 Mb has to date been limited to Sanger methods or combinations of Sanger and next-generation methods. Here we describe the de novo sequencing, assembly, and annotation of the cucumber genome (Cucumis sativus) exclusively using the next-generation FLX Titanium technology from 454 Life Sciences. A high quality draft was generated using shotgun reads along with paired-end reads with span distances of 3 kb and 20 kb to allow scaffolding over longer repeats. Near-complete assembly of the euchromatic portion of the genome is demonstrated by the extremely high mappability (>98%) of over 2 million ESTs, also generated using FLX Titanium. Comparable results were obtained between Celera Assembler and gsAssembler (newbler). The ease of generating a near-complete reference genome reconstruction has the potential to be very useful for ongoing crop improvement breeding programs for cucumber and other cucurbits with similar genomes.

NATURAL GENETIC VARIATION CAUSED BY ENDOGENOUS HUMAN RETROTRANSPOSONS

Rebecca C Iskow[1,2], Micheal T McCabe[3], Ryan E Mills[1], Spencer Torene[1], Erwin G Van Meir[3], Paula M Vertino[3], Scott E Devine[1,2,3,4]

[1]Emory University School of Medicine, Biochemsitry, 1510 Clifton Rd, Atlanta, GA, 30329, [2]Emory University School of Medicine, Graduate Program in Genetics and Molecular Biology, 1510 Clifton Rd., Atlanta, GA, 30329, [3]Emory University, Winship Cancer Institute, 1400 Clifton Rd, Atlanta, GA, 30329, [4] University of Maryland School of Medicine, Institute for Genome Sciences, Dept. of Medicine, Greenebaum Cancer Center, 801 W. Baltimore Street, Baltimore, MD, 21201

Two abundant classes of mobile elements, namely Alu and L1 elements, continue to generate new retrotransposon insertions in human genomes. In fact, some estimates suggest that these elements have generated millions of new germline insertions in personal human genomes worldwide. Unfortunately, current technologies are not capable of detecting most of these young insertions, and the true extent of germline mutagenesis by endogenous human retrotransposons has been difficult to examine. Here, we describe new technologies for detecting these young retrotransposon insertions and demonstrate that such insertions indeed are abundant in human populations. We also found that new somatic L1 insertions occur at high frequencies in human lung cancer genomes. Genome-wide methylation analysis suggests that altered DNA methylation may be responsible for the high levels of L1 mobilization observed in these tumors. Overall, our data indicate that transposon-mediated mutagenesis is extensive in human genomes, and is likely to have a major impact on human biology and diseases.

THE GENOME OF THE ANOLIS LIZARD: A REPTILE IN A MAMMALIAN WORLD

Federica Di Palma[1], Jessica E Alfoldi[1], Manfred Grabherr[1], Lesheng Kong[2], Andreas Heager[2], Craig Lowe[3], Anolis Genome Sequencing Consortium[1], David Haussler[3], Chris Ponting[2], Kerstin Lindblad-Toh[1]

[1]Broad Institute of MIT and Harvard, Genome Biology, 5 Canbridge Center, Cambridge, MA, 02142, [2]MRC Functional Genomic Unit, University of Oxford, South Park Road, Oxford, OX1 3QX, United Kingdom, [3]Center for Biomolecular Science & Engineering, UCSC, 1156 High St., Santa Cruz, CA, 95064

Anolis carolinensis, the green anole lizard, is the first non-avian reptile to have its genome sequenced. Now that two avian genomes have also been sequenced (chicken and zebra finch), we have a great opportunity to understand both reptilian and mammalian evolution via synteny and gene-based comparisons of these sequenced genomes. The 250 million-years between *Anolis* and birds, and the 310 million-years between *Anolis* and mammals add an extra challenge to our analysis, for which we have developed new sensitive genome comparison tools. Especially interesting is the comparison of bird and lizard microchromosomes. Even though lizard microchromosomes do not share the distinguishing characteristics of bird microchromosomes, their syntenic relationships have revealed a great deal about reptilian genome evolution and the nature of these remarkable tiny chromosomes.

Here, we present our analysis of the genome of *Anolis carolinensis*, including our studies of microchromosomes, gene orthology and the amniote evolution of the immunologically-significant MHC region. We highlight several areas in which the *Anolis* genome structure is more reminiscent of mammalian genome structure than that of the more compact avian genomes. Finally, we also thoroughly examine the multitude of *Anolis* repeats that have been exapted as conserved non-coding and sometimes as exons in mammalian genomes.

THE CICHLID MODEL SYSTEM: UNDERSTANDING HOW SMALL RNAS REGULATE PLASTIC AND REVERSIBLE CHANGES IN SOCIAL BEHAVIOR.

Rosa Alcazar[1], Karen Maruska[1], Federica Di Palma[2], Kerstin Lindblad-Toh[2], Poornima Parameswaran[3], Russell D Fernald[1]

[1]Stanford University, Department of Biology, 450 Serra Mall, Stanford, CA, 94305, [2]Broad Institute of MIT and Harvard, Genome Biology, 5 Cambridge Center, Cambridge, MA, 02142, [3]Stanford University, School of Medicine, 300 Pasteur Dr, Stanford, CA, 94305

Natural populations of *Astatotilapia burtoni* live in the river estuary system of Lake Tanganyika in east Africa. In this lake-like social system, only a few males in a given population have access to critical resources of food, spawning sites, and consequently, females. Males have evolved two distinct, reversible phenotypes in adapting to their dynamic social environment: dominant, reproductively competent territorial (T) males and submissive, reproductively suppressed non-territorial (NT) males. T males display bright coloration, aggressively defend territories, and court females, while NT males display dull coloration, mimic females, and limit their behavior to schooling and fleeing.

The transition between social states regulates multiple changes in the brain-pituitary-gonadal (BPG) axis including dramatic changes in GnRH-containing cell size as well as testis size and cell composition, but little is known about how spermatogenic plasticity is controlled in adult vertebrates. In collaboration with the Broad Institute, as part of the ongoing cichlid genome sequencing efforts, we have sequenced small RNA libraries (Illumina) from *A. burtoni* testes of both aggressive territorial and submissive non-territorial males. Here we identify and characterize differentially expressed small RNAs which may modulate changes in spermatogenic activity during social transition.

THE PROMOTER OF THE *IGF2* IMPRINTED GENE IN THE OPOSSUM, *MONODELPHIS DOMESTICA*, SIMULTANEOUSLY EXHIBITS MUTUALLY EXCLUSIVE HISTONE MODIFICATIONS

Kory C Douglas, Paul B Samollow

Texas A&M University, College of Veterinary Medicine, Department of Veterinary Integrative Biosciences, VMA Building Room 107, College Station, TX, 77843

Genomic imprinting, or parent-of-origin-specific gene expression, is an epigenetic phenomenon that occurs in therian mammals, flowering plants, and arguably in some insects, but has not been found in prototherian mammals or other vertebrates. Its failure in humans can result in developmental diseases including Beckwith-Wiedemann syndrome, Prader-Willi/Angelman syndrome, and others. Analyses of fundamental epigenetic signals of imprinting in mammals, such as differential DNA methylation, histone modification, and the presence of certain characteristic genomic elements have been limited primarily to a small number of eutherian (placental) mammals and have not been conducted extensively in non-eutherian species. To further elucidate the evolution and molecular mechanisms of genomic imprinting, it is also vital to examine this phenomenon in metatherian (marsupial) mammals, the only non-eutherian vertebrates in which imprinting is known to occur.
It has been shown that certain post-translational modifications of histone proteins differentially mark the promoter regions of paternally and maternally derived copies of imprinted genes in eutherians and are correlated with active and inactive transcriptional states. More specifically, histone 3 lysine 4 trimethylation (H3K4me3) and histone 3 lysine 9 trimethylation (H3K9me3) have been shown to be mutually exclusive marks for the active and inactive states, respectively, at imprinted gene loci. However, the enrichment of these marks at imprinted genes in metatherians has not been characterized. Utilizing chromatin immunoprecipitation (ChIP), we show the localization and concurrence of H3K4me3 and H3K9me3 epigenetic marks at the promoter of the *IGF2* imprinted gene in Monodelphis domestica. We also analyze the correlation of specific genomic elements with these modifications. This study is the first to show that specific histone modifications mark an imprinted gene in a metatherian mammal and furnishes the basis for designing genome-wide scans for undiscovered, as well as marsupial-specific, imprinted genes.

LARGE-SCALE HUMAN GENOME SEQUENCING AND HAPLOTYPING FOR ADVANCED DISEASE STUDIES

Rade Drmanac

Complete Genomic, Inc., Research, 2071 Stierkin Court, Mountain View, CA, 94043

We have developed a novel DNA nanoarray-based complete human genome sequencing technology providing high accuracy (99.999%) at a cost of less than $1500 in reagents-- Science 327, 78 (2010). In 2009 we sequenced over 50 high quality (40x) human genomes using this technology. Multiple disease genes and many novel cancer mutations were discovered in these genomes, including rare disease variants from a four-member family. Based on further nanoarray and instrument improvements that allow for an average throughput of over 1,000 Gb per run, resulting in 1 genome per instrument-day, we are building a human genome sequencing center with the capacity to analyze thousands of genomes per year. We have also successfully sequenced multiple human genomes using only 10-20 cells (66-120 pg of DNA) by employing a novel sample preparation process allowing separate sequence assembly for each parental chromosome. The Mb-sized haplotype contigs obtained were confirmed by sequencing personal BACs and parental genomes, and compared to HapMap data. Over two million heterozygote SNPs (several times more than found in the HapMap) were haplotyped in the NA19240 genome. Our large-scale human genome sequencing center and these new genome analysis capabilities will allow for advanced molecular studies of both common and rare human diseases by complete sequencing and haplotyping of a large number of patient and control genomes.

INTEGRATION OF LARGE-SCALE FUNCTIONAL GENOMICS DATASETS IN ENCODE AND ENSEMBL

Ian Dunham, Nathan Johnson, Damian Keefe, Daniel Sobral, Steven Wilder, Ewan Birney

EMBL-EBI, PANDA, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Ultra high throughput sequencing methods have been rapidly adopted for analysis of the chromatin state and transcriptome of cells. A number of large- scale initiatives are now under way to produce systematic profiles of chromatin state in a variety of normal and diseased cells and cell lines. In particular, the ENCODE consortium has generated data by ChIP-seq, DNase-seq, RNA-seq, methyl-seq and other high-throughput methods for a broad range of representative human cell lines. Key tasks in maximising the utility of these data include methods for handling large and diverse datasets, strategies for data integration and analysis, and developing appropriate mechanisms for data presentation. We will present the approaches that we have adopted within the ENCODE consortium and in Ensembl.

As part of the ENCODE Data Analysis Centre, we have been involved in both data summarising and integration. Following an ENCODE data freeze in January, we have analysed the status of ENCODE data production as part of a progress review process. In particular, we have developed approaches for assessing the status towards completion of the project with respect to its stated goals. This has included summarising ENCODE productivity, analyses of the diversity of ENCODE cell lines within transcriptome space, and analysis of coverage of the universe of regulatory elements, either within an individual cell line or across all cell lines, to estimate "saturation".

Within the context of the Ensembl Functional Genomics database, we are developing methods to provide access to functional genomics data across multiple species. This has been initiated with incorporation of key datasets form both large-scale consortium projects and smaller laboratory based analyses. We are aiming to provide access to these data, both in their raw form with innovative data visualisations, but also by novel processing that aims to present a unified view of chromatin state by cell type. One approach is the Regulatory Build that provides a set of functional elements across the genome and categorises these elements by their chromatin marks and transcription factor occupancy, either by cell type or as a union set.

THE ROLE OF GENOMIC METHYLATION ABNORMALITIES IN BREAST CANCER

Jeffrey F Hiken[1], Anne H O'Donnell[2], Timothy H Bestor[2], John R Edwards[1]

[1]Washington University School of Medicine, Center for Pharmacogenomics, 660 S. Euclid Ave., Box 8220, St. Louis, MO, 63110, [2]Columbia University, Department of Genetics and Development, 701 W. 168th St., HHSC 1602, New York, NY, 10032

One of the earliest epigenetic abnormalities observed in cancer was extensive genomic hypomethylation. This phenomenon has been implicated in genomic instability and possibly the reactivation of oncogenes; however, studying hypomethylation has been difficult since many of these changes are concentrated in interspersed repeats. The agouti and axin loci in mouse provide clear evidence that demethylation of retrotransposons can lead to ectopic expression of nearby genes, but a systematic screen for such events in cancer genomes has not been performed. To better understand the consequences of wide spread demethylation in repeat regions, we have employed Methyl-MAPS (Methylation Mapping Analysis by Paired-end Sequencing). In this method, genomic DNA is digested with methylation-sensitive and –dependent enzymes, and Next-Gen sequencing is used to map the locations of digested and undigested CpG sites and, thus, map their methylation state. This technique detects methylation status at not only promoter and other single copy regions of the genome, but also at interspersed repeats. A new computational pipeline has been developed to quickly detect differentially methylated features genome-wide. Using this method, we have assessed methylation status of nearly the entire genome in several breast cancers and adjacent normal breast tissues and ER+ and ER- breast cancer cell lines. Combining whole-transcriptome and genome-wide methylation data is allowing us to better understand how hypomethylation events can lead to transcriptional abnormalities in cancer. In particular, we are beginning to understand the effects of repeat hypomethylation on transcription.

GENOME-WIDE ASSOCIATION STUDY FOR MYOCARDIAL INFARCTION IN SOUTH ASIANS: THE INTERHEART STUDY

James C Engert[1], Ron Do[1], Changchun Xie[2], Alexandre Montpetit[1,3], Sonia S Anand[2]

[1]McGill University, Human Genetics, Royal Victoria Hosp., Montreal, H3A 1A1, Canada, [2]McMaster University, Population Health Research Institute, 237 Barton St. E., Hamilton, L8L 2X2, Canada, [3]McGill University, McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield, Montreal, H3A 1A4, Canada

Genome-wide association studies (GWAS) have identified genetic variants that influence the risk of coronary heart disease (CHD) or myocardial infarction (MI). However, most of these studies have been performed in European populations. While some of the discovered variants may be universal in their effect, we believe that others will have different effects in different ethnicities. This can be due to differing gene-gene and gene-environment interactions. In the present study, we performed a GWAS in South Asians from the INTERHEART study. In order to increase the chances to identify novel genetic regions, we compared early onset MI with older controls. We believe that this may also increase the chances for identifying less common variants. In order to assess the association of less common variants we used extended haplotype sharing comparisons as implemented in the software plink to identify regions and loci that contain multiple independent haplotypes that are shared identical by descent (IBD) between unrelated individuals. A single locus with multiple rare variants can be identified by demonstrating increased sharing among case pairs as compared to control pairs or case/control pairs. We have assessed and prioritized regions identified with this method by looking for overlapping signals from homozygosity mapping as well as Hardy Weinberg disequilibrium. We note one region in particular on chromosome 8p22 that showed the greatest signal of association. This region has been previously identified in a genome-wide linkage study of heart disease in the Quebec population.

UBIQUITOUS MIRNA VARIANTS (ISOMIRS) IN CONTROL AND HUNTINGTON'S DISEASE BRAIN REGIONS DETECTED BY MASSIVELY PARALLEL SEQUENCING

Eulalia Martí[1], Monica Bañez-Coronel[1], Lorena Pantano[1], Franc Llorens[2], Isidre Ferrer[3], Xavier Estivill[1,4]

[1]Centre for Genomic Regulation (CRG) and CIBERESP, Genes and Disease program, Dr. Aiguader, 88, Barcelona, 08003, Spain, [2]University of Barcelona and Institute of Bioengineering of Catalonia and CIBERNED, Department of Cell Biology, Josep Samitier, 1-5, Barcelona, 08028, Spain, [3]IDIBELL-Hospital Universitari de Bellvitge, Universitat de Barcelona, Feixa llarga, s/n, Barcelona, 08907, Spain, [4] Pompeu Fabra University, Experimental and Health Sciences department, Dr. Aiguader, 88, Barcelona, 08003, Spain

Huntington disease (HD) is a neurodegenerative disorder that predominantly affects neurons of the forebrain. We have applied the Illumina massive parallel sequencing to deeply analyze the small RNA populations of two different forebrain areas, the frontal cortex (FC) and the striatum (ST) of healthy individuals and individuals with HD. More than 80% of the small-RNAs were annotated as miRNAs in all samples. Deep sequencing revealed length and sequence heterogeneity (IsomiRs) for the vast majority of miRNAs. Around 80-90% of the miRNAs presented modifications in the 3'-terminus mainly in the form of trimming and/or as nucleotide addition variants, while the 5'-terminus of the miRNAs was specially protected from changes. Expression profiling showed strong miRNA and isomiR expression deregulation in HD, most being common to both FC and ST. The putative targets of the seed-region of deregulated miRNAs/isomiRs strongly suggest that their altered expression contribute to the aberrant gene expression of HD. Many REST (RE1-Silencing Transcription Factor) modulated miRNAs were downregulated in HD, suggesting that repressed REST target miRNAs play a role in aberrant gene expression in HD. Our results show that miRNA variability is a ubiquitous phenomenon in the adult human brain, which may influence the mechanism of gene expression modulation in physiological and pathological conditions. The charcaterization of different brain regions in other neurodegenerative disorders should uncover a myriad of IsomiRs involved in regulation of brain genes.

IDENTIFICATION OF AN IMPORTANT EGRESS DEFECT PHENOTYPE IN THE EUKARYOTIC PARASITE *TOXOPLASMA GONDII* USING WHOLE-GENOME MUTATIONAL PROFILING

Andrew Farrell, Keith Eidell, Marc-Jan Gubbels, Gabor Marth

Boston College, Biology, 140 Commonwealth Ave, Chestnut Hill, MA, 02467

The most effective way to identify genes involved in a specific biological process is to create mutant strains that express the desired phenotype and identify the disabling mutation. Traditionally, this is accomplished using molecular techniques (e.g. complementation) or by linkage mapping. Such methods can be extremely time consuming, expensive, and are often not successful. High-throughput sequencing now provides a promising alternative: relatively inexpensive, complete re-sequencing of the entire mutant genome can quickly reveal the genes in which the mutations occurred.

We have successfully applied this approach to the *Toxoplasma gondii* mutant strain F-P2, which exhibits a temperature-sensitive egress defect phenotype. The F-P2 mutant was originally isolated in 2001 and traditional attempts to identify the mutation since then have failed. We shotgun sequenced both the parent and the F-P2 mutant using Illumina sequencers, at the Broad Institute, with 75 bp paired end reads. We aligned the reads to the reference using the Mosaik aligner. Our alignment covered 97% of the genome at 5x coverage or greater, with an average coverage of 48x in the F-P2 mutant and 38x in the parent. Using our Bayesian SNP caller program, GigaBayes, we identified 35 SNPs. Using Sanger sequencing to verify the SNPs 25 have confirmed, 4 did not, and 6 are still in the processes of being sequenced. The SNPs were prioritized by possible biological significance and attempts to complement the mutation using wild type cosmids were begun. On our first attempt we identified a gene encoding a C2 domain-containing protein that successfully complemented the mutation, confirming that this gene is responsible for the phenotype.

Our study demonstrates the power of whole-genome sequencing for the identification of phenotypically important mutations that cannot be isolated with traditional genetic methods. Our pipeline is readily applicable for whole-genome mutational profiling for a range of genetically inaccessible organisms including many human parasites.

SNP IDENTIFICATION AND VALIDATION IN RHESUS MACAQUE USING NEXT-GEN SEQUENCING

Gloria L Fawcett[1,2], Muthuswamy Raveendran[1,2], David Rio Deiros[1,2], David Chen[1,2], Jeff G Reid[1,2], Donna M Muzny[1,2], David A Wheeler[1,2], Kim C Worley[1,2], Ronald A Harris[2], Aleksander Milosavljevic[2], Richard A Gibbs[1,2], Jeff A Rogers[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, TX, 77030

Rhesus macaques (*Macaca mulatta*) are the most widely used non-human primate in biomedical research. While the existing rhesus genome assembly is remarkably useful, the identification and validation of polymorphisms within this species, especially SNPs, will add significantly to its utility. Furthermore, it is advantageous to develop efficient strategies for identifying SNPs in draft genomes of additional species. More than 4.3 million heterozygous nucleotide positions were identified as potential polymorphisms in the original reference assembly of the rhesus genome. We validated rhesus SNPs by re-sequencing the original reference animal and additional unrelated individuals using Next-Gen methods. Two strategies were used to bioinformatically validate SNPs: comparison of potential SNPs found in the same individual using two different sequencing chemistries and comparison of potential SNPs in different individuals regardless of chemistry. We validated ~1.3 million SNPs by comparing the assembly reference sequence (Sanger data) either to SNPs called by corona_lite from SOLiD re-sequencing of the reference animal or to the raw SOLiD reads analyzed using e-genotype (http://e-genotype.com/). We validated ~0.7 million additional SNPs by comparing SOLiD data from unrelated animals to the Sanger and SOLiD reference data. Approximately 500,000 novel polymorphisms not observed in the reference animal were validated by comparing the SOLiD data from the two resequenced non-reference animals. Altogether, we validated about 2.5 million SNPs in this genome, whereas previous work had validated fewer than 800 SNPs for this species. These data will be valuable for genetic analyses of complex phenotypes, population genetics, and management of captive breeding colonies and are being made available through appropriate on-line databases. Future work will investigate other forms of genomic variation in rhesus macaques, and apply this SNP identification methodology to new on-human primate sequences (e.g. baboons, *Papio hamadryas*).

COMPARATIVE EPIGENOMICS – TOWARDS ANCESTRAL GERMLINE METHYLOME RECONSTRUCTION

Lars Feuerbach[1], Rune B Lyngso[2], Thomas Lengauer[1], Jotun Hein[2]

[1]Max Planck Institute für Informatik, Computational Biology and Applied Algorithmics, Campus E1.4, Saarbrücken, 66123, Germany, [2]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom

One of the key objectives of comparative genomics is the characterization of the forces that shape genomes over the course of evolution. In the last decades evidence has been accumulated that for vertebrate genomes also epigenetic modifications have to be considered in this context. Especially, the elevated mutation frequency of 5-methylcytosine (5mC) is assumed to facilitate the depletion of CpG dinucleotides in species that exhibit global DNA methylation. For instance, the underrepresentation of CpG dinucleotides in many mammalian genomes is attributed to this effect, which is only neutralized in so called CpG islands that are preferentially unmethylated and thus partially protected from rapid CpG decay.
It has been shown that CpG decay is 10-fold stronger than any other genomic point substitution processes. Still, the number of evolutionary models that consider the methylation state of CpG dinucleotides is limited and substitution rates of 5mCpG and CpG dinucleotides are usually convoluted in the models.
To bridge this gap, we here propose a comparative epigenomics approach that applies Markov models to the prediction of the germline methylation state of homologous genome regions from their multiple alignments. In a simulation study, we show that under the assumption of bimodal methylation states the model produces reliable predictions for homologous regions with divergence times up to 140 mya. Preliminary results on genomic DNA indicate that the model is applicable to comparative studies of closely related mammalian genomes.

Key words: DNA methylation, CpG islands, comparative epigenomics, CpG depletion, germline, mammalian genome evolution, deamination

EXPLORING THE CONFINEMENT OF DSDNA IN BACTERIOPHAGE VIA MOLECULAR SIMULATION.

Gordon S Freeman[1], David C Schwartz[2], Juan J de Pablo[1]

[1]UW-Madison, Chemical and Biological Engineering, Madison, WI, 53706, [2]UW-Madison, Department of Chemistry, Laboratory for Molecular and Computational Genomics, Madison, WI, 53706

Confinement of genomic sequences plays a crucial role in biological processes. However, this presents a conundrum: dsDNA is a semiflexible polyelectrolyte with a persistence length on the same order of magnitude as the dimensions into which it is routinely confined. The result is a 1000-fold reduction in the volume occupied by the genome within the capsid relative to that of the unconfined genome. Bacteriophage have been shown to package their genome to such extremely high densities (as high as 55 volume percent in many cases) through an active packaging "motor" that translocates dsDNA into the viral capsid. However, the mechanism by which the motor accomplishes this translocation has not been elucidated. Additionally, the conformations dsDNA takes on within the viral capsid are poorly understood at best. The present work employs a coarse-grained model for DNA (the so-called "Three Sites Per Nucleotide" Model) and molecular simulation techniques to probe the role played by the motor in the confinement of phage dsDNA. Specifically, the role that twisting of the genome during active packaging plays in the confinement process is interrogated over a range of twisting rates. Additionally, we use our model to study the response of dsDNA to the high curvatures and associated stresses encountered within the viral capsid. Local dehybridization is observed in regions with the highest curvature leading to a reduction in the local stresses encountered in those regions thus providing a viable mechanism for the extreme confinement of dsDNA observed in bacteriophage.

THE VERVET SYSTEMS BIOLOGY PROJECT

Nelson B Freimer

UCLA, Center for Neurobehavioral Genetics, 695 Charles Young Drive S, LA, CA, 90095

Non human primates (NHPs) are ideal animal models for systems biology, where the distinctions between human and rodent biology are particularly apparent. Systems approaches in humans – in contrast to NHPs – are limited by the difficulties of large-scale longitudinal investigations of multiple phenotypes and of collecting various tissues. The development of reagents for genome-level NHP investigations is now removing the major obstacle to NHP systems biology. I will discuss progress in establishing the vervet monkey as a model for genomic investigation, as well as planned projects of the newly formed Vervet Systems Biology Consortium (VSBC).

Vervets, the most abundant natural hosts of Simian Immunodeficiency Virus (SIV), range throughout Sub-Saharan Africa, and their unusual population history makes them ideal for genetic investigations. Europeans brought small numbers of West African vervets to three Caribbean islands in the 17^{th}-18^{th} Century, where feral populations are now estimated to include ~10^5 monkeys. The Vervet Research Colony (VRC) was initiated in the 1970s with a few dozen Caribbean vervets, has since been maintained as a single extended pedigree, and now includes ~ 500 members. The development of a first generation vervet genetic map has stimulated genetic investigation of neurobehavioral and metabolic quantitative traits, and the establishment of the VRC Tissue Repository has facilitated large scale gene expression profiling studies, including pedigree-wide genetic mapping of expression quantitative trait loci (eQTL). Among the samples collected by the Repository are brain and peripheral tissues representing multiple developmental time points between birth and adulthood.

Several coordinated initiatives are now increasing the scale of genome-level investigations in vervets. The vervet genome sequencing project is now generating a genome assembly and identifying genetic variants which can be assayed for genome-wide mapping of vervet phenotypes. To this end, NCRR has funded the collection of >2500 phenotyped samples from wild vervet populations in the Caribbean and Africa. These samples may be particularly informative for identifying host variants that protect against SIV-disease. Systems biology programs face the challenge of assembling diverse expertise and coordinating the multiple phenotyping and genomics efforts that provide their data. I will discuss our attempt to address this challenge in the VSBC.

SCREENING OF COMPLEMENTING LONG-INSERT CLONES AND SEQUENCES TO THE HUMAN REFERENCE SEQUENCE THROUGH WHOLE-GENOME OR WHOLE-CHROMOSOME APPROACH

Asao Fujiyama[1,2], Yoko Kuroki[3], Shinji Kondo[3], Yuichiro Nishida[3], Atsushi Toyoda[1]

[1]National Institute of Genetics, Copmparative Genomics, 1111 Yata, Mishima, 411, Japan, [2]National Institute of Informatics, Bioinformatics, 2-1-2 Hitotsubashi, Chiyodaku, Tokyo, 101, Japan, [3]RIKEN, Advanced Computational Sciences, 2-1 Hirosawa, Wako, 351, Japan

Since the completion of the initial finished human genome sequencing, we have been trying to fill-in existing gaps or unmapped regions in the human genome reference sequence. Our major target chromosomes are Chr11, Chr18, Chr19, Chr21 and ChrY those we contributed in sequencing/finishing or have special biological interest. During the Human Genome Project, we constructed chromosome-enriched sheared fosmid libraries to each or set of chromosomes those isolated through chromosome sorting technology. In addition to screening of those libraries using traditional chromosome walking, we are also using short-read shotgun sequences obtained from whole genome or isolated chromosomes to design primers to fish out candidate clones that might fill or extend those gaps. The latter screening pipeline includes purification and mapping of IlluminaGAIIx reads onto the reference sequence; assembly then mapping onto other human genomes, characterization of hits against known annotations, screening of BAC and/or fosmid libraries if possible to isolate candidate clones for the gaps.
Although our contribution might be small, we believe that continuing effort to improve the human genome reference sequence is very important and we wish to add genomic information and resources especially on sub-telomeric regions and un-cloned short arm regions of our target chromosomes. The progress of our effort will be presented.

VARIANT VALIDATION AND SCREENING AT THE GENOME CENTER AT WASHINGTON UNIVERSITY SCHOOL OF MEDICINE

Robert S Fulton[1], Li Ding[1], Vincent Magrini[1], Michael D McLellan[1], Daniel Koboldt[1], Heather Schmidt[1], Michelle O'Laughlin[1], Rachel M Abbott[1], Timothy J Ley[2], Elaine R Mardis[1], Richard K Wilson[1]

[1]Washington University School of Medicine, The Genome Center, 4444 Forest Park Avenue, St. Louis, MO, 63108, [2]Washington University School of Medicine, Department of Medicine, Division of Oncology, 660 South Euclid Avenue, St. Louis, MO, 63110

The emergence of next generation sequencing technologies has resulted in an explosion of data production capabilities and the capability to study entire human genomes on a daily basis. With this massive sequencing capacity expansion, the ability to call, validate, and screen additional samples for putative variants is essential. Although variant detection methods are ever improving, with greater sensitivity and specificity, the investment in generating the primary data, as well as the need to understand each genome's exact mutation spectrum, means that validation of putative variants is still a necessary component. The Genome Center at Washington University has developed a robust, multi-platform, cost-effective validation platform, with the capability of validating variants at a comparable pace as the raw data production capabilities. The validation process is capable of leveraging Illumina, Roche 454, and ABI 3730 sequencing instruments. Template targeting is provided by either PCR or hybrid-capture. Both platform selection and template generation methods are determined based on scale, timeline, and data output requirements. The composition of the variant lists can include single nucleotide variants (SNV), small (< 20-30bp) insertion or deletion events, as well as structural variations including large (> 20-30bp), insertions or deletions, translocations, and inversions. Validation can be performed on single samples, panels of individuals that maintain individual genotypes, or pools of individuals providing pooled genotypes. Pooling strategies and sample barcodes are particularly useful when working with sample pools or panels, although they also can be applied for small sample sets such as tumor and normal samples. By enabling sample specificity using barcodes, more efficient use of the immense sequence throughput per run can be realized. In addition to current technologies, new methods and instrumentation are in development with new validation and screening procedures coming soon. Without these developments and continued confirmation of the detection methods, improvements and significant findings will not likely be produced as quickly, accurately, or effectively. This presentation will highlight many of the current methods and ongoing work to improve these processes.

GENOME SEQUENCING OF STRAINS OF *SALMONELLA* TYPHIMURIUM IN HONG KONG

Yinwan Wendy Fung[1], Tik Wan Patrick Law[1], Chun Hang Au[1], Kai Man Kam[2], Hoi Shan Kwan[1]

[1]Chinese University of Hong Kong, Faculty of Science, CUHK, Shatin, Hong Kong, [2]Center for Health Protection, Department of Health, 382 Nam Cheong Street, Kowloon, Hong Kong

In Hong Kong, foodborne infectious disease is a common public health issue and cases of foodborne disease outbreaks due to bacterial causative agents are frequent. There are 500 to 800 cases of food poisoning reported annually due to bacterial infection, with thousands of people affected. Therefore it is essential to have an effective surveillance system to screen for foodborne pathogens and to ensure safe food consumption. *Salmonella enterica* Typhimurium (ST) is a common causative agent of food poisoning in Hong Kong. Although symptoms (vomiting, diarrhea, abdominal pain, occasional fever) are self-limiting, death could occur from serious complications such as dehydration and septicemia. Traditional routine tests currently used for detecting ST take at least two days and rapid detection methods such as PCR- and microarray-based methods require prior knowledge of the ST genome sequence. Therefore this study aims at obtaining whole genome sequence information to identify ST isolates found locally in Hong Kong. The whole genome sequences from local isolates will provide tools for rapid and accurate identification of pathogens and allow the development of markers for identification. The study will also assist the epidemiological investigations of ST strains relevant to Hong Kong. The genome sequencing of a few Hong Kong ST isolates has been completed using the 454 pyrosequencing technology. The draft sequence will be used as a blueprint for the development of markers for rapid identification. The long-term aim is to use the genome sequence information to monitor the trend of foodborne diseases and to ensure that foods available in the community are safe for human consumption.

PREDICTION OF REGULATORY SNPS IN THE HAPMAP LCLS BY INTEGRATING INFORMATION FROM MULTIPLE EXPERIMENTAL SOURCES.

Daniel J Gaffney, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, Jonathan Pritchard

University of Chicago, Human Genetics, 58th Street, Chicago, IL, 60637

Expression QTL mapping is a potentially powerful means of dissecting the genetic basis of gene regulation. The availability of full genotype information suggests that we can now localize the true causal variants which drive differences in gene expression between individuals. However, even with full genotype data it is often difficult to distinguish true from closely linked variants by the strength of association with expression alone. We present an approach which integrates information from multiple different sources, such as DNase accessibility, histone modification state, predicted transcription factor binding sites, ChIP-seq and evolutionary conservation, to identify the most likely regulatory SNP for all human genes simultaneously. Our approach also provides an unbiased assessment of the ability of a large variety of methods to identify gene regulatory regions. By integrating information from a large variety of regulatory annotations, we identify almost 1400 variants with high posterior probability of being the causal regulatory SNP in the HapMap lymphoblastoid cell lines (LCLs). Our results demonstrate that DNase accessibility, some histone marks of gene activation and, to a lesser extent, certain computationally-predicted sites are useful predictors of eQTL location. Interestingly, binding sites for a number of key viral-response transcription factors identified using both ChIP-seq and DNase foot-printing are among the most highly eQTL-enriched regions in the genome, suggesting that challenge with Epstein-Barr virus may be a key factor in determining the landscape of expression in LCLs. We suggest that our approach is a promising step towards dissecting the biology underlying human gene expression variation.

ANTISENSE EXPRESSION INCREASES GENE EXPRESSION VARIABILITY

Zhenyu Xu*, Wu Wei*, Julien Gagneur*, Milosz Smolik, Wolfgang Huber,Lars M Steinmetz[1]

European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstr. 1, Heidelberg, 69117, Germany, *equal contributions

Transcription antisense to coding genes represents the majority of the stable unannotated transcripts in yeast and is prevalent also in humans. However, the genome-wide regulatory effects of antisense expression remain to be elucidated. We have performed a systematic analysis of sense-antisense expression in response to genetic and environmental variations in yeast. Antisense expression appears as an amplifier of gene expression response resulting in larger expression variability between cells in a clonal population, across genetic and environmental variations, and across species. Our results suggest pervasive antisense expression as a possible ancient mechanism for short-term and long-term adaptation by enhancing gene expression response to changes for genes with condition-specific functions.

ASSOCIATION OF *CD226* WITH SLE THROUGH IMPAIRED MRNA PROCESSING IN T CELLS

S E Löfgren*[1], A Delgado-Vega*[1], C J Gallant[1], E Sánchez[2], J Frostegård[3], L Truedsson[4], S D'Alfonso[5], B A Pons-Estel[6], T Witte[7], B Lauwerys[8], E Endreffy[9], L Kovacs[9], C Vasconcelos[10], J Martin[2], M E Alarcón-Riquelme[1], S V Kozyrev[1]

[1]Uppsala U, Dept. Genetics & Pathology, Uppsala, 75185, Sweden, [2]CSIC, Inst. Biomedicina y Parasitología López-Neyra, Granada, 18100, Spain, [3]Karolinska Inst., Dept. Medicine, Stockholm, 14186, Sweden, [4] Lund U, Clinical Microbiology & Immunology, Lund, 22362, Sweden, [5]U of Eastern Piedmont, Dept. Medical Sciences & IRCAD, Novara, 13100, Italy, [6]Sanatorio Parque, Dept. of Rheumatology, Rosario, 2000, Argentina, [7]Hannover Medical School, Immunology & Rheumatology, Hannover, 30625, Germany, [8]U catholique de Louvain, Cliniques Universitaires Saint-Luc, Brussels, B-1200, Belgium, [9]U of Szeged, Dept. PediatricsSzeged, 6720, Hungary, [10]Largo Abel Salazar, Hospital Santo Antonio, Porto, 4099, Portugal

Systemic lupus erythematosus (SLE) is a prototypic autoimmune disease. SLE has a complex polygenic inheritance and susceptibility genes are shared with other autoimmune diseases. A GWAS for type 1 diabetes identified a nsSNP rs763361 in *CD226* and the finding was replicated in other autoimmune diseases. We aimed to test if *CD266* is associated with SLE and to dissect the mechanisms underlying the putative association. Twelve SNPs spanning *CD226* were genotyped in 1194 SLE patients and 1454 controls from Europe. A risk haplotype (P=1.69 x10^{-7}) tagged by the rs727088-C allele and including rs763361 was detected in the 3'-UTR region. It was correlated with decreased levels of *CD226* transcripts and protein levels. Reporter plasmids containing different risk alleles were employed to narrow the risk haplotype. Together, our data strongly support rs727088 as the functional variant responsible for the association with SLE.

GENOMIC IDENTIFICATION OF FUNCTIONAL RNA EDITING SITES AND RECURRENT SINGLE NUCLEOTIDE POLYMORPHISMS

Nandita R Garud[1,2], Jakob S Pedersen[2]

[1]Stanford University School of Medicine, Department of Genetics, 300 Pasteur Drive, Stanford, CA, 94305, [2]University of Copenhagen, The Bioinformatics Centre, Ole Maaloes Vej 5, Copenhagen N, 2200, Denmark

RNA editing, the conversion of one base pair to another, has been found to play an important role in increasing transcriptome diversity. Several forms of editing have been discovered, such as adenosine-to-inosine (A-to-I) and cytosine-to-uracil (C-to-U) editing. Unlike many genomic scans for RNA editing sites, which rely on sequence from a single species, we used a Bayesian graphical model to search for conserved amino acid changing edits in multiple species. The model evaluates the probability of observing a conserved edit site in human, mouse, rat, dog, cow, and chicken given all available mRNA and EST data in GenBank. We did not limit our search to A-to-I and C-to-U candidates, but rather, looked for all possible types of base-pair substitutions in coding sequences.

The top 13 A-to-I edits are all known, classical amino acid changing RNA editing sites. The only known C-to-U amino acid changing editing site ranked fourth in our list of predicted C-to-U sites. Consistent with the classically known cases of A-to-I editing, among the novel candidates there is a Gene Ontology (GO) enrichment for those genes that play a role in the plasma membrane and ion channel formation.

Among our list of 214 predicted editing sites, there are 22 sites that are in the major histocompatability complex (MHC). Indeed, there is also a GO enrichment for the MHC class of genes. Most of the candidates in the MHC are recurrent single nucleotide polymorphisms in human and at least one other species. The MHC is well known for its high level of recurrent polymorphic sites in multiple species. Given the high number of MHC candidates, some of the other sites in our list may potentially be recurrent polymorphisms too.

Identification of new RNA editing sites and novel forms of RNA editing will contribute to our understanding of the mechanisms of post-transcriptional modifications. Furthermore, the discovery of new candidates of recurrent polymorphisms among highly divergent species potentially identifies sites under balancing selection and cases of trans-species polymorphisms.

ANALYSIS OF DIVERSE REGULATORY NETWORKS IN A HIERARCHICAL CONTEXT: CONSISTENT TENDENCIES FOR COLLABORATION IN THE MIDDLE LEVELS

Mark B Gerstein, Nitin Bhardwaj

Yale University, Molecular Biophysics and Biochemistry, 266 Whitney Ave, New Haven, CT, 06511

Gene regulatory networks have been shown to share some common aspects with commonplace social governance structures. Thus, we can get some intuition into their organization by arranging them into well-known hierarchical layouts. Furthermore, these hierarchies can be placed between the extremes of autocratic ones, with well-defined levels and a clear chain of command, and democratic ones, without such well-defined levels and with more co-regulatory partnerships between regulators. In general, the presence of these partnerships decreases the variation in information flow amongst nodes within a level, more evenly distributing 'stress' across the network. Here we study a wide range of regulatory networks (transcriptional, modification and phosphorylation) in a hierarchical context for five evolutionarily diverse species, *E. coli* to Human. We specify three levels of regulators -- top, middle and bottom -- which collectively regulate the non-regulator targets lying in the lowest fourth level, and we define quantities for nodes, levels and entire networks that measure their degree of collaboration and autocratic or democratic character. We show that individual regulators have a range of partnership tendencies: some regulate their target genes in combination with other regulators in local instantiations of a democratic structure whereas others regulate mostly in isolation, in a more autocratic fashion. Overall we show that in all the networks studied, the middle level has the highest collaborative propensity and that co-regulatory partnerships occur most frequently amongst mid-level regulators, an observation that has parallels in efficient corporate settings where middle managers need to interact most to ensure organizational effectiveness. There is, however, one notable difference between networks in different species: the amount of collaborative regulation and democratic character increases markedly with overall genomic complexity.

PTCHD3 IS A NON-ESSENTIAL GENE IN HUMANS; BREAKPOINT MAPPING AND POPULATION FREQUENCY OF A RARE DELETION VARIANT.

Mohammad M Ghahramani seno, Christian R Marshall, Sherylin Bell, Anath Lionel, Stephen W Scherer

Hospital for Sick Children, The Centre for Applied Genomics - Program in Genetics and Genome Biology, 101 College Street, Toronto, M5G 1L7, Canada

In a genomic screen investigating structural variation in Autism Spectrum Disorder (ASD) we detected a heterozygous deletion on chromosome 10p12.1 at a frequency of ~1.4 % (6/427) spanning Patched-domain containing 3 (*PTCHD3*). Screening of another 177 autism probands yielded two additional deletions bringing the total to 8/604 (1.3%) in our ASD cohort. The deletion was found at a frequency of ~0.73% (27/3695) in combined controls from North America and Northern Europe predominately of European ancestry. Screening of the human genome diversity panel (HGDP-CEPH) yielded deletions in 7 (of 1043) unrelated individuals that were confined to those of European/Mediterranean/Middle Eastern ancestry. Breakpoint mapping yielded an identical 102,624 bp (chr10:27,643,753; 27,746,377; NCBI Build 35) deletion in all carriers tested suggesting a single ancestral event. RT-PCR and Northern blot analyses detected *PTCHD3* expression in several tissues, with high level of expression in lymph node, testes and tongue. A novel shorter isoform for *PTCHD3* was also characterized. Expression in transfected COS-7 cells showed PTCHD3 isoforms colocalize with calnexin in the endoplasmic reticulum. The presence of a patched (Ptc) domain suggests a role for PTCHD3 in various biological processes mediated through the Hedgehog (Hh) signaling pathway. However, further investigation yielded one individual with no obvious abnormal phenotype harbouring a homozygous deletion (*PTCHD3* null). Exon sequencing of *PTCHD3* in families with deletions showed compound point mutations also resulting in a null state. Taken together, these data indicate that *PTCHD3* is a non-essential gene in humans.

ANALYZING AND MINIMIZING PCR BIAS AGAINST EXTREME BASE COMPOSITIONS IN ILLUMINA SEQUENCING LIBRARIES

Daniel Aird[1], Wei-Sheng Chen[1,2], Michael G Ross[1], Carsten Russ[1], Sheila Fisher[1], David B Jaffe[1], Chad Nusbaum[1], Andreas Gnirke[1]

[1]Broad Institute, Sequencing, 320 Charles Street, Cambridge, MA, 02141, [2]Rindge and Latin High, School C, 459 Broadway, Cambridge, MA, 02138

Although Illumina shot-gun reads cover most genomes almost completely, sequences with extreme base compositions and idiosyncratic motifs (*e.g.*, runs of G's or C's and AT dinucleotides) are often under-represented or missing. This unevenness in coverage hampers re-sequencing and *de novo* genome sequencing projects alike.

To dissect the laboratory process and find the root cause of biases we developed qPCR assays for several categories of troublesome sequences in the human genome and tracked their abundance relative to well-behaved control sequences throughout the Illumina library preparation protocol. We also developed a panel of 36 qPCR assays for loci ranging from 6% to 90% GC that work well in a pool of three DNA samples of different base composition: *P. falciparum* (19% GC), *E. coli* (51% GC) and *R. sphaeroides* (69% GC).

Following the standard protocol, we saw no significant GC bias up to and including the size-selection. However, GC-rich sequences were dramatically depleted during the subsequent PCR-enrichment when we applied the recommended conditions (Phusion DNA polymerase, 10s denaturation at each cycle). Simply prolonging the denaturation step (to 80s) and/or adding betaine (2M) or using different polymerases (*e.g.*, Accuprime Taq HiFi) significantly improved the representation of GC-rich sequences with little adverse effect on AT-rich loci. We found that PCR-amplified Illumina libraries can be made that display essentially no systematic bias between 15% and 80% GC.

The optimized PCR-amplified libraries are still somewhat more biased than PCR-free library preparations (especially at the very extreme ends of the %GC spectrum). Nonetheless, we chose the PCR protocol as our default process. In our experience, it is more robust and less wasteful than PCR-free protocols and therefore a reasonable choice for high throughput processes handling a wide range of genomic DNA samples that vary in quality and purity and are often limited in quantity.

A ROLE FOR SMALL RNAS IN EPIGENETIC REGULATION DURING STEM CELL DIFFERENTIATION

Loyal A Goff[1,2], Ahmad Khalil[2,3], Mavis Swerdel[4], Jennifer Moore[4], Ronald P Hart[4], John L Rinn[2,3], Manolis Kellis[1,2]

[1]MIT, CSAIL, Cambridge, MA, 02139, [2]Broad Institute, Comp Bio, Cambridge, MA, 02139, [3]Harvard Medical School, Pathology, Boston, MA, 02115, [4] Rutgers University, Stem Cell Research Center, Piscataway, NJ, 08854

small RNAs (smRNAs) are established regulators of gene expression. These tiny molecules, once exclusively believed to affect post-transcriptional regulation, are emerging as important modulators of transcriptional rates as well. However, the mechanisms by which smRNAs regulate transcription of target genes remain unclear. We propose that smRNAs direct the action of histone-modifying enzymes, modulating interactions with promoters and/or non-coding RNAs present in promoter-proximal regions inducing long-term effects on gene expression. This is a novel model that would tie together endogenous smRNAs, noncoding RNAs, and the regulation of histone modifications, leading to long-term stabilization of gene expression. We have begun to define the relationship between small, non-coding RNA sequences and histone modifications through genome-wide analyses, and to explore the mechanisms by which epigenome-modifying complexes may be recruited by smRNAs to their target DNA sequences. Comparison of smRNAs identified in hESC by deep sequencing to previously published epigenomic data has identified a striking association between these two previously independent pathways. We have identified and characterized classes of smRNA sequences associated with specific chromatin modifications in human ES cells during neural differentiation. Using this characterization, we can begin to predict as yet unobserved smRNA-mediated epigenetic effects, and experimentally confirm the functional interactions between smRNAs and histone modifications in hESC. RNA immunoprecipitation of known histone-modifying complexes has identified a subset of smRNAs that may be bound in complex. We are currently exploring the physical interaction between these smRNAs and chromatin. smRNA-directed epigenetic modifications represent a novel mechanism by which cells are capable of managing patterns of transcriptional regulation and these modifications are predicted to drive the stabilization of cell phenotypes during differentiation.

ASSESSING THE FUNCTIONAL IMPACT OF SHORT INDELS IN INDIVIDUAL HUMAN GENOMES

David Goode[1], Dmitri Petrov[2], Arend Sidow[1,3]

[1]Stanford University, Genetics, 300 Pasteur Drive, Stanford, CA, 94305-5120, [2]Stanford University, Biological Sciences, 371 Serra Street, Stanford, CA, 94305-5020, [3]Stanford University, Pathology, 300 Pasteur Drive, Stanford, CA, 94305-5324

While SNVs and large structural variants in the human genome have been extensively studied, little is known about the functional impact of short insertions and deletions in the human genome, particularly at the individual genome level. Using evolutionary constraint as a proxy for function, we assessed the potential functional impact of the short (<11 bp) indels ascertained from four individual genomes (a Chinese, a Korean, a Yoruba and a Caucasian American). Our analysis reveals that there are as many as 25,000 functional small indels in a single human genome. Despite differences between individuals in indel ascertainment, we observe consistent and strong selection against small indels in constrained sequences within each individual, demonstrating their functional importance. We use this signature of selection to investigate the evolutionary processes influencing selection on derived small human indels. We find that, collectively, putatively functional small indels could have a significant impact genome-wide on regulation of transcription and transcript processing. Our results provide insights into an important component of an individual's genetic variation and suggest that the functional impacts of small indels deserve consideration during human genome sequencing studies.

USING THE NEANDERTAL GENOME SEQUENCE TO DETECT POSITIVE SELECTION EARLY IN HUMAN EVOLUTION

Richard E Green[1,2], Michael Lachmann[1], Svante Paabo[1], Neandertal Genome Consortium[1]

[1]Max Planck Inst. for Evol. Anthropology, Genetics, Deutscher Platz 6, Leipzig, D-04103, Germany, [2]Univ. of California, Santa Cruz, Biomolecular Engineering, 1156 High Street, Santa Cruz, CA, 95064

We recently completed one-fold coverage sequencing of the genome of Neandertals, the closest extinct relative of modern humans. These data open a new avenue for revealing the instances of positive selection that have acted on our ancestors since they diverged from Neandertals approximately 300,000 years ago.

We have developed and implemented a new method for detecting instances of positive selection in ancient modern humans using present-day human polymorphism and Neandertal sequence data. This method relies on the convenient fact that much of the genetic variation within humans today was present at the time Neandertals diverged. Therefore, in most regions of our genome, we expect and find that Neandertals sometimes have derived alleles at human polymorphisms. However, in regions that have been subjected to positive selection, the variation within humans is not expected to be shared with Neandertals. By scanning the human genome for such regions we have identified several strong candidates for positive selection in genes involved in diet, cognitive traits, and skeletal morphology.

We show, by simulating sequence evolution under both positive selection and neutral evolution scenarios, that our method can reliably detect instances of positive selection. The resulting signal can be further interpreted to indicate both the timing and intensity of selection. In contrast to previous scans for selection in humans, the Neandertal-based method has power to detect the occurrence of sweeps all the way back to the Neandertal-modern human population divergence. Thus, it offers an unparalleled opportunity to identify the last adaptive changes shared by all current humans. Finally, we demonstrate how Neandertal sequence data can be used in a secondary analysis to identify candidate causal genetic changes for driving these selective sweeps.

A METHOD FOR INFERRING ANCESTRAL POPULATION SIZES AND SPLIT TIMES FROM WHOLE-GENOME SEQUENCE DATA IN THE PRESENCE OF MIGRATION.

Ilan Gronau, Adam Siepel

Cornell University, Department of Biological Statistics and Computational Biology, Weill Hall, Ithaca, NY, 14853

Whole-genome sequence data is a promising source of information about population history, yet it poses a serious computational challenge. MCMCcoal (Rannala and Yang, *Genetics*, 2003) is a Bayesian Markov chain Monte Carlo algorithm for inferring ancestral population sizes and split times, which is efficient enough for use with complete genomes. A major drawback of this algorithm, however, is its strict assumption of no gene flow between populations -– an assumption that can lead to biased estimates of the parameters of interest. We present a modification of this algorithm, called **MCMCcoal-mig**, that explicitly models inter-species migration in a general and flexible manner. Unlike existing tools that accommodate migration, our algorithm maintains the efficiency of MCMCcoal, allowing tens of thousands of loci to be analyzed. We present preliminary results based on simulated data that demonstrate the ability of MCMCcoal-mig to produce accurate parameter estimates in the presence of gene flow.

MCMCcoal samples values for all model parameters (effective population sizes and split-times) together with a specific gene tree for each locus, under the assumption of no inter-population gene flow. In MCMCcoal-mig, gene flow is allowed by introducing migration into the coalescent model. In particular, we adapt the sampling procedure of MCMCcoal by introducing, removing, and perturbing migration events along each gene tree. These migration events draw complex dependencies between populations (which were otherwise assumed to be independent), necessitating a major revision in the way the likelihoods of gene trees are computed. The changes we implemented in the likelihood computation method actually ended up improving the efficiency of MCMCcoal. We performed several tests of MCMCcoal-mig on simulated datasets, demonstrating both an ability to detect inter-species migration, and an improved accuracy in parameter estimation in the presence of inter-species migration. We also examine the effect of allowing for migration in an analysis of the early history of human population evolution based on several individual human genomes, including the recently published Khoisan and Bantu genomes.

MAPPING COMPLEX TRAITS USING A MULTI-INTEGRATED "OMICS" APPROACH IN TWINS: THE MUTHER STUDY

Kerrin Small*[1,2], Elin Grundberg*[1,2], Asa Hedman*[3], Alexandra C Nica[2,4], Daniel Glass[1], James Nisbett[2], Alicja Wilk[2], Amy Barrett[3], Mary Travers[3], Tsun-Po Yang[2], So-Youn Shin[2], Krina Zondervan[3], Nicole Soranzo[1,2], Kourosh Ahmadi[1], Emmanouil T Dermitzakis[4], Mark I McCarthy[3], Timothy D Spector[1], Panos Deloukas[2]

[1]King's College London, Dept of Twin Research and Genetic Epidemiology, London, SE17EH, United Kingdom, [2]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, CB101SA, United Kingdom, [3]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, OX37BN, United Kingdom, [4] University of Geneva Medical School, Dept of Genetic Medicine and Development, Geneva, 1211, Switzerland

* Authors contributed equally

Integrative approaches are needed to disentangle the genetic and environmental contributions to complex trait susceptibility and to identify the underlying molecular mechanisms. To this end we have initiated the MuTHER project (www.muther.ac.uk) which aims to develop a resource of detailed genomic (SNP-genotyping, re-sequencing), transcriptomic (expression profiling), epigenomic (methylation) and metabolomic (metabolite profiling) data from a range of tissues (fat, skin, muscle, lymphocytes and LCL) collected from up to ~900 (1/3 MZ and 2/3 DZ) deeply-phenotyped twins from the TwinsUK resource. We find from initial analysis of the 826 fat biopsies analyzed that ~30% of expressed adipose transcripts have a heritability (h2) greater than 0.3, and that the most highly heritable subset (h2 > 0.6) show a significant enrichment for gene functions related to lipid metabolism. In addition, we find clear correlations of adipose gene expression traits with clinically relevant phenotypes such as DXA-derived fat mass (whole-body and abdominal), r= -0.59–0.62. Preliminary results from cis-eQTL analysis of a subset of heritable (h2 > 0.3), fat mass-correlated transcripts (r>|0.4|) highlight strong cis-regulatory effects of numerous genes, including several genes known to play a role in adiposity. These will shortly be tested for association with obesity-related traits in the complete TwinsUK study. On-going efforts include extending the analyses to include transcriptomic/eQTL data from the additional tissues in combination with related clinical phenotypes as well as utilizing the twin design for accurate estimate of the genetic contributions to both gene expression and clinical phenotypes. This together with epigenomic and metabolomic data gathered within the project will greatly facilitate mapping the genetic architecture of complex traits.

ANALYSIS OF RESTORED FFPE SAMPLES ON HIGH-DENSITY SNP ARRAYS

Dmitry K Polkholok, Jennie Le, Frank J Steemers, Mostafa Ronaghi, Kevin L Gunderson

Illumina, Inc., Advanced Research, 9885 Towne Centre Dr., San Diego, CA, 92021

Millions of Formalin Fixed Paraffin Embedded (FFPE) cancer tissue samples currently present in the US and worldwide provide an enormous, invaluable repository to the discovery of biomarkers in cancer research, drug development, diagnosis and treatment of diseases. However, the processes of fixation and storage of FFPE samples lead to degradation and base modification with chemical residues. This degraded FFPE DNA is inefficiently amplified in the Infinium® whole genome amplification (WGA) step leading to poor performance in the Infinium® assay. In this study, FFPE DNA (100 ng input) was restored through a treatment with DNA polymerase, DNA repair enzyme, ligase, and modified Infinium® WGA reaction conditions. Canonical genotype cluster files optimized for analysis of FFPE samples were generated from Covaris sheared Coriell DNA samples subjected to the restoration process and Infinium® assay. Results of genotyping on high density HumanCytoSNP-12 BeadChip®, featuring approximately 300k SNPs indicated that even moderately degraded FFPE samples with initial call rates of 40-70% prior to restoration can be successfully restored to a state with call rates greater than 90%, and concordance greater than 99.0%. In general, over 80% of FFPE samples, which passed a real time QC assay metric, were successfully restored and genotyped. Infinium® genotyping of restored DNA from a collection of tumor FFPE samples confirmed the presence of genomic aberrations around known cancer genes. Finally, we further demonstrated that paired sample analysis between normal and tumor samples provided improved precision in CNV and LOH analysis compared to single sample analysis.

IDENTIFICATION OF GENE FUSION TRANSCRIPTS AND ISOFORM VARIANTS IN BRCA1-MUTATED BREAST CANCERS BY TRANSCRIPTOME SEQUENCING

Kevin C Ha[1], Emilie Lalonde[1], Lili Li[1,2], Jacek Majewski[1], William D Foulkes[1,2]

[1]McGill University, Department of Human Genetics, 740 Dr Penfield Avenue, Montreal, H3A 1A4, Canada, [2]McGill University, Department of Oncology, 3799 Cote St. Catherine, Montreal, H3T 1E2, Canada

The use of next-generation RNA sequencing (RNA-Seq) has facilitated the global expression profiling of cancer transcriptomes at a resolution previously unattainable by other methods. In particular, RNA-Seq is able to reveal gene fusion transcripts arising from genomic rearrangements, and novel transcript isoform variants arising from alternative splicing, both of which have been implicated in cancer pathogenesis. Here, we sought to identify gene fusion and isoform transcripts expressed in *BRCA1*-mutated breast cancers which may contribute to the observed phenotype. Mutations in *BRCA1* are known to contribute to early onset breast and ovarian cancers. Using RNA-Seq on the Illumina Genome Analyzer platform, we have characterized and compared the transcriptomes of four breast cancer cell lines and one primary tumour with known *BRCA1* mutations. Applying previously described as well as novel strategies, we were able to identify several candidate gene fusions, including the known NFIA-EHF fusion in the HCC1937 cell line identified by Stephens et al. (*Nature*, 2009). *EHF* is a member of the *ETS* transcription factor family that is frequently rearranged in cancers such as the prostate. We have also identified novel splicing events that are present in more than one sample. We will present findings from our comparative analysis thus far and demonstrate the utility of RNA-Seq for investigating cancer transcriptomes.

A PAN-GENOMIC SURVEY OF *STREPTOMYCES*: DIVERSE SECONDARY METABOLISM AND HIGH AUXILLARY GENE CONTENT WITHIN DYNAMIC CHROMOSOME ARMS ARE CHARACTERISTIC OF THIS GENUS.

Brian J Haas[1], Michael A Fischbach[2], Paul Godfrey[1], Mike J Koehrsen[1], Dirk Gevers[1], Jason Holder[3], Jeremy Zucker[1], Aaron Brandes[1], Bruce Birren[1]

[1]Broad Institute, Genome Sequencing and Analysis Program, 7 Cambridge Center, Cambridge, MA, 02142, [2]UCSF, Department of Bioengineering and Therapeutic Sciences, 1700 4th Street, San Francisco, CA, 94158, [3]MIT, Department of Biology, 77 Massachusetts Ave, Cambridge, MA, 02139

Bacteria of the genus *Streptomyces* are known as producers of natural products, including clinically used antibiotics, antifungals, anticancer agents, antiparasitics, immunosuppressants, and anti-obesity drugs. We have generated high-quality draft sequences of twenty new actinomycete genomes, of which 16 are *Streptomyces* species. Comparisons between the *Streptomyces* linear genomes revealed that they are partitioned into a central region enriched in core functions flanked by chromosome arms enriched in auxillary functions, including natural product biosynthetic genes, which vary greatly in type and size between strains. The core genome is well conserved, with localized gene order conserved across the most distantly related species.

Our pan-genomic survey of *Streptomyces* allowed for genus-level comparisons with alternate pan genomes, allowing insights into properties that are characteristic of this genus. For comparative analysis, we selected nine comparator Eubacterial genera, requiring at least eight members with sequenced genomes and intra-genus phylogenetic distances comparable to that of the streptomycetes analyzed. Based on pan genome comparisons, we show that streptomycetes are exceptional in three ways: they have a larger fraction of auxiliary genes than other genera; they encode large repertoires of secondary metabolite biosynthetic gene clusters; and their dynamic chromosome arms appear tuned to specialized biological functions.

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900006C

A COMPUTATIONAL FRAMEWORK TO IDENTIFY FUSION TRANSCRIPTS FROM PAIRED-END RNA-SEQ DATA

Andrea Sboner[1,2], Lukas Habegger[1], Dorothee Pflueger[3], Stephane Terry[3], David Z Chen[1], Joel S Rozowski[2], Ashutosh K Tewari[4], Naoki Kitabayashi[3], Benjamin J. Moss[3], Mark S Chee[5], Francesca Demichelis[3,6], Mark A Rubin[3], Mark B Gerstein[1,2,7]

[1]Yale University, Program in Computational Biology and Bioinformatics, 266 Whitney Ave, New Haven, CT, 06520, [2]Yale University, Molecular Biophysics and Biochemistry Department, 266 Whitney Ave, New Haven, CT, 06520, [3]Weill Cornell Medical College, Department of Pathology & Laboratory Medicine, 1300 York Ave, New York, NY, 10065, [4] Weill Cornell Medical College, Department of Urology, 1300 York Ave, New York, NY, 10065, [5]Prognosys Biosciences, Sequensys NextGen Sequencing, 505 Coast Blvd. South, Suite 103, La Jolla, CA, 92037, [6]Weill Cornell Medical College, Institute for Computational Biomedicine, 1305 York Ave, New York, NY, 10065, [7]Yale University, Department of Computer Science, 51 Prospect Street, New Haven, CT, 06511

Deep sequencing approaches can interrogate genomes and transcriptomes at unprecedented resolution to reveal major molecular alterations. Transcriptome profiling with next-generation sequencing (RNA-Seq) has dramatically changed our understanding of the extent and complexity of eukaryotic transcription. RNA-Seq provides more accurate measurements of gene expression and enables the quantification of the relative abundance of transcript isoforms. More recently, RNA-Seq has been applied to discover chimeric transcripts, where mRNAs of two genes are joined together. Such chimeric transcripts arising from genomic rearrangements have shown to be driving the molecular events in cancer.

Here we describe a computational framework, FusionSeq, to process and analyze paired-end (PE) RNA-Seq data to detect fusion transcripts. This framework consists of three modules: a fusion transcript detector, a filtration cascade, and a junction-sequence identifier. The fusion-transcript-detection module finds candidate chimeric transcripts from PE reads joining two genes. The filtration-cascade module discards candidates with artifacts such as misalignment due to homology or random pairing of transcript fragments. The junction-sequence identifier module unravels the exact sequences at the breakpoints. Furthermore, to prioritize experimental validation, FusionSeq ranks the candidates by several statistics, including SPER (supportive-PE-reads-per-million mapped reads) and the comparison between the observed SPER and various expectations. To calibrate this method, we deeply sequenced six cancers harboring known genomic rearrangements, one sample with no known rearrangements and one normal HapMap cell line. We demonstrate that FusionSeq is able to identify known fusions as well as characterize different isoforms of these events. Moreover, we show that no fusions are detected in the samples known not to have rearrangements. Finally, FusionSeq was able to detect two novel fusions in the cancer samples that were subsequently experimentally validated.

DNA METHYLATION PROFILING OF NORMAL HUMAN CEREBRAL CORTEX

Yurong Xin[1], Anne O'Donnell[2], Benjamin Chanrion[1], Maria Milekic[1], Yongchao Ge[3], Fatemeh Haghighi[1]

[1]Columbia University, Department of Psychiatry, New York, NY, 10032, [2]Columbia University, Department of Genetics and Development, New York, NY, 10032, [3]Mount Sinai School of Medicine, Department of Neurology, New York, NY, 10029

DNA methylation may play an important role in the etiology of neuropsychiatric disorders, perhaps as equally important as genetics and the environment. However, In order to better understand both the wild type genomic DNA methylation patterns and aberrant methylation events that occur in disease states, we first examined DNA methylation profiles within the normal human brain. We have developed a cost-effective, unbiased, whole-genome methylation profiling technique that can assay the methylation state of more than 80% of the CpG sites in the human genome. This method, methylation mapping analysis by paired-end sequencing (Methyl-MAPS) couples advances in next generation sequencing with enzymatic fractionation of DNA by methylation state.

In this large-scale study we have mapped the methylation state of 36% of CpG sites in the human cerebral cortex of 10 normal non-psychiatric subjects (including 6 prefrontal and 4auditory cortical samples). We focused on the prefrontal cortex (PFC) due to converging evidence from neuroimaging and functional studies implicating this region in both depression and schizophrenia. Secondarily, we also examined the auditory cortex, because schizophrenia disorder includes defects in sensory perception and processing. With these data we are for the first time able to explore DNA methylation profiles within two distinct brain regions with differing neurodevelopmental trajectories; the evolutionarily conserved auditory temporal cortex developing early as compared to the prefrontal cortex which undergoes maturation well into early adulthood. Our data reveal that DNA methylation is significantly more conserved in the auditory cortex then the PFC ($P<10^{-15}$). Despite this significant difference, DNA methylation signatures in the cortex are highly conserved, with >25% of the total CpG sites in cortex showing less than 20% difference in methylation state across the 10 samples examined. Cross-species analysis of DNA methylation conservation between human and mouse brains show that DNA methylation is not correlated with sequence conservation. Instead, increase in cross-species DNA methylation conservation is correlated with increasing CpG density. Genomic regions with significant human-mouse DNA methylation conservation (correlation >80%) typically have greater than 5 CpG dinucleotide in 100bp window. Although enriched in gene promoters, these regions also cover gene bodies, as well as repeat sequences that represent both methylated and unmethylated states. These data provide insight in studies of neuropsychiatric disorders, in identifying genomic regions that are developmentally and evolutionarily conserved that when aberrantly methylated may confer increase risk for disease.

GENOME-WIDE MAPPING AND ASSEMBLY OF STRUCTURAL VARIANT BREAKPOINTS IN THE MOUSE GENOME

Ira M Hall[1,2], Aaron R Quinlan[1,2], Royden A Clark[1], Svetlana Sokolova[1], Mitchell L Leibowitz[1], Yujun Zhan[3], Mathew E Hurles[3], Joshua C Mell[4]

[1]University of Virginia, Biochemistry & Molecular Genetics, 1340 Jefferson Park Ave, Charlottesville, VA, 22908, [2]University of Virginia, Center for Public Health Genomics, 6111 West Complex, Charlottesville, VA, 22908, [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom, [4] University of British Columbia, Rm. 2551 Life Sciences Centre, Rm. 2551 Life Sciences Centre, Vancouver, V6T 1Z3, Canada

Structural variation (SV) is a rich source of genetic diversity in mammals, but due to the challenges associated with mapping SV in complex genomes, basic questions regarding their genomic distribution and mechanistic origins remain unanswered. We have developed an algorithm (HYDRA) to localize SV breakpoints by paired-end mapping, and a general approach for the genome-wide assembly and interpretation of breakpoint sequences. We applied these methods to two inbred mouse strains: C57BL/6J and DBA/2J. We demonstrate that HYDRA accurately maps diverse classes of SV, including those involving repetitive elements such as transposons and segmental duplications, however our analysis of the C57BL/6J reference strain shows that incomplete reference genome assemblies are a major source of noise. We report 7196 SVs between the two strains, more than two-thirds of which are due to transposon insertions. Of the remainder 59% are deletions, 26% are insertions of unlinked DNA, 9% are tandem duplications, and 6% are inversions. To investigate the origins of SV we characterized 3316 breakpoint sequences at single nucleotide resolution. We find that approximately 16% of non-transposon SVs have complex breakpoint patterns consistent with template switching during DNA replication or repair, and that this process appears to preferentially generate certain classes of complex variants. Moreover, we find that SVs are significantly enriched in regions of segmental duplication, but that this effect is largely independent of DNA sequence homology and thus cannot be explained by non-allelic homologous recombination (NAHR) alone. This result suggests that the genetic instability of such regions is often the cause rather than the consequence of duplicated genomic architecture.

HIGH-THROUGHPUT IMMUNOREPERTOIRE ANALYSIS BY MULTIPLEX PCR AND 454 SEQUENCING

Chunlin Wang[1], Catherine Sanders[2], Qunying Yang[2], Elijah Wang[1], Jian Han[2]

[1]HudsonAlpha Institute of Biotechnology, 601 Genome Way, Huntsville, AL, 35806, [2]Stanford Genome Technology Center, 855 California Ave, Palo Alto, CA, 94304

An immunorepertoire is the sum of functionally diverse B and T cells in circulation at any given moment. The antigen receptors of T and B cells are produced by somatic recombination of a limited, but large number of gene segments: V, D, and J, which determine the specificity of antibodies and T cell receptors. Because the list of possible V, D, J arrangements is so extensive, studying immunorepertoires is challenging. Additional mechanisms such as somatic hypermutation in B cells, N addition, and nibbling before insertions contribute to an even higher level of diversity. Another challenge is the scarcity of samples. Each V(D)J arrangement may only be represented by a few cells in a peripheral blood sample. Therefore, to study the immunorepertoire comprehensively, a method to amplify the V(D)J signal from a limited amount of sample is crucial. Even more challenging is the requirement for such a method to be inclusive and semi-quantitative, so the sample's original V(D)J diversity and distribution can be evaluated without any bias associated with the amplification process. However, such a multiplex PCR assay is difficult to develop, mainly because of incompatibility among the primer sets used in the amplification system. Other difficulties encountered when using conventional multiplex PCR include high background caused by primer dimers and non-specific amplifications, poor reproducibility, and inefficient amplification of V(D)Js in B cells due to somatic hypermutations. Amplicon rescued multiplex PCR incorporates hundreds of VDJ specific primers in one reaction, semi-quantitatively amplifying all the expressed VDJs in B and T cells in an RNA sample. The amplicon mixture is then subjected to high-throughput sequencing with the next-generation Roche 454 platform. In one run, more than 174,495 and 176,095 unique V(D)Js were obtained from B and T cells, respectively.This novel technology can be used to uncover disease mechanisms associated with abnormal immunorepertoires, to evaluate vaccine efficiency, to identify new biomarkers, and to develop new therapeutics. An online dababase and unique analytical tools have been developed for data sharing, deposition, and mining.

DEVELOPMENT OF A NEXT GEN ANALYSIS PIPELINE FOR IDENTIFICATION AND ANNOTATION OF VARIANTS FROM WHOLE EXOME SEQUENCE

Nancy F Hansen[1], Pedro Cruz[1], Jamie K Teer[1], Praveen F Cherukuri[1], Alice Young[2], Robert Blakesley[2], Gerard G Bouffard[2], Eric Green[1], James C Mullikin[1,2]

[1]National Human Genome Research Institute, Genome Technology Branch, 5625 Fishers Lane, Rockville, MD, 20852, [2]NIH Intramural Sequencing Center, NHGRI, 5625 Fishers Lane, Rockville, MD, 20852

In recent years, the NIH Intramural Sequencing Center (NISC) has been performing large scale, PCR-based Sanger sequencing of genes of interest. To date, we have sequenced over 4.2Gb of genomic sequence in more than 5,000 individuals, depositing over 3.8 million Sanger traces into the NCBI trace repository. Recently, we have begun to perform whole exome capture and Next Gen sequencing on some of these same samples using solution hybridization selection (SHS) and sequencing by synthesis (SBS) technologies. We plan to complete more than 200 whole exomes this year. To analyze this data, we have developed an informatics pipeline that performs paired, gapped alignments of paired-end short reads using Eland and cross_match, detection and genotyping of single nucleotide and deletion/insertion variants using our own Bayesian genotype caller, "MPG", and functional annotation of protein-coding variants with CDPred, which uses evolutionary conservation in protein domains in a position-specific manner to predict severity of observed amino acid substitutions.

Comparison of these second generation-based variants with our rich data set of Sanger derived genotypes and SNP genotyping chip based assays shows our genotypes to be 99.8% accurate. Called variants are stored in a production scale Oracle database, along with functional annotation of coding and splice-site variants, allowing for quick identification of novel variants of interest that conform to expected Mendelian inheritance models. These automated predictions and annotations are then displayed through a wiki, allowing researchers to make notes and commentary suggesting further investigations.

EPIGENOMIC LANDSCAPE OF ERYTHROID GENE REGULATION

R Hardison[1], W Wu[1], Y Cheng[1], C K Capone[1], S A Kumar[1], C Morrissey[1], K-B Chen[1], G Crawford[2], F Chiaromonte[1], J Taylor[3], G Blobel[4], M Weiss[4]

[1]Penn State Univ, CCGB, University Park, PA, 16802, [2]Duke Univ, IGSP, Durham, NC, 27708, [3]Emory Univ, Biol, Atlanta, GA, 30322, [4] Children's Hospital, Peds, Philadelphia, PA, 19104

The commitment of multipotential progenitor cells to a particular lineage and the differentiation and maturation of that lineage are driven by induction of lineage-specific genes and repression of others. Gene expression is regulated by occupancy of cis-regulatory modules by transcription factors, recruitment of co-activators and co-repressors and modifications of the chromatin structure, but a full picture of these epigenomic features and how they result in induction versus repression of specific genes is not yet available. We explore these questions in a mouse cell line model of erythroid maturation. A genetic knockout of the gene *Gata1* in the G1E cell line arrests differentiation at the proerythroblast stage, and rescue by expressing the transcription factor GATA1 in a subline allows the cells to mature to erythroblasts. We have measured comprehensively changes in gene expression during this GATA1-dependent maturation and concomitantly, genome-wide occupancy by the transcription factors GATA1, GATA2, TAL1, and CTCF, as well as chromatin accessibility and histone modifications in the chromatin, using Illumina sequencing technology for ChIP-seq. The changes in protein occupancy and histone modification levels after restoration of GATA1 should reveal features that can account for the expression patterns of the genes and their differential response to GATA1. Surprisingly, we find that only limited changes in histone modification status accompany the substantial changes in gene expression, and rather than altering chromatin structure upon restoration, GATA1 binds to genomic sites that already have activating histone modifications. However, the occupancy of CRMs by specific transcription factors and changes in H3K4me3 at the transcription start site do explain a considerable amount of the largest responses in gene expression.

POPULATION SEQUENCING OF TWO ENDOCANNABINOID METABOLIC GENES IDENTIFIES RARE AND COMMON REGULATORY VARIANTS ASSOCIATED WITH EXTREME OBESITY AND METABOLITE LEVEL.

O. Harismendy[1], V. Bansal[2], G. Bhatia[1], M. Nakano[1], M. Scott[2], X. Wang[1], C. Dib[3], E. Turlotte[3], J. C Sipe[2], S. S Murray[2], J.-F. Deleuze[3], V. Bafna[1], E. J Topol[2], K. A Frazer[1]

[1]UCSD, Moores UCSD Cancer Center, 9500 Gilman Dr, La Jolla, CA, 92093, [2]The Scripps Research Institute, MEM, 10550 N Torrey Pines Rd, La Jolla, CA, 92037, [3]Sanofi-Aventis, Genetics Center, 2, rue Gaston Cremieux, Evry, 91057, France

Targeted sequencing of candidate genes in individuals at the extremes of a quantitative phenotype distribution is a method of choice to analyze the contribution of rare variants to disease susceptibility. The endocannabinoid (EC) system mediates signaling in the brain and peripheral tissues involved in the regulation of energy balance, is highly active in obese patients, and thus a strong candidate pathway to examine genetic association with BMI. We sequenced two intervals (188kb) encoding the EC metabolic enzymes, FAAH and MGLL, in 147 normal controls (BMI≤30) and 142 extremely obese cases (BMI≥40). After applying quality filters, we called 1393 SNVs, 55% of which are rare (MAF<0.01), and 143 indels. We used two different approaches, single marker tests and collapsed markers tests, to identify 4 intervals associated with BMI: the FAAH promoter, MGLL promoter, MGLL intron 2, and MGLL intron 3. Two of these intervals are composed of rare (MAF<0.01) variants, one of rare and low frequencies (MAF≈0.03) variants and one of common (MAF≈0.15) variants. The majority of the associated variants are located in promoters or in predicted transcriptional enhancers, suggesting a regulatory role. Additionally, we demonstrate that the set of rare variants in the FAAH promoter associated with BMI are also associated with increased level of FAAH substrate AEA, further implicating a functional role in obesity. Our study provides insights into study design and analysis approaches and demonstrates the importance of examining regulatory elements rather than exon sequences.

RGASP: RNASEQ GENOME ANNOTATION ASSESSMENT PROJECT

J Harrow, F Kokocinski, J Abril, G Williams, A Mortazavi, R Guigo, T Hubbard

RGASP steering Committee, WTSI, Hinxton, Cambridge, CB10 1HH, United Kingdom

RNAseq is revolutionizing eukaryotic transcriptomics and highlighting the extent different loci are alternatively spliced. Following the successful format of the EGASP workshop in 2005, the RNAseq genome annotation assessment project was launched to assess the current progress of automatic gene building using RNAseq as its primary dataset. The goals of the analysis are to assess the success of computational methods to correctly map RNAseq data to the genome, assemble transcripts and quantify their abundance in particular datasets. Data from different sequencing platforms (illumina, SOLID and Helicos) were analysed as well as transcriptional data from three different organisms Human, Drosophila and C.elegans, since the genome and reference annotation is of high quality.Eighteen groups submitted transcript predictions which were evaluated against annotation datasets for the three organisms. In addition human chromosomes 3 and 4 were newly manually annotated by the GENCODE consortium as an additional “blind” test. Quantification predictions on transcript level are being compared against a gold standard subset of 100 loci selected for each organism. Probes are being designed against at least two transcripts per loci and nanostring experiment performed, since it is a non-RTPCR derived method, to quantify the transcripts. The first round analysis proved challenging because of the volume of data submitted and the variation in data sets (ie pair-end, single and directional data from human K562 cell line). Actually defining the gold standard gene set expressed in particular tissues also raised much debate. Therefore the outcome of the workshop was to rerun the prediction algorithms on a more standardize dataset (75bp pair end read set,at similar depth of sequencing and limit submission to 3 per organism). In addition the submitter were asked to supply BAM alignment files so the read alignments against the genome and annotation could be analysed. Closing date for resubmission was 22nd Jan 2010 and 12 groups resubmitted. The results of the second phase of RGASP will be presented at the meeting including the results of experimental validation of novel predictions.

NOVEL MICRORNAS IN HUMAN EMBRYONIC STEM CELLS AND NEURAL PRECURSORS

Ronald P Hart[1], Cynthia Camarillo[1], Mavis R Swerdel[1], Jonathan L Davila[1], Jennifer C Moore[1], Loyal A Goff[1]

[1]Rutgers University, Stem Cell Research Center, 604 Allison Rd, Piscataway, NJ, 08854, [2]MIT, CSAIL, 32 Vassar Street, Cambridge, MA, 02139

Since microRNAs are required for both maintenance of stem cell pluripotency and differentiation, and since many more microRNAs have been computationally predicted in genome than have been found, there are likely to be many previously unknown microRNAs expressed early in differentiation. We used deep sequencing of small RNAs from undifferentiated human embryonic stem cells (hESC) and neural-restricted precursors to computationally predict 818 novel microRNA genes. These predicted genomic loci are relatively unique to humans or primate mammals and are found to be associated with patterns of modified histones consistent with gene expression. A subset of 146 of the predictions was found to be enriched in Ago2-containing complexes, consistent with RISC-associated microRNAs. Expression of these novel microRNAs parallels the patterns of developmental regulation of other microRNAs and some 30% of the novel microRNAs share seed elements with other microRNAs. To demonstrate function of these microRNAs, we have devised a strategy to detect bioactivity of all microRNAs in cultured cells using synthetic response elements built into a library of GFP-expressing plasmids. The presence in hESC of a broad collection of developmentally-regulated novel microRNAs, lacking broad species conservation but associated with Ago and having detectable bioactivity, predicts a role in modulating gene expression during differentiation.

BARCODING DNA IN POOLED LIBRARIES IMPROVES VARIANT SENSITIVITY OVER A POOLED PCR APPROACH

C. Hartl, A. Kernytsky, K. Garimella, M. Rivas, J. Flannick, M. DePristo, S. Gabriel
Broad Institute, GSA, 5 Cambridge Ctr, Cambridge, MA, 02142

A major goal of medical sequencing is the identification of rare variants across many individuals. One approach is to sequence DNA libraries formed by pooling DNA from multiple individuals, optionally ligating the DNA fragments at their 5' ends with 6-base oligonucleotide "barcodes" identifying the DNA originator. Without barcodes, only population frequencies can be estimated for each variant, whereas barcoding allows direct computation of genotypes for each sample in the library.

We evaluated these two approaches with data from the Framingham Heart Study (FHS). Sequencing for FHS is approached in two phases: a pooled phase using PCR; and a barcoded hybrid-capture phase. Associated with each phase is a pilot study, using DNA from individuals with known HapMap genotypes.

The pilot for unbarcoded pooled PCR libraries uses DNA from 117 CEU individuals, combined into 3 pools of 40 individuals (3 appear in two pools). The barcoded library, comprised of 24 individuals, was prepared using solution hybrid selection. Comparisons of recovery rates were performed on the intersection of the target regions of the two pilots.

Library	Unbarcoded	Barcoded
Samples	117	24
Genes	163	163
Coverage	2000x	1000x
Calls	1221	898
HapMap Variants	502	291
HapMap Singletons	111	54
Sensitivity	80%	98%
Sensitivity to Singletons	24%	98%

The vast increase in barcoded hybrid selection's sensitivity over that of pooled PCR may be due to the fact that sequencing artifacts at a rate of 1-2% are often indistinguishable from a 1-2% allele. Associating reads with individuals increases the tolerance to sequencer errors, as true variation is expected to appear at 50% or 100% for each individual. These results inform the design of FHS production sequencing, which will use barcoded hybrid selection to maximize power to detect rare variants.

DYNAMICS AND DIVERSITY OF THE *DROSOPHILA MELANOGASTER* TRANSCRIPTOME

Brenton R Graveley, Michael Duff, C. Joel McManus, Sara Olson, Li Yang, the ModENCODE Transcriptome Group, Peter Cherbas, Thomas Kaufman, Michael Brent, Tom Gingeras, Roger Hoskins, Brian Oliver, Susan Celniker

University of Connecticut Health Center, Genetics and Developmental Biology, 263 Farmington Avenue, Farmington, CT, 06030-3301

Drosophila melanogaster is perhaps the most widely studied metazoan organism, yet the genome still contains numerous unannotated genes, non-coding RNAs, alternatively spliced isoforms, and RNA editing sites. Our understanding of the temporal and spatial patterns in which these genes are expressed, spliced, and edited to direct development remains limited. I will describe several ongoing projects that are aimed at understanding the dynamics and diversity of the *Drosophila* transcriptome. As part of the modENCODE project, we have generated over 2 billion uniquely mapped single and paired-end RNA-seq reads using the Illumina platform from poly(A)+ RNA isolated from 30 time points throughout development. In parallel, we performed strand-specific sequencing of 12 total, rRNA-depleted embryonic RNA samples using the SOLiD platform. These data have identified many new genes and thousands of alternative splicing events resulting in the discovery of new mRNA and inferred protein isoforms. Second, using paired-end RNA-Seq of RNA isolated from *Drosophila* interspecies hybrids, we have identified ~80 genes that undergo *trans*-splicing, in which the mRNAs are synthesized by splicing exons from two different pre-mRNAs. Finally, we have been working to understand the true diversity of transcripts that are expressed from the *Drosophila Dscam* gene, which has the potential to generate up to 38,000 isoforms. We have developed a three read sequencing method using the Illumina platform and have found that over 92% of the predicted isoforms are indeed expressed, and that these are regulated throughout development. Together, these studies provide tremendous new insight into the extraordinary diversity and regulation of the *Drosophila* transcriptome.

HERITABLE INDIVIDUAL-SPECIFIC AND ALLELE-SPECIFIC CHROMATIN SIGNATURES IN HUMANS

Ryan McDaniell[1], Lingyun Song[2], Michael Erdos[3], Laura Scott[4], Mario Morken[3], Katerina Kucera[2], Francis Collins[3], Huntington Willard[2], Jason Lieb[6], Terrence Furey[2], Gregory Crawford[2], Vishwanath Iyer[1], Ewan Birney[5]

[1]UT, Biology, Biology, Austin, TX, TX 7871, [2]Duke, IGSP, Genomics, Durham, NC, NC 27708, [3]NHGRI, Genome, Genomics, Bethesda, MD, MD 20892, [4] U.Michigan, Genetics, Statistics, Ann Arbor, MI, MI 48109, [5]EBI, PANDA, Services, Hinxton, CB10 1SD, United Kingdom, [6]UNC, Centre for Genomics, Genomics, Chapel Hill, NC, NC 27599

Little is currently known in how variation in individuals of chromatin structure and transcription factor binding influence phenotypes. We have catalogued both individual variation and allele-specific variation in chromatin structure, using DNaseI, FAIRE and Chip assays in cell lines from the 1,000 genomes project. We used high coverage trios for in-depth analysis and selected loci for sampling more individuals. ~10% of active chromatin sites were individual-specific, and ~11% were allele-specific. We specifically confirmed some allele specific signals using a mass spectroscopy assay. Both individual-specific and allele-specific sites were commonly transmitted from parent to child, showing that they are heritable. Using the structure of the families compared to signals on unrelated individuals we can show that there is a local, cis-based genetic component to this heritability.

We examined the functional role of this individual and allele specific variation by comparing to known molecular phenotypes, such as RNA expression and to X-inactivation in the female individuals. As expected there is a strong signal of allele specific chromatin data on X. Interestingly variation in RNA expression has a different pattern of correlation to variation in either DNaseI or to CTCF individual expression. This study shows we have the ability to accurately measure such chromatin variation and its downstream impact on phenotypes, providing opportunities to correlate this information with other more complex phenotypes in the future.

GENOME-WIDE MAPPING OF LONG-RANGE CHROMATIN INTERACTIONS AND TRANSCRIPTION REGULATORY NETWORKS IN HUMAN CELLS

Yijun Ruan

Genome Institute of Singapore, Genome Technology and Biology, 60 Biop0lis Street, Singapore, 138672, Singapore

Genomes are known to be organized into 3D structures in vivo through interactions with protein factors for nuclear process such as transcription, and DNA elements separated by long genomic distances are known to functionally interact. This view has been further emphasized by recent observations that many transcription factors bind remotely to gene promoters. However, it is still largely unknown to us how and to what extent chromatin interactions are involved in transcription regulations on a whole genome scale. To study these questions, we have developed the Chromatin Interaction Analysis using Paired-End-Tag sequencing (ChIA-PET) strategy for de novo detection of genome-wide chromatin interactions, and demonstrated this approach through the comprehensive mapping of chromatin interactions involved in transcription regulations mediated by estrogen receptor α (ERα) in a human genome (Nature 2009 462: 58-64). In order to map all chromatin interactions involved in all transcription regulation networks in the human genome, we have applied the ChIA-PET strategy to active transcriptional marks such as RNA polymerase II (RNAPII) and trimethylation of lysine 4 on histone H3 (H3K4me3) as analysis targets in a number of human cells. Our results have shown that both RNAPII and H3K4me3 are excellent targets for ChIA-PET experiments to detect long-range chromatin interactions between gene promoters and distal regulatory elements, as well as to identify co-localization of remote genes (intra-chromosome and inter-chromosome) in close proximity of nuclear space. Through comprehensive mapping of chromatin interactions and transcriptional activities, we have revealed that a large proportion of actively transcribed genes are involved in extensive chromatin interaction looping structures. The most abundant gene-centric chromatin interactions are appeared to be within local range of megabase genomic span, and nearby genes such as gene family members are organized to share common transcription factories. In addition, we have identified many hot spots of interaction hubs, in which clusters of genes crossing large megabase distance and different chromosomes are co-localized in close proximity. Collectively, our data suggests that long-range chromatin interaction is a primary mechanism for transcription regulation in human genomes. Further analyses of the chromatin interaction and transcription maps will provide deep insights to advance our understanding of transcription regulatory networks and the human genome biology.

INTEGRATIVE ANALYSIS OF GENOMIC AND EPIGENOMIC DATASETS IN THE DROSOPHILA MODENCODE PROJECT

Manolis Kellis for The modEncode Consortium

MIT, Computer Science / Broad Institute, 32 Vassar St. #32D-530, Cambridge, MA, 02139

The goal of the modENCODE project (model organism ENCyclopedia of DNA Elements) is a systematic identification of all functional elements in fly and worm and their dynamics across development and in multiple cell lines, through large-scale data generation and integration, including gene expression, transcription factor binding, chromatin modifications, histone variants, microRNAs.

In this talk, we describe the data integration efforts of the Drosophila consortium. We have used supervised and unsupervised probabilistic methods for combining multiple datasets to define active and repressed promoter regions, tissue-specific and stage-specific enhancers, and a map of insulator elements and regulatory units. We have used comparative genomics methods to define conserved regulatory motif instances within these regions that are associated with functional binding of transcriptional regulators. We have combined these datasets to gain insights into gene regulation, including defining candidate tissue-specific regulators of active and repressed regions, identifying combinations of regulatory motifs predictive of transcription factor binding by their synergistic and antagonistic effects, and combinations of transcription factors predictive of distinct chromatin states at specific stages of development.

Overall, our integrative analysis provides a general framework for studying the logic of transcriptional regulation in metazoan genomes from systematic large-scale genomic and epigenomic datasets, assembling the building blocks of gene regulation, towards a systems-level mechanistic understanding of regulatory control.

NEXT-GENERATION MENDELIAN GENETICS BY EXOME SEQUENCING.

Sarah B Ng[1], Emily H Turner[1], Mark J Reider[1], Michael Bamshad[1,2], Deborah A Nickerson[1], Jay Shendure[1]

[1]University of Washington, Department of Genome Sciences, 1705 NE Pacific St, Seattle, WA, 98195-5065, [2]University of Washington, Department of Pediatrics, 1959 NE Pacific St, Seattle, WA, 98195-6320

Rare monogenic diseases have been of incredible value to biomedical research, as the identification of the genes underlying phenotypes of interest has yielded fundamental, medically relevant insights into human biology. However, many rare disorders have yet to be solved. In many cases, this is because only a small number of cases/families are available, limiting the power of traditional gene mapping strategies. As most disease-causing variants affect coding sequences, comprehensive sequencing of all genes in affected individuals has the potential to serve as a genome-wide scan for the underlying cause of a rare monogenic disease. Consequently, we have sought to apply second-generation methods for targeted sequencing of the human exome as a cost-effective, genome-wide scan for coding mutations underlying monogenic diseases that is independent of linkage data or assumptions inherent to a candidate gene approach. Our progress includes the proof-of-concept demonstration that Freeman-Sheldon syndrome, an autosomal dominant disorder previously determined to be caused by mutations in MYH3 through a candidate gene approach, could be solved directly by exome sequencing. More recently, we applied exome sequencing to identify DHODH as the causative gene for Miller syndrome, a previously unsolved, autosomal recessive disorder. We are currently extending this strategy to additional genetic diseases of unknown etiology. Low-cost, high throughput technologies for exome resequencing have the potential to rapidly accelerate the discovery of candidate gene(s) and mutations that underlie rare monogenic diseases that have been resistant to conventional analysis. This strategy may also be applicable to the identification of rare variants contributing to diseases with more complex genetics through larger sample sizes, better filters against common variants, and appropriate weighting of nonsynonymous variants by predicted functional impact.

"CALLING CARDS" FOR DNA BINDING PROTEINS

Haoyi Wang[1,2], David Mayhew[1], Xuhua Chen[1], Mark Johnston[1,3], Rob Mitra[1]

[1]Washington University in St. Louis, Department of Genetics, 4444 Forest Park, St Louis, MO, 63108, [2]Massachusetts Institute of Technology, Department of Biology, 77 Massachusetts Avenue, Boston, MA, 02215, [3]University of Colorado, Denver, Department of Biochemistry and Molecular Genetics, Mail Stop 8101, P.O. Box 6511, Aurora, CO, 80045

The transcriptional networks that control organism development are precise, highly coordinated, and complex. Our ability to analyze these networks is limited because existing methods cannot trace transcription factor binding throughout a cell lineage, making it impossible to correlate DNA-binding events in progenitor cells to the final cell fates of their progeny. We report on our progress towards developing transposon “Calling Cards”, a technology that combines aspects of ChIP-Seq and lineage tracing. The method entails fusing the transposase of a transposon to a transcription factor, thereby causing it to direct the insertion of transposon DNA into the genome near where it binds. The transposon becomes a "Calling Card" that permanently marks the transcription factor's visit to that place in the genome. By recovering these Calling Cards along with some of the genomic DNA that flanks them and then determining their DNA sequences, it is possible to map the genome-wide binding history of the transcription factor. As a proof-of-principle, we have implemented Calling Cards in yeast, and multiplexed the method, mapping the binding of 8 transcription factors in a single expeirment. Recently, we have ported the method into a mammalian system and benchmarked our results to ChIP-Seq.

DNA METHYLOME MAP REVEALS CONSERVED ROLE OF GENE BODY METHYLATION IN REGULATING ALTERNATIVE PROMOTERS

Ting Wang[1], Alika Maunakea[2], Raman Nagarajan[2], Steve Jones[3], Tracy Ballinger[4], David Haussler[4], Marco Marra[3], Martin Hirst[3], Shaun Fouse[2], Brett Johnson[2], Chibo Hong[2], Joseph Costello[2]

[1]Washington University, Genetics, 4444 Forest Park ave, St. Louis, MO, 63108, [2]UC San Francisco, Neurosurgery, 1450 3rd Street, San Francisco, CA, 94158, [3]BC Cancer Agency, Genome Sciences Centre, 675 W. 10th Avenue, Vancouver, V5Z 1L3, Canada, [4] UC Santa Cruz, Center for Biomolecular Science and Engineering, 1150 High Street, Santa Cruz, CA, 95064

We present two complementary approaches to detect methylated and unmethylated genomic DNA. The first, methyl DNA immunoprecipitation and sequencing (MeDIP-seq), uses antibody-based immunoprecipitation of 5-methylcytosine and sequencing to map the methylated fraction of the genome. In the second method, unmethylated CpG sites are identified by sequencing size-selected fragments from parallel DNA digestions with the methyl-sensitive restriction enzymes (MRE-seq). We generated a genome-wide, high-resolution methylome map of human brain tissue, and a second map of human ES cell H1. These maps on average interrogate close to 90% of all CpGs (25 million of 28 million total) and 98% of CpG islands in the human genome, at the modest expense of relatively small amount specimen and a few lanes of Illumina flowcell.

We investigated the role of DNA methylation in gene bodies with these methylome maps. From high-resolution coverage of CpG islands, the majority of methylated CpG islands were revealed to be in intragenic and intergenic regions, while less than 3% of CpG islands in 5' promoters were methylated. The CpG islands in all three locations overlapped with RNA markers of transcription initiation, and unmethylated CpG islands also overlapped significantly with trimethylation of H3K4, a histone mark enriched at active promoters. The general and CpG-island-specific patterns of methylation are conserved in mouse tissues. These and other results support a major role for intragenic methylation in regulating cell context-specific alternative promoters in gene bodies.

GENOME-WIDE MAPS OF NUCLEOSOME ORGANIZATION IN THREE PRIMARY HUMAN CELL TYPES IDENTIFY SPECIFIC MECHANISMS THAT GOVERN NUCLEOSOME ORGANIZATION.

Anton Valouev[1], Steven Johnson[2], Scott Boyd[1], Cheryl Smith[1], Andrew Fire[1,2], Arend Sidow[1,2]

[1]Stanford University School of Medicine, Pathology, 300 Pasteur Dr, Palo Alto, CA, 94305, [2]Stanford University School of Medicine, Genetics, Mail Stop-5120, Stanford, CA, 94305

To gain insights into the global and local organizaiton of nucleosomes in primary human cells, and understand mechanisms that govern this organization, we performed high-resolution mapping of genome-wide nucleosome organization in three primary human cell types. To understand sequence-dictated rules that govern nucleosome organization, we reconstituted human genomic DNA with recombinantly derived nucleosome particles in vitro, and sequenced nucleosome bound DNA fragments to high depth. Across these experiments, we have obtained 2.2 billion mapped SOLiD sequence and additional 240 M RNA-seq reads to determine expression levels of these cells. Our analysis of in vivo and in vitro nucleosome organization revealed highly unexpected results that help us better understand organization of chromatin. Human chromatin contains a large number of sites with stereotypically positioned nucleosomes that are also evenly spaced relative to each other (nucleosome phasing). Our *in vitro* experiment informed that many sites throughout the human genome encode stereotypically positioned nucleosomes by their underlying DNA, but this positioning signal does not account for phasing of nucleosomes observed in vivo. Sites that encode stereotypic positioning of nucleosomes *in vitro*, contain a novel sequence signal that is not based on 10 bp dinucleotide periodicities, as was previously thought, but rather contains a specific sequence motif. The same DNA signal positions nucleosomes in vivo across the three primary cell types.
We observe a significant deviation from the sequence-dictated rules of nucleosome organization in vivo at CpG islands, binding sites of transcription factors and promoters of genes. CpG islands have a dramatic reduction of nucleosome occupancy, transcription factor binding sites contain distinct and cell-type specific organization of positioned and phased nucleosome arrays with binding sites localizing to the linker DNA. Promoters of actively transcribed genes adopt distinct organization of nucleosomes that is not present at silent promoters. There are substantial global differences in the chromatin structure. The average distance between the nucleosomes is 11 bps shorter in granulocytes compared to T-cells. We found that inter-nucleosome distance varies across the genome and depends on the local transcriptional levels within the occupied genes. Significantly more compact packing of nucleosomes was observed within the genes with high transcriptional activity. These results demonstrate that multiple forces shape nucleosome organization of the human chromatin.

BIASED GENE CONVERSION AND THE EVOLUTION OF HUMAN GENOMIC LANDSCAPES

Laurent Duret

CNRS, Université Lyon 1, Biométrie et Biologie Evolutive, 43 Bd du 11 novembre, Villeurbanne, 69100, France

Recombination is typically thought as a symmetrical process resulting in large-scale reciprocal genetic exchanges between homologous chromosomes. Recombination events, however, are also accompanied by short-scale, unidirectional exchanges in the neighborhood of the initiating double-strand break: gene conversion. A large body of evidence suggests that gene conversion is GC-biased in many eukaryotes, including mammals and human. AT/GC heterozygotes produce a larger amount of GC- than AT-gametes, thus conferring a population advantage to GC-alleles in high-recombining regions. This apparently unimportant feature of our molecular machinery has strong evolutionary consequences. Structurally, GC-biased gene conversion explains the spatial distribution of GC-content in mammalian genomes - the so-called isochore structure. Functionally, GC-biased gene conversion promotes the "undesired" segregation and fixation of deleterious AT->GC mutations, thus increasing our genomic mutation load. I will review the recent evidence for a GC-biased gene conversion process in mammals, its consequences on genomic landscapes, molecular evolution, and human functional genomics.

ANALYSIS OF 1000 GENOMES EXON CAPTURE PILOT DATA

Amit R Indap[1], Wen Fung Leong[1], Christopher L Hartl[2], Kiran V Garimella[2], Fuli Yu[3], Richard A Gibbs[3], Gabor T Marth[1], 1000 Genomes Project Exon Sequencing Group[4]

[1]Boston College, Dept of Biology, 140 Commonwealth Ave, Chestnut Hill, MA, 02467, [2]Broad Institute, Medical and Population Genetics, 320 Charles Street, Cambridge, MA, 02141, [3]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030, [4] 1000 Genomes Project

The 1000 Genomes Pilot 3 Project generated high coverage sequence data primarily in the coding regions of approximately 8,300 exons totaling 1.43 Mbp from 697 individuals sampled from seven different populations. The data was collected using multiple DNA capture technologies combined with two different next generation sequencing platforms (Illumina and 454). Boston College (BC) and the Broad Institute (BI) independently implemented informatics pipelines for data analysis that includes read mapping; duplicate filtering (i.e. the removal of reads from duplicated DNA fragments); base quality score recalibration; and SNP calling with similar multi-sample Bayesian algorithms.

Median per-individual sequence coverage within the seven populations ranged from 30 to 67X. Analysis of all 697 samples BC and BI yielded 19,890 total SNPs, 64% of which were in consensus between the BC and BI callsets. 30% of these consensus calls overlapped known SNPs in dbSNP v129, in line with population genetic expectations. The number of segregating sites within the different population samples ranged from 3,729 (Tuscany) to 6,370 (Luhya). The number of sites within African populations was significantly higher than in European or Asian samples. Initial analysis, based on comparison to experimental validation results originally collected for Pilot 1 (many samples, whole-genome, 4x-coverage), indicates that the accuracy of the SNP calls is well above 90%.

The considerable sequence depth makes it possible to detect low-frequency variants with very high sensitivity, and therefore, ascertain the low-frequency end of the site frequency spectrum with much better accuracy than achievable in low-coverage sequence data. The ability to detect rare alleles in genomic regions of interest make capture-sequencing approaches attractive for medical re-sequencing studies.

DE NOVO ASSEMBLY OF RNASEQ FOR TRANSCRIPTOME RECONSTRUCTION AND CHARACTERIZATION FROM YEASTS TO HUMAN

Moran Yassour[1,2], Manfred Grabherr[1], Joshua Z Levin[1], Mike Berger[1], Pamela Russell[1], Jessica Alfoldi[1], Andi Gnirke[1], Federica Di Palma[1], Kerstin Lindblad-Toh[1,3], Nir Friedman[3,4], Aviv Regev[1]

[1]The Broad Institute of MIT and Harvard, Genome Sequence and Analysis program, 7 Cambridge Ctr, Cambridge, MA, 02142, [2]The Hebrew University, School of Engineering and Computer Science, Givat Ram, Jerusalem, 91904, Israel, [3]Uppsala University, Department of Medical Biochemistry and Microbiology, P.O. Box 256, Uppsala, SE-751 05, Sweden, [4] The Hebrew University, Alexander Silberman Institute of Life Sciences, Givat Ram, Jerusalem, 91904, Israel

Experimentally defining the complete transcriptome of eukaryotic organisms has traditionally been a challenging task, but advances in sequencing RNA (RNAseq) offer new and powerful approaches to the study of transcriptomes. Recent studies have used RNAseq to quantify the expression levels of known genes, identify splice isoforms and refine gene boundaries. However, many studies depend on existing annotation, limiting the ability of discovering novel transcripts, and most require mapping to an available genome sequence, limiting their applicability to organisms without a sequenced genome, complex environmental samples, and cancer. Here, we present a novel approach for *de novo* assembly of a transcript catalog from read data alone. At the heart of our approach is a novel algorithm that takes read data and generates a host of assembly graphs, each one ideally corresponding to a single transcript. Our algorithm then extracts from each graph one or more transcript isoforms, quantifies their levels, and scores their confidence. These transcripts can then be mapped to a reference genome, from the same organism or a related species.
We show how these approaches scale to organisms from yeasts to vertebrates, helping in genome annotation of newly discovered organisms from the *Schizosaccharomyces* clade, transcriptome analysis in the *Anolis grahami* lizard for which the genome sequence is not available, and for the discovery of novel fusion transcripts in human cancers.

DISCOVERY OF HUMAN HETEROPLASMIC SITES ENABLED BY AN ACCESSIBLE INTERFACE TO CLOUD-COMPUTING INFRASTRUCTURE

Enis Afgan[1], Hiroki Goto[2], Ian Paul[3], Kateryna Makova[2], Anton Nekrutenko[2], James Taylor[1]

[1]Emory University, Biology, Atlanta, GA, 30322, [2]Penn State University, Center for Comparative Genomics and Bioinformatics, University Park, PA, 16802, [3]Penn State University, College of Medicine, Hershey, PA, 17033

Proliferation of DNA sequencing instruments has enabled any investigator to produce enormous amounts of sequence data. However, working with these raw sequences presents significant problems. For an experimental group with no computational expertise, simply running a data analysis program is a barrier, let alone building a compute and data storage infrastructure capable of dealing with this scale of data data. A computational model, "Cloud computing", has emerged that is ideally suited to the analysis of large-scale data. However, formidable challenges need to be addressed to make these resources available to individual investigators. We have developed a solution that allows experimentalists to perform large-scale analysis using cloud-computing resources with nothing more than a web browser. A user without computational expertise can instantiate an analysis environment on a cloud, adding storage and compute resources to this environment as needed. Popular tools and analysis workflows are built-in, ready to run.

Using this analysis solution, we analyzed a mitochondrial genome re-sequencing experiment representing three mother and child pairs. For each mother and child pair the DNA was collected from cheek swab specimen and from blood by our clinical collaborators at Penn State Medical School and mtDNA was amplified with PCR. To control for possible PCR-induced errors, each amplification was performed twice. In total we generated 24 Illumina datasets (eight for each mother and child pair - two mtDNA amplification for each cheek swab and blood samples). All analysis steps from data pre-processing to polymorphism calling were performed using a Galaxy instance instantiated on the cloud. Within Galaxy we use a variety of analysis tools to process this data and identified a number of somatic mutations and heteroplasmic sites. This is the first practical demonstration that cloud-computing resources can be made available to researchers with no computational infrastructure to successfully perform complex large-scale analyses.

INFORMATICS CHALLENGES IN HUMAN MICROBIOME RESEARCH

Jennifer Russo Wortman

University of Maryland School of Medicine, Institute for Genome Sciences and Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, Baltimore, MD, 21201

The Human Microbiome Project (HMP) was launched by the National Institutes of Health (NIH) Roadmap for Medical Research and is designed to fuel research into the microbes that live in the various environments of the human body (http://nihroadmap.nih.gov/hmp/). A major goal of the HMP is to look for correlations between changes in the microbiome and human health. The Institute for Genome Sciences at the University of Maryland School of Medicine is involved in multiple HMP projects including an investigation of gut microbiota and obesity in the Amish population, the structure of the gut microbiome in Crohn's disease, and the microbial ecology of bacterial vaginosis. In addition, we provide the Data Analysis and Coordination center for the HMP, providing a specialized data management and analysis infrastructure to support the collection, integration and standardization of HMP data to facilitate research (http://hmpdacc.org). The intersection of second generation sequencing technologies and the field of metagenomics is driving an explosion of data and presenting unique informatics challenges. I will discuss a number of these challenges and how we are attempting to tackle them in the context of our HMP projects.

BUILDING PHYLOGENIES WITH METAGENOMIC SEQUENCE READS

Samantha J Riesenfeld[1], Thomas J Sharpton[1], Steven W Kembel[2], Jessica L Green[2], Katherine S Pollard[1]

[1]Gladstone Institutes, University of California, San Francisco, 1650 Owens St, San Francisco, CA, 94158, [2]University of Oregon, Center for Ecology and Evolutionary Biology, 335 Pacific Hall, Eugene, OR, 97403

Metagenomic shotgun sequencing permits the study of genetic material recovered from environmental samples, without isolation or culturing. This technique has identified millions of previously unknown genes in microbial communities in situ, ranging from the ocean to the human gut. Phylogenetic trees enable powerful analyses of taxonomy, community diversity, and evolutionary patterns. It is not obvious, however, how to quantify evolutionary distance between fragmentary, non-overlapping metagenomic sequence reads. Our aim is to establish whether metagenomic phylogenies can be reliably constructed for individual gene families.

Most molecular phylogenetic methods depend upon a multiple sequence alignment. We exploit complete gene sequences from fully sequenced microbial reference genomes to compute probabilistic profiles for gene families. To build alignments from metagenomic data, we match and align reads and reference sequences to the family profiles. From the alignments, we infer phylogenies that enable us to assess phylogenetic relationships between reads that may not overlap in the alignment.

To validate our profile alignment approach and compare performance of tree-building algorithms, we developed a well-parameterized simulation pipeline that considers, in particular, the size and complexity of the simulated community and reference database. As a command-line tool, it can be run in batch to generate large sets of simulated metagenomic data. It leverages existing databases of protein marker genes and 16S rRNA, as well as existing software including MetaSim and HMMER.

We compared phylogenies inferred in the simulated metagenomic context to those inferred from corresponding full-length gene sequences. Initial results indicate that many measures of performance are strongly influenced by the size and breadth of the reference database. We also evaluated the robustness of downstream analyses that rely on trees, such as identifying taxa and measuring diversity within and between metagenomic samples. Our results highlight the potential and limitations of phylogeny-based metagenomic analyses.

MINING 1000 GENOMES DATA TO IDENTIFY THE CAUSAL VARIANT IN REGIONS UNDER POSITIVE SELECTION

Shari Grossman[1,2], Ilya Shlyakhter[1,2], Elinor Karlsson[1,2], Mitch Guttman[2], John Rinn[2], Eric Lander[2], Steve Schaffner[2], Pardis Sabeti[1,2], 1000 Genomes Project Consortium[3]

[1]Harvard University, Center for Systems Biology and Department of Organismic and Evolutionary Biology, 52 Oxford St., Cambridge, MA, 02138, [2]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02142

The human genome contains hundreds of regions whose patterns of genetic variation indicate recent positive natural selection, yet for most the underlying gene and the advantageous mutation remain unknown. We recently reported the development of a method, Composite of Multiple Signals (CMS), that combines tests for multiple signals of natural selection and increases resolution by up to 100-fold.

Applying CMS to candidate selected regions from the International Haplotype Map, we localized several hundred signals to ~50-100kb, identifying individual gene and polymorphism targets of selection. These regions included genes involved in processes known to be targets of selection, such as infectious disease, skin pigment, metabolism, and hair and sweat. We further identified many candidates that are like regulatory elements. In several regions we identified variants that are significantly associated with expression of nearby genes in the selected population. Moreover nearly half of the ~200 regions we examined localized to regions with no genes at all. Thirteen of contain long non-coding RNAs that have been shown to often regulate nearby genes, suggesting that variation within the RNAs may have functional consequences.

With preliminary data now available from the 1000 Genomes Project, we are beginning to explore full sequence data which should contains most if not all of the causal selected polymorphisms. We extended the CMS method to the preliminary dataset, validating our previously identified candidates and identifying many new intriguing coding and regulatory variants.

EPIGENOMIC TRIANGULATION OF HUMAN METHYLOMES REVEALS GERMLINE-SPECIFIC METHYLATION DESERTS ASSOCIATED WITH GENOMIC INSTABILITY

Jian Li, Ronald A Harris, Cristian Coarfa, Zuozhou Chen, Zachary M Franco, Aleksandar Milosavljevic

Baylor College of Medicine, Molecular and Human Genetics, One Baylor Plaza, Houston, TX, 77030

Methylomes and genome-wide maps of histone marks are being produced at an increasing pace and, with the advance of massively parallel sequencing technologies, at an increasing level of resolution. As part of the NIH Epigenomics Roadmap Initiative, we are developing and validating methods for comparative epigenome analysis and the construction of a Human Epigenome Atlas. The human germline methylome was reconstructed based on an evolutionary CpG mutability index and was compared to the H1 (embryonic stem cell line) and IMR90 (fetal fibroblast cell line) methylomes using the newly developed Epigenomic Triangulation method. Cell-lineage-specific demethylation is detected in both IMR90 and the germline. Germline methylation deserts, regions that are extremely hypomethylated specifically in the germline, comprise a total of 1.5% of the human genome and contain 15% of structural rearrangements that occurred in the human genome since the branching of human and chimpanzee. More than 20% of human-specific structural rearrangements are attributable to hypomethylation between extreme and moderate levels in the germline. These results point to a likely role of chromatin structure and the epigenome in mediating structural mutability of the human genome. The results also validate the utility of the Epigenomic Triangulation method for identifying specific biologically significant epigenomic characteristics of individual cell types and for the eventual construction of a Human Epigenome Atlas.

DISEASE MODEL DISTORTION IN ASSOCIATION STUDIES

Eliana Hechter[1,2], Damjan Vukcevic[1,2], Chris Spencer[1,2], Peter Donnelly[1,2]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom, [2]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom

Most findings from genome-wide association studies (GWAS) are consistent with the simplest disease model, in which each additional copy of the associated allele increases disease risk by a multiplicative factor, in contrast to recessive or dominant effects. Using theoretical and empirical results we show that imperfect linkage disequilibrium (LD) markedly distorts these effects, with the power to detect dominant or recessive effects dropping off dramatically. For example, power to detect departures from the simplest model decays as a function of r^4, where r^2 is the usual correlation between the causal and marker loci. Compared to the well-known result that power to detect a multiplicative effect decays as a function of r^2, this is striking. Similar results apply to the detection of interactions among distinct GWAS loci. Disease model distortion may explain the relative paucity of observed non-multiplicative signals of association, and we show that it can also account for some of the "missing" heritability in common diseases.

COMPARATIVE ANALYSIS OF TRANSCRIPTION FACTOR REPERTOIRES IN THE *ASCOMYCOTA*

Jaqueline Hess, Nick Goldman

EMBL, EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, United Kingdom

The nature of the evolutionary processes underlying the generation of macroevolutionary changes has been a long-standing debate in evolutionary biology. Genome-sequencing quickly established that gene content is a poor explanation of complexity of the organism and changes in gene expression have since been a popular explanation for the organismal complexity unaccounted for by differences in gene content.

The exact mechanisms underlying gene expression divergence and their adaptive significance have been the subject of intensive debate in the recent past. The last few years have seen a wave of studies starting to address those questions on a systematic level and we are beginning to understand some of the relationships between sequence evolution and gene expression. Cross-species studies of *cis*-regulatory DNA and conservation of transcription factor (TF) binding suggest extensive evolutionary plasticity of transcriptional regulatory networks. A systematic study of the evolutionary dynamics experienced by TFs, being the protein components of transcriptional regulatory networks, has yet to be published.

In order to examine the evolutionary pressures experienced by TFs and speculate on the role they play in the evolution of transcriptional regulation, we have collected TF repertoires from 15 species of yeasts whose most recent common ancestor existed about 300 mya. These repertoires comprise all known sequence-specific DNA-binding proteins in those species. Overall, the TFs we recovered contained 50 families of DNA-binding domains and accounted for about four to five percent of the annotated protein-coding genes in each genome.

We found the overall composition of these repertoires to be similar across the species examined. Species- and clade-specific patterns of duplications and losses however indicate extensive rewiring of regulatory pathways even at intermediate evolutionary range. This is also reflected in the substitution rates experienced by the TFs themselves. The DNA-binding domains tend to be highly conserved whereas the remainder of the protein is evolving fast. Here we will discuss the patterns of duplications and losses observed within their topological context of the regulatory network as well as the results of the evolutionary rate analysis and their implications towards the mutational mechanisms governing the rewiring of transcriptional regulatory networks in yeast.

COMPREHENSIVE PAIRED-END-TAG MAPPING REVEALED CHARACTERISTIC PATTERNS OF STRUCTURAL VARIATIONS AND AMPLIFICATION MECHANISMS IN CANCER GENOMES

Axel M Hillmer, Yao Fei, Koichiro Inaki, Wah-Heng Lee, Pramila N Ariyaratne, Hao Zhao, Leena Ukil, Audrey S Teo, Xing Y Woo, Wan T Poh, Kelson F Zawack, X Ruan, Atif Sahab, Valere Cacheux-Rataboul, Guillaume Bourque, Wing K Sung, Edison T Liu, Yijun Ruan

Agency of Science Technology & Research, Genome Institute of Singapore, 60 Biopolis Street, Singapore, 138672, Singapore

Somatic genome rearrangements are thought to play important roles in cancer development. However, our ability for thorough characterization of cancer genome architecture is still very limited. We optimized the paired-end-tag (PET) sequencing approach for analyzing large genomic DNA fragments to study human genome structural variations, and applied this approach for comprehensive characterization of 2 normal, 8 breast cancer and two other cancer genomes. Our analyzes revealed that most deletions, inversions, and insertions are germ line structural variations, whereas tandem duplications, inverted orientations, isolated translocations, and complex rearrangements are over represented as somatic events in breast cancer genomes. Large tandem duplications are probably among the initial rearrangement events that trigger the genome instability for extensive amplification in breast cancer genomes.

FOSMID-BASED MOLECULAR MHC HAPLOTYPE SEQUENCING

Eun-Kyung Suk[1], Jorge Duitama[2,1], Sabrina Schulz[1], Stefanie Palczewski[1], Britta Horstmann[1], Gayle McEwen[1], Stefan Schreiber[3], Roger Horton[1], Thomas Huebsch[1], Margret Hoehe[1]

[1]Max Planck Institute for Molecular Genetics, Vertebrate Genomics, Ihnestr. 63-73, Berlin, 14195, Germany, [2]University of Conneticut, Dept. Computer Science & Engineering, 371 Fairfield Rd.,Unit 2155, Storrs, CT, 06269-2155, [3]University Kiel, Institute for Clinical Molecular Biology, Arnold-Heller-Str. 3, Kiel, 24105, Germany

The use of next generation sequencing approaches has expanded the analysis of genetic variation in human genomes. The amount of SNPs and extent of structural variations have not yet reached a plateau. In addition, an individual pair of chromosomes may differ in gene and sequence content, thus, mixed diploid sequencing may not be sufficient to fully resolve the underlying molecular haplotype sequences. A prime example illustrating these limits is the human major histocompatibility complex (MHC), recognized as the most important genomic region related to common diseases.
We have established key resources and technologies to sequence and assemble molecular haplotypes separately: 1) A unique 'Haploid Reference Resource' of 100 human fosmid libraries from a representative German population cohort, each library with 1.44 million fosmids pooled in units of 5000 cfu (~ 5% of the human genome), and 2) a SOLiD next generation sequencing (NGS) and data analysis pipeline including fosmid specific detection modules. MHC haplotype sequence information is generated by use of targeted enrichment of MHC sequences and complemented by NGS of selected MHC haplotype-informative fosmid pools.
Up to now, we have generated more than 60 GB of unique data, providing the basis for a fosmid-based phasing of the ~4 Megabase MHC region. Preliminary results demonstrate major advantages of this approach as compared with data obtained on genomic DNA.

GENOMICS OF PIGMENT PATTERNS: FROM AKITAS TO ZEBRAS

Lewis Hong[1], Chris Kaelin[2], Greg Barsh[1,2]

[1]Stanford University School of Medicine, Department of Genetics, 279 Campus Dr, Stanford, CA, 94305, [2]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL, 35806

Stripes and spots in multicellular animals are a prominent feature in nature that contributes to camouflage, species recognition and morphologic diversity. The molecules and pathways that give rise to different types of pigment have been well-characterized from the genetics of mouse coat color, but patterned control of those pathways in animals such as cheetahs and zebras is still a mystery. However, the scale of next gen sequencing technology and the increasing availability of mammalian genome sequences has allowed us to investigate the basis of pigment patterns using a genomic approach.

We developed a highly sensitive and robust methodology—**E**coP15I-tagged **D**etection of **G**ene **E**xpression, or EDGE—that is suitable for detecting and comparing gene expression among tissues from animals for whom fully assembled and annotated genomes do not yet exist. Our method is adapted for the Illumina Genome Analyzer platform, and involves building and sequencing a cDNA library where each transcript is represented by a unique 27 bp sequence tag. Consequently, the frequency at which a particular cDNA tag appears in a library serves as a proxy for quantifying and comparing transcript abundance. Because there is a one-to-one correspondence between transcript and tag, the method allows gene expression differences to be assayed in organisms where genome sequence or transcript annotation is not yet available.

In a pilot study, we applied EDGE to a mouse model for red hair, fair skin, and skin cancer susceptibility, and discovered a set of co-regulated genes that lie downstream of the melanocortin 1 receptor (Mc1r) but which control non-pigmentary phenotypes. We then carried out an EDGE analysis of patterned skin from dogs (yellow and black brindled stripes), cheetahs (yellow and black spots), and zebras (white and black stripes). In brindled dogs, the striping pattern is caused by a segmental duplication that leads to gene silencing, and our results suggest that epigenetic alterations in gene expression are confined to the duplicated segment. In zebras, the striping pattern is limited to hair rather than skin, and our results indicate that alterations in hair color are accompanied by alterations in hair structure.

MODENCODE: PROMOTER ARCHITECTURE IN THE *D. MELANOGASTER* EMBRYO

Roger Hoskins[1], Jane Landolin[1], Ben Brown[2], Jeremy Sandler[1], Nathan Boley[2], Thomas Kaufman[3], Brenton Graveley[4], Joseph Carlson[1], Piero Carninci[5], Susan Celniker[1]

[1]LBNL, LSD, Berkeley, CA, 94720, [2]UC, Dept of Statistics, Berkeley, CA, 94720, [3]IU, Dept of Biology, Bloomington, IN, 47405, [4] UCHC, Dept of Genetics & Dev Bio, Farmington, CT, 06030, [5]RIKEN, Omics Science Ctr, Yokohama, 230-0045, Japan

Core promoters are primary sites of gene regulation, and transcription start sites (TSS) are primary indicators of promoter activity. To identify and characterize active promoters, we determined TSSs within promoters of long capped transcripts expressed in the *D. melanogaster* embryo using integrative analysis of four data types: Cap Analysis of Gene Expression (CAGE [42M 27nt reads]), RNA Ligase Mediated Rapid Amplification of cDNA Ends (RLM-RACE [2.1M ~130nt reads]), cap-trapped ESTs [71K ESTs] and RNA-seq (SOLID [100M 50nt stranded reads]). We modeled the biochemical background in the CAGE assay as proportional to the signal RNA-seq signal. Thresholding under this model reduced mapping of the CAGE reads from 5M to 300K sites, which cluster into 45K distinct peaks. Integration with EST and RACE data supports ~10K of the strongest CAGE peaks, which account for >80% of the signal. We have discovered hundreds of unannotated promoters.

These data also resolve promoter boundaries and TSS distributions. About 20% of promoters have a peaked TSS distribution, usually < 8 bp wide. The other 80% have broad TSS distributions. TF binding motifs are differentially enriched in peaked and broad promoters, indicating different regulatory modes.

We have also discovered CAGE peaks in annotated 3' UTRs, many of which are supported by RNA pol II peaks from ChIP assays, some of which are supported by ESTs. Sequencing of cDNAs corresponding to these ESTs defines transcripts of ~1 kb that are polyadenylated at the annotated 3' end of the parent protein-coding transcript. Similar transcripts have been described in mammals. These CAGE peaks may mark promoters of primary transcripts of miRNAs, piRNAs or TASRs, or they may represent a previously reported re-capping phenomenon.

GENOTYPE IMPUTATION WITH THOUSANDS OF GENOMES: NEW METHODOLOGY AND APPLICATIONS IN AFRICA

Bryan N Howie[1], Jonathan Marchini[2], Matthew Stephens[1]

[1]University of Chicago, Human Genetics, 920 E. 58th St, Chicago, IL, 60637, [2]University of Oxford, Statistics, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom

Genotype imputation has helped discover many putative disease loci in people of European ancestry, but it has been less successful in other populations. This challenge can be addressed by better imputation algorithms and better reference panels; our recent work spans both of these approaches.

On the algorithmic side, we have developed an extension of IMPUTE2 that can handle large and diverse reference panels. By modeling local genealogies, our method allows every individual to choose custom reference panels in different parts of the genome. This leads to many practical advances, such as the ability to impute genotypes in modern African populations using African American reference data.

To assess the prospects for new reference panels like those being generated by the 1,000 Genomes Project, we applied our method to three African datasets collected through the Malaria Genomic Epidemiology Network (MalariaGEN) and genotyped on the Illumina 650Y array: 666 trios from Gambia, 608 trios from Ghana, and 122 trios from Malawi. After removing the children, we constructed a panel of 100 individuals from each country to mimic three target panels of the 1,000 Genomes Project. We used the remaining parents to assess imputation accuracy in cross-validation experiments with: (1) a well-matched reference panel of 100 individuals; (2) a "cosmopolitan" reference panel of 300 individuals; (3) a HapMap 3 panel containing 2,022 haplotypes from around the world. For comparison, we also performed a number of cross-validations within HapMap 3.

We found that imputation quality in African populations remains low, even when large and well-matched reference panels are available. Nonetheless, the 1,000 Genomes project is likely to make meaningful improvements over HapMap 3, even at the relatively common SNPs considered in this study. IMPUTE2 can be applied now to make good inferences from the HapMap 3 data, and it will be even more effective when the full 1,000 Genomes dataset is finished. Despite previous findings that strong selection pressures and population structure can hinder the transfer of reference data between African populations, we find that it is usually better to use a large, cosmopolitan reference panel than a smaller, well-matched panel.

THE GENOMES OF THE ARGENTINE AND RED HARVESTER ANTS

Hao Hu[1], Aleksey Zimin[2], Jay Kim[3], Juergen Gadau[4], Hugh Robertson[5], Andrew V Suarez[5], Christopher R Smith[6], Neil Tsutsui[7], Mark Yandell[1], Christopher D Smith[3]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, 84112, [2]University of Maryland, Genome Assembly Group, College Park, MD, 20742, [3]San Francisco State University, Department of Biology, San Francisco, CA, 94132, [4] Arizona State University, School of Life Sciences, Tempe, AZ, 85287, [5]University of Illinois at Urbana-Champaign, Department of Entomology, Urbana, IL, 61801, [6]Earlham College, Department of Biology, Richmond, IN, 47374, [7]UC Berkeley, Department of Environmental Science, Policy and Management, Berkeley, CA, 94720

Ants represent a significant portion of the biomass in terrestrial ecosystems and play critical roles in nutrient cycling. Invasive ants can drastically alter native arthropod biodiversity and protect agricultural pests. Ants are renowned for their behavioral complexity and division of labor, yet few molecular or genomic resources are available to facilitate study of ant biology. Here we present the draft genome assembly and annotation for the first two ant species, *Linepithema humile* (Argentine ant, 251Mb) and *Pogonomyrmex barbatus* (Harvester ant, ~260Mb). Both species were sequenced using next-generation pyrosequencing technologies, assembled, and annotated in only a few months for less than $100,000 each. The assembled haploid genomes are 215 Mb and 235 Mb and represent 86% and ~90% of the estimated total genome size for *L. humile* and *P. barbatus*, respectively. Scaffolds were assembled using the Celera assembler and annotated using the MAKER genome annotation pipeline using transcriptome data. Preliminary analyses reveal that both ants have over 98% of conserved CEGMA protein functional domains. Preliminary annotations support the presence of complete *de novo* DNA methylation and RNA interference gene systems as well as numerous chemosensory proteins. Both ants have a rich diversity of transposable elements and repetitive sequence profiles more similar to solitary wasps than to the only other sequenced social insect, the honeybee. These two ant genomes provide the basic information necessary for future molecular studies in other ants, comparison to other social and solitary hymenopterans, and to provide insights into the genetic components of behavioral complexity.

HIGH THROUGHPUT SEQUENCING AND APPLICATIONS AT ILLUMINA

Sean Humphray[1], Vincent Smith[1], Klaus Maisinger[1], Stephen Rawlings[1], Carolyn Tregidgo[1], Francisco Garcia[2], Mark Wang[2], Geoff Smith[1], Kevin Hall[1], David Bentley[1]

[1]Illumina Inc, R&D, Chesterford Research Park, Cambridge, CB10 1XL, United Kingdom, [2]Illumina Inc, R&D, 9885 Towne Centre Drive, San Diego, CA, 92121

The Illumina Genome Analyzer has been instrumental in transforming a wide range of genomic, genetic and functional studies in many laboratories. The principle for sequencing by synthesis (SBS) on this platform is well-established: genomic DNA fragments are attached to the modified surface of an eight-channel flowcell to form high-density random arrays, amplified to form clusters and then sequenced using reversible terminator chemistry by cycles of incorporation, imaging and deblock of deoxynucleotide triphosphates that are fluorescently labelled in a base-specific manner. During the past twelve months we have introduced chemistry, hardware and software developments which have resulted in progressive improvement of the performance and utility of the system. Improved image analysis software enables delineation of clusters at higher resolution, leading to increases in data density and yield. Image processing has also been fully automated for near real-time analysis (RTA) to allow deletion of images after each cycle. The sequencing chemistry has been optimized to improve reaction efficiency and cycle time, yielding improved accuracy and enabling runs of 60 or 95 Gb (2 x 100 or 2x150 base paired reads). The SBS system has also been transferred to a new platform, the HiSeq2000, which employs a 4-camera line-scanning epifluorescent detection system for fast sequencing (25 Gb/day). Data from 1 billion clusters are collected by imaging both top and bottom surfaces of two independently operated flowcells during a single run of paired 100 base reads and generates ~200 Gb of purity filtered data. The recent technology developments on both platforms are being implemented in production sequencing and used in multiple applications including cancer genome sequencing and whole transcriptome expression studies.

AGE-DEPENDENT RECOMBINATION EVENTS IN HUMAN PEDIGREES.

Julie Hussin, Marie-Helene Roy-Gagnon, Gregor Andelfinger, Philip Awadalla

Ste Justine Research Centre, University of Montreal, Pediatrics, 3175, Chemin de la Côte-Sainte-Catherine, Montreal, H3T1C5, Canada

Recent studies have demonstrated that the number and location of meiotic recombination events influences the likelihood of meiotic non-disjunction in humans. For reasons that remain unknown, increasing maternal age is the only factor indisputably known to modulate the risk of meiotic non-disjunction. Furthermore, in rodents, the frequency of recombination events of oocytes is reported to decrease with age. We focus here on determining whether recombination rate is related to the age of the mother in humans. We present a study of French-Canadian multi-generation pedigrees. This study focuses on a very dense genome-wide SNP survey (6.0 Affymetrix platform) genotyped among all 478 individuals from 89 pedigrees. We localized crossovers at high spatial resolution and observed similar variation in fine-scale recombination rates and patterns as previously observed in Hutterites families (Coop et al. Science. 2008). However, we observed that viable offspring of older mothers tend to have significantly reduced recombination rates, and the most pronounced effect is seen between mothers, particularly for mothers over the age of 30. The observation is a genome-wide effect but, among submetacentric chromosomes, the effect was significantly far more pronounced, suggesting a subtelomeric effect. This result contradicts a previous study that found a positive correlation between maternal recombination counts of an offspring and maternal age (Kong et al. Nature Genetics. 2004) based on ~1000 markers. Furthermore, we observed that both males and females exhibited significant increases in recombination moving distally from the centromere, although males exhibit more pronounced elevated rates of recombiniation in sub-telomeric regions, as previously reported. We postulate that, in humans as in other species, recombination rates decrease with maternal age, and that age-related altered recombination patterns may be implicated in age-dependant meiotic non-disjunction leading to aneuploidy.

EASY, ACCURATE GENOME-WIDE DETECTION OF GENE FUSIONS WITH THE SOLID SYSTEM USING BIOSCOPE SOFTWARE

Fiona C Hyland, Onur Sakarya, Heinz Breu, Liviu Popescu, Paolo Vatta, Asim Siddiqui

Life Technologies, Biological Information Systems, 850 Lincoln Center Drive, Foster City, CA, 94404

Chromosome aberrations, especially gene fusions, are implicated in the initiation of tumorigenesis. Gene fusions are important diagnostic and prognostic indicators in leukemia, sarcomas, and other solid tumors.
The high throughput of massively parallel sequencers (up to 1 billion mapped reads on a single run of the SOLiD™ System 4.0) enables genome-wide hypothesis-free detection of gene fusions. The availability of DNA barcoded paired-end reads facilitates cost-effective concurrent sequencing of many samples on a single slide.
We developed a novel algorithm to detect exon junctions with strand-specific SOLiD transcriptome sequencing, to predict fusion transcripts and to facilitate identification of alternative splicing. For paired-end whole transcriptome sequencing experiments, we require two types of evidence; the two tags must map uniquely to two different exons, and a single tag must span the exon junctions. Single-read placement is done with a novel suffix array single read junction finder algorithm that allows fast detection of reads split between any pair of exons. Gene fusions are called if there is sufficient non-redundant evidence of both types, mapping to exons in two different genes.
We validated the algorithm with real and simulated data. We use real reads spanning existing exon boundaries, and simulate gene fusions. We detect >80% of these simulated human gene fusions at exons having sufficient unique coverage.
We sequenced UHR (Universal Human Reference sample), and we predicted 36 fusions using initial thresholds. We validated these predictions with TaqMan. 44% of our predictions were correct, including the three previously annotated gene fusions in UHR; two inter-chromosomal (BCR-ABL1 and BCAS4-BCAS3) and GAS6-RASA3 on chr13. We validated an additional 13 novel gene fusions in this sample, many of which are putative read-throughs from genes in close proximity. Testing these candidate fusions in the breast cancer cell line MCF7 reveals that 13 of these gene fusions are also present in MCF7.
Easy and low-cost genome-wide detection of novel gene fusions allows interrogation of large numbers of tumor samples and the discovery of biologically important gene fusions.

THE CARTAGENE GENOMICS PROJECT: SYSTEMS BIOLOGY OF HUMAN FUNCTIONAL VARIATION

Youssef Idaghdour, Julie Hussin, Philip Awadalla

University of Montreal, Sainte-Justine Research Center, 3175 Chemin de la Côte Sainte-Catherine, Montreal, H2V 2V3, Canada

An area of fundamental biomedical research that merges population and quantitative genomics is the identification of transcriptional and other intermediate biomarkers for disease susceptibility and disease status. These biomarkers are quantitative traits whose architecture is modulated through mechanisms that can incorporate genetic and environmental cues. But what are the relative magnitudes of these effects, and how gene expression profiles correlate with other haematological and clinical phenotypes? To address these questions, we generated gene expression profiles from peripheral blood samples from a random aging cohort in Québec, Canada coming from a number of different geographic locations and lifestyles and for which clinical material and medical information spanning a wide range of medically relevant phenotypes from a well characterized and extended pedigree have been collected. We documented latent structure in gene expression profiles and identified genes and networks of genes that best correlate with haematological and metabolic syndrome phenotypes. The analysis is being extended to incorporate genotypic data in order to identify the genetic control points of the transcriptional and clinical traits studied.

A JOINT-GENOME GRAPH OF THE 1000 GENOMES PROJECT DATA REVEALS MANY HIGHLY DIFFERENTIATED GENOMIC REGIONS

Zamin Iqbal[1], Gil McVean[1,2], The 1000 Genomes Project[1]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom, [2]University of Oxford, Department of Statistics, South Parks Road, Oxford, OX1 3TG, United Kingdom

A fundamental goal in population genetics is to understand the extent to which natural selection has shaped evolution, and to understand the nature and function of variants that have been selected for in the origin and diversification of a species. Population-scale genome sequencing offers the potential to identify genetic variation in a manner that is not biased by the type of variant or the completeness of the reference sequence. However, most current approaches to the analysis of genome-sequence data require the mapping of sequence reads to an assembled reference genome and in doing so inevitably leads to the loss of sequence that is novel or unassembled in the current reference.

To overcome these problems we have developed a novel approach to the analysis of genomic variation in which we de novo assemble a coloured de Bruijn graph from the sequence reads of a population. Applying the method to 1.9 Terabases of data from 179 individuals sequenced within the 1000 Genomes Project we can identify variants through searching for characteristic motifs within the graph and identify regions of strong population differentiation through the presence of nodes in the graph with strong frequency differences (greater than 80% difference) between populations. This approach reveals a remarkable density of high-differentiation variants of at least 1 per Mb in the human genome. Such differentiation is extremely unlikely to occur by chance alone and is most likely the result of strong local adaptation. In addition to SNP variants known to have experienced strong selection we find many indel and structural variants with strong differentiation that map to genomic features of diverse function, from antigen recognition to centromere formation.

HMM-SEG – A NOVEL FRAMEWORK FOR IDENTIFICATION OF COPY NUMBER CHANGES IN CANCER FROM NEXT-GENERATION SEQUENCING DATA

Sergii Ivakhno[1], Keira R Cheetham[2], Tom Royce[3], Dirk Evers[2], David R Bentley[2], Simon Tavaré[1]

[1]CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, United Kingdom, [2]Illumina Cambridge, Chesterford Research Park, Little Chesterford, CB10 1XL, United Kingdom, [3]Illumina Inc., Corporate Headquarters, 9885 Towne Centre Drive, San Diego, CA, 92121

Copy number abnormalities (CNAs) represent an important type of genetic mutation that can lead to abnormal cell growth and proliferation. Microarray technologies have been used successfully to identify CNAs in cancer, but their resolution is limited by the location of probes on the array. New high-throughput sequencing technologies promise comprehensive characterization of CNAs. In contrast to microarrays, where probe design follows a carefully developed protocol, reads represent a random sample from a sequence library and may be prone to representation biases due to GC-content and other factors. With the deep coverage achievable from decreasing sequencing costs, the discrimination between true and false positives becomes an important issue.

Here we present a novel approach, called HMM-seg, to identify CNAs using next-generation sequencing data. It uses depth of coverage to estimate copy number states and flowcell-to-flowcell variability in cancer and normal samples to control the false positive error rate. The algorithm comprises several steps: the GC-normalized read counts are de-noised using a discrete wavelet transform, segmented with an HMM, and followed by a segment merging step, where inter-lane variability in cancer and control samples is used to estimate the merging threshold. It uses the Skellam distribution to compare read depth in tumour and control samples, which allows smaller window size for count estimation.

We tested the method using COLO-829 melanoma cell line sequenced to 40-fold coverage on Illumina GAII [1]. An extensive simulation scheme was developed to recreate different scenarios of copy number changes and depth of coverage based by altering a real dataset with spiked-in CNAs. Comparison to alternative approaches using both real and simulated datasets showed that HMM-seg achieves superior precision and improved sensitivity estimates.

[1] Pleasance, E., et al. (2010) Nature, 463, 191-196

FROM GWAS TO FUNCTION USING SENSITIZED STRAINS, TRANSGENIC RESCUE AND GENE KO

Howard J Jacob[1.2.3], Aron M Geurts[1,2,3,5], Rebecca Schilling[1], Angela Lemke[1], Shawn Kalloway[1.6], Jamie Foeckler[1,6], Jason Klotz[1], Hartmut Weiler[1,6], Jozef Lazar[1,6], Melinda R Dwinell[1,6], Carol Moreno[1,3], for GO Grant Team[1]

[1]Medical College of Wisconsin, 1. Human & Molecular Genetics Ctr. 2. Physiology 3. Pediatrics 4. Dermatology 5. Cardiovascular Ctr 6. Blood Research In, Milwaukee, WI, 53226

GWAS and other types of genetic studies have nominated genes contributing to hypertension and renal failure. Functional studies must now be undertaken for these genes. The rat is the dominant model for the physiological assessment of the cardiovascular system, but has lacked the ability to target genes. Using zinc finger nucleases, which takes a fraction of the time of ES cell technology, we have KOed 30 genes in 6 months. An additional 70 genes will be KOed over the next 18 months. As hypertension is caused by genes, multiple risk factors and environmental factors, we are using a sensitized screening strategy whereby we deploy candidate gene KO on top of a genetically susceptible, but not necessarily hypertensive strain. For select genes we also use transgenic rats over expressing a gene of interest.

Phenotyping of the first few strains using our high throughput physiology program for BP, renal function, vascular function and response to the top three classes of anti-hypertensive drugs is just beginning. We have studied Rab38 a gene responsible for a QTL for renal failure and a host of other traits associated with protein trafficking using both transgenic rescue and site directed gene KO, and replicated the entire range of phenotypes originally found in the congenic animal. Finally, we KOed the renin gene (the renin-angiotensin system is the dominant pathway used to treat hypertension) in a low renin (common in African Americans) model of hypertension. Such a KO would not be expected to modify BP. Mean Arterial Pressure in these KO animals was a remarkably low 60 mmHg demonstrating that low renin hypertension does not mean this system is not playing a role in blood pressure maintenance, but rather that it is not contributing to the development of low renin hypertension, which is a novel finding.

UNDERSTANDING AND REMEDIATING SEQUENCING BIAS

David B Jaffe, Michael G Ross, Sean Sykes, Dan Aird, Aaron M Berlin, Kristen Connolly, Jim Meldrim, Sarah K Young, Sheila Fisher, Andreas Gnirke, Carsten Russ, Chad Nusbaum

Broad Institute, Sequencing, 320 Charles St., Cambridge, MA, 02142

Each sequencing technology performs poorly on certain motifs, for example extremes of base composition, homopolymer runs or dinucleotide runs. As a result of this bias, coverage of a genome by reads is uneven, and coverage by high quality bases is generally even more uneven. Increasing data density on the Illumina sequencing instrument (to increase coverage) can actually exacerbate bias, further reducing coverage of difficult motifs.

Sequencing bias has a deleterious effect on many applications:
• Variants cannot be discovered if there is no coverage, and uncovered loci can be biologically important. For example, in some datasets, as a result of their proximity to CpG islands, thousands of human promoter regions have essentially no coverage.
• De novo assemblies can be shredded by loss of coverage at recalcitrant motifs, and this appears to be the single greatest challenge for genome assembly. The impact of bias varies widely among genomes. Notably challenging genomes include medically important pathogens at GC extremes such as *P. falciparum* (19% GC) and *M. tuberculosis* (66% GC). Bias can also have dramatic effects on assembly of genomes whose overall GC content is intermediate, but which are punctuated by difficult regions.

We developed computational assays to facilitate comparisons of bias between and across genomes, sequencing technologies, and datasets, and to drive laboratory experiments that could mitigate the problem. These assays take as input the community standard BAM files. Using these assays we compared and contrasted bias in reads from eight technologies.

Here we describe four key categories of challenging sequence motifs, hypothesize their physical causes, and outline a plan of attack to improve performance. These categories are high GC, G runs, AT runs, and very low GC. We demonstrate a reduction in sequencing bias obtained via assay-driven improvements to laboratory protocols.

EFFICIENT MAPPING OF BRAIN EQTL USING PERIPHERAL BLOOD AS SURROGATE TISSUE.

Anna Jasinska[1], Susan Service[1], Lynn Fairbanks[1], Matthew Jorgensen[4], David Jentsch[3], Roger Woods[2], Nelson Freimer [1]

[1]Univ. of California, Center for Neurobehavioral Genetics, 695 Ch. E. Young Dr. South, Los Angeles, CA, 90095, [2]Univ. of California, Dept. of Neurology, Brain Mapping, Los Angeles, CA, 90095, [3]Univ. of California, Dept. of Psychology, 8441B Franz, Los Angeles, CA, 90095, [4] Wake Forest Univ. Health Sciences, Dept. of Pathology, Medical Center Blvd, Winston-Salem, NC, 27157

Non-human primate (NHP) species are important models to study molecular mechanisms underlying neurobehavioral processes relevant to human neuropsychiatric phenotypes, but which cannot be modeled in species more distantly related to humans such as rodents. Here we use an extended pedigree of vervet monkeys (Chlorocebus aethiops sabaeus) phenotyped for a wide range of brain-related traits to present an approach for mapping expression quantitative trait loci (eQTL) acting in brain, using peripheral blood as a surrogate tissue. Although blood is frequently used as a surrogate for tissues that are difficult to obtain, its utility has often been limited by insignificant or conflicting results, as only a fraction of the transcriptome is readily comparable between tissues. We developed a strategy to maximize such comparability, utilizing data from individuals in whom we obtained gene expression profiles in both brain and blood, to select a subset of transcripts (a) in which brain expression levels are well reflected in peripheral blood and (b) which display greater inter-individual variation than intra-individual variation. We have confirmed our hypothesis that these selection criteria would identify good candidates for eQTL mapping. We previously found significant heritability in 29 out of 32 transcripts selected by these criteria, based on gene expression levels in blood in almost ~350 pedigree members. We now have genetically mapped eQTL for 12 of the 29 heritable candidate transcripts, using microsatellite genotype data from 261 STRs in the pedigree. We identified 10 cis eQTL and 2 trans eQTL at a genome wide significance threshold (>LOD 3). Next steps will include refinement of the linkage regions and studies of correlations between eQTL transcripts and other complex traits investigated in the vervet pedigree.

CHARACTERIZATION OF COPY NUMBER CONSTANT REGIONS

Anna C Johansson, Lars Feuk

Uppsala University, Department of Genetics and Pathlogy, Rudbeck Laboratory, Uppsala, 75185, Sweden

Over the last few years there has been a major drive to identify submicroscopic structural variations in the human genome, ranging from a few hundred base pairs to some Mb. The main type of variations found are copy number variations (CNV), which have been shown to have a significant impact on gene expression and genome architecture. As more CNVs are identified, the portion of the human genome that is known to be variable increases. Highly variable regions and the genes within them have been carefully characterized, but less effort have been put into characterizing the non-variable regions. Our aim is to define a list of genes and elements that are dosage sensitive in humans, which may facilitate interpretation of novel CNVs in patient data. Here we attempted to define these Copy Number Constant regions and the elements and genes that are associated with them.

We used CNV data from the The Database of Genomic Variants, which contains experimentally defined structural variations > 1kb, identified in healthy controls. To exclude false positives and rare variants, only regions reported in at least two different studies were included. In the resulting dataset 92% of the nucleotides were considered as non-variable. To more easily characterize the non-variable regions, the genome was divided into 500kb bins analyzed independently and the non-variable bins comprise 64% of the genome. In addition, we identified the longest stretches of non-variable regions in the genome and regions located at a maximum distance away from any variation. We then characterized these regions in correlation to a large number of genomic features. All results were compared to the average for the entire genome, the highly variable regions and randomized datasets.

Our results show that non-variable regions have a lower gene-density and contain longer genes than variable regions. The genes within these regions also contain a higher fraction of OMIM-genes, cancer related genes and lethality-phenotype associated genes identified through knock-out experiments in mouse. These correlations are stronger in longer non-variable segments and regions located farther from the closest variations. The SNP density is decreased in these regions and important functional elements such as enhancers, miRNA and ultra-conserved regions are more abundant.

Genes within the non-variable regions are enriched in GO-terms associated with developmental processes and transcriptional control. We use these features to identify a list of genes that are highly dosage sensitive, causing large regions around them to also be structurally conserved.

A MODEL OF REGULATORY PROGRAM DIFFERENTIATION IN IMMUNE CELL DEVELOPMENT

Vladimir Jojic*[1], Tal Shay*[2], The Immunological Genome Project Consortium[3], Aviv Regev[2,4,5], Daphne Koller[1]

[1]Stanford University, Computer Science, 353 Serra Mall, Stanford, CA, 94305, [2]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02142, [3]UCSD, Fox Chase, U. Mass, UCSF, HMS, Mount Sinai, DFCI, Harvard, MA,[4] MIT, Biology, 1 Ames Street, Cambridge, MA, 02139, [5]Howard Hughes Medical Institutes, 4000 Jones Bridge Road, Chevy Chase, MD, 20815

* Equal contribution

We introduce a novel model of gene expression aimed at uncovering the differential regulation that underlies immune cell development. This model builds on a body of work in module network reconstruction and extends it by permitting within module regulatory program variation. This new framework trades off two competing desiderata. First is the preference for preserving regulatory programs among related stages in cell development. Second is the flexibility to capture sudden and substantial changes in regulator roles, for example, a switch from activator to repressor. Importantly, the fitting of the model is achieved in a statistically consistent manner which prevents the introduction of spurious regulators or regulator role change events. We demonstrate the utility of this model by analyzing ImmGen compendium data. This data consists of gene expression measurements from 118 distinct cell types of the mouse immune system. In this analysis, the model leverages a developmental tree to encourage conservation of the regulatory programs between daughter and parent cells. Thus, regulatory programs preserved throughout a particular lineage are recovered, in addition to deviations specific to sublineages. We present examples of differential regulation specific to T-cells, B-cells and NK lineages.

ASSESSING THE RELATIONSHIP BETWEEN FREQUENCY AND RISK IN COMPLEX DISEASE

Luke Jostins, Jeffrey C Barrett

Wellcome Trust Sanger Institute, Human Genetics, Wellcome Trust Genome Campus, Hinxton, CB10 1HH, United Kingdom

For evolutionary reasons, it is believed that rare genetic variants that contribute to complex disease should have larger effect sizes (relative risks or odds ratios) than more common variants. The results of Genome-Wide Association Studies (GWAS) and candidate resequencing studies support this hypothesis, and have lead many to suspect that a glut of rare large-effect variants will be discovered by case-control resequencing. However, a number of potentially confounding effects on the observed relationship between frequency and risk throw the usefulness of these observations into doubt.

Firstly, the lower power to detect disease association in rare variants means that they are more likely to be near the P-value threshold, and thus more likely to be affected by the Winner's Curse. Secondly, there is a larger sampling variance associated with risk estimates for rare variants; we show that this can cause drastic inflation of mean risk estimates. Thirdly, the lower power to detect rare variants means that only those with large effect sizes are picked up in GWAS, compared to a less skewed effect size range for common variants; this introduces a sampling-induced frequency-risk bias.

We demonstrate how these effects can be compensated for in order to detect the existence of an underlying frequency-risk relationship. The Winner's Curse can be compensated for using established statistical methods, or by estimating risk in independent cohorts, and the effect of increased sample variance can be estimated in a logistic regression framework.

By combining expressions for the power to detect disease association, multiplicative models of genetic risk, and a model of risk prevalence, we show how the observed relationship between frequency and risk in sampled data can be predicted, conditional on there being no underlying relationship. These predictions can be used to compensate for sampling-induced frequency-risk bias.

We apply these techniques to both simulated and empirical case-control data, and evaluate the remaining power to detect frequency-risk relationships after correcting for confounding effects. Our general conclusion is that the majority of the observed frequency-risk relationship at intermediate effect sizes (OR < 2) is due to confounding effects; however, there is evidence that very large effect sizes (OR > 2) are truly enriched in low-frequency variants.

DEEP EXPRESSION PROFILING THROUGH MULITIPLE LIBRARIES GENERATED BY ILLUMINA GA□

Sotaro Kanematsu, Kosuke Tanimoto, Suzuki Yutaka, Sumio Sugano

the University of Tokyo, Medical Genome Sciences, 5-1-5 Kashiwahoha, kashiwa-shi, 277-8562, Japan

Although recent outstanding progress of deep sequencing technology has enabled us to obtain enormous amount of data about human each transcript, we are still facing great difficulties in understanding interaction among transcripts. In order to reveal the molecular basis of gene expression regulation, we employed multiple methods such as RNA-seq, small RNA-seq and RNA Immuno-precipitation (RIP) -seq. First, we used the small RNA-seq method to observe small RNA expression changes in DLD1 (colon cancer cell line) and TIG3 (normal lung fibroblast), where both cultured in normoxic and hypoxic conditions. The total tag number mapped to genome uniquely were: 8623614 tags in TIG3-Normoxic condition, 9506308 tags in TIG3-Hypoxic condition, 8309609 tags in DLD1-Normoxic condition, and 2083635 tags in DLD-Hypoxic condition. 7234226 tags out of 8623614 tags (84%) were mapped to known miRNA positions in TIG3-Normoxic condition, 8520106 tags out of 9506308 tags (90%) in TIG3-Hypoxic condition, 7215951 tags out of 8309609 tags (86%) in DLD1-Normoxic condition, and 186816 tags out of 2083635 tags (9 %) in DLD1-Hypoxic condition. Next, we compared small RNA expression levels between normoxia and hypoxia. Surprisingly, in TIG3 cells cultured in hypoxic conditions, only three miRNAs,(has-mir-200b, has-mir-210, and has-mir-770) showed elevated levels of expression, While in DLD1cells cultured hypoxic condition There were 35 miRNAs with elevated expression.were elevated expression. We are currently constructing Target RNA library by RIP, hoping that the combination of libraries constructed by multiple methods will allow us to better understand our transcriptome.

DESIGN AND VALIDATION OF A NEW, HIGH DENSITY CANINE SNP ARRAY

Elinor K Karlsson[1], Matthew T Webster[2], Snaevar Sigurdsson[1], Catherine Andre[3], Cindy Taylor Lawley[4], Gerli Rosengren-Pielberg[2], Danika L Bannasch[5], Hannes Lohi[6], Merete Fredholm[7], Mark S Hansen[4], Mike Thompson[4], Christophe Hitte[3], Kerstin Lindblad-Toh[1,2]

[1]Broad Inst., Genome Biol, 7 Cambridge Center, Cambridge, MA, 02142, [2]Uppsala U, IMBIM, Box 582, Uppsala, 751 23, Sweden, [3]U de Rennes, Medicine, 2 ave Pr. Léon Bernard, Rennes, 35043, France, [4] Illumina, 9885 Towne Center Drive, San Diego, CA, 92121, [5]UC Davis, Vet Med, 1114 Tupper Hall, Davis, CA, 95616, [6]U of Helsinki, Vet Med, P.O. Box 66, Helsinki, 00014, Finland, [7]U of Copenhagen, Life Sciences, Bülowsvej 17, Frederiksberg, 1870, Denmark

We have developed a new ~170,000 SNP canine array, CanineHD, based on Illumina's High Density BeadChip technology, in collaboration with the LUPA consortium. This array offers more uniform genome coverage and exceptional call accuracy compared to earlier resources, and can assay 12 samples simultaneously. The SNPs were selected from among the 2.5 million discovered in the Dog Genome Project, as well as through targeted resequencing of sparsely covered regions (1696 SNPs). Of those chosen, 93% validated, resulting in a final set 173,662 markers with mean spacing of 13kb and just 21 gaps larger than 200kb. To validate the panel, we genotyped: 1) 352 dogs representing 26 breeds; 2) two families, several trios and replicates; 3) the CanFam2.0 reference boxer; 4) buccal whole genome amplified (WGA) samples. Call rates and reproducibility exceeded 99% for all breeds genotyped, with blood samples yielding marginally higher call rates (99.8%) than the buccal WGA samples (99.1%). The number of polymorphic markers per breed ranged from 85,193 in Greenland Sledge Dogs, a relatively isolated breed, to 126,387 in Jack Russell Terriers. With even, dense coverage across the genome, high call rates and exceptional accuracy, this array is a significant step forward for canine gene mapping, especially for mulitgenic, complex diseases such as cancer, neurological disease, and autoimmune disease.

AN ALGORITHM TO INFER HAPLOTYPES OF COPY NUMBER VARIATIONS FROM GENOME-WIDE HIGH-THROUGHPUT DATA

Mamoru Kato[1], Naoya Hosono[2], Anthony Leotta[1], Tatsuhiko Tsunoda[2], Michael Q Zhang[1,3,4]

[1]Cold Spring Harbor Laboratory, Cancer Center, 1 Bungtown Road, Cold Spring Harbor, NY, 11724, [2]RIKEN, Center for Genomic Medicine, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan, [3]Tsinghua University, Department of Automation, Beijing, 100084, China, [4] New address: University of Texas at Dallas, Department of MCB, Center for Systems Biology, 800 West Campbell Road, Dallas, TX, 75080

Accurate and complete information on haplotypes and individuals' diplotypes is required for population-genetic analyses; however, microarrays do not provide data on a phased diplotype (*e.g.*, 1 copy in one haplotype and 2 copies in the other) at a copy number variation (CNV) locus, but only provide data on a diploid number, which is the total number of copies/bases over a diplotype (3 copies in the above example). Moreover, the data often include large errors on the estimated diploid numbers when the signal intensities deriving from different diploid numbers are not clearly separated due to noises.

Here we report an algorithm and computational tool to infer CNV haplotypes and individuals' diplotypes from noisy microarray data, incorporating uncertainty due to noises into the expectation-maximization procedure. The uncertainty is represented by likelihoods that a given observed signal intensity may be derived from different underlying diploid numbers. This tool can handle a combination of integer copy numbers, single nucleotide variations in CNV regions (*e.g.*, AAB, unlike AB of single nucleotide polymorphism), and single nucleotide polymorphisms.

We performed simulation studies based on the known diplotypes of 600 individuals in a European population as well as an error model obtained from real microarray data. The new algorithm can outperform previous algorithms not using likelihood information: the accuracy of estimated haplotype frequencies is increased from 78-92% to 99-99.6%. In addition, our algorithm can also correct diploid numbers wrongly determined due to noises: the rate of wrong diploid numbers is decreased from 17-40% to 4-10%. Our algorithm thus enables one to accurately analyze the population-genetic nature of CNVs and to attain a greater statistical power in disease association studies.

A COMPREHENSIVE WHOLE-GENOME MAP OF ENDOGENOUS RETROVIRAL ELEMENTS AND THEIR FUNCTIONAL EFFECTS ACROSS 17 LABORATORY MOUSE STRAINS

Thomas M Keane[1], Kim Wong[1], Jim Stalker[1], Richard Mott[2], Jonathan Flint[2], Wayne Frankel[3], David J Adams[2]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom, [2]Wellcome Trust Centre for Human Genetics, Oxford University, Oxford, Oxford, OX3 7BN, United Kingdom, [3]The Jackson Laboratory, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME, 04609

It has been estimated that endogenous retroviral elements (ERVs) are a significant source (~10%) of spontaneous germline mutations among laboratory mouse strains. Two high-copy families of ERVs in particular, IAP and ETns, have been found to be responsible for the vast majority of these mutations. We have recently completed deep illumina sequencing of 17 key mouse laboratory strains to between 20-35x depth. In this study, we report the first use of new sequencing technologies data to catalogue the full repertoire of ERVs across these strains.

Our strategy utilises paired-end information to build a highly accurate map of several types of ERVs across the strains. Our method finds clusters of mate-pairs where only one end maps well to the reference genome and the other to a set of ERV reference sequences. From the initial set 1-2kb call regions, we then further refine the calls by using the read depth information to identify the exact breakpoints. To identify insertions that are shared between the reference sequence and the strains, we search for clusters of bridging mate-pairs to already annotated insertions. We assess the accuracy of our calls by comparing the set of calls we make in the C57B6NJ strain sequenced by us compared to the insertions annotated in the C57B6J reference genome. We find extremely high concordance between these two sets with just a few differing insertions. We have also compared our findings to a large set of manually PCR validated insertions falling into introns across five strains and find that our illumina predictions show extremely high accuracy. We have focused on young ERVs that occur in just a single strain and compared this set of ERVs found across the strains to existing expression data finding several examples of gene expression being affected by the occurrence of ERVs. Work is on-going to map these ERVs to QTL regions across the strains.

THE HUMAN MUTATION RATE ESTIMATED USING PROBABILISTIC GENOME-WIDE DE NOVO MUTATION DISCOVERY AND VALIDATION USING HIGH-THROUGHPUT SEQUENCING OF FAMILIES WITHIN THE 1000 GENOMES PROJECT.

Jonathan Keebler[1,2], Donald Conrad[3], Matthew Hurles[3], Reed Cartwright[1,4], Ferran Casals[2], Youssef Idaghdour[2], Eric Stone[1], Philip Awadalla[2], The 1000 Genomes Consortium[3]

[1]North Carolina St. Univ, Bioinformatics Research Center, 840 Main Campus Dr, Raleigh, NC, 27606, [2]Univ of Montreal, Ste-Justine Research Center, 3175 chemin de la cote-Sainte-Catherine, Montreal, H3T 1C5, Canada, [3]Wellcome Trust Sanger Institute, Human Genetics, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, United Kingdom, [4] Univ of Houston, Dpt of Biology&Biochemistry, 369 Science&Research Building 2, Houston, TX, 77204

Two methods were applied to discovering *de novo* mutations within the 1000 Genomes Project Pilot 2 trios (CEU and YRI), producing a direct estimate of the human mutation rate. One approach involves a probabilistic framework that leverages the relatedness between family members to produce a probability for the entire pedigree rather than genotypes at each site. A drawback of jointly inferring genotypes in this manner is exposed by short-read sequencing datasets, which may be enriched with sequencing and alignment errors; errors at a site within one individual can percolate through the pedigree and affect genotype calls for closely related individuals. A second approach uses the MAQ genotype consensus model to independently call genotypes for each individual, potentially avoiding this issue. Through collaboration between two groups, the two methods were applied in tandem, comparing results and uniformly applying filters to create a list of candidate mutations. With the intention of capturing as many of the true mutations as possible, we liberally considered 3,286 sites for the CEU trio and 2,757 sites for the YRI trio as possible mutations. We attempted to validate these candidate *de novo* mutations using targeted genomic DNA sequence capture and massively parallel oligo-ligation sequencing (SOLiD) of available samples within the two pedigrees, allowing inference of the germline or somatic status of each mutation. In addition to simulation results for the pedigree-based approach, we present these mutation findings.

HUMAN POPULATION DIFFERENTIATION IS STRONGLY CORRELATED WITH LOCAL RECOMBINATION RATE

Alon Keinan[1], David Reich[2]

[1]Cornell University, Department of Biological Statistics & Computational Biology, 102A Weill Hall, Ithaca, NY, 14853, [2]Harvard Medical School, Department of Genetics, 77 Avenue Louis Pasteur, Boston, MA, 02115

Allele frequency differences across populations can provide valuable information both for studying population structure and for identifying loci that have been targets of natural selection. Building on the expectation that regions with low recombination rate are more likely to be indirectly affected by natural selection (hitchhiking or background selection), we used the variability of recombination rate across the genome as a tool to learn about how substantially the forces of natural selection have shaped patterns of allele frequency differentiation across human populations. We studied two genome-wide data sets in which SNPs have been uniformly ascertained across the genome, in a manner that is independent of local recombination rate: Perlegen "class A" SNPs and the uniformly-ascertained subsets of HapMap that we previously reported.

We found that in both data sets population differentiation as assessed by inter-cotinental F_{ST} shows a strong inverse correlation with recombination rate, with F_{ST} decreasing by an average of 4% for every 1 cM/Mb increase in recombination rate. As we studied ascertained polymorphisms, the mutagenic effect of recombination is not a confounder as it is for the study of the correlation between *nucleotide diversity* and recombination rate. Instead, the negative correlation between F_{ST} and recombination rate reflects the impact of selection in the last 100,000 years since human continental populations split, which is further supported by a stronger correlation that we observed in coding regions.

Interestingly, the relationship between recombination rate and population differentiation is qualitatively different for F_{ST} between African and non-African populations and for F_{ST} between European and East Asian populations. While the former relationship is pretty linear, the latter exhibits an inverse U-shaped relationship, with a dip in differentiation at very low recombination rate loci, and a significant quadratic term when regressing F_{ST} as a function of recombination rate. These results suggest that varying levels or types of selection have prevailed in different epochs of human history. Studying the differences in the patterns across different epochs should make it possible to gain new insight into the mixture of selective pressures that have shaped human genetic variation.

THE IDENTIFICATION OF REGULATORY MOTIF INSTANCES AND THEIR CHARACTERIZATION IN RELATION TO CHROMATIN MARKS AND TRANSCRIPTION FACTOR BINDING

Pouya Kheradpour[1], Jason Ernst[1,2], Christopher Bristow[1,2], Rachel Sealfon[1], Manolis Kellis[1,2]

[1]MIT, CSAIL, 32 Vassar St, Cambridge, MA, 02139, [2]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02142

We and others have found that conservation can be used to identify functional regulatory motif instances. Here we describe our work in using diverse experimental data in order to characterize these motif instances. We have applied our methods to chromatin modification and TF binding data both from the literature and in the context of the White and Karpen modENCODE groups and the Bernstein ENCODE group.

We find that annotations of chromatin modifications complement conservation in identifying motif instances that are likely to be bound by the corresponding factors. Active modifications from conditions matching when factor binding was surveyed appear to be the datasets most useful in finding bound motif instances. Considering dips in the chromatin signal, which may arise from nucleosome exclusion in the footprint of bound factors, leads to even higher precision. We also find that dips specific to a cell type were preferentially enriched in the motifs of factors expressed in that cell type.

In order to understand the dynamics of factor binding, we investigate motif enrichment patterns in regions unique to a condition. Through this analysis, we identify activator and repressor signatures by examining the relationship between motif enrichments and expression. For example, we predict that STAT5 is a likely activator in K562 cells as its motif is enriched in chromatin states associated with enhancers in concordance with its expression. By focusing specifically on activators, we are able to use this analysis to produce an activator association for the modifications which is consistent with their known roles.

Moving beyond single factors acting in isolation, we have developed signatures to identify cooperation including (1) over-representation of instances near one another, (2) a mutual increase in conservation, (3) and biases in the strands and ordering of their instances. We find that all three of these metrics correlate with known interacting factors. We also find that a small number of motif instances very likely to be bound can be identified by combining conservation with homotypic clustering of motif matches.

PATHWAY ANALYSIS OF INTEGRATED DATASETS SUPPORT SIGNIFICANT GENETIC HETEROGENEITY IN AUTISM

Helena Kilpinen[1,2], Karola Rehnstrom[1,2], Juha Saharinen[2], Dario Greco[2], Teppo Varilo[2], Iiris Hovatta[2], Leena Peltonen[1,2]

[1]Wellcome Trust Sanger Institute, Hum.Genetics, Hinxton, Cambridge, CB10 IHH, United Kingdom, [2]Institute for Molecular Medicine, Finland, FIMM, Tukholmank.8, HKI, 00290, Finland

We have previously used linkage and LD mapping to analyze an extended pedigree from Central Finland (CF) to search for rare, potentially high impact autism susceptibility variants enriched in this isolate. Here, we have used high-density SNP data and an extended family set from CF (n=51) to follow up this study by investigating allelic sharing at previously identified loci, hypothesizing that these genealogically connected individuals would share a common predisposition to their phenotype.

We also analyzed global gene expression in an available subset of ten ASD cases and controls from the CF extended pedigree and combined this information with SNP data through pathway and promoter analysis. We proceeded to replicate our findings in a larger Finnish dataset and multiple other autism datasets, such as the AGRE. However, bearing in mind the family-specificity of most autism-predisposing variants identified so far, we sought primarily for replication on the level of biological pathways.

Based on the SNP data, risk factors of autism show considerable locus heterogeneity even within genealogically connected individuals, and no significantly enriched haplotype sharing was detected in the isolate. CNV analysis revealed two large (~2Mb) duplications at 1q42 and 9q21 in two nuclear families. Pathway analysis was performed using a non-parametric GO-based algorithm which searches for the optimal regulated pathway compositions without a priori criteria for significance of individual genes, and accounts for bias introduced by variable gene length by extensive permutations. No single biological theme connected all datasets, but interesting overlap in the identified pathways and involved genes was found between the datasets from the CF isolate, further supporting family-specificity of predisposing factors. Preliminary results of promoter analysis identified two TFs linked to the CF-specific pathways, which implicated angiogenesis-related processes and highlighted multiple previously reported autism susceptibility genes.

ARTEFACTS AND DATA ANALYSIS CHALLENGES FOR THE ILLUMINA GENOME ANALYZER

Martin Kircher, Janet Kelso

Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Deutscher Platz 6, Leipzig, 04275, Germany

As with earlier sequencing technologies, the new high-throughput sequencing technologies have specific limitations and problems. Such problems either require consideration during project design, or need to be handled during data analysis. We focused our analysis on the Illumina Genome Analyzer I/II system. This platform provides parallel fluorescence-based readout of millions of immobilized sequences by iterative detection using fluorescent-dye reversible-terminator chemistry.
In addition to creating a high-quality sequencing library, handling as well the sequencing chemistry has a strong impact on the quality of a run. Particles like chemistry lumps, dust and lint in the sequencing chemistry can cause pseudo-sequence signals which result in the analysis of artificial reads not belonging to the library sequenced or distorted dye readouts. Additionally, reflections, air bubbles, uneven application of oil and an imperfectly calibrated instrument may cause the quality of the data to vary between sequencing runs and even between lanes of the same run. Identification of adapter and chimera sequences is not part of the standard processing and is often hampered by higher error rates at the ends of reads as well as by short reads showing only a few adapter bases. Finally, short read lengths and high error rates can complicate downstream analyses, like mapping and assembly.
We present principles for good analysis practice that facilitate handling of these problems, or at least their identification in a sequencing run. First, indexing/tagging as well as filtering low complexity sequences is efficient in removing most non-adapter related sequencing artefacts. Second, we suggest that PHRED-like base quality scores should be used for quality-based read filtering on the complete sequence rather than on only the first bases. Third, we identify protocol-specific library artifacts like adapter chimeras. We recommend using these for filtering library artefacts and trimming starting adapters. Further contamination from other DNA sources (e.g. enzyme DNA contamination) is expected and should be filtered before data analysis.

SEQUENCING, ASSEMBLY AND ANALYSIS OF THE CASSAVA GENOME

Chinnappa Kodira[1], Simon Prochnik[2], Brian Desany[1], Mohammed Mohiuddin[1], Cynthia Turcotte[1], Todd Arnold[1], James Knight[1], Michael Egholm[1], Tim Harkins[1], Dan Rokhsar[2], Steve Rounsley[3]

[1]Roche 454, Research and Development, 20 Commercial St, Branford, CT, 06405, [2]DOE Joint Genome Institute, Genomics and Development, 2800 Mitchell Drive, Walnut Creek, CA, 94598, [3]University of Arizona, School of Plant Sciences & BIO5 Institute, 1657 E. Helen St, Tucson, AZ, 85719

Cassava (*Manihot esculenta*) is a root crop that serves as the primary food source for more than 750 million people each day. However, it has many limitations as a food – particularly poor nutritional content, and it is susceptible to many pathogens, particularly in Africa, where one third of the continental harvest is lost each year to viral diseases. Through a partnership between Roche 454, DOE Joint Genome Institute and Steve Rounsley at the University of Arizona, we produced a draft sequence of the cassava genome primarily using 23 fold sequence coverage from 454's FLX Titanium sequencing technology. Despite significant repetitive content, the draft genome that we assembled using the Newbler assembler appears to contain 95% of known cassava genes.

Since then, we have made significant improvements to the preliminary draft assembly by incorporating an additional 6 fold sequence coverage from Roche's unreleased 1K GS FLX Sequencing Kits and upgraded instruments. With the addition of long reads, we now have an improved assembly of 573 Mb in size containing more than 98% of the gene space, thus demonstrating the importance of high quality long reads for generating de novo reference assemblies of complex non-model genomes. The updated genome assembly is a useful substrate for genome annotation, gene discovery and identification of useful SNPs. We present here the results from the assembly, annotation and analysis of our latest version of the Cassava genome.

Genomic information generated by this project will serve as a great resource that will benefit key breeding programs throughout Africa by facilitating discovery of useful markers, fine mapping in the region of a candidate resistance gene for Cassava Brown Streak Disease and crop improvement in Cassava.

SEARCH OF EMT-RELATED GENES BY ANALYZING NCI-60 PANELS.

Kensuke Kojima, Toshio Ota

Kyowa Hakko Kirin Co., Ltd., Drug Discovery Research Laboratories, 1188, Shimotogari, Nagaizumi-cho, Sunto-gun, Shizuoka, 411-8731, Japan

A characteristic of anticancer resistant cells and/or metastasizing cells is their transition from an epithelial state to a mesenchymal phenotype, a process known as epithelial-mesenchymal transition (EMT). We've already known that several molecules' expressions are changed according to EMT or mesenchymal-epithelial transition (MET), and some of those molecules are related to the mechanism of EMT or MET. Identification of genes that show specific expression in epithelial or mesenchymal type cells leads not only to the elucidation of EMT-mechanism but to the identification of molecular target of novel anticancer drugs. Now we tried to identify EMT-related genes by analyzing comprehensive gene expression data of NCI-60 cell lines.

ACCELERATED EVOLUTION OF *PAK3*- AND *PIM1*-LIKE KINASE GENE FAMILIES IN THE ZEBRA FINCH, *TAENIOPYGIA GUTTATA*

Lesheng Kong[1], Peter V Lovell[2], Andreas Heger[1], Claudio V Mello[2], Chris P Ponting[1]

[1]University of Oxford, Department of Physiology, Anatomy and Genetics, MRC Functional Genomics Unit, South Parks Road, Oxford, OX1 3QX, United Kingdom, [2]Oregon Health & Science University, Department of Behavioral Neuroscience, 3181 Sam Jackson Park Road, Portland, OR, 97239

Genes encoding kinases tend to evolve slowly over evolutionary time, and only rarely do they appear as recent duplications in sequenced vertebrate genomes. Consequently, it was a surprise to find two families of kinase genes that have greatly and recently expanded in the zebra finch (*Taeniopygia guttata*) lineage. In contrast to other amniotic genomes (including chicken) that harbour only single copies of PAK3 and PIM1 genes, the zebra finch genome appeared at first to additionally contain 67 PAK3-like and 51 PIM1-like protein kinase genes. An exhaustive analysis of these gene models however revealed most to be incomplete, owing to the absence of terminal exons. After re-prediction, 31 PAK3-like genes and 10 PIM1-like genes remain, and all but three are predicted, from the retention of functional sites and open-reading frames, to be enzymatically active. PAK3-like, but not PIM1-like, gene sequences show evidence of recurrent episodes of positive selection, concentrated within structures spatially adjacent to N- and C-terminal protein regions that have been discarded from zebra finch PAK3-like genes. At least seven zebra finch PAK3-like genes were observed to be expressed in testis whilst two sequences were found transcribed in the brain, one broadly including the song nuclei, the other in the ventricular zone and in cells resembling Bergmann's glia in the cerebellar Purkinje cell layer. Two PIM1-like sequences were also observed to be expressed with broad distributions in the zebra finch brain, one in both the ventricular zone and the cerebellum and apparently associated with glial cells, the other showing neuronal cell expression and marked enrichment in midbrain/thalamic nuclei. These expression patterns do not correlate with zebra finch-specific features such as vocal learning. Nevertheless, our results show how ancient and conserved intracellular signalling molecules can be co-opted, following duplication, thereby resulting in lineage-specific functions, presumably affecting the zebra finch testis and brain.

TARGETED CAPTURE METHODS FOR NEXT GENERATION SEQUENCING: AMPLICON TILING VS ARRAY CAPTURE

Melissa Kramer, Magdalena Gierszewska*, Jianchao Yao*, W. Richard McCombie

Cold Spring Harbor Laboratory, Genome Research Center, 500 Sunnyside Blvd, Woodbury, NY, 11797

*Authors contributed equally

While the cost of whole genome sequencing is rapidly decreasing with next-generation technologies, capture of specific target regions of the genome for re-sequencing is often necessary to enable sampling of large numbers of individuals. Two of the most widely used target capture methods are amplicon tiling and array capture. Here we present a comparison of these methods for targeted gene capture and Illumina sequencing.

MOBILE ELEMENT INSERTION DETECTION FROM THE 1000 GENOMES PROJECT PILOT DATA REVEALS HIGH VARIATION BETWEEN INDIVIDUALS

Deniz Kural[1], Michael P Strömberg[1], Chip Stewart[1], Jerilyn A Walker[2], Miriam K Konkel[2], Adrian Stuetz[3], Alexander E Urban[4], Fabian Grubert[4], Mark A Batzer[2], Jan Korbel[3], Gabor T Marth[1], 1000 Genomes Project Structural Variation Subgroup[1,2,3,4]

[1]Boston College, Dept. of Biology, 140 Comm. Ave., Chestnut Hill, MA, 02467, [2]Lousiana State University, Dept. of Biological Sciences, 202 Life Sciences Building, Baton Rouge, LA, 70803, [3]European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstraße 1, Heidelberg, 69117, Germany, [4] Yale University, MB&B Dept., 266 Whitney Ave., New Haven, CT, 06520

Mobile elements comprise ~50% of the human genome. Polymorphism arising from the genomic insertions of these elements is a major contributor to human genetic variation, but the precise extent of this contribution is not fully elucidated. We developed algorithms for detecting insertion sites of three of the most active mobile element types (Alus, L1s, and SVAs) from high-throughput sequencing data: a method based on paired-end fragments spanning the insertion breakpoint, and a split-read approach enabling the direct, single-base accuracy characterization of breakpoints.

We analyzed the 1000 Genomes Project Pilot 1 (150 low-coverage samples) and Pilot 2 (6 high-coverage samples), and identified a total of 5,364 mobile element insertion loci. Experimental validation indicates a very high accuracy (false positive rate <5%). The detection efficiency is estimated at 90 % for Alu insertions in the high coverage set (6 samples).

The events discovered represent the largest catalogue of polymorphic mobile element insertions to date. Initial analysis of these events within the populations they were discovered indicates segregation properties that are similar to SNP markers. Based on the number of events detected within the high-coverage samples we estimate that roughly 3,000 Alu, 350 L1, and 40 SVA insertion polymorphisms occur between two individuals, significantly higher than expected the number of insertion events previously reported from Human-Chimpanzee comparisons.

COMPARISON OF SHORT-READ ALIGNMENT SOFTWARE

Sendu Bala[2], Ahmet Kurdoglu[1], James Long[1]

[1]TGen, Neurogenomics, 445 N 5th St., Phoenix, AZ, 85004, [2]WTSI, Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, United Kingdom

Sendu Bala*, Ahmet Kurdoglu*, James Long*, Heng Li, Gerton Lunter, Richard Durbin, Thomas Keane, Goncalo Abecasis, David W Craig and Mapper Comparison subgroup of the 1000 Genomes Project (* contributed equally)

Identifying genetic variation from short-read genome-wide sequencing is fundamentally dependent upon the alignment strategy. In order to assess which alignment software would be suitable to the objectives of the 1000 Genomes Project we compared several different software packages on simulated data. Two sets of reads were simulated. The first set of simulated data was generated by adding SNPs and indels into a reference genome to create an ancestor, with respective rates of 0.0005 and 0.0001. Two haplotypes were then generated by adding further SNPs and indels to the ancestor at an independent set of sites, effectively creating SNPs that were 1/3rd homozygous and 2/3rds heterozygous. The program STAMPY was then used to generate reads with a defined insert size and introduce sequencing error based on representative data sets from the 1000 genomes production centers. Quality scores were introduced by sampling representative runs. Read names were provided indicating the genetic variant and its location. In the second set of simulated data structural variants were introduced within individual reads. Reads containing 200 bp deletions, 100 bp deletions, 20K insertions and 1-30 bp indels were generated using STAMPY under a similar procedure to the first dataset. Simulated sets for SOLiD included both paired 25mers with 1.5Kb inserts. Simulated datasets for Illumina included paired 37, 54, 76 and 108 bp reads on 400 bp fragments. Only alignment software producing the SAM format was considered. SOLiD aligners included BWA, BFAST, iMAP, Corona_lite, Mosaic, and Karma. Illumina aligners included BFAST, Bowtie, BWA, Eland2, Karma, MAQ, Novoalign, Soap, Smalt, Srprism, and Stampy. Each software package was assessed for memory footprint, input/outputs, and alignment speed on clusters at TGen (SOLiD datasets) or Sanger (Illumina datasets). Alignment accuracy was assessed for each type of read class (no variants, SNPs, indels, etc). These are reported either categorically, as a relation to mapping quality, or through an ROC curve comparing mapped reads to incorrectly mapped reads.

CLOUD-SCALE STATISTICAL ANALYSIS OF MULTIPLE RNA-SEQ DATASETS

Ben Langmead, Kasper D Hansen, Jeffrey T Leek

Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, 615 North Wolfe Street, Baltimore, MD, 21205

As next-generation sequencing gains increasing use as a tool for studying variation and disease, improvements in sequencing continue to decrease cost and increase dataset size. For example, Illumina reports that its HiSeq instrument generates 25 gigabases of sequence per day, and the cost of a gene-expression experiment is $200 per sample. Lower costs and greater adoption and throughput create an urgent bioinformatics challenge: can computing infrastructure keep up with this ever-growing data deluge? Recent studies [1, 2] propose Cloud Computing as a paradigm that addresses challenges faced by academic researchers working with sequence data. Cloud Computing combines a parallel software framework with a resource model whereby large clusters can be rented over the Internet for a per-computer-per-hour fee.

We present Myrna, a Cloud Computing pipeline for computing differential expression of genes from large RNA-seq datasets. Myrna is scalable, flexible in the choice of statistical and gene models used, and readily handles input data consisting of many experiments, samples, and sequencing runs and lanes. Myrna is built on Hadoop, a parallel software framework providing scalability and fault tolerance. At its core, Myrna uses Bowtie [3] to align short reads to the reference transcriptome, then uses Bioconductor [4] to assign alignments to exons and genes, normalize alignment counts, calculate observed and null statistics, and report final per-gene P values.

We present results from re-analyzing multiple large, publicly-available RNA-seq datasets with Myrna, including timing results demonstrating that it scales efficiently to millions of input reads. More significantly, we show how this cloud-scale analysis sheds new light on the appropriateness of often-used statistical assumptions about technical and biological artifacts.

[1] Schatz, MC: CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 2009, 25:1363-1369
[2] Langmead, B, et al. Searching for SNPs with cloud computing. Genome Biol 2009, 10:R134
[3] Langmead, B, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10:R2
[4] Gentleman, R, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004, 5:R80

REGULATION OF HUMAN GENE EXPRESSION AS A TARGET OF NATURAL SELECTION

Tuuli Lappalainen[1], Antigone S Dimas[1,2], Stephen B Montgomery[1], Eugenia Migliavacca[1], Barbara E Stranger[3], Emmanouil T Dermitzakis[1]

[1]University of Geneva Medical School, Department of Genetic Medicine and Development, Rue Michel-Servet 1, Geneva, 1211-4, Switzerland, [2]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom, [3]Harvard Medical School/Brigham and Women's Hospital, Division of Genetics/Department of Medicine, 77 Avenue Louis Pasteur, Boston, MA, 02115

In this study, we have analyzed evolutionary patterns of expression quantitative trait loci (eQTLs) in different populations and in different tissues in order to shed light on the importance of natural selection in gene expression differentiation between human populations as well as systemic targets of selection in humans. A combination of population genetic statistics and conservation scores was used to study both recent and ancient natural selection, and these analyses will gain additional power from the emerging genomic and RNA sequencing data.

We investigated global patterns of natural selection on gene expression by analyzing eQTLs detected from LCLs in a total of 792 samples from 8 populations of the HapMap3 sample set, genotyped for 1.5 million SNPs. Comparison of eQTL SNPs with similar variants without proven regulatory potential showed an interesting combination of purifying and positive selection. The patterns eQTL sharing between populations suggest that also other processes than genetic drift affect differences in gene expression between populations.

Natural selection is expected to target genes and regulatory regions active in different tissues in a distinct manner. We studied systemic targets of natural selection from a dataset of gene expression levels in fibroblasts, T-cells and B-cells of 75 Caucasian individuals with genome-wide SNP data. Comparison of signs of natural selection in eQTLs and in the entire gene regions specific to different tissues suggested stronger purifying selection in B- and T-cell specific than fibroblast specific genes and eQTLs.

Altogether, this study sheds light on the tissue-specific mechanisms of natural selection, and supports the important role of regulation of gene expression in the recent evolution of human populations.

INFLUENZA EVOLUTION: A MOLECULAR RED QUEEN RACE

Natalja Strelkowa[1], Michael Lässig[2]

[1]Imperial College, Biomedical Engineering, South Kensington Campus, London, SW72AZ, United Kingdom, [2]University of Cologne, Theoretical Physics, Zülpicher Str. 77, Cologne, 50937, Germany

The seasonal influenza A virus undergoes rapid evolution to escape human immune response. This process occurs primarily in the viral epitope, the antibody-binding domain of the surface protein haemagglutinin. Aminoacid changes within the epitope have antigenic effects, whereas changes outside the epitope may affect protein stability. Here, we study the genome evolution of influenza A (H3N2) and quantify its underlying selective forces. Our analysis is based on frequency time-series of single-nucleotide polymorphisms, which are obtained from coalescence trees of influenza sequences over 39 years. We analyze these data in terms of a simple population-genetic model with positive selection for aminoacid changes in the epitope, negative selection for changes outside the epitope, and complete genetic linkage (i.e., absence of recombination) between both classes. This model quantitatively reproduces the statistics of polymorphisms and substitutions of the influenza sequences. Our main findings are:

1. Influenza evolution shows strong clonal interference, which can explain the punctuated pattern of evolution observed in this system: the competition between strains produces recurrent selective sweeps with clusters of simultaneous fixations every 3 to 4 years, without additional assumptions that selection itself is punctuated. Our analysis shows that influenza evolution is governed not only by antigenic adaptation within the epitope, but also by background selection outside the epitope.

2. The high mutation rates of influenza generate a complex competition between strains, which involves the combinatorics of coexisting strongly beneficial mutations (driving mutations). A given driving mutation often arises independently in two or more competing strains. At the same time, a selective sweep has 3 to 4 driving mutations on average.

3. This process produces efficient adaption: of the 80 epitope aminoacid substitutions in 39 years, at least 50 are predicted to be adaptive. The total rate of adaptive epitope changes is about twice the rate expected under neutral evolution.

Thus, the joint evolutionary forces of positive and negative selection, genetic linkage, and high mutation rates in the viral haemagglutinin determine the red queen race between viral strains and set the speed of their turnover.

DNA REARRANGEMENTS IN CANCER: SIXTY TUMOR GENOMES AND THEIR SOMATIC STRUCTURAL ALTERATIONS

Michael S Lawrence, Yotam Drier, Michael F Berger, Michael Chapman, Robb Onofrio, Kristian Cibulskis, Carrie Sougnez, Wendy Winckler, Levi A Garraway, Eric S Lander, Todd R Golub, Stacey B Gabriel, Matthew L Meyerson, Gad Getz

Broad Institute of MIT and Harvard, Cancer Program, 7 Cambridge Center, Cambridge, MA, 02142

DNA rearrangements are known to drive certain cancers; for example, discovery of the BCR-ABL fusion in chronic myelogenous leukemia (CML) led to the development of imatinib, the classic example of a targeted cancer therapeutic. Complete understanding of a tumor will require study of its genomic rearrangements.

We report the deep-coverage whole genome shotgun (WGS) sequencing of more than 60 tumor/normal pairs: 26 multiple myeloma, 10 ovarian, 8 prostate, 5 glioblastoma multiforme (GBM), 2 chronic lymphocytic leukemia (CLL), 4 melanoma, and 6 colorectal cases. We describe the analysis of this data using dRanger, an automated method for detecting somatic structural alterations, by identifying clusters of paired reads bridging distant genomic loci, and BreakPointer, which assembles the basepair-resolution fusion sequences. Somatic rearrangements were validated with a rate of 90% by PCRing and sequencing the chimeric DNA in the tumor, and confirming its absence in the normal.

The results reveal astonishing diversity between and within tumor types: some cases have hundreds of rearrangements, others only a few. Some have many interchromosomal events, others extremely complex locally scrambled regions. Most of the GBM cases have EGFR amplified to hundreds of copies and rearranged with other chromosomes. We have detected novel PTEN and NF1 fusions, and confirmed a known TMPRSS2-ERG fusion. RNA-Seq evidence has confirmed expression of some fusions. A prostate case is unique in the data set, having a large set of mutually balanced translocations, with many of the breakpoints located in important cancer genes, including TP53 and ABL1. Our findings emphasize the prevalence and importance of DNA rearrangements as potential cancer drivers.

A LARGE PUBLIC HEALTH EFFECT OF A COMMON VARIANT ON CHROMOSOME 11Q13 (RS7927894) ON ECZEMA, ASTHMA, AND HAY FEVER

Ingo Marenholz[1,2], Anja Bauerfeind[2], Jorge Esparza-Gordillo[1,2], Tamara Kerscher[1,2], Raquel Granell[3], John Henderson[3], Young-Ae Lee[1,2]

[1]Charité, Pediatric Pneumology and Immunology, Augustenburger Platz 1, Berlin, 13353, Germany, [2]Max-Delbrück-Centrum (MDC) for Molecular Medicine, Medical Genomics, Robert-Rössle-Str. 10, Berlin, 13092, Germany, [3]University of Bristol, Department of Community-based Medicine, Oakfield Grove, Bristol, BS8 2BN, United Kingdom

Eczema (atopic dermatitis) is a chronic inflammatory skin disorder and a major manifestation of allergic disease. In a genome-wide association study for eczema, a common variant on chromosome 11q13.5 (rs7927894) has been identified as a susceptibility locus in four study populations from Central Europe.
The aim of this study was to determine the effect of this risk variant on various allergic diseases on the population level. More than 7,400 individuals from a large English birth cohort born in 1991 and 1992, were genotyped for rs7927894. Association analyses were performed for eczema, asthma, hay fever, allergic sensitization, and combinations of these phenotypes. The population attributable risk fractions were estimated. Beyond replicating the association with eczema, we found that the effect of this risk variant was restricted to individuals with the allergic subtype of eczema (OR, 1.31; 95% CI, 1.07-1.60). In contrast, no association of rs7927894 with non-atopic eczema was observed. Moreover, we detected association with concomitant asthma (OR, 1.50; 95% CI, 1.14-1.97) and hay fever (OR, 1.54; 95% CI, 1.15-2.06). The population attributable risk fraction for atopic eczema was estimated to be 23.8%.
We conclude that the rs7927894 risk allele is a common variant that confers a moderate risk for atopic disase, but carries a large public health effect on atopic eczema and allergic airways disease. Furthermore, the association with atopic eczema as well as concomitant asthma and hay fever points to a key role in the atopic march.

GENOME-WIDE DETECTION OF TARGET GENES OF LONG-RANGE CIS-REGULATION

Altuna Akalin[1], David Fredman[2], Xianjun Dong[1], Gemma Danks[1], Boris Lenhard[1]

[1]University of Bergen, CBU - Bergen Center for Computational Science, and Sars Centre for Marine Molecular Biology, Thormøhlensgate 55, Bergen, 5008, Norway, [2]University of Vienna, Department for Molecular Evolution and Development, Althanstrasse 14, Vienna, 1090, Austria

Highly conserved non-coding elements (HCNEs) are one of the most intriguing features of Metazoan genomes. HCNEs cluster around key developmental genes, and many were also shown to be enhancer that driving the expression of those genes. The need to keep HCNEs as enhancers in cis to their target gene puts constraints on gene rearrangements and genome organization, leading to strong synteny conservation. These findings lead us formulate the genomic regulatory block (GRB) model, in which one or more target genes are located in an HCNE-dense area, often along with other "bystander" genes that do not respond to regulatory input from HCNEs.

In this work we present innovative computational approaches for the characterization of genomic regulatory blocks and their target genes. First we demonstrate that non-coding sequence conservation in GRBs reveals a characteristic and unusual pattern of purifying selection on GRB sequences even between closely related species, enabling the detection of large lineage-specific developmental regulatory changes. Second, we present a method for the automated genome-wide detection of target genes of GRBs. While these genes are among the most important and most well studied genes in developmental biology, their determination has been a manual and insufficiently well defined process until now. We have shown that GRB target genes have specific features, most notably long and multiple CpG islands and broad transcription initiation regions, as well as a distinct set of epigenetic markers. We now present an expanded list of genomic features that make GRB target genes distinguishable from other genes in their neighborhood and the rest of the genome. Using these features, we have developed a predictive method based on random forests to aid in cataloguing genome-wide sets of all possible GRBs and their target genes. We uncover an exhaustive list of target genes of long-range developmental regulation and the regulatory blocks around them, and show compelling evidence for their complex regulation and high long-range regulatory potential.

GENOME-WIDE RECONSTRUCTION OF IDENTICAL-BY-DESCENT HAPLOTYPES SHARED BY FIRST DEGREE RELATIVES USING WHOLE GENOME RESEQUENCING

Denis M Larkin[1], Miri Cohen-Zinder[2], Michael E Goddard[3], Alvaro G Hernandez[4], Chris L Wright[4], Lorie A Hetrick[4], Lisa Boucek[4], Sharon L Bachman[4], Mark R Band[4], Tatsiana Akraiko[4], Jyothi Thimmapuram[4], Tim Harkins[5], Jennifer E McCague[6], Ben Hayes[7], Iona Macleod[3], Hans Daetwyler[7], Harris A Lewin[1,2]

[1]University of Illinois, Department of Animal Sciences, Urbana, IL, 61801, [2]University of Illinois, Institute for Genomic Biology, Urbana, IL, 61801, [3]University of Melbourne, Melbourne, Australia, [4] University of Illinois, The W. M. Keck Center for Comparative and Functional Genomics, Urbana, IL, 61801, [5]Roche, Indianapolis, IN, 46250, [6]454 Life Sciences, Branford, CT, 06405, [7]Department of Primary Industries, Australia

The genomes of Holstein bull “Chief” and his son “Mark” were sequenced to 7x and 13x coverage, respectively, using 454 Titanium chemistry. We identified ~1.4 million SNPS that were useful for reconstruction of IBD segments inherited by Mark from Chief. The precision of SNP genotyping and phasing of alleles in Mark haplotypes were validated by genotyping Mark and 92 offspring using the Illumina Bovine SNP50 array. Using SNP50 genotypes as a reference, the precision of allele phase reconstruction of the Mark genome was 97%. The shared IBD segments distinguishing the alternative haplotypes of the two first order relatives were resolved to a median spacing of 800 bp. The Mark genome shows runs of homozygosity, indicating inbreeding due to the reduced effective population size following the domestication of cattle, and regions of dense heterozygosity, indicating a large effective population size pre-domestication. The approach used will facilitate the simultaneous identification of the candidate QTL regions and QTNs for several economically important traits segregating among the offspring of these two champion bulls.

COMBINED EVIDENCE FROM EVOLUTIONARY, POPULATION-GENETIC AND DISEASE STUDIES POINTS TO A ROLE OF THE EPIGENOME IN MEDIATING STRUCTURAL MUTABILITY OF THE HUMAN GENOME

Jian Li[1,2], Pawel Stankiewicz[1], Ronald A Harris[1], Sau Wai Cheung[1], Ankita Patel[1], Sung-Hae Kang[1], Chad A Shaw[1], Craig Chinault[1], Lisa White[1], Tomek Gambin[3], Anna Gambin[4], James R Lupski[1,5,6], Aleksandar Milosavljevic[1,2]

[1]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, 77030, [2]Baylor College of Medicine, Program in Structural and Computational Biology and Molecular Biophysic, Houston, TX, 77030, [3]Warsaw University of Technology, Institute of Computer Science, Warsaw, 00-661, Poland, [4] University of Warsaw, Institute of Informatics, Warsaw, 02-097, Poland, [5]Baylor College of Medicine, Department of Pediatrics, Houston, TX, 77030, [6]Texas Children's Hospital, Houston, TX, 77030

Submicroscopic structural variation is highly non-randomly distributed in the human genome. Non-Allelic Homologous Recombination (NAHR) mediated by Low Copy Repeats (LCRs) can account for only a small fraction of it. Prompted by the recently discovered association of hypomethylation and structural instability in white-cheeked gibbons, we examined the relative contribution of LCR-mediated NAHR and hypomethylation to the submicroscopic structural mutability in humans. We combined evidence from human genome evolution, structural polymorphisms in the human population, and disease studies. Combined evidence from multiple sources, including our study of association of hypomethylation in human germline with human-specific structural rearrangements, copy number variation from published disease and HapMap studies, and our own study of 400 patient samples conducted using custom oligonucleotide arrays specifically designed to probe NAHR-susceptible regions consistently points to a stronger association of germline hypomethylation with genomic instability than LCR-mediated NAHR, indicating a major role for the epigenome in mediating structural mutability of the human genome.

DISCOVERY AND CHARACTERIZATION OF CODING INSERTIONS AND DELETIONS IN 1000 EXOMES BY *DE NOVO* ASSEMBLY

Yingrui Li and Jun Wang

BGI, Shenzhen, Main Building, Beishan Industrial Zone,Yantian, Shenzhen, 518083, China

Exome capturing and sequencing have been utilized to characterize polymorphisms in exonic regions in human with ever-highest resolution. Nonetheless, current mapping-based approaches still have limited in detection of insertions and deletions, which in general have more potential functional impact than SNPs. Here we present an unprecedented discovery of coding insertions and deletions in 1000 Danish Caucasians by de novo assembly of captured exome sequences with approximately ~10,000-fold coverage in total. Experimental and computational validations prove the detected indels have high accuracy and they are much less biased to short indels (1-3bp in length) than previous studies. The data enables us to characterize the population pattern of coding indels with minor allele frequency >0.02 in over 90% of the exon region. Investigations in genomic distribution and frequency spectrum have unveiled novel population and selection aspects of coding indels that are different from SNPs. Our results provide an abundant source and new insights of this important variation type.

ACCURATE CNV GENOTYPING FROM MASSIVELY PARALLEL SEQUENCING DATA

Yun Li[1], Robert E Handsaker[2], Gonçalo R Abecasis[3], Steven A McCarroll[2,4]

[1]University of North Carolina, Genetics, Biostatistics, 120 Mason Farm Road, Chapel Hill, NC, 27599, [2]Broad Institute, MIT and Harvard, Cambridge, MA, 02142, [3]University of Michigan, Biostatistics, 1420 Washington Heights, Ann Arbor, MI, 48103, [4]Harvard Medical School, Genetics, 77 Avenue Louis Pasteur, Boston, MA, 02115

A core goal of next generation sequencing (NGS)-based studies is to provide information on a full spectrum of genetic variants including SNPs and CNVs. Sequencing based studies pose analytical and computational challenges due to the sheer volume of data generated, short read lengths and high sequencing error rates. These problems become even more challenging when a low-coverage sequencing design is used to cost-effectively analyze genetic variation in as many individuals as possible.
Here we describe an approach for integrating numerous features of NGS data to obtain highly accurate CNV genotypes in population-sequencing studies. The approach builds upon our hidden Markov model proposed earlier that generates accurate SNP calls by leveraging LD information from a population sample of individuals, using either genotype or sequencing data. Here we have extended the method to analyze SNPs and CNVs jointly. We applied our methods to NGS data with an average coverage ~5X from the 1000 Genomes Project. We obtained preliminary CNV likelihoods by analyzing the local distribution of read depth across a population together with the presence of read pairs that have unexpected spacing when aligned to the genome. We obtained preliminary SNP likelihoods using read depth, base and mapping quality. We then integrated the two sets of first round likelihoods to obtain refined CNV genotypes. Comparing these refined CNV genotypes with experimental counterparts, our preliminary results show that a large class of human CNVs – those containing at least a kilobase of unique, mappable sequence – can be genotyped with a call rate of 99.30% and an error rate of 1.2%. We thus demonstrate that joint analysis of CNV and SNP data from massively parallel sequencing can generate more complete and accurate CNV genotypes. Our method is sufficiently flexible to handle low-coverage sequencing data from many individuals in a population sample. Our imputation method is implemented in C/C++ and is at www.sph.umich.edu/csg/abecasis/mach/ for free downloading.

TRANSCRIPTOME ALTERATION IN HIPPOCAMPUS AND SPLEEN UNDER THE TREATMENT OF REGULATIVE PEPTIDE SELANK AND SOME OF ITS FRAGMENTS

Svetlana A Limborska[1], Timur A Kolomin[1], Maria I Shadrina[1], Stanislav I Shram[2], Petr A Slominsky[1], Nikolay F Myasoedov[2]

[1]Institute of Molecular Genetics Russian Academy of Sciences, Department of Human Molecular Genetics, 2 Kurchatov Sq., Moscow, 123182, Russia, [2]Institute of Molecular Genetics Russian Academy of Sciences, Department of chemistry of Physiologically Active Compounds, 2 Kurchatov Sq., Moscow, 123182, Russia

Selank is a synthetic peptide, an analogue of tuftsin (the short fragment Thr–Lys–Pro–Arg of the human immunoglobulin G heavy chain), elongated at the C terminus with the tripeptide Pro–Gly–Pro. Selank showed dual influence on organism: on the one hand it has a nootropic and anxiolytic characteristics, on the other – exhibits an antiviral effect.
Taking into account the bipartite action of Selank we have carried out two lines of analysis. In the first, we have analyzed the expression profiles of 12000 genes in rat hippocampus using the SBC–R–RC–100–13 microtemplate (Biochip^{TM}). Single and course Selank administration caused a more than 2,5-fold change in the expression of 18 and 15 genes, respectively. In both cases, under Selank administration the most expression changes were observed for five genes *Actn1*, Cx3cr1, Fgf7, Ptprn2 and Xtrp1. In the second place we studied the effect of Selank on the expression of these genes in the spleen. Selank showed much stronger influence on the gene expression in this organ than in the hippocampus. After a single Selank administration the most significant increase of the expression was observed for three genes: *Ptprn2*, Actn1 and Cx3cr1.
Moreover we have carried out detailed study of the Selank and its fragments (Gly-Pro, Arg-Pro-Gly-Pro and tuftsin) action on the gene expression. We analyzed the expression of 84 genes (the panel of genes RT^{2}Profile^{TM} PCR Array SABiosciences) involved in processes of inflammation in the mouse spleen 6 and 24 hours after single intraperitoneal injection of peptides. There was a significant alteration in the expression of 34 genes. In most cases the maximal response was observed after 6 hours of peptides injection. It was revealed that each peptide gives an individual pattern of changes in the expression of genes studied.

REGULATORY NETWORK THAT ORCHESTRATES THE DEVELOPMENT OF THE B CELL LINEAGE

Yin C Lin[1], Suchit Jhunjhunwala[1], Christopher Benner[2], Sven Heinz[2], Robert Mansson[1], Mikael Sigvardsson[3], James Hagman[4], Celso A Espinoza[5], Christopher K Glass[2], Cornelis Murre[1], [1]

[1]University of California, San Diego, Department of Molecular Biology, 9500 Gilman Dr, La Jolla, CA, 92093, [2]University of California, San Diego, Department of Cellular and Molecular Medicine, 9500 Gilman Dr, La Jolla, CA, 92093, [3]Linkoping University, Department of Biomedicine and Surgery, Lab 1 Level 13, Linkoping, 58185, Sweden, [4] National Jewish Health, Integrated Department of Immunology, 1400 Jackson St, Denver, CO, 80206, [5]University of California, San Diego, Ludwig Institute for Cancer Research, 9500 Gilman Dr, La Jolla, CA, 92093

It is now established that E2A, FOXO1 and EBF play critical roles in B cell development. However, how they act in concert to induce lineage-specific programs of gene expression and how their DNA binding patterns relate to architectural proteins like CTCF is unknown. We demonstrate that E2A, FOXO1 and EBF, primarily associate with H3K4 monomethylated regions. In contrast, CTCF occupancy and H3K4 monomethylation are mutually exclusive. We further demonstrate that: (1) E2A binds coordinately with EBF and FOXO1 to a vast spectrum of B-lineage specific enhancers, (2) these associations are dynamic during developmental progression, (3) E47 directly modulates the pattern of H3K4 monomethylation, (4) the induction of H3K4me patterning correlates well with B-lineage specific gene expression and enhancer activity. From these data, a regulatory network is generated that shows how the combined activities of E2A, FOXO1 and EBF orchestrate B cell development.

Keywords: ChIP-seq, HOMER de novo motif finding algorithm, Ward's hierarchical clustering, B cell development, regulatory network, tumor suppression

EPIGENOMIC REPROGRAMMING DURING INDUCTION OF A PLURIPOTENT STATE IN HUMAN CELLS

Ryan Lister [1], Yasuyuki Kida[2], Shigeki Sugii[2], Mattia Pelizzola[1], Michael Downes[2], Ruth Yu[2], Ronald M Evans[2], Joseph R Ecker[1]

[1]The Salk Institute for Biological Studies, Genomic Analysis Laboratory, 10010 North Torrey Pines Road, La Jolla, CA, 92037, [2]The Salk Institute for Biological Studies, Gene Expression Laboratory, 10010 North Torrey Pines Road, La Jolla, CA, 92037

Induced pluripotent stem cells (iPSCs) are capable of differentiating into diverse cell types, and thus hold enormous potential for personalized regenerative therapies. We have recently developed an iPSC system whereby readily available adipocyte progenitor cells are rapidly reprogrammed to a pluripotent state at very high efficiency in a completely feeder-free environment, enabling efficient and convenient iPSC generation. However, central questions remain regarding epigenetic regulatory processes in the iPSC system. What are the epigenetic changes that take place during the reprogramming process? How closely does the iPSC epigenome resemble that of embryonic stem (ES) cells? Do residual epigenetic marks characteristic of the pluripotent state remain in somatic cells derived from iPSCs? To address these questions we have characterized the DNA methylomes and transcriptomes throughout the transition of a differentiated cell into an iPSC, and subsequent differentiation back into the original cell type. Using the Illumina HiSeq2000 sequencer, which yielded over 200 gigabases per run, genome-wide single-base resolution mapping of DNA methylation sites (MethylC-Seq) and strand-specific whole transcriptome (RNA-Seq) profiling were performed for progenitor adipocytes, adipocyte-derived iPSCs, and progenitor adipocytes that were derived from differentiation of the iPSCs. Integration of these data and comparison to an extant human ES cell DNA methylome has enabled for the first time a high-resolution characterization of the epigenomic and transcriptional changes during iPSC generation, and interrogation of the epigenomic patterns within differentiated cells derived by iPSC technology.

ANALYSIS OF COPY NUMBER VARIATIONS AMONG CATTLE BREEDS

George E. Liu, Yali Hou, Bin Zhu, Maria F. Cardone, Lu Jiang, Angelo Cellamare, Apratim Mitra, Lee J. Alexander, Luiz L. Coutinho, Lou C. Gasbarre, Michael P. Heaton, Robert W. Li, Lakshmi K. Matukumalli, Dan Nonneman, Luciana C. de A. Regitano, Tim P. Smith, Jiuzhou Song, Tad S. Sonstegard, Curt P. Van Tassell, Mario Ventura, Evan E. Eichler, Tara G. McDaneld and John W. Keele

USDA-ARS, Bovine Functional Genomics Lab., ANRI, 10300 Baltimore Ave, Building 200 Rm 124B, Beltsville, MD, 20705

Genomic structural variation is an important and abundant source of genetic and phenotypic variation. Here we describe the first systematic and genome-wide analysis of copy number variations (CNVs) in the modern domesticated cattle using array comparative genomic hybridization (array CGH) and quantitative PCR and fluorescent in situ hybridization (FISH). Our panel includes 90 animals from 11 Bos tarus, 3 Bos indicus and 3 composite breeds for beef, dairy or dual purposes. We identified over 200 candidate CNV regions (CNVRs) in total and 177 of which are within known chromosomes, which harbor or are adjacent to gains or losses. These 177 high-confidence CNVRs cover 28.1 mega bases, ~1.07% of the genome. Over 50% CNVRs (89/177) were found in multiple animals or breeds and analysis of them reveals breed-specific frequency differences and reflects aspects of the known ancestry of these cattle breeds. Selected CNVs were further successfully validated by independent methods using qPCR and FISH. About 67% CNVRs (119/177) completely or partially span cattle genes and 61% CNVRs (108/177) directly overlap with segmental duplications. CNVRs span about 400 annotated cattle genes that are significantly enriched for specific biological functions such as immunity, lactation, reproduction and rumination. For gene families like ULBP which have gone through ruminant lineage-specific gene amplification, we detected and confirmed marked differences in CNV frequencies across diverse breeds, demonstrating the evolutionary contributions of CNVs to cattle domestication and breed formation. Our results provide a valuable resource beyond microsatellites and single nucleotide polymorphisms to explore the full dimension of genetic variability for the future cattle genomic research.

THE GENOME OF THE MAN OF THE FOREST

Devin Locke on behalf of The International Orangutan Genome Analysis Consortium

Washington University School of Medicine, The Genome Center, 4444 Forest Park Ave, St. Louis, MO, 63108

The highly endangered orangutan species, Pongo abelii (Sumatran orangutan) and Pongo pygmaeus (Bornean orangutan), are the most phylogenetically distant great apes with respect to humans. The orangutan genome sequence thus provides a unique perspective from which to investigate great ape and human evolution. We present here the draft genome assembly of a Sumatran orangutan, as well as genome sequence from 10 additional Bornean and Sumatran orangutans. Our analysis indicates that the structural evolution of the orangutan genome proceeded more slowly than that of the chimpanzee and human genomes. We observe a lower rate of chromosomal rearrangements, segmental duplications, gene duplications and Alu retroelement transposition. Other notable observations include a polymorphic neocentromere, numerous lineage-specific gene inactivation events among great apes, and suggestive positive selection signals from six components of a single metabolic pathway. From a population perspective, Bornean and Sumatran orangutans exhibit pronounced divergence, including several megabases of species-specific duplication. Demographic analysis of our SNP dataset (13 million SNPs) estimates the Bornean-Sumatran split time at 400,000 years ago, with subsequent reduction in Bornean effective population size and exponential expansion of Sumatran effective population size since the split. These data, considered together, present a unique view into the evolutionary history of the biologically rich orangutan species and provide extensive resources for conservation efforts.

NOVEL GENE REGULATORY NETWORK RECONSTRUCTION IN MULTIPLE HAPMAP POPULATIONS

Benjamin A Logsdon[1], Stephen B Montgomery[2], Barbara E Stranger[3], Emmanouil T Dermitzakis[2], Jason G Mezey[1,4]

[1]Cornell University, Department of Biological Statistics and Computational Biology, 1198 Comstock Hall, Ithaca, NY, 14853, [2]University of Geneva Medical School, Department of Genetic Medicine and Development, 1 rue Michel-Servet, Geneva, CH-1211, Switzerland, [3]Harvard Medical School and Brigham and Woman's Hospital, Division of Genetics, Department of Medicine, 75 Francis Street, Boston, MA, 02115, [4] Weill Medical College of Cornell University, Department of Genetic Medicine, 1300 York Avenue, New York, NY, 10065

Discovery of regulatory networks from genome-wide gene expression and genotype data is an idealized goal of systems biology. We have developed a new algorithm, which we have named ANCHoR, that is novel in its capacity to learn unique directed, undirected, and cyclic regulatory relationships among genes. We use our method to identify gene expression networks in cell lines collected from eight populations in HapMap 3 and compared these reconstructed regulatory networks among populations.

ANCHoR solves a major theoretical concern for network learning: unmeasured network components and connections can produce spurious regulatory signals among a set of expressed genes. We address this problem in ANCHoR by requiring a sufficient set of cis-expression Quantitative Trait Loci (cis-eQTL), which act as perturbations of expression, to be present for an analyzed gene set. With sufficient cis-eQTL, a graph that captures both observed directed and undirected regulatory relationships, as well as unobserved relationships, can be efficiently learned from a combination of gene expression and genotype data.

Using ANCHoR, we reconstructed networks for genes expressed in the immortalized lymphoblastoid cell lines for eight populations from HapMap3. For each population, a global undirected expression network was first reconstructed for expression probes and modularized into 10-40 highly connected sub-networks each containing 5-30 genes. A combined directed and undirected sub-network was then inferred for each module. As an example of our results, networks involving immune response genes, (e.g. CD79B and LY86) were conserved across populations, indicating conserved immunological response for certain genes.

PHYLOGENETIC MAPPING OF RNA-SEQ READS USING A GRAPH ALGORITHM

Albert Vilella, Tim Massingham, Ari Löytynoja

EMBL, EBI, WTGC, Hinxton, CB10 1SD, United Kingdom

RNA-seq analyses of non-model organisms suffer from the lack of a reference gene set to map the reads against. We propose to tackle this problem by aligning the reads to multiple alignments of gene phylogenies from related species using the PAPAYA sequence graph aligner. Our simulation study and analyses of real data indicate better performance than current de novo assembly approaches.

Our approach builds profile-HMM models from EnsemblCompara GeneTrees codon alignments, scans the RNA-seq reads using HMMER3 and clusters reads as putatively homologous to the highest scoring gene tree with evalue<1e-3. For each gene tree we then construct a sequence graph representing the ancestral gene model, with alternative splicing if present, using the codon alignment and its phylogeny. We align graphs representing the putatively homologous reads to this ancestral graph, accommodating possible sequencing errors and lowly scoring bases in the graph structure; reads with low overlap are discarded. We finish by merging the reads from each species using the underlying paralogous haplotype structure in duplicating genetrees and indicating the inferred missing site information.

We performed simulations to study the expected coverage of a given RNA-seq run for a non-model species with no close sequenced genome available. For a subgroup of vertebrate species with 5 available reference genomes encompassing a divergence of 200-300 MYA (e.g. Saurias or Teleosts), and given a 30x coverage RNA-seq run of a complex tissue sample for a new ingroup species 100 MYA to the closest reference genome, our comparative genomics approach would be able to reconstruct 60% of the coding gene structures with a median of 80% of the predicted full length.

We applied our approach to reconstruct the brain transcriptome of 12 birds publicly available in the NCBI SRA against the 5 available genomes in Ensembl (chicken, turkey, duck, zebrafinch and lizard), and for these low-coverage runs (1x-4x of 454 Ti sequencing) we already obtained an average coverage of 50% avian genes at 60% median of full length. Considering current NGS technologies, we predict that our phylogenetic approach would be able to deliver comprehensive transcript models for a vertebrate species in a sampled subgroup for less than $10k in sequencing costs.

SMALL INSERTION AND DELETION VARIATION IN THE 1000 GENOMES PROJECT

G A Lunter[1], C A Albers[2], J Marchini[3], S Montgomery[4], R Durbin[2], G McVean[3], 1000 Genomes Project Indel Variation Subgroup*[4]

[1]WTCHG, Oxford, OX3 7BN, United Kingdom, [2]WTSI, Hinxton, CB10 1SA, United Kingdom, [3]U.Oxford, Oxford, OX1 3TG, United Kingdom, [4]Geneva U. Med. School, Geneva, CH-1211, Switzerland

Small insertions and deletions (indels) are a common yet under-studied class of variation. The 1000 Genomes project allows, for the first time, a genome-wide and systematic study of common indel polymorphisms.

Candidate indels of length up to 50bp were called by several groups and methods. From these we called a single high-confidence set of indels with genotypes. Validation is in progress.

We found that indels are often in strong LD with neighbouring SNPs, showing that indels can be imputed from sparse genotype data. As databases of disease-causing variants (eg HGMD) are enriched with indels, they may be enriched for causal variants of common diseases too, making them interesting targets in GWAS studies. Using imputation this is currently being investigated on several existing GWAS datasets, and preliminary results will be discussed.

We were also interested in determinants of indel accrual rates. We found significant variations, e.g. G+C content has a strong and non-linear effect. As expected rates are lower in inter-species conserved regions, UTRs, and particularly in coding exons, while in recombination hot-spots indel accrual rates are up by 25%. No evidence for a previously reported mutagenic effect of heterozygous indels was found (Tian et al., Nature 455).

Indels are particularly prevalent in homonucleotide runs (HRs): e.g. in 12bp HRs, per-bp rates are increased 150-fold, so that about 50% of indels are predicted to occur in the 5% longest HRs by genome coverage. A substantial (~1%) HR indel error rate in Illumina reads make HR indels difficult to call; while we accounted for this, it suggests some caution in interpreting results. Nevertheless, the site-frequency spectrum suggests that calls are reliable, and the inferred rates and HR abundance are consistent with an equilibrium model of HR evolution.

* E Banks, K Chen, D Craig, M DePristo, M Gerstein, M Hurles, F Hyland, T Keane, R Li, G Marth, Z Ning, A Siddiqui, C Sougnez, K Walter, K Ye, Z Zhang, authors

STUDYING THE EXTENT AND FUNCTION OF EPIGENETIC VARIATION IN TWINS

Kristina Gervin[1], Gregor Gilfillan[1], Håkon Gjessing[3], Jennifer Harris[3], Dag Undlien[1,2], Robert Lyle[1]

[1]Oslo University Hospital, Department of Medical Genetics, Kirkeveien 166, Oslo, 0374, Norway, [2]University of Oslo, Institute of Medical Genetics, Kirkeveien 166, Oslo, 0374, Norway, [3]Norwegian Institute of Public Health, Division of Epidemiology, Post Box 4404 Nydalen, Oslo, 0403, Norway

The phenotypic differences between individuals are an outcome of genetic and epigenetic variation. Whereas variation at the sequence level (SNPs and CNVs) has been studied extensively, much is still unknown regarding the extent and function of epigenetic variation. Disturbance in DNA methylation leading to aberrant gene expression has been shown to be involved in many diseases, and variation in DNA methylation may contribute to the risk of common disease.

The aim of this study is to explore variation and patterns of epigenetic variation using twins. Since each cell type has its own epigenome, we have isolated different lymphocyte subpopulations (CD19+, CD8+, CD4+ and CD4+CD25+) from more than 350 twin pairs.
We have used both region-specific and genome-wide analyses. First, we studied DNA methylation in the classical human major histocompatibility complex (MHC). The MHC is a gene dense and highly polymorphic region on human chromosome 6p21.3, containing genes with a broad range of functions within the innate and adaptive immune systems. We performed extensive bisulphite sequencing of 1670 individual CpG sites distributed in 176 regions in the classical human Major Histocompatibility Complex (MHC) in 49 monozygotic (MZ) and 40 dizygotic (DZ) healthy Norwegian twin pairs. Regions of interest include CpG islands, the 5'end of genes and non-coding conserved regions. We observed significant variation in DNA methylation both between and within regions. Interestingly, the heritability of this variation is low, ~6% for individual CpGs and ~11% for amplicons, suggesting DNA methylation variation is not under strong genetic control. Second, we are using genome-wide methods to examine disease discordance in monozygotic twins. We have performed genome-wide methylation analyses using both Infinium arrays and bisulphite sequencing using the Illumina platform. In parallel, covalent histone modifications were examined by chromatin immuno-precipitation (ChIP).

SEQUENCING OF FOUR TYPE I DIABETES SUSCEPTIBILITY LOCI IN 1000 SAMPLES

Aaron J Mackey[1], Shom N Paul[1], Roderick V Jensen[2], Aaron R Quinlan[3], Benjamin J Boese[4], Neil M Walker[5], Helen Stevens[5], Chris Wallace[5], Ira M Hall[3], Timothy T Harkins[6], Suna Onengut-Gumuscu[1], Patrick J Concannon[1], John A Todd[5], Stephen S Rich[1]

[1]University of Virginia, Center for Public Health Genomics, PO Box 800717, Charlottesville, VA, 22908, [2]Virginia Polytechnic Institute and State University, Department of Biological Sciences, Washington Street, MC 0477, Blacksburg, VA, 24061, [3]University of Virginia, Department of Biochemistry and Molecular Genetics, PO Box 800733, Charlottesville, VA, 22908, [4] Roche Diagnostics, 454 Life Sciences, 20 Commercial Street, Branford, CT, 06405, [5]University of Cambridge, JDRF/WT Diabetes and Inflammation Laboratory, Addenbrooke's Hospital, Cambridge, CB2 0XY, United Kingdom, [6]Roche Diagnostics, Roche Applied Science, 9115 Hague Road, Indianapolis, IN, 46250

Genome-wide association and genetic linkage analyses have identified almost 50 genetic loci that modify risk of type 1 diabetes. In an effort to identify the underlying candidate genes and sequence variants responsible, four regions were selected for resequencing. These regions, on human chromosomes 7, 10, 16, and 19, were identified as having genome-wide levels of significance, and were not known to be the target of ongoing sequencing efforts. The regions were each ~250 kb, for a total of 1 Mb of sequence. As part of the Type 1 Diabetes Genetics Consortium (T1DGC), a custom Nimblegen DNA capture array targeting the four regions was designed. DNA samples were pooled in groups of 10 T1D cases (drawn randomly from the T1DGC WTCCC-UKGrid) or 10 controls (drawn to match the geographic location of cases). Pooling, DNA capture, and 454 sequencing were performed on 100 pools (50 case, 50 control), with an initial 1/4-plate of sequencing per pool. DNA samples were not barcoded in this 100 pool study; however, samples from two pools (one case and one control) were barcoded, captured, pooled and resequenced at one plate per pool, to provide a reference sample for comparison. While barcodes for the 100 pool sequences were not available, donor genotype information from genome wide SNP arrays for all 1000 DNA samples were obtained from the WTCCC. These data provided expected allele frequencies across 131 SNP loci contained in the four targeted regions, which were used to estimate the final donor sequence composition of each pool. The known genotypes were also used to assess our ability to discover and genotype the rarest possible variants in each pool. A mixed linear model was used to test associations between common allele frequencies to disease, adjusting for pool compositional bias. We were able to replicate known T1D associations to marker SNPs in the Affy panels, and to identify nearby SNPs exhibiting increased significance. Many novel rare variants were also seen. We have also interrogated the unaligned 454 reads for mosaic sub-alignment patterns indicative of large-scale structural variation, identifying a number of candidates for further study.

POLYMORPHISM DISCOVERY IN LOW-PASS POPULATION SEQUENCING

Jared R Maguire[1], Eric Banks[1], Manuel Rivas[1], Mark DePristo[1], David Altshuler[1,2], Stacey Gabriel[1], 1000 Genomes Project Analysis Group[3], Mark J Daly[1,2]

[1]The Broad Institute of MIT and Harvard, Medical and Population Genetics, 7 Cambridge Center, Cambridge, MA, 02142, [2]Massachusetts General Hospital, Center for Human Genetic Research, 185 Cambridge St., Boston, MA, 02114, [3]Massachusetts General Hospital, Department of Molecular Biology, 185 Cambridge St., Boston, MA, 02114, [4] The 1000 Genomes Project

Many medical research topics require a comprehensive catalog of variations in a group of related individuals. Deep (30x) sequencing of whole genomes is still too expensive for most studies. However, since many variations are shared among individuals, the amount of sequencing per individual can be drastically reduced.

We describe a process for detecting variations in low-coverage (4x) sequencing of a population. Our method was developed to discover SNPs in the 1000 Genomes Project Pilot 1, This method is one of the three approaches that have been combined for the official project release.

Our method consists of three steps:

1. Compute (by evaluating in all samples sequenced) the odds in favor of a variant being present at each base in the genome. Emit these odds along with other annotation such as depth and strand bias.

2. Collect all sites having non-zero odds of being a SNP, and pick optimal annotation thresholds via a grid search.

In 60 CEU individuals at 4x depth, 90% of discoveries have an estimated true positive rate of 97%, another 5% have a TP rate of 80% and a final 5% have a TP rate of 67%.

In replication of a small number of sites, we meet a >90% true positive rate for SNP discovery.

3. For the sites that pass optimization, refine genotype likelihoods by leveraging inferred haplotype information using the program BEAGLE by Brian and Sharon Browning.

After refinement with BEAGLE, we attain a 5% non-reference genotyping error rate (GER) per individual when compared with genotype calls from a custom microarray. Without BEAGLE our average GER was 20%.

The complete tools for this pipeline are publically available in the Broad's GATK and in the BEAGLE software by Browning & Browning.

COMBINATION OF DIFFERENTIAL ALLELIC EXPRESSION IN NORMAL BREAST WITH GWAS DATA FOR IDENTIFICATION OF BREAST CANCER SUSCEPTIBILITY LOCI

Ana-Teresa Maia[1,2], Roslin Russell[1], Martin O'Reilly[1], Mark Dunning[1], Don Conroy[3], Caroline Baynes[3], SEARCH Team[3], Kerstin Meyer[1], Bruce Ponder[1,2,3]

[1]Cambridge Research Institute, CRUK, Robinson Way, Cambridge, CB2 0RE, United Kingdom, [2]University of Cambridge, Dept of Oncology, Addenbrooke's Hospital, Hills Road, Cambridge, CB20XZ, United Kingdom, [3]University of Cambridge, Dept of Oncology, Strangeways Research Laboratories, Cambridge, CB1 8RN, United Kingdom

We and others have performed genome-wide association studies (GWAS), which have identified a significant number of new loci associated with increased risk of developing breast cancer. Predictions are that most of the unidentified underlying variants have a regulatory function, as most are located in intergenic regions or in "gene deserts". We performed differential allelic expression (DAE) analysis, a robust technique to identify cis regulatory loci, in normal breast tissue to prioritise loci from the GWAS for further follow-up analysis.

Sixty-four samples were genotyped and assessed for allelic expression using Illumina Exon510S-Duo BeadChips. In 665 SNPs located in exons and UTR regions, 240 (36%) displayed preferential allelic expression ($p<0.05$) with 22 of these (3%) displaying highly significant DAE ($p<10^{-5}$). We have also identified genes known to be imprinted in other tissues, which also show allele-specific expression in breast tissue.

We are combining our data with the UK GWAS1 (Easton et al Nature 2007) and GWAS2 (unpublished) studies for breast cancer susceptibility, to generate a list of genes that show regulatory variation for further evaluation as candidates. We have initiated genotyping of these in a set of 14,000 cases and controls, to determine their association with susceptibility to breast cancer.

This is the first genome-wide differential allelic expression study in normal breast tissue, and has revealed a global cis regulatory map of breast tissue. We predict that this will be a powerful tool to combine with GWAS data to identify breast cancer susceptibility loci.

COMPUTATIONAL AND EXPERIMENTAL DEFINITION OF A MICROSATELLITE BASED UPON MUTATIONAL BEHAVIOR

Kateryna D Makova[1], Yogeshwar D Kelkar[1], Noelle Strubczewski[2], Suzanne E Hile[2], Francesca Chiaromonte[3], Kristin A Eckert[2]

[1]Penn State University, Department of Biology, 305 Wartik Lab, University Park, PA, 16803, [2]Penn State College of Medicine, Department of Pathology, 500 University Drive, Hershey, PA, 17033, [3]Penn State University, Department of Statistics, 505A Wartik Lab, University Park, PA, 16803

Microsatellite sequences are abundant in eukaryotic genomes and have high germline rates of strand slippage-induced repeat number alterations. However, the minimal number of repeats required to constitute a microsatellite has been debated, and a functional definition of a microsatellite that takes into account their mutational behavior has been lacking. To address the definition of a microsatellite, we investigated slippage dynamics for a range of repeat sizes, using two approaches. Computationally, we assessed size polymorphism at repeat loci in ten ENCODE regions resequenced in four human populations, assuming that the level of polymorphism reflects strand slippage rates. Experimentally, we determined the in vitro DNA polymerase-mediated strand-slippage error rates as a function of repeat number. In both approaches, we compared strand slippage rates at tandem repeats to the genome-wide background slippage rate at two-unit monitor loci. We observed two distinct modes of mutational behavior. At small repeat numbers, slippage rates were low, and indistinguishable from background measurements. A marked transition in slippage rates was observed as the tandem repeat array lengthened, such that at large repeat numbers, slippage rates were significantly higher than at monitor loci. For both mononucleotide and dinucleotide microsatellites studied, the transition length corresponded to a similar number of nucleotides: 9-10 repeats for [A/T]n and 5-6 repeats for [GT/CA]n. Our results argue for the existence of microsatellite threshold that is determined not by the presence/absence of strand slippage at repeats, but by an abrupt alteration in slippage rates and directionality, relative to background. These findings have implications for understanding microsatellite mutagenesis, for standardization of microsatellite analyses in completely sequenced genomes, and for predicting polymorphism levels of individual microsatellite loci.

ANCIENT AND MULTIPLE ORIGINS OF PRZEWALSKI'S HORSES

Kateryna D Makova[1], Hiroki Goto[1], Wen-Yu Chung[2], Oliver Ryder[3], Anton Nekrutenko[4]

[1]Penn State University, Department of Biology, 305 Wartik Lab, University Park, PA, 16803, [2]Penn State University, Department of Computer Science and Engineering, 505 Wartik Lab, University Park, PA, 16803, [3]San Diego Zoological Society, Beckman Center for Conservation Research, 15600 San Pasqual Valley Road, Escondido, CA, 92027, [4] Penn State University, Department of Biochemistry and Molecular Biology, 505B Wartik Lab, University Park, PA, 16803

Unraveling the genetic relationship between domestic and Przewalski's horses is critical for understanding the domestication history of the former and for formulating conservation strategies for the latter. Indeed, the endangered Przewalski's horse is the only true wild horse surviving today and is the closest relative of the domestic horse. Once found throughout the Eurasian steppe, Przewalski's horse had become virtually extinct, but was later bred in captivity and reintroduced to the wild. The present Przewalski's horse population originated from a mere 12 horses.The question of whether Przewalski's and domestic horses form distinct genetic clades is still contentious. To resolve this controversy, we used the next generation sequencing technology to determine the sequences of complete mitochondrial genomes as well as of substantial portions of the nuclear genomes in four Przewalski's horses representing all four surviving mitochondrial lineages. Two markedly distinct mitochondrial haplotypes among Przewalski's horses were observed, which was unanticipated, since their population went through a severe genetic bottleneck. The two mtDNA haplotypes discovered did not form a separate clade on a horse phylogenetic tree, thus rejecting the monophyly of Przewalski's horses, and were estimated to split 120,000-180,000 years ago, simultaneously with the divergence of the major modern horse lineages and significantly preceding horse domestication. The phylogenetic network analysis indicated that, among horse haplotypes analyzed, one of the Przewalski's horse mtDNA haplotypes might be ancestral to the extant horse mtDNA haplotypes. Thus, Przewalski's horses have ancient origins and are more genetically varied than previously realized.

CHARACTERIZATION OF HUMAN-SPECIFIC DELETIONS THAT HAVE BEEN FIXED IN OUR LINEAGE

Tomas Marques-Bonet, Lin Chen, Jarrett Egertson, Jeffrey M Kidd, Peter Sudmant, Gregory M Cooper, Carl Baker, Orangutan Genome Consortium, Evan E Eichler

University of Washington, Department of Genome Sciences, 1705 Pacific St, Seattle, WA, 98105

The less-is-more hypothesis has argued that some aspects of human evolution may have emerged by the loss of gene functions that are common among all other primates. Since the human genome shows extensive copy-number variation, we set out to identify and characterize segmental deletions (> 1kbp) that emerged specifically within the human lineage of evolution. To avoid potential assembly artifacts from non-human primate genomes, we developed a method to use chimpanzee and orangutan whole genome-shotgun datasets in combination with end-mapping, to identify and map segments that are missing from the human reference genome. Using aCGH, we experimentally validated 1,291 human-specific deletions when compare to all other great-ape species. The events ranged in size from 1 to 42 kbp and corresponded to 3.54 Mbp of sequence. Based on RefSeq data and gene models present in orangutan, we detected 65 genes that have been affected by these deletions and identified more than 500 Kb (~150 sites) of deleted ultra-conserved sequence among mammals, suggesting the possible loss of functional important elements during the course of human evolution. We assessed experimentally the polymorphism of these sequences in 60 human individuals from 7 different populations and from several individuals in different species of non-human primates. We show that 95.6% of the events are fixed in all human populations. Of the 56 remaining polymorphic segments that are present in at least one individual, we estimate that 30% of these sites restricted to one human continental group while 41% are fixed as a deletions in a particular population but still polymorphic in others. These data allow us to assign different events to different periods of human evolution and provide the first characterization of genes and gene families that have been completely erased from our lineage. Our preliminary data suggest that human specific duplications have affected 3-fold more sequence to our genome when compared to human specific deletions.

RNA-SEQ ANALYSIS OF GENE REGULATORY DIVERGENCE IN *DROSOPHILA*

Joel McManus[1], Joseph D Coolon[2], Michael O Duff[1], Jodi E Mains[1], Patricia J Wittkopp[2], Brenton R Graveley[1]

[1]University of Connecticut Health Center, Department of Genetics & Developmental Biology, 263 Farmington Ave, Farmington, CT, 06030, [2]University of Michigan, Department of Ecology & Evolutionary Biology, 830 North University, Ann Arbor, MI, 48109

The regulation of gene expression is biologically essential, and differences in gene expression are an important source of phenotypic variations between species. Regulatory divergence can result from changes in cis-acting sequences (cis) or in trans-acting factors (trans). Changes in cis are predicted to affect the regulation of single genes, while changes in trans are more likely to be pleiotropic. Subsequently, trans-regulatory divergence is predicted to be less common than cis-regulatory divergence.

We used RNA-seq to quantify total mRNA levels in two closely related Drosophila species (D. melanogaster and D. sechellia) and allele-specific gene expression in their F1 hybrid. By comparing these results, we dissected the genome-wide cis and trans contributions to regulatory divergence. 78% of expressed genes are differentially expressed between species, and 51% and 66% of expressed genes are affected by cis- and trans-regulatory divergence, respectively. This is a relatively larger trans contribution than expected, which may result from the historically small population size of D. sechellia. Genes with cis-regulatory divergence have lower rates of sequence conservation in their promoter regions. In addition, our results support numerous correlations between gene regulatory divergence mechanisms and the inheritance of gene expression.

We are currently investigating the regulatory divergence of alternative isoform expression. Our preliminary results suggest that regulatory divergence also contributes to species-specific differences in RNA-processing. For example, the spinster gene contains a set of mutually exclusive cassette exons that are known to modulate its function in ovarian and neuronal development. We identified and validated cis-acting regulatory divergence of pre-mRNA splicing that has resulted in differential exon utilization at spinster. In summary, our results illustrate the power of mRNA-seq for investigating the genetic changes underlying interspecific differences in gene expression and pre-mRNA splicing.

SINGLE NUCLEOTIDE VARIANTS ASSOCIATED WITH THERAPY-RELATED ACUTE MYELOID LEUKEMIA SUBTYPES

Megan E McNerney, Christopher D Brown, Kevin P White

University of Chicago, Institute for Genomics and Systems Biology, 900 E. 57th St., Chicago, IL, 60637

Therapy-related acute myeloid leukemia/myelodysplastic syndrome (t-AML) is a uniformly fatal complication that occurs in up to 5% of cancer survivors treated with chemotherapy or radiation. As more cancer patients achieve longer remissions, t-AML has increased in incidence, yet remains refractory to current therapy, with a median survival of 8-10 months. T-AML is thought to be a direct product of genetic mutations induced by alkylating agents, topoisomerase II inhibitors, and radiation. Two main subtypes of t-AML have been described that differ in their morphology, cytogenetic karyotype, and prognosis: alkylator-agent-induced and topoisomerase-II inhibitor-induced. Further understanding of t-AML will provide insight into non-therapy related, de novo AML, the most common acute leukemia of adults, with 12,000 new cases per year in the U.S. alone. Additionally, t-AML may serve as a model for the development of other cancers, which are often the result of the interaction between environmental exposures, such as carcinogens, and induced genetic changes. To identify genetic mutations associated with leukemogenesis, we have sequenced the leukemic mRNA of 26 t-AML patients to identify genetic abnormalities, including gene fusions, single nucleotide variants (SNVs), small insertions and deletions, and alternative mRNA isoforms, using Illumina paired-end sequencing. We have found over 50,000 SNVs per sample. Some of the affected genes have been previously described in t-AML, supporting our protocol and computational pipeline. Phenotypically relevant SNVs are prioritized by population-level allele frequency estimates, evolutionary constraint estimates, and the frequency of the aberrant gene or biological pathway within the patient samples. Putative disease-relevant SNVs are being compared between the two main subtypes of t-AML, as well as between therapy-related and de novo AML. Findings are being validated by constrast to germline DNA genotypes, as well as in a larger cohort of 100 t-AML leukemia samples and 50 de novo AML leukemia samples. This information will allow the identification of pathways altered in t-AML, correlation of mutations with clinical outcome, identification of potential biomarkers, and generation of candidate targets for leukemia therapeutics.

CANCER GENOME SEQUENCING OF PRIMARY TUMORS, XENOGRAFTS AND DERIVED CELL LINES.

John D McPherson[1], OICR/UHN/PMH Pancreatic Cancer Xenograft and Sequencing Team[2]

[1]Ontario Institute for Cancer Research, Cancer Genomics, 101 College St., Suite 800, Toronto, M5G0A3, Canada, [2]OICR/UHN/PMH, Cancer Genomics/Medical Biophysics/Oncology, University Ave., Toronto, M5G2C4, Canada

One in 75 men and women will be diagnosed with pancreatic ductal adenocarcinoma (PDAC) during their lifetime. Most will die of their disease within one year following diagnosis with an overall 5-year relative survival rate of only 5% - an outcome that has not appreciably changed since the 1970's. Unlike many other cancers, PDAC has remained recalcitrant to advanced treatments with little benefit seen from new therapeutics and targeted strategies. Identifying the molecular mechanistic undercurrent behind the disease will provide opportunities for developing strategies for prevention, treatment and early detection of PDAC.

As a contributing member of the International Cancer Genome Consortium, the OICR will generate a comprehensive catalogue of genomic abnormalities found in >350 PDAC. To obtain sufficient samples newly consented for large-scale sequencing, the OICR has established collaborations locally, at the Mayo Clinic, Rochester and three Boston-area Hospitals (Massachusetts General and Beth Israel Hospitals and Dana Farber Cancer Institute).

Our program utilizes xenografts generated from each tumor analyzed to augment the primary tumor and matched normal control. The xenografts provide ample material for analysis as well as reagents for functional analyses and pre-clinical models for target validation. In addition, establishment of cell lines from all xenografts is attempted with ~50% success rate. Initial efforts have focused on the deep characterization of early matched sets of normal, primary tumor, xenograft and associated cell line to establish the extent to which the xenografts and cell lines are representative of the initial tumor state. Data sets which are compared include copy number variation, miRNA, transcriptome, exome and whole genome obtained through next-generation sequencing. Xenografts derived from PDAC have proven to vary greatly in their human tumor content requiring the investigation of methods for enriching the tumor content prior to sequencing.

Comparison of somatic variants across sample types and enrichment methods will be presented as well as preliminary analyses of identified somatic events at this early stage of the ICGC project.

COMPARATIVE ANALYSIS OF THE CENTROMERE TANDEM REPEAT SEQUENCES IN EUKARYOTES

Daniël P Melters[1,2], Keith Bradnam[1], Simon Chan[2], Ian Korf[1]

[1]University of California, Davis, Genome Center, 1 Shields Dr, Davis, CA, 95616, [2]University of California, Davis, Plant Biology, 1 Shields Dr, Davis, CA, 95616

The centromere is a chromosomal locus that is essential for proper chromosome segregation during cell division. Despite their conserved, essential function, centromeres are characterized by the rapid evolution of both centromeric DNA sequence and centromeric structure. Most animal and plant species studied thus far have fast-evolving high copy tandem repeats in their centromeric regions. Centromere tandem repeat arrays are difficult to manipulate, so we are taking a comparative genomics approach to study their function. To gain a better understanding how widespread this feature is, we developed bioinformatics tools to identify candidate centromere sequences from whole genome shotgun sequences, and have looked at hundreds of eukaryote genomes. Tandem repeat sequences were identified with Tandem Repeats Finder. One theory states that the centromere repeat unit is about the size of one nucleosome; this is appealing because centromere function requires a centromere-specific form of histone H3. A large survey of centromere repeat unit size contradicts this theory. One model to explain the rapid evolution of centromeres is the ‘library’ hypothesis. This model predicts the presence of various tandem repeat arrays in a given genome, which, through stochastic amplification, compete to become the functional centromere. Homologies between centromeric sequences in species A and non-centromeric sequences in species B may be a signature of this type of evolution. By identifying similarity between centromeric and non-centromeric tandem repeats of different species we can test this ‘library’ hypothesis.

IDENTIFICATION OF A YY-1 BINDING SITE AS A CAUSAL VARIANT AT ONE OF 8Q24 PROSTATE CANCER PREDISPOSITION LOCI

Kerstin B Meyer, Ana-Teresa Maia, Maya Ghoussaini, Martin O Reilly, Radhika Prathalingam, Jason Carrol, Bruce A Ponder

Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinsons Way, Cambridge, CB2 0RE, United Kingdom

GWAS have identified at least nine independent cancer susceptibility loci upstream of the *c-myc* oncogene at chromosome 8q24. One of these, an approximately 80kb haplotype block, confers risk for both prostate and breast cancer. Genetic mapping studies have suggested that this haplotype block contains distinct causative variants for the two cancer types. We argued that candidate causative SNPs are likely to reside in open chromatin and have therefore mapped DNase I hypersensitive sites (DHSs) as means of prioritising regions for further functional analysis. Within the risk region, we have identified a strong, ubiquitous DHS that maps to a 500bp region known to carry H3K4me2, H3K4me3 and H3K9ac modifications in K562 cells. Within a 1.2kb fragment overlapping the DHS only few common SNPs exist, one of which, rs378854, is in perfect linkage disequilibrium with rs620861, the SNP most highly associated with prostate cancer within this risk interval. For this SNP the common allele carries increased risk, while the minor allele is protective. Using electrophoretic mobility shift assays we demonstrate that the minor allele of rs378854 generates a site with increased binding affinity for YY-1, a protein known to be able to act as transcriptional repressor. Two closely spaced Sp-1 binding sites are unaffected by the SNP. Using transient transfection assays we demonstrate that in the context of a 395bp fragment overlapping the DHS the minor allele of the rs378854 functions as a potent transcriptional repressor in the prostate cell lines PC3. In chromatin conformation capture experiments the DHS surrounding rs378854 was found to interact both with the *c-myc* and the *pvt-1* promoters, suggesting that both of these are potential targets of the identified repressor element. In conclusion, we have identified a likely causative variant for one of the 8q24 prostate cancer susceptibility loci, unusually the minor allele is protective and generates a site of transcriptional repression.

DISCOVERING FUNCTIONAL MODULES RELEVANT FOR CANCER PROGRESSION BY IDENTIFYING PATTERNS OF RECURRENT AND MUTUALLLY EXCLUSIVE MUTATIONS IN TUMOR SAMPLES.

Christopher A Miller[1], Stephen H Settle[2], Erik P Sulman[2], Kenneth Aldape[3], Aleksandar Milosavljevic[1]

[1]Graduate Program in Structural & Computational Biology & Molecular Biophysics and, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, TX, 77030, [2]Department of Radiation Oncology, the University of Texas M. D. Anderson Cancer Center, Houston, TX, 77030, [3]Department of Pathology, the University of Texas M. D. Anderson Cancer Center, Houston, TX, 77030

Recurrent somatic aberrations can implicate individual genes in cancer progression. Gaining insight into the roles of these genes often requires analyzing them in the context of pathways, interactions, and functional modules. Such analyses are typically limited by the incompleteness of pathway and interactome databases. To overcome this key problem, we develop a method that identifies functional modules without any information other than patterns of recurrent and mutually exclusive aberrations (RME patterns) that arise due to positive selection for key cancer phenotypes.

These RME patterns exist between sets of altered genes, and may be explained by the presence of functional relationships. Specifically, an aberration in one of the genes may result in the development of a key tumorigenic phenotype, removing selective pressure for mutation of the others.

We use this algorithm to do an integrative analysis of 145 glioblastoma tumors, using copy-number alteration and somatic mutation data from the TCGA. Without any knowledge of the interactome, we are able to extend known pathways by adding new members and discover new functional modules. After discovering that EP300 may play a role in glioblastoma, in the context of the TP53 pathway, we validate this hypothesis in two independent data sets by demonstrating that expression of EP300 predicts survival independent of age at diagnosis and tumor grade. We also demonstrate that the method is useful for generating hypotheses on very large data sets, like those that will be produced by current tumor characterization projects.

Though we apply this algorithm to tumor data, we suggest that it will be useful in a variety of other contexts, like identifying patterns of epigenetic marks that alter gene expression or finding specific sequence motifs.

HELMINTH GENOMICS: THE IMPACT ON GLOBAL HEALTH

Makedonka Mitreva

Washington University School of Medicine, The Genome Center, 4444 Forest Park Ave, St. Louis, MO, 63108

Helminth parasites are parasitic worms, from the phyla Nematoda (roundworms) and Platyhelminthes (flatworms). They infect one-third of humanity, and the collective burden of the common helminth diseases rivals that of the major high mortality conditions such as HIV/AIMDS or malaria. The diseases caused by helminth infections range from the dramatic sequelae of elephantiasis and blindness to the more subtle effects on child development, pregnancy and productivity, resulting in maintenance of poverty in the developing countries.

Recent transcriptome and genome projects have dramatically expanded the biological data available across the phylum Nematoda. At The Genome Center, we have continued development of molecular information, bioinformatics tools, and reagents for the study of parasitic nematodes, which is crucial to accelerating basic research and developing new diagnostics and therapeutics.

Here we present the genome of a basal nematode, the zoonotic parasite Trichinella spiralis that provided a bridge between basal and crown nematode species, thus spanning the phylum Nematoda. The analysis revealed genomic and molecular details of evolutionary relationships among this organism, and other lineages of parasitic and free-living nematodes, and the next closest relatives, the arthropods. Intrachromosomal rearrangements persisted along the evolution of the phylum, and circumscribed extensive genome plasticity in the Nematoda with more numerous gene loss and gain events than in the phylum Arthropoda. Within the Nematoda, protein family deaths outnumbered births for the parasitic relative to the free-living species, consistent with host exploitation. Finally, comparison of the basal nematode to other sequenced nematodes including Caenorhabditis elegans identified archetypical genes of potential evolutionary importance along with molecular signatures exclusive to all nematodes. These molecular determinants are of great value and provide foundations for developing broad new strategies to treat and/or eradicate global parasites that infect over 2 billion humans, and impede production of food animals and crops that provide basic nutrition for the global poor. Furthermore, delineation of these basic characteristics is essential for elucidation of principles underlying the evolution of metazoan animals.

DIRECT SINGLE CELL METHYLOME PROFILING IN MEMORY CIRCUITS: RAPID AND MASSIVE DNA DEMETHYLATION INDUCED BY NEUROTRANSMITTERS

Leonid L Moroz[1,2], A Kohn[1], M Citarella[1], E Bobkova[3], M Lyons[3], E Levandowsky[3], H Peckham[3], K McKernan[3]

[1]Univ Florida, Neuroscience, 100 Newell Dr, Gainesville, FL, 32610, [2]Univ Florida, Marine Bioscience, 9505 Ocean Shore Blv, St. Augustine, FL, 32080, [3]Life Technologies, AB, 500 Cummings Cnt, Beverly, MA, 01915

Persistent cell memory and plasticity are results of interactions between multiple transcriptional and epigenetic modifications in a cell's genome. These events are cell-specific and largely unknown. We (i) developed an unbiased single-neuron assay of genome-wide DNA methylation complemented by RNA-seq analysis from the same cell, and (ii) identified target genes involved in the control of unique neuronal identities and learning processes within one of the simplest memory-forming circuit of the mollusc Aplysia, a powerful model in neuroscience. For single cell methylome, we employed of a novel enrichment strategy (for selective capture of 5-methylated cytosines) and massive parallel sequencing (SOLiD) & designed sequencing libraries for directional RNA-seq. We achieved >40x coverage of the cellular genome in each of 6 neuronal types tested. Our data suggest that >45% of neuronal transcripts are differentially expressed among tested cells. We unexpectedly revealed very rapid (<2hrs) and global DNA demethylation in postmitotic neurons in response to application of serotonin (5HT; 9% vs 20% in control). 5HT is known as the inductor of long-term plasticity and the described process releases transcriptional repression (e.g. gene silencing via DNA methylation of promoter regions) leading to coordinated activation of multiple genes underlying learning and memory. The examples of methylated targets are the promoter regions of the CREB 2 cAMP-response element binding and Ubiquitin Hydrolase – 2 "master genes" essential for the initiation of coordinated gene expression cascades. DNA methylation also occurs outside of CpG islands, expanding the scope of regulatory mechanisms and the need for genome-wide correlation of methylome and transcriptional outputs from the same cell at different functional states that leads to cell fate specification including transcription dependent changes in synaptic efficacy in memory circuits.

THE GENOME OF THE CTENOPHORE PLEUROBRACHIA BACHEI - MOLECULAR INSIGHTS INTO INDEPENDENT ORIGINS OF NERVOUS SYSTEMS AND COMPLEX BEHAVIORS

Leonid L Moroz, F Yu, M Citarella, A Kohn

Univ Florida, Marine Biosciences, 9505 Ocean Shore Blvd, St. Augustine, FL, 32080

The origin and early evolution of animals is enigmatic, primarily because of the lack of molecular data from the basal Metazoa. The phylum Ctenophora (comb jellies) is one of the earliest lineages of prebilaterian animals having Precambrian fossil records (>560 Mya). They may be descendants of the most ancestral groups of the animals, branching before cnidarians and have the most basal neural & "true" mesoderm-derived muscular systems. The sea gooseberry, Pleurobrachia bachei, has one of the most compact genomes (~150 Mb). These large (1-3 cm) holoplanktonic predators are common in Northern Seas and have highly sophisticated ciliated locomotion, unique glue-based capture mechanisms & distinct development. We constructed paired-end sequencing libraries for 454-Roche/Illumina platforms, and achieved >50x coverage of the genome. In parallel, we performed RNA-seq profiling from major tissues and used these data for the genome annotation. As a result we have predicted ~15,000 genes in the Pleurobrachia genome. Using a subset of evolutionarily conserved genes, we validated the phylogenetic position of Ctenophora within the animal tree as the most basal group (however with a possibility that ctenophores could be a sister group of sponges). Such phylogenetic profiling is supported by comparative analysis of genes encoding homeo-domain transcription factors from basal metazoans (i.e. Sponges, Placozoans, Cnidarians and now Ctenophores). Next, we have characterized a subset of neurogenic genes as well as genes involved in intercellular signaling in Ctenophores. In concert with phylogenetic data, this analysis suggests that the nervous system in ctenophores involved independently from other animals. Surprisingly, we found that many "classical bilaterian neuron-specific" genes are not expressed in neurons of Pleurobrachia. Injury-associated adaptations leading to secretion of signal peptides, regenerative growth and formation of polarized secretory cells can be considered as the major factors driving the appearance of neurons in the first place. Identification of evolutionary predecessors of inter-neuronal signaling and candidate molecules for such ancestral signal peptides are currently under experimental validation using the tools of microanalytical CHEMISTRY.

THE SEQUENCING OF MULTIPLE GENOMES AND TRANSCRIPTOMES TO CHARACTERIZE THE EVOLUTION OF HOST SPECIFICITY IN NEMATODES OF THE GENUS *STEINERNEMA*

Ali Mortazavi[1], Adler Dilman[1,2], Igor Antoshechkin[1], Erich M Schwarz[1], Paul W Sternberg[1,2]

[1]California Institute of Technology, Division of Biology, 1201 East California Blvd, Pasadena, CA, 91125, [2]Howard Hughes Medical Institute, 1201 East California Blvd, Pasadena, CA, 91125

Nematodes are one of the largest invertebrate phyla with an estimated one million species occupying every conceivable niche. Besides the free-living model organism *Caenorhabditis elegans*, many nematode species are parasites of plants, insects, livestock, and humans; these parasites typically demonstrate great specificity to their hosts. We have sequenced and assembled the genome and transcriptomes of two very different nematode species as a pilot of the applicability of Illumina-only sequencing to a wider exploration of the phylum: an 85 Mb assembly of *Caenorhabditis* sp. 3 PS1010 with an N50 of 9.4 kb and an 84 Mb assembly of the beneficial entomopathogen (insect parasite) *Steinernema carpocapsae* with an N50 of 32 kb. We are using our developmentally staged RNA-seq transcriptomes of *S. carpocapsae* to make a quantitative, paired comparison of gene expression levels during development with the corresponding stages in *C. elegans* to identify which genes are most likely to enable nematodes to infect their hosts. We are currently sequencing *Steinernema scapterisci* and *Steinernema intermedium*, each of which have distinct host-specificity, to identify the evolution of non-coding cis-regulatory elements within the *Steinernema* clade, and to compare these elements to those conserved throughout the *Caenorhabditis* clade. The in-depth, systematic genomic exploration of parasitic nematode clades should give us better insight into how changes in gene content and regulation adapt different nematodes to their environmental niches.

NUCLEOTIDE-RESOLUTION ANALYSIS OF STRUCTURAL VARIANTS USING BREAKSEQ AND A BREAKPOINT LIBRARY

Hugo Y Lam[1], Xinmeng J Mu[1,2], Adrian M Stütz[3], Andrea Tanzer[4], Philip D Cayting[5], Michael Snyder[2], Philip M Kim[6], Jan O Korbel[3], Mark B Gerstein[1,5,7]

[1]Yale University, Program in Computational Biology and Bioinformatics, 266 Whitney Ave, New Haven, CT, 06520, [2]Yale University, Department of Molecular, Cellular and Developmental Biology, 219 Prospect St, New Haven, CT, 06520, [3]European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstrasse 1, Heidelberg, 69117, Germany, [4] University of Vienna, Institute for Theoretical Chemistry, Währingerstrasse 17/3/303, Vienna, A-1090, Austria, [5]Yale University, Molecular Biophysics and Biochemistry, 266 Whitney Ave, New Haven, CT, 06520, [6]University of Toronto, Terrence Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Toronto, M5S 3E1, Canada, [7]Yale University, Department of Computer Science, 51 Prospect St, New Haven, CT, 06520

Structural variants (SVs) are a major source of human genomic variation; however, characterizing them at nucleotide resolution remains challenging. Here we assemble a library of breakpoints at nucleotide resolution from collating and standardizing ~2,000 published SVs. For each breakpoint, we infer its ancestral state (through comparison to primate genomes) and its mechanism of formation (e.g., nonallelic homologous recombination, NAHR). We characterize breakpoint sequences with respect to genomic landmarks, chromosomal location, sequence motifs and physical properties, finding that the occurrence of insertions and deletions is more balanced than previously reported and that NAHR-formed breakpoints are associated with relatively rigid, stable DNA helices. Finally, we demonstrate an approach, BreakSeq, for scanning the reads from short-read sequenced genomes against our breakpoint library to accurately identify previously overlooked SVs, which we then validate by PCR. We have recently extended our BreakSeq approach to the 1000 Genomes Project pilot phase and compiled a breakpoint library of ~15,000 SVs thereof.

USING PROTEOGENOMICS TO VALIDATE AND REFINE GENOME ANNOTATION.

Jonathan M Mudge, Markus Brosch, Gary Saunders, Jennifer Harrow, Adam Frankish, Mark O Collins, Lu Yu, Jyoti S Choudhary, Tim Hubbard

The Wellome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

Recent advances in protein mass spectrometry are of great excitement to genome annotators, offering the chance to marry high-throughput peptide sequencing to predicted coding sequences (CDS), and to identify new coding loci. Here, we discuss the incorporation of tandem MS data into the HAVANA annotation pipeline, utilising the new software package Mascot Perculator which gives reliable confidence values to peptide assignments.

The HAVANA group have been manually annotating the gene content of the human and mouse genomes for 10 years. The intron / exon structures of our existing gene loci are already well described on the basis of EST / mRNA libraries, although additional alternative transcripts will be discovered as RNAseq data is collected from increasing numbers of cell types and developmental stages. In contrast, we have less confidence in the annotation of CDSs, simply due to the lack of direct evidence for translation. In particular there is a lack of experimental validation for alternative protein isoforms predicted to arise through alternative splicing. As such, estimations regarding mammalian proteome size can vary dramatically.

One million peptide fragments (generated in-house or taken from Peptide Atlas) were searched against a superset of Ensembl, Vega and IPI annotation, and the resulting alignments were subjected to manual analysis. These fragments represent a fraction of the proteomics data likely to be available in the near future. Even so, we could validate 32% of known protein coding loci whist identifying 29 completely novel putative genes. Furthermore, we observe translation from several pseudogene loci, potentially indicating their resurrection into coding forms. Our data also allow us to confirm putative protein isoforms generated by alternative splicing, refine existing splice junctions, and extend existing CDS predictions at the 5' and 3' ends. In all, this analysis demonstrates the potential of peptide to gene mapping in validating and refining genome annotation.

THE INFLUENCES OF CHROMATIN STRUCTURE ON TARGET SITE SELECTION BY THE *HERMES* TRANSPOSON

Loris Mularoni*[1], Sunil Gangadharan*[2], Nancy Craig[3], Sarah Wheelan[1]

[1]Johns Hopkins University School of Medicine, Oncology Biostatistics and Bioinformatics, 550 North Broadway, suite 1103, Baltimore, MD, 21205, [2]Johns Hopkins University School of Medicine, Molecular Biology & Genetics, 725 North Wolfe street, Baltimore, MD, 21205, [3]Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Molecular Biology & Genetics, 725 North Wolfe street, Baltimore, MD, 21205

Transposons are mobile genetic elements, present in virtually every known genome, that have great impact on chromosome structure by their sheer numbers and propensity to recombine. Transposons are also useful tools for genome exploration by insertional mutagenesis, and eventually they may be vectors for gene therapy. Analysis of large datasets of transposon insertion sites is useful for understanding the molecular mechanisms that underlie target selection by eukaryotic transposases. We have applied Solexa high-throughput sequencing methods and a ligation-mediated PCR approach to map the sites of insertion of the eukaryotic DNA transposon *Hermes* in the genome of the bakers' yeast *Saccharomyces cerevisiae*. We induced transposon insertions into naked DNA in vitro as well as into chromatin in vivo using both haploid and diploid yeast genome. We found that despite the relatively relaxed target site sequence preferences of the *Hermes* transposon, insertions are not uniformly or randomly distributed across the yeast genome but are clustered in specific regions and in certain locations. Here we illustrate the experimental strategy as well as bioinformatics methods used to study the targeting behavior of the autonomous *Hermes* transposable element in *Saccharomyces cerevisiae*; we have extended the consensus target site sequence and analyzed insertion preferences for specific genomic features. *Hermes* shows a strong bias for intergenic regions, and more specifically, for regions that flank transcriptional units and are free of nucleosomes.

* authors contributed equally

DNA SEQUENCE ANALYSIS OF A CLINSEQ PARTICIPANT USING WHOLE GENOME AND WHOLE EXOME SEQUENCING STRATEGIES

James C Mullikin, Hatice O Abaan, Jamie K Teer, Praveen F Cherukuri, Pedro Cruz, Nancy F Hansen, Daniel A King, Stephen C Parker, Gerard G Bouffard, Robert W Blakesley, David Ng, Eric G Green, Elliott H Margulies, Leslie G Biesecker

National Human Genome Research Institute and NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, MD, 20892

ClinSeq is a pilot project to investigate the use of large-scale medical sequencing as a tool for clinical research. Since 2007, this prospective study has recruited more than 800 of its targeted 1000 participants, who are consented broadly for research on all traits and for whole genome sequencing (WGS). With recent rapid advances in sequencing technology, we are now converting from a candidate gene approach focused on cardiovascular disease to whole exome sequencing (WES) and WGS. One participant was selected for WGS after analysis of 251 candidate genes did not identify a genetic etiology for his high coronary calcification level. We generated over 50X sequence coverage of this individual's genome using Illumina GAiiX instruments. Reads were aligned to the human genome reference sequence using a hybrid eland/cross_match process, followed by genotype calling using a Bayesian method called Most Probable Genotype. Initial sequence coverage analysis indicated a 1.4Mb hemizygous deletion that is known to cause hereditary neuropathy with liability to pressure palsies (HNPP, OMIM: 162500); a result, which after CLIA validation was returned to this participant. However, this is not the cause of his elevated coronary calcification. Thus we have inspected the heterozygous non-synonymous variants after removing those seen in other datasets and ranking the severity of amino acid changes to prioritize the review process. We have recently generated WES for this individual, and interestingly, the WES yields higher coverage of the CCDS than the 50X WGS, thus we will combine both data sets to increase our completeness across the coding portion of the genome in search of the causative mutation.

PATTERNS OF INCOMPLETE LINEAGE SORTING AND ANCESTRAL POPULATION GENETICS AMONG THE GREAT APES

Kasper Munch, Thomas Mailund, Asger Hobolth, Julien Y Dutheil, Mikkel H Schierup

Aarhus University, Bioinformatics Research Centre, C.F. Møllers Allé 8, Aarhus C, 8000, Denmark

Most species and subspecies of apes have now been sequenced and the genomes of these closely related species are easily aligned. Along such an alignment, divergence times between species differ due to segregating polymorphism in the ancestral species. For some species the population size of the ancestral species is sufficiently large and the time span between speciation events is sufficient small that ancestral polymorphism may lead to gene trees with a topology different from the species tree. This phenomenon is termed incomplete lineage sorting (ILS) and implies that segments of the genome will display a closer relation to species other than the sister species. ILS is well established between human, chimpanzee and gorilla and has also been proposed between chimpanzee, bonobo and humans.

We have developed a theoretical framework that allows for inference of population genetic parameters and patterns of ILS. The framework is based on a hidden Markov model where the hidden states along the alignment represent gene trees with separate topologies and separate coalescent times. We apply the model to the bonobo genome and demonstrate ILS between bonobo, chimpanzee and humans. We also present results from the gorilla genome with focus on the human, chimpanzee speciation time and ancestral population size. We describe how the occurrence of ILS correlates with gene annotation and recombination rate and with the X/autosome contrast. We find evidence for more ILS in regions of high recombination and in regions of low gene density consistent with widespread selection.

IMPROVED MICROBIAL ASSEMBLY AND FINISHING USING 8KB 454 LIBRARIES

Donna Muzny, Christian Buhay, Yuan-Qing Wu, Shannon Dugan, Xiang Qin, Irene Newsham, Sarah Highlander, Joseph Petosino, Richard Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030

The BCM-HGSC is one of four major centers performing the Human Microbiome Project. To date, 104 microbial genomes have been assembled as part of the jumpstart portion of this project. Of these assembled microbes, 33 have been finished. As we have continued to evaluate additional microbial genomes, we have identified some key issues related to assembly and finishing. A major hurdle in assembly and finishing of microbial genomes has been the initial assembly of large repeats such as rRNA operons. These regions are about 5kb in length and cause inconsistencies in the draft assembly and scaffolding due to the presence of multiple copies of this highly similar sequence within each genome. Ordering and orienting these regions within the assembly requires extensive work with multiplex and long range PCR. Here we show that larger 454 insert paired end libraries of 8-10Kb significantly improve the quality of the draft assembly and scaffolding, providing a mechanism to sequence and finish highly repetitive genomes. These longer 8Kb inserts have been applied to 13 organisms and sequenced along with the traditional 3Kb inserts for assembly comparison using N50 scaffold and N50 contig metrics. As part of the evaluation, 4 organisms of high GC content (63 to 72)% and demonstrated poor contiguity were included. Using a normalized 20 fold sequence coverage the traditional 3Kb libraries produced 183Kb to 1.3Mb scaffold N50, while the 8Kb libraries increases the scaffold N50 as much as 8 fold (1.2-5.2 Mb). The contig N50 statistics remained constant at an average of 80Kb. Most significantly, scaffolding statistics showed that for 63% of these organisms, the genome was found in one large scaffold obviating the need for optical mapping. Success with this new technique indicates the majority of microbial genomes will be ordered and oriented upon initial assembly, providing greatly enhanced draft assemblies that require significantly less finishing efforts. This strategy has been implemented for all reference strain sequence assemblies and finishing.

TECHNOLOGY ADVANCEMENTS FOR WHOLE EXOME AND WHOLE GENOME SEQUENCING

Donna Muzny[1], Jeff Reid[1], Mark Wang[1], Yuan-Qing Wu[1], Irene Newsham[1], Huyen Dinh[1], Matthew Bainbridge[1], Thomas Albert[2], Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030, [2]Roche NimbleGen, Inc., Advanced Research, 1 Science Court, Madison, WI, 53719

Current NexGen technologies have now become the vehicle by which genome analysis will transform medicine. For the HGSC SOLiD pipeline, capacity is projected to increase with the use of higher density slides to yield 90-100Gb per run and a total center capacity of 6 Tb/month. Internal activities have developed library automation, bulk emPCR processes and application of read mapping tools such as BFAST to match pipeline flow with instrument capacity.

Development of whole exome capture protocols using NimbleGen arrays and solution-based reagents have been a major pipeline focus. Laboratory advances include optimizations of adaptor ligation efficiency, input DNA quantity, and pre-and post-capture LM-PCR cycle number as well as automation of the capture process. Whole exome liquid phase reagents (NimbleGen SeqCap EZ Exome) have been tested for SOLiD sequencing. A second, larger custom whole exome design (VCROME) was developed at the HGSC and includes 20K genes. Over 200 whole exome capture samples have been completed for cancer and functional mutation discovery projects including TCGA- glioblastoma, Hepatocellular carcinoma (HCC), Autism and two projects (Diabetes and Hypertension) in collaboration with Eric Boerwinkle and the Atherosclerosis Risk in Communities Study. Current capture metrics for whole exome studies achieve 80% of the target bases at 20X coverage.

Whole genome sequencing activities have also expanded with over 30 deep coverage (30X) whole genomes now completed for medically related diseases including TCGA- ovarian and glioblastoma, breast cancer, HCC, Charcot-Marie-Tooth neuropathy (CMT) and a new Marfan-related disease. Here our functional mutation detection program has succeeded in identifying clinically relevant variants in a pedigree afflicted with an inherited neuropathic disease using whole genome sequencing.

A GENOME-WIDE VIEW OF RECOMBINATION, SELECTION, AND DRUG RESISTANCE IN A SOUTHEAST ASIAN POPULATION OF *PLASMODIUM FALCIPARUM*

Rachel A Myers[1,2,3], Jianbing Mu[4], Xin-zhuan Su[4], Philip Awadalla[1,2]

[1]University of Montreal, Department of Pediatrics, 3175, chemin Côte Sainte-Catherine, Montreal, H3T 1C5, Canada, [2]University of Montreal, CHU Ste Justine Research Center, 3175 chemin de la cote-Sainte-Catherine, Montreal, H3T 1C5, Canada, [3]North Carolina State University, Bioinformatics Research Center, 840 Main Campus Drive, Raleigh, NC, 27606, [4] National Institutes of Health, Laboratory of Malaria and Vector Research, 12735 Twinbrook Parkway, Bethesda, MD, 20892

Plasmodium falciparum is the most virulent agent of malaria; affecting 300-500 million people a year. Due to the parasite's unique relationship with the human and mosquito hosts, coupled with rapid evolution of anti-malarial drug resistance, the evolutionary history of adaptive changes mediating drug resistance has been the focus of recent research. Genome-wide association studies provide the framework to test for candidate loci mediating drug response, and with newly available genotyping arrays, researchers are able to use this approach in *P. falciparum*. The first genome-wide scan of recombination, positive selection, and anti-malarial drug resistance associations in a world-wide collection of samples yielded several results; e.g. *pfcrt* had strong signatures for both positive selection and association with chloroquine response. Among other results from this study, we noted the population structure analysis of the samples collected from Asia revealed separate populations corresponding with samples from Thailand, Cambodia, and admixture between the two countries. We refined our analysis of these three sub-populations; Thai, Cambodian, and Thai-Cambodian admixed, to characterize the admixture, estimate recombination events and rates, and identify signals of natural selection shared and specific to each sub-population. We related these findings to those of drug resistance measurements, finding differences between the admixed group and the remaining two populations for key anti-malarial drugs. This gives us new insight to occurrence and spread of drug resistance in *P. falciparum*.

GALAXY - FROM SAMPLE TRACKING TO SNP CALLING: AN INTERACTIVE POSTER

Ramkrishna Chakrabarty[1,4], Greg Von Kuster[1,4], Mark Chee[2], James Taylor[3,4], Anton Nekrutenko[1,4]

[1]Penn State, CCGB, 505 Wartik, University Park, PA, 16802, [2]Prognosys Biosciences, PrognosysBio, 505 Coast Blvd, La Jolla, CA, 92037, [3]Emory, Biology, 1510 Clifton Road NE, Room 2006, Atlanta, GA, 30322, [4]galaxyproject.org

A new generation of DNA sequencing technologies has enabled a variety of novel genome-scale experimental techniques. What is perhaps most unique about this recent data explosion is that it is distributed – relatively inexpensive instruments allow any lab or institution to produce enormous amounts of data. Yet the infrastructure upstream and downstream of sequencing instruments is largely undeveloped. In addition to the instrument cost labs, core facilities and sequencing service providers are forced to spend thousands on commercial LIMS systems and sequence analysis packages, which are in-turn based on tools from the public domain.

Galaxy (http://usegalaxy.org) provides a robust open-source alternative. Its lightweight sample tracking system is aimed at helping small labs and core facilities managing requests for sequencing runs. It allows one to track the entire "life-cycle" of sequencing request from the initial sample submission to the resulting dataset. Once the run is complete the user can apply a variety of NGS tools including format converters, mappers, ChIP-seq and transcriptome utilities. Results of these analyses can be visualized, shared, and published.

In this interactive poster we will demonstrate sample tracking functionality from the moment of sample submission to the sequencing facility, through the sequencing run, until the sample becomes a dataset and can be analyzed with a variety of NGS tools.

THE ARCHITECTURE OF THE REGULATORY LANDSCAPE ACROSS MULTIPLE TISSUES: THE MUTHER STUDY

Alexandra C Nica[1,2], Leopold Parts[1], Stephen B. Montgomery[2], Antigone Dimas[2,3], James Nisbett[1], Magdalena Sekowska[1], Amy Barrett[3], Mary Travers[3], Simon Potter[1], Tsun-Po Yang[1], Josine Min[3], Elin Grundberg[1,4,] Kerrin Small[1,4], Åsa Hedman[3], Daniel Glass[4], Krina T. Zondervan[3], Kourosh Ahmadi[4], Richard Durbin[1], Panos Deloukas[1], Mark I. McCarthy[3], Timothy D. Spector[4], Emmanouil T. Dermitzakis[2] for the MuTHER consortium

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; [2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; [4]Department of Twin Research, King's College London, London, UK

Regulation of gene expression is a crucial cellular function determining a significant proportion of the phenotypic variance. In the current study we thoroughly explore the complexity of the human regulatory variation landscape and how regulation is differentially variable in three tissues (LCLs, fat and skin) derived from a subset (N ~ 200) of phenotypically well-characterized twins from the MuTHER resource (http://www.muther.ac.uk/). We discover an abundance of expression quantitative trait loci (eQTLs) per tissue using standard methods and we apply factor analysis to remove effects of latent variables and further increase the power to detect larger numbers of eQTLs. Our discovery rate is similar to previous studies for such sample size, however the factor analysis doubles the number of eQTLs in each tissue at the same threshold of significance. The study design permits immediate replication between twins of the same family and hence validation of the substantial gain in eQTL discovery. With both strategies we observe and confirm extensive tissue specificity in all tissues, two of which are of clinical importance and one has not been previously characterized. As current threshold based estimates of tissue specificity are highly dependent on power, we explore and propose alternative methods to measure tissue specificity. We map the discovered eQTLs to historically independent genomic regions being thus able to provide a refined description of the tissue dependent particularities of regulatory variants. We observe that the majority of the genes are regulated by single independent eQTLs and we describe their spatial properties and distribution around basic gene structure landmarks. These discoveries, combined with the thorough phenotypic information available, will substantially aid the project's further efforts to understand the mechanisms behind complex trait susceptibility.

HIGH-RESOLUTION LANDSCAPE OF *UBE3A* ALLELIC EXCLUSION REVEALED BY USING HIGHLY PARALLEL SNP TYPING

Koji Numata[1], Chihiro Kohama[2], Kuniya Abe[1,2], Hidenori Kiyosawa[2,3]

[1]RIKEN BRC, Team for Mammalian Cellular Dynamics, Koyadai, Tsukuba, 305-0074, Japan, [2]University of Tsukuba, Graduate School of Life and Environmental Sciences, Tennoji, Tsukuba, 305-8577, Japan, [3]RIKEN BRC, Team for BioSignal Program, Koyadai, Tsukuba, 305-0074, Japan

Recent transcriptomic studies in mammals have identified many examples of natural antisense transcript (NAT), which is transcribed from the opposite strand of an annotated gene. A possible regulatory role of NAT is the regulation of monoallelic gene expression. Many monoallelically expressed genes have been identified thus far, and some of them have been shown to possess an antisense counterpart. These NATs are often broadly transcribed in the vicinity of 100–1000 kb, suggesting that NAT-mediated regulation of the loci is achieved in a locus-wide manner.
We have studied the locus-wide landscape of *Ube3a* (ubiquitin protein ligase E3A) expression and its antisense expression along with the strand- and allele-specificity. We have initially identified 1,420 single nucleotide polymorphism (SNP) sites between C57BL/6J (B6) and MSM/Ms, which is an inbred strain derived from the Japanese wild mouse (*M. musculus molossinus*), within the *Ube3a-Snurf/Snrpn* region according to the nucleotide sequences of MSM BAC clones and Illumina GoldenGate genotyping assay.
We then performed RNA-targeted highly parallel SNP typing by obtaining RNA from the brain and liver tissues of two reciprocal F1 hybrids. This provided a detailed view of the transcriptional landscape of *Ube3a* allelic exclusion. Because *Ube3a* is biallelically expressed in the liver in the absence of antisense transcription, antisense expression may play a crucial role in monoallelic gene expression. Moreover, the present study revealed that (1) brain-specific maternal expression of *Ube3a* is restricted in the second half of the locus but not in the first half, and (2) the brain-specific antisense of *Ube3a* is significantly decreased in the upper region of the *Ube3a* transcription start site. Although these findings do not reflect the conditions in a particular cell type of the tissue, our result reveals many aspects of the locus-wide landscape of sense and antisense transcriptional competition.

AN EFFICIENT AND ROBUST ALGORITHM FOR INFERRING ANCESTRY IN ADMIXED GENOMES

Larsson Omberg[1], Ronald Crystal[2], Andy Clark[3], Jason Mezey[1,2]

[1]Cornell University, Biological Statistics and Computational Biology, Biotechnology Building, Ithaca, NY, 14853, [2]Weill Cornell Medical College, Genetic Medicine, 1305 York Ave, New York, NY, 10065, [3]Cornell University, Molecular Biology and Genetics, Biotechnology Building, Ithaca, NY, 14853

Accurate determination of population structure and ancestral origin of individual haplotypes in admixed individuals can be used to detect recent historical events and can be leveraged to avoid spurious associations between disease and genetic loci in genome-wide association studies. Most proposed methods for mapping admixture require unlinked markers or phased data, and most are computationally inefficient for analyzing entire genomes when considering high coverage marker data. We present a novel method using support vector machines to classify segments of the genome based on the alleles of ancestral or approximately ancestral genomes. Support vector machines are extremely efficient classifiers that we use over windows of the genome. By simultaneously testing the ability of the support vector machine to classify the ancestral populations we are able to dynamically assign window sizes that corresponds to the amount of information present in the genome over the window. We show that this method is almost independent of prior knowledge of recombination rates, generations since admixture and tunable parameters in the algorithm. The algorithm also efficiently scales to millions of markers and hundreds of genomes. To demonstrate the value of this method, we analyze a population from Qatar of varying degrees of admixture between populations from Northern Africa, Persia and the Arabian peninsula.

EXOME AND CNV “HOTSPOT” RESEQUENCING IN AUTISM

Brian J O'Roak, Akash Kumar, Sarah B Ng, Ian Stanaway, Santhosh Girirajan, Choli Lee, Emily H Turner, Evan E Eichler, Jay Shendure

University of Washington School of Medicine, Genome Sciences, 1705 NE Pacific St., Seattle, WA, 98195-5065

Autism is the most genetic of all neuropsychiatric syndromes. However, identifying the genetic factors involved has proved difficult. We are using exome and targeted resequencing as complementary approaches to identity novel risk loci. Our targeted approach is focused on 17 “hotspot” regions of the genome where recurrent copy number variants (CNVs) have been associated with autism, including: 1q21.1, 16p11.2 and others (Itsara et al. 2009). Most of these regions are large and involve >10 genes, making the identification of the underlying risk loci difficult. We aim to identify the causative genes by resequencing of all coding exons within each CNV interval using molecular inversion probes (MIPs) in a large set of autistic individuals. Early studies using MIPs have demonstrated extensive multiplexing, high specificity, low DNA input requirements, and simplified library construction (Porreca et al. 2007, Turner et al. 2009). We have designed a 13k probe MIP set to simultaneously target 450 genes within our hotspot intervals and 65 autism candidate genes (1Mb). We have begun a pilot screen, with the goal of increasing to a final set of 1,320 samples. Initial findings and performance of the MIP assay will be presented. We are comparing the efficacy of this high-throughput targeted approach with complete exome sequencing (Ng et al. 2010). We have selected 20 autism trios showing a strong likelihood of sporadic autism where deleterious CNVs and point mutations have not been identified, with the goal of identifying *de novo* disruptive coding mutations. By analyzing complete trios we can determine the global rate of *de novo* events in these families and prioritize those genes carrying disruptive mutations. Furthermore, we will search for evidence of genic overlap with rare CNVs in a database of over 20k patients with developmental delay, autism, or intellectual disability. The development of methods to rapidly and economically screen large numbers of individuals over large number of loci is a critical complement to next-generation sequencing.

TRANSCRIPT ASSEMBLY AND ABUNDANCE ESTIMATION FROM RNA-SEQ REVEALS THOUSANDS OF NEW TRANSCRIPTS AND SWITCHING AMONG ISOFORMS

Cole Trapnell[1,2,5], Brian A Williams[3], Geo Pertea[2], Ali Mortazavi[3], Gordon Kwan[3], Marijke J van Baren[4], Steven L Salzberg[1,2], Barbara J Wold[3], Lior Pachter[5]

[1]University of Maryland, Computer Science, Biomolecular Sciences Bldg, College Park, MD, 20742, [2]University of Maryland, Center for Bioinformatics and Computational Biology, Biomolecular Sciences Bldg, College Park, MD, 20742, [3]California Institute of Technology, Division of Biology and Beckman Institute, 1200 California Blvd, Pasadena, CA, 91125, [4] Washington University, Genome Sciences Center, 1 Brookings Drive, St. Louis, MO, 63130, [5]University of California, Berkeley, Mathematics & Molecular and Cell Biology, Evans Hall, Berkeley, CA, 94720

Large-scale sequencing of mRNA should reveal, with unprecedented quantitative detail, the complete pattern of gene expression for any cell type. However, this potential has not yet been fully realized because of computational and experimental hurdles. One difficulty is that genes have multiple isoforms that complicate the assembly of short reads into complete transcripts and confound abundance estimation. We introduce an approach to transcript discovery coupled with a statistical model for RNA-Seq experiments that produces estimates of transcript abundances. Our algorithms are implemented in a freely available and open source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed more than 430 million paired 75bp RNA-Seq reads from a mouse myoblast cell line representing a differentiation timeseries. We detected 13,689 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Analysis of transcript expression over the timeseries revealed complete switches in the dominant transcription start site (TSS) or splice-isoform in 330 genes, along with more subtle shifts in a further 1,304 genes. These dynamics suggest substantial regulatory flexibility and complexity in this well-studied model of muscle development.

PIGGYBAC-ING ON A PRIMATE GENOME: NOVEL ELEMENTS, RECENT ACTIVITY AND HORIZONTAL TRANSFER.

Heidi J Pagán[1], Jeremy D Smith[1], Robert M Hubley[2], David A Ray[1]

[1]Mississippi State University, Biochemistry and Molecular Biology, 402 Dorman, Mississippi State, MS, 39762, [2]Institute for Systems Biology, Computational Biology, 1441 North 34th Street, Seattle, WA, 98103

Early surveys of Class II transposable element (TE) activity in mammals revealed a widespread shutdown around 40 mya in the genomes of model organisms. The first deviation from these observations appeared in the little brown bat, *Myotis lucifugus*, which was found to harbor multiple highly and recently active Class II elements. Further analyses of recent activity from the *hAT* superfamily suggested reasons to question the extent of Class II TE extinction in mammals. Anthropoid genomes analyzed thus far showed no signs of recent activity. Thus, we chose to probe the genome of the gray mouse lemur (*Microcebus murinus*) in order to determine if Class II TEs have also become extinct in strepsirrhine primates. Several elements were found to have been active within the last 40 my. Most notable were members from the *piggyBac* superfamily, which included an autonomous element which is very similar (Blastn E-value = 0) to *piggyBac2_ML* in the genome of *M. lucifugus* yet absent from other strepsirrhine primates and from the genomes of other surveyed mammals. Age analyses and comparative analyses with other primates indicate recent, lineage-specific activity. In combination with a second, novel *piggyBac* family, these results suggest that horizontal transfer has played an important role in the continued success of Class II TE elements in the genome of *Microcebus murinus* and in avoiding a complete shutdown of Class II activity in primates.

STRONG PURIFYING SELECTION AT GENES ESCAPING X CHROMOSOME INACTIVATION

Chungoo Park[1,2], Laura Carrel[1,3], Kateryna Makova[1,2]

[1]The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, Wartik Lab, University Park, PA, 16802, [2]The Pennsylvania State University, Department of Biology, Mueller Lab, University Park, PA, 16802, [3]The Pennsylvania State University College of Medicine, Department of Biochemistry and Molecular Biology, Hershey Medical Center, Hershey, PA, 17033

To achieve dosage balance of X-linked genes between mammalian males and females, one female X chromosome becomes inactivated. However, approximately 15% of genes on this inactivated chromosome escape X chromosome inactivation (XCI). Here, using a chromosome-wide analysis of primate X-linked orthologs, we test a hypothesis that such genes evolve under a unique selective pressure. We find that escape genes are subject to stronger purifying selection than inactivated genes and that positive selection does not significantly affect the evolution of these genes. The strength of selection does not differ between escape genes with similar vs. different male-female expression levels. Intriguingly, escape genes possessing Y homologs evolve under the strongest purifying selection. We also found evidence of stronger conservation in gene expression levels in escape than inactivated genes. We hypothesize that divergence in function and expression between X and Y gametologs is driving such strong purifying selection for escape genes.

NEW FAMILIES OF HUMAN REGULATORY RNA STRUCTURES IDENTIFIED BY COMPARATIVE ANALYSIS OF VERTEBRATE GENOMES.

Brian J Parker, Ida Moltke, Jakob S Pedersen

University of Copenhagen, Department of Biology, BINF, Ole Maaloes Vej 5, Copenhagen, DK-2200, Denmark

Regulatory RNA structures in both cis-regulatory elements and ncRNAs are often members of families with multiple paralogous instances spread across the genome. Based on genome-wide alignments of 41 vertebrate genomes produced by the 2x mammalian sequencing consortium, we developed a computational method, EvoFam, to identify new families of regulatory RNAs in human.
Method: First, the EvoFold program, which identifies stem-pairing regions from their characteristic substitution pattern, was used to predict structural RNAs genome-wide. We modeled both the sequence and structural information of these predictions in a position-specific manner using profile stochastic context-free grammars. We defined an inter-model similarity measure used to cluster the EvoFold predictions into families, correcting for model-dependent false positive rates, using a graph-based approach that detects densely connected subgraphs and is robust in the presence of noise.
Results: We detected 220 families, with an estimated FDR of < 5% (for families of size > 3). The method correctly identifies known families of cis-regulatory structures, e.g., iron responsive elements in TFRC, and translational control elements found in collagen gene 5' UTRs. Amongst ncRNAs, the method identifies known miRNAs, snoRNAs and a recently characterized family including the linc-RNA MALAT1.
More excitingly, it identifies multiple new candidate families, e.g., a family of six long hairpins in the 3'UTR of the key metabolic gene MAT2A, producer of the primary methyl donor, that we hypothesize is involved in transcript stability regulation in response to substrate concentration. We have also identified families of short six base-pair long hairpins enriched in immunity-related genes, e.g., TNF, FOS and CTLA4. Individually, these hairpins would be extremely difficult to detect with high confidence, but taken together the family is supported by strong evidence, and includes known transcript destabilizing elements. Another family includes a tRNA-like structure in the intron of the tRNA maturation gene POP1, potentially involved in auto-regulation. The identification of these structural families has generated hypotheses testable by directed experiments, which may further elucidate human post-transcriptional regulation.

WHOLE-GENOME SEQUENCING AND ANALYSIS OF MULTIPLE PAIRS OF PATIENT-MATCHED MELANOMA TUMOR AND NORMAL SAMPLES

Stephen C Parker[1], Isabel Cardenas-Navia[1], Hatice Ozel Abaan[1], Jamie K Teer[1], Praveen F Cherukuri[1], Pedro Cruz[1], Nancy F Hansen[1], Subramanian S Ajay[1], Andrew L Young[1], James C Mullikin[1], Steven A Rosenberg[2], Yardena Samuels[1], Elliott H Margulies[1]

[1]National Institutes of Health, National Human Genome Research Institute, 5625 Fishers Lane, Bethesda, MD, 20892, [2]National Institutes of Health, National Cancer Institute, 6116 Executive Boulevard, Bethesda, MD, 20892

Melanoma is the most common lethal skin cancer. Despite years of research, the median patient survival is six months following metastatic diagnosis, with less than 5% surviving five years. As a pilot towards cataloging all genetic variations associated with melanoma, we present the sequence and analysis of multiple pairs of melanoma tumor and patient-matched normal genomes.

Using the Illumina GAIIx and HiSeq 2000 platforms coupled with new bioinformatics methods, we generated 30X coverage genome builds. For each pair of tumor and normal genomes, we identified copy number and single nucleotide variants (CNVs and SNVs). Variant calls were verified with an array platform and with Sanger sequencing technology, allowing us to develop a robust algorithm for identifying variants from the whole genome builds.

The vast majority of tumor-specific variants occur in non-coding regions and it is unclear how these SNVs are partitioned among passenger and driver classes of mutations. Since DNA shape can be important for protein binding specificity, we searched for SNVs that disrupt potentially functional shapes. Evolutionarily conserved non-coding regions accumulate significantly less mutations than expected. A previous study reported the DNA mutational signature of a melanoma genome and how it reflects ultraviolet light exposure. Here, we examine this signature from the perspective of local DNA structure.

We provide results that help identify the underlying genetic components of melanoma. Future plans include sequencing additional melanoma genomes **TO BUILD TUMOR-SPECIFIC SIGNATURES AND BETTER CLASSIFY DISEASE STATE.**

APPLYING ARRAY CAPTURE, ILLUMINA SEQUENCING, AND NEUROBIOLOGY TO THE INVESTIGATION OF BIPOLAR DISORDER

Jennifer S Parla[1], Melissa Kramer[1], Ivan Iossifov[1], Fernando S Goes[2], James B Potash[2], W. Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Genomics and Quantitative Biology, 1 Bungtown Road, Cold Spring Harbor, NY, 11724, [2]Johns Hopkins School of Medicine, Psychiatry and Behavioral Sciences, 600 N. Wolfe Street, Baltimore, MD, 21287

The high heritability of bipolar disorder emphasizes a significant genetic component of the illness, and several genetic studies have been conducted in order to characterize its etiology. However, linkage and genome-wide association studies have had limited success in identifying common variants and particular genes consistently associated with bipolar disorder, suggesting that rare genetic variants and multiple genes may play significant roles in the bipolar phenotype. To continue the efforts to identify genes and variants that contribute to bipolar disorder, we have initiated a targeted re-sequencing approach that focuses on genes that encode synaptic proteins as well as genes that were identified through genome-wide association studies of bipolar disorder. The synapse is a critical component of brain and neurological function, and studies continue to associate psychiatric and cognitive disorders with mutations in synapse genes. Our approach involves the use of custom Agilent 244K arrays that are designed to target the exons and promoters of up to 1,124 genes. Currently, our largest data set consists of Illumina sequence data generated from array captures performed with 30 pilot study samples and a synaptome array design that targets 618 genes and a total of 3.9 Mb of genomic space. Our best captures with our pilot study samples have produced over 99% target coverage and 78% of the targets covered at ≥ 20X depth when evaluated using our original ELAND-based analysis method. Subsequent data analyses using BWA and strict requirements for unique mapping coordinates at the base level per sequence read have revealed that significant improvements in the capture procedure are necessary in order to obtain proper and adequate coverage of our targets. We have also started classifying the variants that we have found based on the results of genetic studies performed by other groups in order to formulate a strategy to validate these variants.

HUMAN CIS-REGULATORY SNPS (CIS-RSNPS) ALTERING TRANSCRIPTION

Tony Kwan[1], Dominique Verlaan[1], Manon Ouimet[2], Bing Ge[1], Vincent Gagné[2], Kevin Lam[1], Vonda Koka[1], Kevin Gunderson[3], Daniel Sinnett[2], Tomi Pastinen[1]

[1]McGill U, Hum Gen, 740 Dr Penfield, Mtl, H3A1A4, Canada, [2]U de Mtl, HSJ, 3175 Côte-Ste-Catherine, Mtl, H3T1C5, Canada, [3]Illumina, Adv Res, 9885 Towne Centre Dr, San Diego, CA, 92121

Direct mapping of cis-rSNPs in human cells using genome-wide allelic expression (AE) measurements on Illumina 1M BeadChips shows a high prevalence of common cis-rSNPs in CEU LCLs, explaining >50% of population variance in AE (Ge et al., NG 2009). Follow-up with allele-specific validation tools allows for isolation of causal cis-rSNPs (Verlaan et al., AJHG 2009).

Our systematic approach to isolate causal cis-rSNPs involves expanding AE-mapping to YRI LCLs and intersecting with catalogs of common SNPs (1000 Genomes) and functional non-coding elements (wgENCODE). For 570 genes, we fine-mapped the effect to <5 SNPs, with bias near the TSS and with significant enrichment (5-10x) of regulatory elements (e.g. RNAPII, DHS). Reporter gene assays validated 62% of promoter cis-rSNPs. We tested ~2500 RefSeq genes with cis-rSNPs altering transcription in YRI for enrichment of functional sites by inclusion of published eQTL data (Stranger et al., NG 2007). Conditioning AE-data on converging cis-eQTLs (n=402) showed minor increase in overlap with functional sites (~1.2x). Conditioning with positive AE-data (n=294) revealed a more pronounced (~2x) improvement.

Comparison of top AE-associations in fibroblasts of Caucasian origin (36 trios) to top cis-rSNPs replicated in both LCL panels among transcripts expressed in both cell types show 50% overlap. Tissue-independent effects are concentrated to TSS, whereas tissue-specific cis-rSNPs are enriched 50-100kb 5' of TSS. Finally, tissue-specificity of chromatin activity and cis-rSNPs are correlated.

This indicates that AE-mapping in primary transcripts: 1) reveals non-coding SNPs altering transcription and 2) captures heritable variation in transcription not detected by eQTL mapping. Tissue-specific cis-rSNPs yield additional functional sites. Combination of allele-specific tools will allow comprehensive cataloging of causal cis-SNPs.

ELUCIDATING THE CHROMATIN ARCHITECTURE OF LOCI ASSOCIATED WITH BLOOD TRAITS AND CORONARY ARTERY DISEASE

Dirk S Paul[1], Sylvia Nürnberg[2], Nicole Soranzo[1,3], Willem H Ouwehand[1,2], Panos Deloukas[1]
[1]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom, [2]Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Long Road, Cambridge, CB2 0PT, United Kingdom, [3]Department of Twin Research and Genetic Epidemiology, King's College London, Westminster Bridge Road, London, SE1 7EH, United Kingdom

Genome-wide association studies (GWAS) have identified loci harbouring common genetic variants associated with diseases, including coronary artery disease (CAD) and its main complication, myocardial infarction (MI). These loci have usually small effects, whereas the underlying causative variant(s) remain typically unknown. In parallel, GWAS on quantitative haematological traits, e.g. the count and volume of blood cells, have started to shed some light on the biological processes through which such disease-associated loci are likely to act.

As most association signals are intra- or intergenic, causal variants are likely to impact regulation of gene expression. Our aim was to map regulatory elements at loci associated with CAD/MI and blood traits to aid the identification of functional variants. We performed FAIRE-chip (formaldehyde-assisted isolation of regulatory elements), a robust method for identifying accessible chromatin and regulatory regions in human chromatin, in a megakaryocytic (MK) and an erythroblastoid (EB) cell line. Our array comprises all replicated and/or meta-analysed GWA loci (n=61, PubMed query, October 2009) associated with CAD, MI, and blood traits (385K tiling array, 9.6 Mb). We found 259 and 267 FAIRE peaks in MK and EB cells respectively, of which 148 (57.1%) were common to both.

In seven of the 61 tested loci, we identified sequence variants located within a FAIRE peak in strong LD with the GWA lead SNP ($r2 \geq 0.8$, HapMap2, CEU). For these loci, FAIRE peaks were either shared among both cell types (n=3), unique to EB cells (n=2), or MK cells (n=2). As a proof of principle, we investigated a 65-kb locus associated with mean platelet volume and function at chr7q22.3, which harbours a MK-specific FAIRE peak containing the lead SNP. Resequencing of this nucleosome-depleted region in 882 healthy individuals from the Cambridge BioResource did not reveal additional sequence variation, strongly suggesting that the lead SNP is most likely the functional variant. Preliminary results of the functional annotation of this region by ChIP-seq, showed the presence of a binding site for the MK-specific transcription factor MEIS1 overlapping the lead SNP, and was contained within the nucleosome-depleted region, further corroborating our findings.

We are in the process of following up this variant through *in vitro* tests, such as electrophoretic mobility shift assays (EMSA), to unravel its function and targets.

ASSAYING DNA METHYLATION WITH REDUCED REPRESENTATION BISULFITE SEQUENCING

Florencia Pauli, Katherine E Varley, Jason Gertz, Timothy E Reddy, Kevin M Bowling, Stephanie L Parker, Rebekka O Sprouse, Richard M Myers

HudsonAlpha Institute for Biotechnology, Myers Lab, 601 Genome Way, Huntsville, AL, 35806

The methylation of cytosines in CpG dinucleotides, which is usually accompanied by gene silencing, plays a crucial role in cellular differentiation and cancer progression. To study DNA methylation in human cell lines, we adapted Reduced Representation Bisulfite Sequencing (RRBS; Meissner et al. Nature 2008), a method that involves next-generation sequencing of size-selected bisulfite-treated Msp1 restriction fragments. With this method, we measure more than 500,000 CpGs with greater than 10X coverage throughout the human genome in each sample. Data from more than 60 biological samples, including primary and cancer cell lines, lymphocytes from a three generation family, and some primary tissues, indicates that RRBS is highly reproducible ($R^2 = 0.94$), that non-cancer cell lines are very similar and distinct from cancer cell lines, and that CpGs in the body of transcribed genes downstream from the first exon exhibit high overall methylation. We observe many instances of allele-specific DNA methylation, and we are using these in the family to determine whether methylation events are inherited or are a consequence of imprinting. We are analyzing the association of different sequence features (repeats and DNA motifs) with levels of DNA methylation. Overall, these data provide detailed pictures of genome-wide DNA methylation patterns and how these patterns vary over a large number of cell lines and primary tissues from different people.

FITNESS DETERMINANTS ASSOCIATED WITH COPY NUMBER CHANGES

Celia Payen, Anna C Brosius, Maitreya J Dunham

University of Washington, Genome Sciences, 1705 NE Pacific St, Seattle, WA, 98195

The ability to evolve is a fundamental feature of organisms. Changes in gene and chromosome copy number are widely observed in cancers and closely associated with genome evolution. Despite the importance of aneuploidy to many phenomena, little work has been done to determine how cells adapt to such extreme changes in gene dosage. Prior work from the Dunham lab using the eukaryotic model yeast, has found that aneuploidy can be detrimental to cells at high growth rates, but can be beneficial to cells growing in poor environments. A series of experiments using the well-defined system of yeast experimental evolution in the chemostat in different nutrient environments could be done to begin to understand how aneuploidy affects the fitness in a rigorous, systematic, and genome-wide manner. To accomplish this, we took advantage of strain collections consisting of every yeast gene deleted in one copy and every gene amplified to multiple copies. By competing these strains against each other and measuring their abundance via high- throughput sequencing, we can determine the fitness effect associated with changes in copy number for every gene simultaneously. For the collections carrying barcodes (MoBY and the deletion collection), each strain is tagged with two unique 20mer sequences that can be amplified with common primers. The change in strain abundance can be measured using Solexa sequencing, also known as “Bar-seq”. The quantification of the non-barcoded plasmids is done by sequencing the insert-vector junctions. By comparison with the time zero sample, a relative frequency of each plasmid is calculated and from this, a fitness measurement.

These data, which are in progress, will be compared to the fitness of aneuploid strains recovered in experimental evolution. One particularly interesting question that will be answered by these experiments is whether the observed copy number changes that are recovered from experimental evolutions are actually the best possible mutations. An alternate possibility is that certain events are predisposed by virtue of their mutability, for example, their proximity to repeat elements. This is the first genome-wide attempt to dissect the precise molecular causes of the fitness changes, positive and negative, associated with aneuploidy.

FLUCTUATIONS OF THE GASTROINTESTINAL MICROBIOME ASSOCIATED WITH DIABETES MELLITUS AND TRAVELERS' DIARRHEA

Joseph F Petrosino[1,2], Matthew C Ross[1], Bonnie Youmans[1], Sarah K Highlander[1,2], Susan P Fisher-Hoch[3], Richard A Gibbs[1]

[1]Baylor College of Medicine, Department of Molecular Virology and Microbiology, One Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030, [3]University of Texas, Houston School of Public Health, Division of Epidemiology, 80 Fort Brown, S1.330, Brownsville, TX, 78520

We are examining how fluctuations in the GI microbiota associate with various chronic and acute diseases, including Travelers' Diarrhea and Diabetes Mellitus. **Diabetes:** We are identifying the fluctuations in gut microbiota in a cohort of Mexican Americans having high rates of obesity and diabetes (52% and 20% respectively). Phylogenetic and sequence alignment-based analytical methods are being used to compare the microbial populations in these groups. We are also correlating microbiome composition to plasma levels of adipokines, cytokines and other molecules known to link the microbiome with obesity and diabetes using a limited systems biology analytic approach. This will be a critical first step to understanding the role of the microbiome in obesity and diabetes in influencing important cross-talk between metabolic and inflammatory pathways within cells and between adipocytes and immune cells. **Travelers' Diarrhea:** Secretory diarrhea is one mechanism by which the GI community can be altered, and up to 60% of individuals traveling from industrialized countries to developing countries acquire a form of secretory diarrhea know as Travelers' Diarrhea (TD). Enterotoxigenic Escherichia coli (ETEC) is the leading cause of TD and produces two toxins: heat labile toxin (LT) and heat stable toxin (ST). Predisposition to and the specific associations of TD-ETEC and its toxins on the GI microbiota have not been studied. 16S metagenomic sequencing technologies are examining the GI microbiome of three groups of individuals: healthy, TD-ETEC positive and non-ETEC diarrheal control. This research details the differences in the GI populations between these three groups of samples and aims to determine if LT or ST have a similar effect on the GI community. These data will provide a more complete understanding of the alterations caused by TD-ETEC and aid in development of preventative treatments.

GENOME SEQUENCING OF ELEPHANT ENDOTHELIOTROPIC HERPESVIRUS 1A FROM INFECTED HEART TISSUE

Joseph F Petrosino[1,2], Jeffrey G Reid[2], David Deiros[2], Yi Han[2], Jeffrey Stanton[1], Paul D Ling[1], Richard A Gibbs[2]

[1]Baylor College of Medicine, Molecular Virology and Microbiology, One Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030

Elephant endotheliotropic herpesvirus (EEHV) causes significant morbidity and mortality in both captive and wild juvenile Asian elephants, which are an endangered species. Advancements in the diagnosis, treatment, and prevention of this disease are needed to ensure the survival of this endangered species. All 6 Asian elephant calves born at the Houston zoo in the last two decades have died from EEHV infection. The most recent death of Mac, a highly charismatic 2-year old elephant calf in November of 2008, prompted the Houston zoo to contact experts at Baylor College of Medicine to provide some help to solve the EEHV problem. *The goal of our research program is to better detect and prevent lethal disease from this virus in both captive and wild Asian elephants through the development of diagnostics and a protective vaccine to EEHV.* Understanding the genetics of EEHV, its mechanisms of pathogenesis, and how elephants generate immune responses to EEHV will facilitate the development of effective vaccines and therapeutics for this virus. Toward this goal, the Human Genome Sequencing Center at Baylor College of Medicine (HGSC-BCM) has sequenced total genomic DNA isolated from infected heart tissue from Mac. The viral to host genome ratio in this tissue has been estimated to be approximately 50:1 using a RT-PCR assay. Over 22 Gb of 75bp read-length sequence was generated on the Illumina platform, and EEHV reads from a previously sequenced ~50kb fragment were mined from these data using a k-mer based, e-Genome-typing strategy. These data suggest that this initial sequencing has generated ~30X coverage of the EEHV genome (estimated to be between 250 and 500 Kb in size). Efforts are underway to mine all EEHV reads from the sequence pool using previously sequenced fragments and homology to other EEHV genomes. Assembly of these reads will provide the gene/protein targets to serve as potential vaccine candidates that will hopefully limit the devastating effects of these viruses in the elephant population.

RAPID QUANTITATION OF MRNA, PROTEINS, AND PTMS APPLIED TO A SYSTEMS-LEVEL ANALYSIS OF HUMAN ES, IPS, AND FIBROBLAST CELLS

Douglas H Phanstiel[1], Brumbaugh Justin[1], Thomson A James[2], Coon J Joshua[1]

[1]University of Wisconsin - Madison, Chemistry, 1101 University Ave, Madison, WI, 53706, [2]Morgridge Institute, Regenerative Biology, P.O. Box 7365, Madison, WI, 53706

We have developed a high mass accuracy-based proteomic strategy for rapid, large-scale, identification and quantitation of proteins and PTMs. This technology was combined with transcriptomic analyses to compare human embryonic stem cells (lines H1 and H9), induced pluripotent stem cells (iPS), and their differentiated precursors – newborn foreskin fibroblasts (NFF).

We identified 7,962 proteins and 15,091 phosphorylation sites. Phosphorylation was identified on 3,714 proteins including many involved in the maintenance of pluripotency including OCT4, SOX2, and LIN28. To facilitate the interrogation of such a large data set we developed software for automated detection and enrichment analysis of proteins that differ between any of the four samples. While ES and iPS cells were nearly identical at all levels of regulation, 35% of proteins differed by two-fold or more between pluripotent and NFF cell lines. Proteins upregulated in pluripotent cells were enriched in GO terms and KEGG pathways that reflected the proliferative nature of the pluripotent cells. Proteins depleted in these cells were enriched developmental function. Phosphorylation sites upregulated in pluripotent cells were enriched in 25 motifs, all containing acidic residues or the combination of a proline and a lysine. Downregulated phosphorylation sites were characterized by a strikingly different set of motifs, most containing an arginine upstream of a phosphorylated serine. Motifs enriched in pluripotent cells pointed to activity of kinases relevant to pluripotency such as ERK1, ERK2, and GSK3. While the protein levels of these kinases were largely unchanged, activating sites of phosphorylation were upregulated in pluripotent cells suggesting that modulation of these kinase activities may be effected by PTM rather protein level. Finally, we detected a subset of genes whose mRNA and protein levels changed in opposite directions. These genes were enriched in developmental functions which shed light on the mechanisms that govern the maintenance of pluripotency of cells grown in culture.

DHP is supported by an NHGRI training grant to the Genomic Sciences Training Program (5T32HG002760).

UNDERSTANDING MECHANISMS UNDERLYING HUMAN GENE EXPRESSION VARIATION WITH RNA SEQUENCING

Joseph K Pickrell[1], John C Marioni[1], Athma A Pai[1], Jacob F Degner[1], Barbara E Engelhardt[3], Everlyne Nkadori[1,4], Jean-Baptiste Veyrieras[1], Matthew Stephens[1,2], Yoav Gilad[1], Jonathan K Pritchard[1,4]

[1]University of Chicago, Department of Human Genetics, 920 E. 58th St., Chicago, IL, 60637, [2]University of Chicago, Department of Statistics, 5734 S. University Ave., Chicago, IL, 60637, [3]University of Chicago, Department of Computer Science, 1100 E. 58th St., Chicago, IL, 60637, [4] University of Chicago, Howard Hughes Medical Institute, 920 E. 58th St., Chicago, IL, 60637

Understanding the genetic mechanisms underlying natural variation in gene expression is a central goal of both medical and evolutionary genetics, and studies of expression quantitative trait loci (eQTLs) have become an important tool for achieving this goal . While all eQTL studies to date have assayed mRNA levels using expression microarrays, recent advances in RNA sequencing enable the analysis of transcript variation at unprecedented resolution. We sequenced RNA from 69 lymphoblastoid cell lines (LCLs) derived from unrelated Nigerian individuals that have been extensively genotyped by the International HapMap Project. Pooling data from all individuals, we generated a map of the transcriptional landscape of these cells, identifying extensive use of unannotated untranslated regions (UTRs) and over 100 novel putative protein-coding exons. Using the genotypes from the HapMap project, we identified over a thousand genes at which genetic variation influences overall expression levels or splicing. We demonstrate that eQTLs near genes generally act via a mechanism involving allele-specific expression, and that variation that influences the inclusion of an exon is enriched within and near the consensus splice sites. Our results illustrate the power of high-throughput sequencing for the joint analysis of variation in transcription, splicing, and allele-specific expression across individuals.

INVERSE MAPPING APPROACH IMPLIES THE ROLE OF LARGE CNVS IN INTELLECTUAL DEFICITS AND LEARNING DIFFICULTIES IN A POPULATION COHORT

Olli Pietiläinen[1,2], Susan Service[3], Marjo-Riitta Järvelin[4], Nelson B Freimer[3], Leena Peltonen*[1,2]

[1]The Sanger Institute, Human Genetics, Genome Campus, Cambridge, CB10 1SA, United Kingdom, [2]Institute for Molecular Medicine Finland, FIMM, Human Genetics, Tukholmankatu, Helsinki, 00290, Finland, [3]UCLA, Center for Neurobehavioral Genetics, Gonda Center, Charles Young Dr. S., Los Angeles, CA, 90095-1761, [4] Imperial College, Department of Epidemiology and Public Health, St Mary's Campus, London, W2 1NY, United Kingdom

*on behalf of the work group

The availability of genome wide data on representative population samples provides an opportunity to apply a strategy of inverse mapping for correlating human traits with genotypes. Whereas the traditional forward mapping aims to define genotypic sharing accounting for a common phenotype among group of individuals, the inverse mapping seeks to discover phenotypic features shared among individuals demonstrating allelic similarity. This approach would avoid the inherent imprecision in phenotype definition that makes it difficult to determine a priori which individuals are sharing a phenotype. Here we provide a proof of principle of this inverse mapping approach by systematically scanning all CNVs >500 kb in a population cohort (N= 4,932). The participants were drawn from a prospective birth cohort of all individuals born in 1966 in North of Finland. Follow ups at different time points have resulted in extensive phenotype battery collected from the study participants. We identified 634 large CNVs observed in 529 individuals. To narrow down our data inquiries to a reasonable number of phenotypes, we focused on a category of traits postulated to relate with previous CNV findings in neuropsychiatric deficits. We observed significantly higher frequencies of cognitive defects defined as 8 among carriers of large deletions (5.0%) compared to non carriers (1.4%) ($p < 0.0024$). Intriguingly, the deletion carriers were also more likely to present with sub clinical learning difficulties than the general population (10% vs. 3.9%; $p=0.00088$). Our results suggest that large deletions confer risk to intellectual defects. The study highlights the opportunity to utilize inverse mapping as a strategy to characterize phenotypic consequences related to genetic variants in an unbiased population sample.

RECOGNITION, CATEGORIZATION, AND CHARACTERIZATION OF TRANSPOSABLE ELEMENTS IN A NON-MUROID RODENT: *SPERMOPHILUS TRIDECEMLINEATUS*.

Roy N Platt II, David A Ray

Mississippi State University, Dept. of Biochemistry and Molecular Biology, 402 Dorman, Mississippi State, MS, 39759

Transposable elements are key drivers of genome evolution. Understanding the dynamics of their mobilization will lead to a more complete understanding of speciation mechanisms, morphological evolution, karyotypic megaevolution, as well as population genetics. The release of genomic scaffolds from the *Spermophilus tridecemlineatus* sequencing project presents the opportunity to study the history of transposable element mobilization in a highly diverse rodent lineage. In general, many questions regarding the evolution of *Spermophilus* exist. A recent mitochondrial phylogeny (Herron et al. 2004) found that prairie dogs (*Cynomys*), marmots (*Marmota*), and the antelope-ground squirrels (*Ammospermophilus*) form a paraphyletic relationship with Spermophilus. A better understanding of transposable elements in *Spermophilus* may lead to a better understanding of this unique phenomenon. Transposable elements were recovered and characterized using various computational approaches. In total the *Spermophilus* genome contains 80 transposable elements not found in other species at the sub-family level. Herein 44 uncharacterized elements are presented. Class I elements, the retrotransposons, dominate the genomic landscape with *Spermophilus* specific SINEs, LINEs, and LTRs occupying 5.44%, 7.98%, and 3.71% of the genome respectively. Class II elements, the DNA transposons, do not appear to be actively amplifying and occupy less than 0.13% of the genome. Elements were dated based on sequence divergence and phylogenetic methods and validated using clustered insertion analyses (TCF; Giordano et al. 2007) to better ascertain correlations between peak mobilization periods and diversification events. Though in an unassembled form, data gleaned from *Spermophilus* is important from a phylogenetic perspective. Additionally, comparisons between *Spermophilus*, *Mus*, and *Rattus* are important to gain a more complete understanding of transposable element dynamics in rodents since *Spermophilus* represents a basal rodent lineage. More work remains to fully elucidate the dynamics of transposable element mobilization in the genome of *Spermophilus*; however data presented herein represents a crucial starting point.

GENOME REPEAT STRUCTURE AND CONTEXT-DEPENDENT EVOLUTION OF GENOMES

David D Pollock, A.P.Jason de Koning, Todd A Castoe, Wanjun Gu

University of Colorado School of Medicine, Biochemistry and Molecular Genetics, 12801 17th Ave, Aurora, CO, 80045

A major focus in our laboratory has been to understand context-dependent evolution in genes and genomes. In particular, we have been developing new methods to better understand the repeat structure of genomes, and to use that information to evaluate evolutionary rates across genomes. Here, I will discuss new results on evaluating the repeat structure of genomes using "element-specific" P-clouds, and present evidence that previous evaluations of repeat structure in genomes are extreme underestimates. Furthermore, I will discuss how sample sequencing of eukaryotic genomes can be used to better understand transposable element evolution. I will then discuss how new rapid context-dependent Bayesian mixture model approaches we have recently developed can be used to analyze substitution processes in a local and genome-wide context-dependent fashion. Such analyses have been shedding light on important molecular evolutionary phenomena such as adaptation, convergence, and coevolution. I will also discuss how the nucleotide substitution processes can then be used to obtain better genome-wide evolutionary models of functional (selected) entities such as proteins and transcription factor binding sites.

RECALIBRATION OF BASE QUALITY SCORES

Ryan Poplin[1], Eric Banks[1], Anthony A Philippakis[2], Andrew Kernytsky[1], Mark Daly[1], David Altshuler[1], Stacey Gabriel[1], Mark DePristo[1]

[1]Broad Institute, Program in Medical and Population Genetics, 7 Cambridge Center, Cambridge, MA, 02142, [2]Brigham and Women's Hospital and Harvard Medical School, Department of Medicine, 75 Francis Street, Boston, MA, 02115

There are now terabytes of next generation sequencing data being produced by sequencing centers all around the world. Sequencing machines provide a confidence metric, or quality score, with every called base in a read which represents the probability that this particular base is erroneous. These base quality scores are often both uninformative -- a substantial percentage of the called bases are assigned the same quality score -- and inaccurate -- the actual probability of mismatching the reference genome is higher than the stated value. We have developed a software package to recalibrate these base quality scores which significantly reduces residual error between the reported base quality score and the empirical quality of each base. Use of this procedure greatly increases the performance of downstream probabilistic models (for example, Bayesian mutation calling). The recalibration tool attempts to correct for covariation in quality with machine cycle and sequence context, among other things, and by doing so provides not only more accurate quality scores but also more widely dispersed, and therefore more informative, scores.

The system works with data generated by the Illumina, SOLiD, 454, and Complete Genomics platforms taking into account the specifics of the chemistry of each platform in order to provide increased accuracy. Additionally, the recalibrator takes full advantage of the color space information found in data produced by the SOLiD platform and substantially reduces reference bias caused by some of the early color space aligners.

The recalibration tool is already being used for all sequencing data in the 1000 Genomes Project and is now an integral part of the sequencing pipeline at the Broad Institute. Full documentation including graphs showing before- and after-recalibration accuracy can be found at: http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration

DETECTION OF COPY NUMBER VARIATIONS IN INDIVIDUALS WITH AUTISM SPECTRUM DISORDERS USING THE AGILENT 1M CGH ARRAY

Aparna Prasad, Dalila Pinto, Christian Marshall, Bhooma Thiruvahindrapduram, Zhuozhi Wang, Stephen W Scherer

The Centre for Applied Genomics, Program in Genetics and Genomic Biology, The Hospital for Sick Children, 101 College Street, Toronto, M5G 1L7, Canada

Autism spectrum disorder (ASD) is a neurodevelopmental condition which has a strong genetic component. Previous studies have indicated that copy number variations (CNVs) are associated with ASD. There are several different platforms available for detection of CNVs including SNP and CGH arrays. Each platform has its own advantages and can complement the other. In the present study, we used Agilent 1M (CGH) arrays for detection of CNV in >200 unrelated ASD probands that were previously run on the Illumina 1M single SNP array.
The Agilent high resolution 1×1M array contains 963,029 distinct probes evenly distributed across the genome. CGH experiments using genomic DNA of cases was competitively hybridized to a pool of fifty sex-matched individuals as a reference. The CNV calling was performed using the Aberration Detection Method-2 (ADM-2) algorithm implemented in the DNA Analytics 4.0.85 software. The calls were compared with the CNV data from the Illumina 1M single SNP array analyzed using the iPattern and QuantiSNP algorithms. We found that only 41 % of the total Agilent 1M call set was detected using the Illumina 1M platform. Conversely, 38% of calls in the Illumina call set were not detected using the Agilent 1M platform. The average number of calls generated for Agilent 1M and Illumina 1M is 66 and 50 respectively, and the average size of these calls is 119 kb and 143 kb respectively. The differences in the number and size of CNVs discovered between platforms are possibly due to distinct probe distribution and sensitivity of the detection algorithms used. In addition, we are performing analysis of the data using both the Genome Alteration Detection Algorithm (GADA) and iPattern for the Agilent 1M experiments to minimize the possible effect of algorithm-specific sensitivity. We conclude that the use of multiple platforms and algorithms is advantageous for both maximizing CNV discovery and validation rates in case cohorts.

AN INTEGRATED APPROACH BASED ON 454-SEQUENCING OF JAZF1 GENE, GENOTYPING AND DATA FROM HAPMAP AND 1000 GENOMES PROJECTS IDENTIFIES NOVEL CANDIDATE SNPS FOR ASSOCIATION WITH PROSTATE CANCER.

McAnthony Tarway[1], Patricia Porter-Gill[1], Wei Tang[1], Yi-Ping Fu[1], Allison Burrel[1], Zuoming Deng[2], Luyang Liu[1], Kevin Jacobs[2], Demetrius Albanes[3], Ryan Divers[4], Michael Thun[4], Gilles Thomas[2], Meredith Yeager[2], Stephen Chanock[1,2], Ludmila Prokunina-Olsson[1]

[1]LTG, DCEG/NCI, 8717 Grovemont Cr, Bethesda, MD, 20892, [2]CGF, DCEG/NCI, 8717 Grovemont Cr, Bethesda, MD, 20892, [3]NEB, DCEG/NCI, 6120 Executive Blvd, Bethesda, MD, 20892, [4] ACS, Department of Epidemiology, 1599 Clifton Rd, Atlanta, GA, 30329

A multi-stage genome-wide association study (GWAS) has identified a SNP rs10486567 within the JAZF1 gene to be associated with prostate cancer (PrCa) in individuals of European ancestry (p=7.79x10-11). Fine mapping based on 106 additional SNPs did not reveal additional independent signals in this region. Now, we aimed to identify all variants within JAZF1 gene in high LD (r2>0.8) with rs10486567. We combined data from 454-sequencing of 127 Kb in 63 samples and data from the HapMap and 1000 Genomes projects. For confirmation we genotyped 27 SNPs found to be in an r2>0.6 with rs10486567 in any of these sets in 60 samples used for 454-sequencing and in 1,000 samples (500 PrCa cases and 500 controls) of European ancestry. Based on genotyping in 1,000 samples, 14 SNPs were found to be in r2>0.8 with rs10486567. The selection of SNPs in an r2>0.8 with rs10486567 based only on 454-sequencing, data from HapMap or 1000 Genomes projects would miss up to 50% of these SNPs. The problematic SNPs were located in repetitive regions (LINE and SINE repeats) that are difficult to genotype and sequence with 454 technology. Among 742 SNPs identified by 454-sequencing of 127 Kb of JAZF1 in 63 individuals, the completion rate for SNPs located within repetitive sequences (n=268) was 51.1%, compared to 75.9% for SNPs located within unique sequences (n= 474), p=7.75x10-23.

ASSESSING THE ACCURACY AND COMPLETENESS OF THE BONOBO GENOME SEQUENCE

Kay Prüfer[1], Susan E Ptak[1], Anne Fischer[2], Jeffrey M Good[3], James C Mullikin[4], Jason Miller[5], Chinnappa D Kodira[6], James R Knight[6], The Bonobo Genome Consortium[1], Janet Kelso[1], Svante Pääbo[1]

[1]Max-Planck-Institute for evolutionary Anthropology, Evolutionary Genetics, Deutscher Platz 6, Leipzig, 04103, Germany, [2]icipe - African Insect Science for Food and Health, Molecular Biology and Biotechnology Department, Thika Road, Nairobi, 00100, Kenya, [3]The University of Montana, Division of Biological Sciences, 32 Campus Drive, HS104, Missoula, MT, 59812, [4] National Human Genome Research Institute, Comparative Genomics Unit, 5625 Fishers Ln, Rockville, MD, 20892-9400, [5]J. Craig Venter Institute, Informatics Group, 9704 Medical Center Drive, Rockville, MD, 20850-3343, [6]454 Life Sciences, a Roche company, Assembly Team, 15 Commercial Street, Branford, CT, 06405

Next generation sequencing holds the promise of facilitating the study of new large and complex genomes in shorter time and at lower cost compared to previous technology. However, the sequences generated by next generation sequencing are generally shorter and differ in the type and amount of sequencing error from sequences generated by the Sanger technology. Here, we present the first *de novo* assembly of a large and complex genome sequence from 454 data.
We analyze the quality of the bonobo genome assembly, generated from 25-fold coverage of 454 Sequencing™ data, encompassing both shotgun and paired end reads. This assembly has a N50 scaffolds size of 9.6 megabases and a N50 contig size of 67 kilobases. The bonobo genome sequence is particularly suited for the analysis of assembly quality since the genome sequences of two closely related species (human and chimpanzee) are available for comparison.
Our analysis – based on whole genome alignments between the genomes of human, chimpanzee and bonobo – shows that the bonobo assembly achieves comparable quality to the chimpanzee assembly in terms of sequence accuracy and genome completeness. Using the high quality of the finished chimpanzee chromosome 21 we are able to estimate a rate of approximately two errors in 10,000 base pairs for the bonobo genome. The quality of the Bonobo X chromosome assembly is as good as the other autosomes since a female individual was selected for sequencing. This places this bonobo's X and all her autosomes in an excellent position for comparative analysis to other primates. We conclude that large and complex genomes can be *de novo* assembled from next generation sequencing data.

DISSECTION OF GENETICALLY COMPLEX TRAITS WITH EXTREMELY LARGE POOLS OF YEAST SEGREGANTS

Ian M Ehrenreich, Noorossadat Torabi, Yue Jia, Jonathan Kent, Stephen Martis, Joshua A Shapiro, David Gresham, Amy A Caudy, Leonid Kruglyak

Princeton University, Princeton, NJ, 08544

Most heritable traits, including many human diseases1, are caused by multiple loci. Studies in both humans and model organisms, such as yeast, have failed to detect a large fraction of the loci that underlie such complex traits2,3. A lack of statistical power to identify multiple loci with small effects is undoubtedly one of the primary reasons for this problem. We have developed a method in yeast that allows the use of much larger sample sizes than previously possible and hence permits the detection of multiple loci with small effects. The method involves generating very large numbers of progeny from a cross between two *Saccharomyces cerevisiae* strains and then phenotyping and genotyping pools of these offspring. We applied the method to 17 chemical resistance traits and mitochondrial function, and identified loci for each of these phenotypes. We show that the level of genetic complexity underlying these quantitative traits is highly variable, with some traits influenced by one major locus and others by at least 20 loci. Our results provide an empirical demonstration of the genetic complexity of a number of traits and show that it is possible to identify many of the underlying factors using straightforward techniques. Our method should have broad applications in yeast and can be extended to other organisms.

GENOMIC HETEROZYGOSITY AND LOSS-OF-HETEROZYGOSITY IN WILD YEAST

Paul M Magwene

Duke University, Department of Biology, Box 90338, Durham, NC, 27708

High rates of inbreeding lead to reduced heterozygosity both within populations and at the level of individual genomes. Reduced heterozygosity in turn contributes to reduced genetic and phenotypic diversity in inbred populations. The life cycle, mating system, and population structure of *Saccharomyces cerevisiae* are thought to promote inbreeding. Using whole genome sequencing of diploid isolates sampled from a variety of environments, I show that many yeast strains have extensive genomic heterozygosity. Clinical isolates in particular have very high levels of heterozygosity, typically in excess of 20,000 heterozygous sites across the genome. The impact of heterozygosity on the proteome and cellular phenotypes is significant; in highly heterozygous strains more than 30% of proteins are represented by two distinct peptide sequences and segregants derived from naturally heterozygous diploids exhibit a wide range of variation for many traits. I further show that heterozygous sites have a non-uniform chromosomal distribution that suggests that loss-of-heterozygosity (LOH) events are frequent. The size and location of LOH regions is consistent with mitotic, rather than meiotic, recombination. These findings suggest that both sex and mitotic recombination play important roles in shaping the genome architecture of this model eukaryote.

FIVE VERTEBRATE CHIP-SEQ REVEALS THE EVOLUTIONARY DYNAMICS OF TRANSCRIPTION FACTOR BINDING

Dominic Schmidt[1,2], Michael D Wilson[1,2], Benoit Ballester[3], Petra C Schwalie[3], Iannis Talianidis[4], Paul Flicek[3], Duncan T Odom[1,2]

[1]University of Cambridge, Oncology, Hills Road, Cambridge, CB2 0XZ, United Kingdom, [2]Cancer Research UK, Oncology, Robinson Way, Cambridge, CB2 0RE, United Kingdom, [3]European Bioinformatics Institute (EMBL-EBI), Vertebrate Genomics, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom, [4] Biomedical Sciences Research Center Al. Flemming, Molecular Biology and Genetics, 34 Fleming Street, Vari, 16672, Greece

Mammalian transcription factor binding evolves rapidly, yet tissue-specific transcription is highly conserved. To explore this apparent paradox, we experimentally determined the genome-wide occupancy of two transcription factors (TF) CEBPα and HNF4α in livers of multiple vertebrates.

Although each TF has highly conserved DNA binding preferences, most binding is species-specific and ultra-shared events are rare. Functional target genes are associated with an enrichment of shared TF binding, yet collectively, the binding events near functional targets show no increase in sequence constraint. Most lineage-specific lost TF binding can be explained by sequence mutations of the binding motif, and only half of the apparently lost binding events appeared to have turned over to a nearby location.

Our results reveal the plasticity of vertebrate TF binding and the complex evolutionary dynamics of transcriptional regulation.

RETROTRANSPOSONS IN THE ORANGUTAN *(PONGO PYGMAEUS)* LINEAGE: A NEW EVOLUTIONARY TALE

Miriam K Konkel[1], Jerilyn A Walker[1], Brygg Ullmer[2], Leona G Chemnick[3], Oliver A Ryder[3], Robert Hubley[4], Arian F A Smit[4], Mark A Batzer[1], for the Orangutan Genome Sequencing and Analysis Consortium[5]

[1]Louisiana State University, Department of Biological Sciences, Biological Computation and Visualization Center, 202 Life Sciences Building, Baton Rouge, LA, 70803, [2]Louisiana State University, Department of Computer Science, Center for Computation and Technology (CCT), 216 Johnston Hall, Baton Rouge, LA, 70803, [3]Beckman Center for Conservation Research (CRES), Zoological Society of San Diego, San Diego Zoo, San Diego, CA, 92112, [4] Institute for Systems Biology, Computational Biology, 1441 North 34th Street, Seattle, WA, 98103, [5]Washington University School of Medicine, Genome Sequencing Center, 4444 Forest Park Ave, St. Louis, MO, 63108

Orangutans (*Pongo pygmaeus*) are the only living Asian ape and are highly endangered. We investigated the mobile DNA composition (mobilome) of the orangutan draft genome sequence derived from a female of Sumatran origin (*Pongo pygmaeus abelii*). Similar to other primate genomes, about half of the orangutan draft genome sequence is comprised of repetitive sequences. L1 and SVA have been highly active within the orangutan lineage and show a mostly linear subfamily evolution. L1 and SVA elements have expanded in the orangutan lineage at a rate comparable to other higher apes. In contrast, *Alu* elements appear to be relatively quiescent and have propagated at a very low rate in orangutans. The identification of polymorphic and population-specific *Alu* insertions indicates that *Alu* retrotransposition may be ongoing albeit at a very low rate. In addition, we investigated the population structure within orangutans. For this purpose, we performed a structure analysis with 37 orangutans (18 Bornean and 19 Sumatran) using polymorphic retrotransposons. These elements were selected from the orangutan draft genome and also from Illumina paired-end reads from a Bornean orangutan. The Bornean orangutans were clearly distinct from the Sumatran population with almost no evidence of ongoing admixture. In addition, Sumatran orangutans showed clear evidence of population substructure. The clear distinction of Sumatran from Bornean orangutans supports the recent suggestion that Bornean and Sumatran orangutans represent separate species.

ADAPTATION IN *DROSOPHILA* IS NOT LIMITED BY MUTATION AT SINGLE SITES

Dmitri A Petrov

Stanford University, Dept. of Biology, Stanford, CA, 94305

Adaptation in eukaryotes is generally assumed to be mutation-limited because of small effective population sizes. This view is difficult to reconcile, however, with the observation that adaptation to anthropogenic changes such as the introduction of pesticides can occur very rapidly. Here we investigate adaptation at a key insecticide resistance locus (*Ace*) in *D. melanogaster* and show that multiple simple and complex resistance alleles evolved fast and repeatedly within individual populations. Our results imply that the current effective population size of modern *D. melanogaster* populations is likely to be substantially larger („100 fold) than commonly believed. This discrepancy arises because estimates of the effective population size are generally derived from levels of standing variation and thus reveal long-term population dynamics dominated by sharp - even if infrequent - bottlenecks. The short-term effective population sizes relevant for strong adaptation, on the other hand, might be much closer to census population sizes. Adaptation in *Drosophila* may therefore not be limited by waiting for mutations at single sites, and complex adaptive alleles can be generated quickly without fixation of intermediate states. Adaptive events should also commonly involve the simultaneous rise in frequency of independently generated adaptive mutations. These so-called soft sweeps have very distinct effects on the linked neutral polymorphisms compared to the standard hard sweeps in mutation-limited scenarios. Methods for the mapping of adaptive mutations or association mapping of evolutionarily relevant mutations may thus need to be reconsidered. We also argue that these results apply broadly to all populations of large census size and specifically to modern human populations.

A NEANDERTAL PERSPECTIVE ON HUMAN ORIGINS

Svante Paabo[1], David Reich[2], Richard E Green[1], and The Neandertal Genome Analysis Consortium

[1]MPI-EVA, Dept Evol. Genetics, Deutscher Platz 6, Leipzig, D-04103, Germany, [2]Broad Institute, Dept of Genetics, Deutscher Platz 6, Cambridge, MA, 02142

The Neandertals are the closest evolutionary relatives of present-day humans. Thus, for any definition of what sets fully anatomically modern humans apart from other hominin forms, the relevant comparison is to Neandertals.

We present a draft sequence of the Neandertal genome composed of over 3 billion nucleotides from three individuals. Through comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world we identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans. These include genes involved in metabolism, cognitive and skeletal development. We also present a catalog of genomic changes that have become fixed or have risen to high frequency in modern humans during the last few hundred thousand years. Finally, we analyze the relatedness of Neandertals and other early hominins to present-day humans in different parts of the world in order to analyze their relationship to each other and present-day humans.

HUMAN-SPECIFIC LOSS OF REGULATORY DNA AND THE EVOLUTION OF HUMAN-SPECIFIC TRAITS

Cory Y McLean[1], Philip L Reno[2,3], Alex A Pollen[2], Abraham I Bassan[2], Terence D Capellini[2], Catherine Guenther[2,3], Vahan B Indjeian[2,3], Bruce T Schaar[2], Douglas B Menke[2,3], Aaron M Wenger[1], Gill Bejerano[1,2], David M Kingsley[2,3]

[1]Stanford University, Computer Science, 279 Campus Drive West, Stanford, CA, 94305, [2]Stanford University, Developmental Biology, 279 Campus Drive West, Stanford, CA, 94305, [3]Howard Hughes Medical Institute, Stanford University School of Medicine, 279 Campus Drive West, Stanford, CA, 94305

The availability of several primate whole genome sequences has spurred great excitement for the prospect of understanding the molecular basis of what makes us human. Recent investigations have discovered conserved non protein coding genomic loci that have experienced accelerated base pair changes in the human lineage, as well as protein coding genes that show similar evidence of positive selection.

Here we expand these studies to look for a type of event particularly likely to produce functional effects: complete deletion in humans of sequences that are otherwise highly conserved in other organisms. By searching for regions of the chimpanzee genome highly conserved over mammalian evolution that are clearly missing in humans, we discover hundreds of human-specific losses of putatively functional ancestral DNA.

PCR and computational validation show that roughly 80% of the deletions are fixed in human populations, while others are polymorphic in different individuals. Most of the deletions removed conserved non-coding sequences rather than protein-coding regions, and many lie in proximity to genes involved in development, neural function, and steroid hormone signaling.

We have functionally tested a subset of these regions in mice, and have found intriguing examples of regulatory alterations in humans that appear to be associated with evolution of specific anatomical differences between humans and other animals.

EMERGING SPECIATION IN OCEAN MICROBES IS DRIVEN BY A FEW ECOLOGICALLY-RELEVANT GENOMIC LOCI

B. Jesse Shapiro[1], Jonathan Friedman[1], Otto X Cordero[2], Sarah Preheim[2], Eric Alm[2,3,4], Martin Polz[2]

[1]Massachusetts Institute of Technology, Computational and Systems Biology, 77 Massachusetts Avenue, Cambridge, MA, 02139, [2]Massachusetts Institute of Technology, Civil and Environmental Engineering, 77 Massachusetts Avenue, Cambridge, MA, 02139, [3]The Broad Institute of MIT and Harvard, Microbial Genomes Program, 7 Cambridge Center, Cambridge, MA, 02142, [4] The Virtual Institute of Microbial Stress and Survival, Berkeley, CA, 94720

Microbes adapt to changing selective pressures in their natural environments, leading to a potentially dynamic process of ecological specialization and speciation, even over short periods of time. Yet little is known about the microevolutionary processes leading to ecological differentiation of microbial populations in the wild. In particular, it is unclear how species emerge and diverge in face of the homogenizing force of recombination.

We sequenced and analyzed complete genomes from 8 closely-related strains of *Vibrio splendidus*, representing two nascent populations, that appear to have recently diversified ecologically: 3 strains found primarily on small particles, and 5 strains found primarily attached to zooplankton in the coastal ocean. To assess patterns of recombination among strains, we developed a dynamic-programming algorithm, and inferred ~2500 recombination breakpoints in the population genome. Although gene-flow between populations in the two habitats is common, we observe a significant excess of recent recombination within habitats, suggesting the emergence of ecologically differentiated populations. Gain and loss of DNA is extensive among these strains (each strain contains ~100-300 kb of strain-specific DNA), and a few recently acquired genes may provide habitat-specific adaptive value. For example, a suite of genes involved in O-antigen and mannose-sensitive hemagglutinin (MSHA) biosynthesis are absent in small-particle strains but present in zooplankton-associated strains, perhaps promoting preferential attachment to zooplankton.

We also identified a few 'core' genomic regions that, while present in all strains, are highly differentiated between the two habitats, and are likely targets of habitat-specific positive selection. These ecologically-associated loci include genes involved in stress response (rpoS), DNA repair (cysteine methyltransferase), and chitin metabolism, leading us to hypothesize that switching between zooplankton-associated (rich in insoluble exoskeletal chitin) and small-particle-associated lifestyles may require fine-tuning of chitin metabolism.

Taken together, these results support a model of microbial genome evolution where habitat-specific genes follow species boundaries, but the majority of other loci are freely recombined across species.

THE SEQUENCE AND ANALYSIS OF 17 LABORATORY AND WILD-DERIVED MOUSE GENOMES, AND HIGH RESOLUTION QTL ANALYSIS IN A HETEROGENEOUS STOCK CROSS

Jonathan Flint[1], Thomas M. Keane[2], Binnaz Yalcin[1], Jim Stalker[2], Kim Wong[2], Xiangchao Gan[1], Petr Danecek[1], Avigail Agam[1], Martin Goodson[1], Guy Slater[1], Ian Jackson[3], Laura Reinholdt[4], Leah Rae Donahue[4], Steve Brown[5], Ewan Birney[6], Allan Bradley[2], Chris Ponting[7], Richard Mott[1], Richard Durbin[1], David J Adams[2].

[1]Wellcome Trust Centre for Human Genetics, Oxford, UK, [2]The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK. [3]MRC-HGU, Edinburgh, UK. [4]The Jackson Laboratory, Bar Harbour Maine, USA. [5]MRC-Harwell, Oxford, UK. [6]European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. [7]MRC Functional Genomics Unit. University of Oxford, UK.

The Mouse Genomes Project (http://www.sanger.ac.uk/mousegenomes/) has sequenced the genomes of 17 inbred mouse strains; NOD/ShiLtJ, A/J, BALBc/J, CBA/J, C3H/HeJ, DBA/2J, CAST/EiJ, AKR/J, LP/J, 129S5, 129P2, SPRETUS/EiJ, C57BL/6N, PWK/PhJ, NZO/HILtJ, WSB/EiJ and 129S1/SvImJ with the aim of generating a complete map of nucleotide and structural variation, and ultimately a de novo genome assembly of each strain. We have sequenced each of these inbred genomes to on average 24x using Illumina paired-end sequencing of 200-600bp fragments.

From the sequence of these genomes we have been able to harvest >125M high-quality SNPs in addition to short indels, SVs, and CNVs providing unparalleled resolution of the variation between these strains of mice. With these data we set out to answer several questions; 1) Are there different mutagenic processes operative in different mouse strains? 2) What effect do mutations have on genes and regulatory regions? and 3) Can the sequence of these genomes help identify causal variants within QTLs in crosses derived from the sequenced strains (including the collaborative cross and heterogeneous stock mice)?

Using the sequence of these strains, we have built an extensive catalogue of all types of variants from single base substitutions and short indels, right up to larger structural events such as large insertions, deletions, inversions, and translocations. From this set, we have assembled a near complete catalogue of coding differences between strains and found thousands of truncating loss of function mutations, and sequence variants predicted to have a deleterious affect on protein function. Many of the mutated genes cluster into pathways possibly reflecting the effects of selective inbreeding for traits of interest during the derivation of these strains. We have also observed significant differences in the type, frequency and distribution of transposable elements between strains of mice with many intragenic transposon insertions being predicted to disrupt genes.
Using these data we have been able to refine the location and assign candidate genes to QTLs identified in crosses between the sequenced strains. Confirmatory experiments using knockout and transgenic models are underway.

RESULTS AND LESSONS FROM THE 1000 GENOMES PROJECT PILOT, AND MOVING ON TO MUCH MORE

Richard Durbin, on behalf of the 1000 Genomes Project

Wellcome Trust Sanger Institute, Informatics, Hinxton, Cambridge, CB10 1SA, United Kingdom

The 1000 Genomes Project pilot generated over 5Tbp sequence in three separate pilot experiments. These assess the effectiveness of low coverage (2-4x) sequencing to find shared variants in human, a high coverage trio design, and targeted sequencing of 1000 genes in 700 samples. Over 15M consensus variant calls from these have now been made and extensive validation has been carried out or is being completed. Much has been learned about data processing, variant calling and sampling properties. A wide variety of analyses are being carried out, including investigations of population differentiation and of polymorphic novel sequence (i.e. sequence present in a subset of individuals but not the reference). The data are being used in a large number of other studies.

Beyond the pilot, data collection for the main project is well under way. We aim to collect genome-wide data at 4x depth from 2000 people from 22 populations in 4 population groups of 500 each, which we project will meet the goals of having ~95% power to find 1% variants in the accessible genome per population group. Currently we have data from several hundred individuals, perhaps approaching 1000 by the time of the meeting, and will present initial analyses showing how these data are developing the information resource from the project as we go to deeper depth.

We will also discuss how the lessons and data from the project can be applied to strategies for population sequencing from phenotyped sample sets that are now being designed or planned.

COMPLETE GENOME SEQUENCING AND ANALYSIS OF DIPLOID AFRICAN-AMERICAN AND MEXICAN-AMERICAN GENOMES: IMPLICATIONS FOR PERSONAL ANCESTRY RECONSTRUCTION AND MULTI-ETHNIC MEDICAL GENOMICS

Carlos D Bustamante[1,2], Jeremiah D Degehnardt[2], Shaila Musharoff[2], Katarzyna Bryc[3], Jeffrey M Kidd[1], Vrunda Seth[3], Sarah Stanley[4], Abra Brisbin[2], Alon Keinan[2], Andrew Clark[2], Francisco M De La Vega[4]

[1]Stanford U., Genetics, Stanford, 94305, [2]Cornell U., BSCB, Ithaca, 14853, [3]Life Technologies, Genetic Systems R&D, Beverly, 01915, [4] Life Technologies, Genetic Systems R&D, Foster City, 94404,

Understanding the contribution of rare and common genetic genetic variants to disease susceptibility will likely require multi- and trans-ethnic sequencing studies that compare the genomes of many individuals with and without a particular disease. Of particular importance will be accounting for the role of population stratification at fine scales both in terms of genomic and geographic location. Here, we present results from sequencing, assembly, and genomic analysis of two diploid genomes from Phase 3 HapMap sequenced to ~20X coverage using SoLiD technology. The donor individuals are of Mexican-American and African-American ancestry and represent the first "admixed" genomes to be sequenced to high coverage. We demonstrate that genomic sequencing provides finer resolution of "admixture breakpoints" based on allele frequency estimates from HapMap and TGP. For each admixed genome, we use the distribution of admixture breakpoints to infer the personal admixture history of the sample and patterns of genomic diversity to reconstruct the demographic history of European, African, and Native American continental populations. Furthermore, we compare the distribution of functional and putatively neutral genetic variation among 12 sequenced genomes and find that difference in demographic history may account for statistically significant, differences in distributions of synonymous vs. benign, possibly damaging, and probably damaging non-synonymous coding variants. Finally, we use the SoLiD comparative personal genomic data sets and TGP data to quantify the relative proportions of private, rare, and common functional and neutral genetic within and among populations.

COMPARATIVE POPULATION GENOMICS: ANALYSIS OF GENOME-WIDE SEQUENCE DATA IN 10 WESTERN CHIMPANZEES

Peter Donnelly[1,3], Adi Fledel-Alon[2], Stephanie C Melton[2], Adam Auton[1], Oliver Venn[1], Susanne Pfeifer[3], Gerton Lunter[1], Zam Iqbal[1], Rory Bowden[3], Simon Myers[3,1], Gil McVean*[3,1], Molly Przeworski*[2]

[1]University of Oxford, Wellcome Trust Centre For Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom, [2]University of Chicago, Department of Human Genetics, 920 East 58 Street, Chicago, IL, 60637, [3]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom

In spite of its evolutionary significance, there is only limited polymorphism data and no genetic map available for our closest evolutionary relative. We have sequenced 10 Western chimpanzees (*Pan troglodytes verus*) to 6-10X coverage using paired end sequencing with 50-bp reads. This resource, which will be publicly released, is the first to document genome-wide sequence variation in the chimpanzee. We identified over 4.5 million novel SNPs, representing a four-fold increase over those previously documented. We show that patterns of variation are correlated between species, but with interesting differences. This resource will enable analyses of polymorphism patterns; inference of selection in the chimpanzee; genome-wide comparative study of human and chimpanzee diversity patterns; comparison of evolutionary patterns in particular genes of interest, for example those involved in immunity; the construction of fine-scale genetic maps and the detection of recombination hotspots; detection and analyses of copy number polymorphism in the chimpanzee; and improvement of the chimpanzee reference sequence.

The talk will focus on comparative population genomics, in assessing and comparing levels of diversity in human and chimpanzee, and on initial results on a chimpanzee genetic map and its implications for the regulation of chimpanzee recombination.

*Joint senior authors.

Population genetic analyses of next-generation sequencing data.

Rasmus Nielsen[1,2], Thorfinn Korneliusen[1,2], Emilia Huerta-Sanchez[1], Nicolas Vinckenbosch[1], Yingrui Li[2,3], Jun Wang[2,3]

[1]UC Berkeley, Departments of Integrative Biology and Statistics, Berkeley, CA, 94720, [2]University of Copenhagen, Department of Biology, KBH O, 2100, Denmark, [3]BGI-Shenzhen, Beijing Genomics Institute, Shenzhen, 518083, China

Next generation sequencing provides a powerful tool for generating genome-wide population genetic data. However, most data is at a relatively low coverage (e.g., <20x). The analyses of such data introduces new inferential problems because the genotypes for each individual will be associated with statistical uncertainty.
We will present methods for addressing these problems, that quantify the statistical uncertainty and incorporates it into methods for population genetic analyses. We will show that accurate Site Frequency Spectra (SFS) can be obtained for data with very low coverage (e.g., <2x), and can be used for accurate and valid population genetic inferences. In many cases, the most cost-effective design is, therefore, to sequence many individuals at low coverage rather than sequencing few individuals at a high coverage. We will discuss a number of applications, including applications to a large sample of humans of European descent and a sample of Tibetan and Han Chinese individuals.

GENETIC VARIATION IN NATIVE AMERICANS

Jeffrey D Wall[1], Rong Jiang[1], Celeste Eng[2], Scott Huntsman[2]

[1]UCSF, Institute for Human Genetics, San Francisco, CA, 94143, [2]UCSF, Dept of Biopharmaceutical Sciences, San Francisco, CA, 94143

Many current human groups, such as Latinos in the United States, have been formed by the recent mixing of different continental source populations over the last several hundred years. The genomes of individuals from admixed populations are mosaics of chromosomal 'chunks' inherited from different ancestral populations, and the boundaries of these chunks can be estimated from dense genotype data. We demonstrate how dense genotyping followed by targeted resequencing of admixed individuals can be used to assay genetic variation in the ancestral source populations, even if these source populations are unknown or no longer exist. We concentrate on genetic variation in Native Americans, inferred from genotype and resequencing data from extant Mexican Americans. We observe low levels of diversity and high levels of linkage disequilibrium in the Native American-derived sequences, consistent with a recent, severe population bottleneck associated with the initial peopling of the Americas. We also estimate the timing and the strength of this bottleneck from the patterns of genetic variation in Native American-derived sequences.

SIGNATURES OF NATURAL SELECTION IN THE FIRST PILOT EXPERIMENT OF THE 1000 GENOMES PROJECT

Ryan D Hernandez[1,2], Joanna L Kelley[2], S. Cord Melton[2], Adam Auton[3], Gil McVean[3,4], Guy Sella[5], Molly Przeworski[2], 1000 Genomes Project[2]

[1]University of California, San Francisco, Department of Bioengineering and Therapeutic Sciences, 1700 4th St, San Francisco, CA, 94158, [2]University of Chicago, Department of Human Genetics, 920 E. 58th St, Chicago, IL, 60637, [3]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom, [4] University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom, [5]The Hebrew University, Jerusalem, Department of Evolution, Systematics and Ecology, Berman building, room 102, Givat Ram, Jerusalem, JR 91904, Israel

Nearly 200 genomes from four populations have been resequenced for the first pilot experiment of the 1000 Genomes Project. These data avoid many of the ascertainment biases that have plagued previous large-scale human data sets, allowing long standing evolutionary questions to be resolved. We used these data to characterize genomic signatures of natural selection. We show that diversity increases with distance from coding regions, with the strongest effect for regions of low recombination, suggesting that selection on coding regions leads to distortionary effects on diversity. In particular, we report a clear footprint of natural selection on diversity patterns around human-specific amino acid substitutions. In addition, we identified regions of the genome with extreme differences in allele frequencies between population samples. While all three findings reflect the action of natural selection, it remains unclear to what extent they are explained by adaptive evolution or purifying selection, with recent reports offering conflicting conclusions in this regard. To disentangle the relative contributions of the two evolutionary forces, we ran extensive simulations of the human genome, incorporating information about functional annotations, fine-scale genetic maps, and realistic demographic models of all four populations.

HIGH RESOLUTION QTL MAPPING BY DEEP SHOTGUN SEQUENCING A SEGREGATING YEAST POPULATION UNDER SELECTION.

Leopold Parts*[1], Gianni Liti*[2], Kanika Jain[2], Francisco Cubillos[2], Edward J Louis[2], Richard Durbin[1]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB101HH, United Kingdom, [2]Institute of Genetics, University of Nottingham, Nottingham, NG72UH, United Kingdom

* These authors contributed equally.

One approach for understanding the complex architecture of quantitative traits is to study offspring of two natural strains that are phenotypically different. However, there are two major limitations to the classical linkage analysis: low resolution of linkage intervals, and large labour requirements for genotyping and phenotyping. We have developed a novel approach to map quantitative trait loci (QTLs) rapidly and in narrow intervals using massively parallel sequencing. We created an F12 population (12 generations of random mating and meiosis) between two phenotypically different wild isolates of baker's yeast *Saccharomyces cerevisiae* with sequenced reference genomes. The resulting haploid intercross pools consist of 10-100 million segregants with an average of ~1000 crossovers per segregant. We then applied selective pressure on the intercross pool by growing it asexually in a restrictive condition to enrich for individuals with alleles that confer a positive fitness effect. Three different stress conditions (heat, paraquat, and caffeine) were used, as well as a control in permissive condition. Genotyping of the intercross pool before and after selection has confirmed a large change in the population allele frequency for a known heat QTL in a subtelomeric region of chromosome XIII, as well as increased resolution. We are currently sequencing DNA from the pool before and after selection to >100X to pinpoint alleles responsible for the increased fitness. We expect to dramatically improve on power and resolution by enriching for alleles that have modest fitness effects, and narrowing the causative locus down to a few hundred bases due to the power of the selection experiment and the increased number of recombination events. This method will provide a rapid and fine scale QTL mapping strategy, improving resolution by up to two orders of magnitude with the potential of pinpointing quantitative trait nucleotides, and may be adapted to other genetic systems like *Drosophila melanogaster* and *Caenorhabditis elegans*.

SINGLE BASE RESOLUTION OF MEDAKA FISH DNA METHYLOMES

Wei Qu[1], Shin-ichi Hashimoto[1], Atsuko Shimada[2], Yutaka Suzuki[3], Sumio Sugano[3], Hiroyuki Takeda[2], Shinichi Morishita[1,4]

[1]The University of Tokyo, Department of Computational Biology, Graduate School of Frontier Sciences, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-0882, Japan, [2]The University of Tokyo, Department of Biological Sciences, Graduate School of Science, 7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, Japan, [3]The University of Tokyo, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, 4-6-1 Shirokanedai, Minato, Tokyo, 108-8639, Japan, [4] Japan Science and Technology Agency (JST), Bioinformatics Research and Development (BIRD), 5-3 Yonbanchou, Chiyoda, Tokyo, 102-8666, Japan

DNA cytosine methylation is essential in biological processes including gene transcription regulation, reprogramming and development. Here we report the first single-base-resolution DNA methylomes of a fish, medaka (*Oryzias latipes*). We have observed methylomes for three representative cell lines, liver, testis and blastula (0.5 day embryo considered to maintain germ-line character) with genome coverage over 70%, using whole genome bisulfite sequencing on Illumina GA. Significant differences among the DNA methylation profiles suggest a different methylation mechanism is used in cells having differentiation potency, which accords with mammalian DNA methylome studies on different cell lines that a different methylation mechanism (non-CG context) may be involved in the gene regulation in mammalian embryonic stem cells. Furthermore, we provide a more direct answer to how the methylation changes relate to chromatin structure by using nucleosome positioning maps, in contrast to previous studies which focus on the relation between DNA methylation and chromatin modifications.

A DELETION IN CHROMOSOME 5 GENERATING A CHIMAERIC GENE IS A PROTECTIVE FACTOR AGAINST ISCHAEMIC STROKE

R Rabionet[1], S Villatoro[1], J Aigner[1], J Jimenez-Conde[2], R Elosua[2], L Armengol[3], I Fernandez-Cadenas[4], J Montaner[4], E Marti[1], J Roquer[2], X Estivill[1]

[1]CRG, Genes and Disease, PRBB, Barcelona, 08003, Spain, [2]IMIM-Hospital del Mar, Neurology Unit - URLEC, PRBB, Barcelona, 08003, Spain, [3]Qgenomics, PRBB, Barcelona, 08003, Spain, [4] FIR-HUVH, Neurovascular Disease, Pg Vall d'Hebron, Barcelona, 08035, Spain

Stroke is a heterogeneous disorder with both environmental and genetic risk factors. Several studies have shown different association peaks for ischemic stroke but none has been consistently replicated. We have looked at structural variants as potential risk factors for stroke. We have screened an initial dataset of 169 hemorrhagic stroke, and 729 ischemic stroke cases, and 477 populational controls, and found an association of an insertion/deletion variant on chromosome 5. The breakpoints of this variant have been identified, and an allele-specific multiplex PCR assay has been designed and used for genotyping. Carriers of the intact allele are more frequently affected by stroke, while carrying one or two copies of the deleted allele confers protection against stroke. Association was significant in both the ischemic stroke (p –val: 0.0005; OR: 0.66) and the hemorrhagic stroke subsets (p –val: 0.002; OR: 0.58). This result has been replicated in an additional dataset with 475 ischemic stroke cases (p –val: 0.0087; OR: 0.71), and further replication in a second hemorrhagic stroke dataset is underway. Surrounding SNPs have been genotyped in order to assess linkage disequilibrium in the region, and identify potential tagging SNPs. The deletion spans 55Kb and affects two genes from the same family, generating a chimeric gene. Expression in EBV transformed lymphocytes of the two intact genes is correlated with copy number state, while the chimeric gene is expressed at a lower mRNA level. Downstream expression analysis from expression array data from HapMap samples carriers of zero, one, or two copies of the deleted allele shows differential expression of several genes, including genes involved in blood vessel formation and angiogenesis.

CONSTRUCTION OF THE FIRST SSR-BASED LINKAGE MAP OF FLAX (*LINUM USITATISSIMUM* L.) AND LOCALIZATION OF QTLS UNDERLYING FATTY ACID COMPOSITION

Raja Ragupathy[1], Scott Duguid[2], Sylvie Cloutier[1]

[1]Cereal Research Centre, Agriculture and Agri-Food Canada, 195 Dafoe Road, Winnipeg, R3T 2M9, Canada, [2]Morden Research Station, Agriculture and Agri-Food Canada, Route 100, Morden, R6M 1Y5, Canada

Flax seed is a rich source (~45-65%) of linolenic acid, an omega-3 fatty acid associated with multiple health benefits in human. To date, only two linkage maps of flax have been published and they were based on RAPD, RFLP and AFLP markers. In this study, a linkage map was generated based on 114 EST-derived SSR markers developed in our lab, as well as 5 SNP markers, 4 fatty acid desaturase genes (*FAD2A*, *FAD2B*, *FAD3A* and *FAD3B*) and seed coat color. A doubled haploid population of 78 individuals, generated from a cross between SP2047 (a yellow seeded Solin™ line with 2-3% linolenic acid) and UGG5-5 (a brown seeded flax line with 60-65% linolenic acid) was used for mapping. This population was grown in the field in five environments and was evaluated for fatty acid profiles. This map consists of 24 linkage groups with 112 markers spanning ~833.8 cM. QTL analysis detected two major QTLs for each of linoleic acid (LIO), linolenic acid (LIN) and iodine value (IOD). The mutant allele of *FAD3A* mapped to the chromosomal segment inherited from the parent SP2047 forms the basis of the QTL on linkage group 7 and was positively associated with high linoleic acid content but negatively associated with linolenic acid and iodine value. This FAD3A locus accounted for ~34%, ~25% and 29% of phenotypic variation observed for LIO, LIN and IOD, respectively. The QTL localized on linkage group 16 accounted for approximately 20%, 25 % and 13% of the phenotypic variation for these traits, respectively. For palmitic acid, a major QTL was localized on linkage group 9 which accounted for ~42% of the phenotypic variation. Being the first SSR based linkage map in flax, this will serve as a resource for mapping additional genes and traits, map-based cloning and molecular breeding programmes.

ONLINE QUANTITATIVE TRANSCRIPTOME ANALYSIS HTTP://GALAXY.TUEBINGEN.MPG.DE

Regina Bohnert, Jonas Behr, Andre Kahles, Geraldine Jean, Gunnar Raetsch

Max Planck Society, Friedrich Miescher Laboratory, Spemannstr. 39, Tuebingen, 72076, Germany

The current revolution in sequencing technologies allows us to obtain a much more detailed picture of transcriptomes. Studying them under different conditions or in mutants will lead to a considerably improved understanding of the underlying mechanisms of gene expression and processing. An important prerequisite is to be able to accurately determine the full complement of RNA transcripts and to infer their abundance in the cell.

We present the first integrative platform for quantitatively analyzing RNA-seq experiments. It is based on the Galaxy-framework [1] and builds on recently developed methods for NGS sequence analysis: a) We extended the alignment method QPALMA [2] and combined it with GenomeMapper [3] to align both spliced and unspliced reads with high accuracy, while taking advantage read quality information and splice site predictions. b) We extended the gene finding system mGene [4] to take advantage of read alignments to more accurately predict gene structures de novo. c) We developed the method rQuant that simultaneously estimates biases inherent in sequencing protocols and determines the abundances of transcripts [5]. It more accurately predicts abundances of alternative transcripts. d) Finally, we developed test techniques that determine significant differences between two RNA-seq experiments to find differentially expressed regions (with or without knowledge of transcripts).

The platform can be used for many purposes, including 1) to (re-)annotate genomes while profiting from NGS data; 2) to identify novel transcripts that are only expressed under certain conditions; and 3) to identify regions or transcripts that are the target of gene or RNA processing regulation. We have tested the system with data from several model organisms and human. Moreover, we have participated in the RGASP competition for an external evaluation of alignment, transcript identification and quantification accuracy.

[1] D. Blankenberg et al., Curr Protoc Mol Biol., Chapter 19, 2010.
[2] F. De Bona et al., Bioinformatics, 24 (16): i174, 2008.
[2] http://1001genomes.org/downloads/genomemapper.html
[4] G. Schweikert et al., Genome Research, 19(11):2133-43, 2009.
[5] R. Bohnert et al., BMC Bioinformatics, 10(S13):P5, 2009.

EXON CAPTURE AND RE-SEQUENCING IN RHESUS MACAQUES FOR IDENTIFICATION OF SNPS IN GENES EXPRESSED IN THE BRAIN

M Raveendran [1], GL Fawcett[1], M Bainbridge[1], F Yu[1], J Yu[1], D Muzny[1], RA Harris[2], A Milosavljevic[2], RA Gibbs[1], J Rogers[1]

[1]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030,
[2]Bioinformatics Research Laboratory, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030

Rhesus macaques (Macaca mulatta) are widely used as a model organism to study neurobiology and behavior, including susceptibility to anxiety disorders, depression or drug addiction among humans. Extensive knowledge of single nucleotide polymorphisms (SNPs) in genes related to neurobiology will facilitate genetic analysis in those studies, but to date, only very few (less than 800) total validated SNPs are currently available in public databases for this species. In order to increase the number of exonic SNPs in rhesus macaque in genes relevant to neurobiology, we carried out a re-sequencing study using NimbleGen capture arrays and Roche 454 sequencing. We selected 540 genes expressed in the human and or rhesus amygdala and 2234 genes expressed in hippocampus (http://www.ensembl.org/index.html). 37,266 exons from these 2,774 genes were printed on the 2.1M high density NimbleGen sequence capture array along with the full length (exon and intronic regions) sequences for 60 of these genes. The total number of target base pairs on the chip is 8.5Mbp. Three unrelated female rhesus macaques were selected for study and their DNA pooled and hybridized to the capture array following NimbleGen protocols. Eluted DNA was sequenced in two runs on a Roche 454 Genome Sequencer FLX Instrument. Duplicate reads were removed from the resulting sequence and SNPs called using the Atlas-SNP program (http: //www.hgsc.bcm.tmc.edu/cascade-tech-software-ti.hgsc). To date we have identified 55,647 potential SNPs in genes expressed in rhesus brain. We are in the process of validating these SNPs by comparing with data from other rhesus macaques using multiple chemistries. This approach can be used to identify SNPs in the exome or any subset of genes chosen due to tissue-specific expression or relevance to disease endophenotypes.

TRANSPOSABLE ELEMENT LANDSCAPE CHARACTERIZATION IN FIVE BAT GENOMES USING 454 SEQUENCE DATA

David A Ray[1,2], Heidi Pagan[1]

[1]Mississippi State University, Department of Biochemistry and Molecular Biology, 402 Dorman Hall, Mississippi State, MS, 39762, [2]Mississippi State University, Life Sciences and Biotechnology Institute, 650 Stone Blvd., Mississippi State, MS, 39762

Transposable elements (TEs) are major players in the evolution of eukaryotic genomes. They contribute to genome mass, recombination processes, the evolution of novel genes and are particularly effective mutagens. Vespertilionid bats appear to be unique among mammals in having been home to multiple waves of Class II TE activity in the recent past as exhibited by one WGS target, *Myotis lucifugus*. It is unclear, however, how far this activity extends taxonomically. Unfortunately, identifying the the unique TE signature of a genome can be a difficult and time-consuming process, especially in taxa that are not targets of whole genome sequencing projects. We have employed next-generation sequencing and a variety of computational tools to screen five vespertilionid bat genomes for TE content. Using a single run of a 454 GS FLX using Titanium chemistry, we generated between 6.8 and 25.3 Mbp for each taxon, corresponding to between 0.76% and 3.36% genome coverage. As would be expected from related taxa, analysis indicates both shared and unique TE histories for each species. For example, while all vespertilionids share common hAT and helitron, landscapes, genus *Myotis* remains the sole playground of piggyBac elements in these and other surveyed chiropterans. This methodology is an effective and relatively inexpensive method to gain an understanding of the unexplored genome landscapes of taxa not targeted for whole genome sequence analysis.

TOOLS TO EXTRACT SYSTEMS BIOLOGY DATA FROM MICROBIAL MODEL SYSTEMS

Chris Armour[1], Yasuhiro Oda[2], Sam Phattarasukol[2], Matt Biery[1], Caroline Harwood[2], Colin Lappala[2], Chris Raymond[1]

[1]NuGEN Technologies, Inc., Research & Development, 19805 North Creek Parkway, Suite 200, Bothell, WA, 98011, [2]University of Washington, Department of Microbiology, 1705 NE Pacific St, Seattle, WA, 98195

Microbes possess a wealth of metabolic capabilities that have enormous potential in the development of environmentally-friendly technologies. We have been using the rich species variation found among natural isolates of the purple, nonsulfur, phototrophic bacteria Rhodopseudomonas palustris as a model system for exploring phenotypic variation in the context of strain-specific differences in genomic architecture. By gathering detailed phenotypic information, corresponding whole transcriptome profiles, and complete genome sequences, we intend to further our understanding of how environmental conditions coupled with genetic variation influence phenotypic outcome.

In this presentation, we will focus on the development of tools that enable sequencing-based whole transcriptome profiling and complete genome sequencing. Genome sequencing is necessary to reveal strain-specific genomic variation and to provide a scaffold onto which transcriptome information can be mapped. The complementary transcriptome profiles provide a complete picture of how genetic variation and diverse growth conditions influence genome wide transcription patterns. This coupled approach is broadly applicable across microbial species. Our goal is to enable the research community to extract systems biology information from a variety of model systems.

IDENTIFYING FUNCTIONAL REGULATORY DNA SEQUENCE VARIANTS IN THE HUMAN GENOME WITH CHIP-SEQ AND RNA-SEQ

Timothy E Reddy[1], Jason Gertz[1], Florencia Pauli[1], Kimberly M Newberry[1], Ali Mortazavi[2], Brian A Williams[2], Georgi Marinov[2], Barbara Wold[2], Richard M Myers[1]

[1]HudsonAlpha Institute for Biotechnology, Myers Lab, 601 Genome Way, Huntsville, AL, 35806, [2]California Institute of Technology, Division of Biology, 1200 E. California Boulevard, Pasadena, CA, 91125

Identifying functional regulatory DNA sequence variants in the human genome is a major challenge. The use of high-throughput next-generation DNA sequencing to measure transcription factor occupancy in chromatin (ChIP-seq) allows us to witness not only the binding sites but also the DNA sequence underlying functional regulatory sites on a genomic scale. Sequencing the transcriptome with RNA-seq reveals the sequence of expressed transcripts. Through the efforts of re-sequencing projects such as the 1,000 Genomes Project, individual diploid genome sequences are now becoming available. By aligning sequence tags from ChIP-seq and RNA-seq experiments to the individual's diploid reference genome, it is possible to identify functional regulatory variants in the human genome.

In this work, we developed an alignment approach and a statistical framework to measure allelic specificity in next-gen, short-read sequencing data. We applied this approach to ChIP-seq data that we generated for RNA Polymerase 2 as well as for more than 30 transcription factors in a lymphoblastoid cell line from one individual. We observe many instances of various transcription factors binding preferentially to a single allele, with allelic specificity ranging from absolute to weak. Many of the allele-specific transcription factor binding events are associated with allele-specific gene expression that we observe in the accompanying RNA-seq data, giving further evidence that the allele-specificity of the binding is functionally important. For a subset of these transcription factors, we also performed ChIP-seq in lymphoblastoid cell lines from both parents of the original individual. By using these datasets, we can trace the inheritance of functional events in a family. In doing so, we are beginning to understand the extent to which inherited regulatory DNA sequence variants direct allele-specific gene expression.

TESTING FOR GENE FLOW BETWEEN NEANDERTALS AND MODERN HUMANS

David Reich[1], Richard E Green[2,3], Svante Paabo[3]

[1]Harvard Medical School, Genetics, NRB, 77 Avenue Louis Pasteur, Boston, MA, 02115, [2]University of California, Santa Cruz, Department of Biomolecular Engineering, 1156 High St, Santa Cruz, CA, 95064, [3]Max Planck Institute, Leipzig, 6 Deutscher Platz, Leipzig, 04103, Germany

The morphological features typical of Neandertals first appear in the European fossil record about 400,000 years ago, followed by the appearance of progressively more distinctive Neandertal forms. The archaeological record shows that Neandertals likely came into contact with anatomically modern humans in the Middle East by about 120,000 years ago, and subsequently overlapped geographically and temporally with anatomically modern human in parts of Europe and western Asia before dying out about 30,000 years ago.

There is controversy about whether Neandertals exchanged genes with modern humans during the period when they overlapped. Morphological features of early modern human fossils as well as some present-day human populations have been interpreted as evidence both for and against genetic exchange. Deep lineages at specific genetic loci have also been interpreted as evidence of gene flow. However, only mitochondrial DNA has been compared in detail between Neandertals and modern humans, and this locus provides no evidence for gene flow.

To perform a high resolution test for evidence of gene flow beween modern humans and Neandertals, we analyze more than 3 billion base pairs of data that we obtained by sequencing three Neandertal bones, and compare it to five diverse present-day humans. We measure how Neandertals are related to different populations of present-day humans both by searching for differences in average relatedness and by searching for haplotype patterns diagnostic of gene flow.

APPLICATIONS OF RAPID SEQUENCE FRAGMENT SCREENING TO LARGE-SCALE SEQUENCING DATA

Jeffrey G Reid, David Rio Deiros, Matthew Bainbridge, Richard A Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030

With the increasingly rapid pace of sequence production, the development of tools to provide sequence comparison results rapidly has become essential to large-scale sequencing. Building on our e-GenoType allele comparison approach (www.e-genotype.com), which screens raw read data for expected alleles (in silico genotyping), we have expanded the sensitivity of our probe sequence screening algorithm and developed a variety of tools that use sequence fragment screening to support a broad spectrum of analysis approaches packaged in a traditional open source software framework.

Beyond improvements that allow for one or two differences between the probe sequence and the read data, we have also generated tools for characterizing and filtering finished genome data. A variety of fragment screening approaches will be discussed, including screening of all (fixed-size) subsequences from a genome against any given probe set to characterize the specificity of probes, screening the genome subsequences against other genomes to identify probes that can serve as unique markers for any given genome relative to a metagenomic background, fragment screening for contamination detection and filtering in de novo sequencing, along with a general discussion of optimized sequence fragment screening. As this analysis is orders of magnitude faster than mapping, it is ideal for QA/QC applications. For example, in tumor/normal pair sequencing it can distinguish the tumor from the normal rapidly, and provides an instant snapshot of the biology to inform downstream analysis.

DETECTION OF RARE POLYMORPHISMS IN HYPERMUTATED IGVH IN CHRONIC LYMPHOCYTIC LEUKEMIA (CLL) CELLS.

Juan L Rodriguez-Flores, Olivier Harismendy, Karen Messer, Bradley Messmer, Thomas J Kipps, Kelly A Frazer

UCSD, Moores Cancer Center, 3855 Health Sciences Dr., La Jolla, CA, 92093

As normal B-cells mature in the germinal center, IgVh sequences become hypermutated. In CLL, patients without hypermutated IgVh have a poor prognosis. Our aim is to quantify the diversity and frequency of IgVh mutations in the population of leukemic cells within a CLL patient. Next generation sequencing has enabled deep sequencing of individual DNA molecules and thus a digital read out of DNA variant frequency. However, the identification of rare variants (<10% frequency) remains challenging due to relatively high sequencing error rates.

In this study, we used Sanger sequencing to obtain a consensus DNA reference sequence of the IgVh locus DNA for 2 CLL tumors (CLL1 indolent, IgVh 91.2% homologous to germline ; CLL2 aggressive, IgVh 99.6% homologous to germline). We then deeply sequenced the transcriptomes of each tumor using an ABI SOLiD instrument. Using the BFAST aligner, we mapped the 50 base-pair reads to the consensus IgVh reference of each sample (120k reads mapped in CLL1; 152k reads mapped in CLL2). Statistically significant variants were called using a script based on the SAMtools library, assuming a Poisson distribution for the number of sequencing errors at each base, and using a 5% family-wise significance level adjusted for multiple tests. To estimate the sequencing error rate, we mapped 28k reads in CLL1 and 30k in CLL2 to GAPDH and CD20, housekeeping genes without known hypermutation in CLL. We applied stringent quality filters to each read and basecall.

In each sample, we were able to successfully identify several high-confidence variants that were present at low prevalence (<10% frequency). We also reduced the expected false-positive rate in comparison to the SAMtools consensus base-caller. In future work we will determine the halplotype phase of the variants and trace their divergence from the consensus reference, to investigate clonal selection in CLL.

NOVEL MULTI-NUCLEOTIDE POLYMORPHISMS IN THE HUMAN GENOME CHARACTERIZED BY WHOLE GENOME AND EXOME SEQUENCING

Jeffrey A Rosenfeld, Anil K Malhotra, Todd Lencz

Zucker Hillside Hospital, Psychiatry Research, 75-59 263rd Street, Glen Oaks, NY, 11004

Genomic sequence comparisons between individuals are usually restricted to the analysis of single nucleotide polymorphisms (SNPs). While the interrogation an analysis of SNPs is efficient, they are not the only form of divergence between genomes. We expand the scope of polymorphism detection by investigating the occurrence of double nucleotide polymorphisms (DNPs) and triple nucleotide polymorphisms (TNPs), in which consecutive nucleotides are altered compared to the reference sequence. We have found such DNPs and TNPs throughout two complete genomes and eight exomes. The overall abundance of DNPs and TNPs in these genomes is approximately 1% of that of SNPs, encompassing tens of thousands of DNPs and thousands of TNPs in each genome. Within exons, these novel polymorphisms are over-represented amongst protein-altering variants; nearly all DNPs and TNPs result in a change in amino acid sequence and, in some cases, two adjacent amino acids are changed. DNPs and TNPs represent a potentially important new source of genetic variation which may underlie human disease and they should be included in future medical genetics studies. The preponderance of pathogenic protein alterations caused by a DNP or a TNP indicates their potential for causing illness.

POST-LIGHT SEQUENCING WITH SEMICONDUCTOR CHIPS

Jonathan M Rothberg

Ion Torrent, Research, 37 Soundview Road, Guilford, CT, 06437

Ion Torrent has developed a DNA sequencing system that directly translates chemical signals into digital information on a semiconductor chip. This approach leverages a trillion dollars of investment from the semiconductor industry taking advantage of existing state-of-the-art chip fabrication technology, and the entire semiconductor design and supply chain. Unprecedented scalability and cost reduction result from decades of Moore's Law advances in semiconductor technologies that are brought to bear within a few years for DNA sequencing.

Ion Torrent sequencing takes place in semiconductor microchips that contain sensors which have been fabricated as individual electronic detectors, allowing one sequence read per sensor. Current configurations have 1.5 million sensors in a 1 cm^2 chip, with proof of principle to enable densities over 100 million sensors per chip.

The sequencing chemistry itself is remarkably simple. Native nucleotides are incorporated into the growing strand by native DNA polymerase. As a base is incorporated, a direct electrical measurement of the incorporation event is made and the sequence is read out directly into the digital domain. Thus, sequencing is direct, efficient, and massively parallel, requiring no specialized reagents and no optical systems. Using native DNA chemistry with real time detection enables run times to be very short, on the order of an hour with a throughput on the order of 100 Megabases per hour. We will present data and describe metrics from the adenovirus and *E. coli* genomes.

The Ion Torrent sequencer provides a powerful tool for driving research, which will be commercially available in 2010. The simplicity of a semiconductor chip that reads itself means that desktop and portable instruments will be available at a fraction of the cost of other next-generation instruments; The use of standard reagents, low reaction volume and high data density keep reagent costs low; Computational infrastructure and staff support requirements are modest; Finally, the short run time supports fast research cycle times and promotes the use of sequencing in everyday research.

SURVEY OF 20,000 HUMAN Y CHROMOSOMES SHOWS THAT TWO DELETIONS WITHIN THE AZFC REGION EXIST AT POLYMORPHIC FREQUENCIES IN DIVERSE POPULATIONS

Steve Rozen[1,2], Janet D Marszalek[2], Katherine Irenze[3], Kristin Ardlie[3,4], David C Page[2]

[1]Duke-NUS Graduate Medical School, NBD, 8 College Road, Singapore, 169857, Singapore, [2]Howard Hughes Medical Institute, Whitehead Institute, and Massachusetts Institute of Technology, Biology, 9 Cambridge Center, Cambridge, MA, 02142, [3]Genomics Collaborative / Bioserve, Global Repository, 9000 Virginia Manor Road, Beltsville, MD, 20705, [4] Broad Institute, Biological Samples Platform, 7 Cambridge Center, Cambridge, MA, 02142

Deletion of most of the Y-chromosome's AZFc region causes severely low sperm count. Other deletions involve only part of AZFc, and many studies have focused on the influence of these deletions on sperm count. However, lack of fundamental knowledge of the population genetics of partial AZFc deletions has hampered these studies. To better understand the population genetics of these deletions, we studied them in a total of 20,884 men in five populations. In contrast to previous studies, the men were unselected for sperm count or fertility. We found that two classes of partial AZFc deletion are present at polymorphic frequencies and vary significantly in prevalence by population. The prevalence of gr/gr-deletions ranges from 2% (US) to 15% (Vietnam). The prevalence of b2/b3 deletions ranges from 0.5% (India and Tunisia) to 2.2% (Poland). A third group of partial AZFc deletions, the b1/b3 deletions, is much rarer, at 0.1% across the five populations combined. This is only three times the prevalence of b2/b4 "complete" AZFc deletions, which cause severely low sperm count and which are almost never inherited. The prevalence of different classes of deletions also varies significantly by Y haplogroup within population, even after excluding the known association of b2/b3 deletions with haplogroup N1. We hypothesize that this variation results mainly from founder effects. However, haplogroup-specific variation in the frequency of mutations causing the deletions is also conceivable. In either case, the variation in deletion prevalence by haplogroup implies that haplogroups must be considered when assessing the effects of partial AZFc deletions on spermatogenesis.

DROSOPHILA CHROMOSOMAL EVOLUTION: THE ROLE OF TRANSPOSABLE ELEMENTS

Alfredo Ruiz[1], Oriol Calvete*[1], Fernando Prada*[1], Alejandra Delprat[1], Josefa González[1,2]

[1]Universitat Autonònoma de Barcelona, Departament de Genètica i de Microbiologia, Bellaterra (Barcelona), 08193, Spain, [2]Stanford University, Department of Biology, Stanford, CA, 94305-5020

Transposable elements (TEs) are intragenomic parasites with the potential to restructure genomes through element-mediated chromosomal rearrangements. However, their actual contribution to chromosomal evolution in the genus Drosophila is still controversial. We have previously found that three *D. buzzatii* polymorphic inversions were generated by ectopic recombination between copies of the transposon *Galileo*. In contrast no evidence for an implication of TEs in the generation of inversions fixed as interspecific differences has been found in the genus at large. In order to assess the contribution of TEs to the generation of Drosophila inversions and contrast different hypotheses to explain inversion breakpoint reuse, we have cloned and sequenced the breakpoints of six second-chromosome inversions fixed in the buzzatii complex (repleta group). The six inversions have different ages as shown by their phylogenetic distribution and include three instances of cytological breakpoint coincidence. We found that in all cases but one the coincident breakpoints fall in the same intergenic region corroborating the cytological observations. Fragments of *Galileo* elements were found in the breakpoint junctures of two inversions sharing one breakpoint whereas in all breakpoint junctures (but one) of the other four inversions, fragments of the transposon *BuT5* were detected, sometimes along with fragments of other TEs. These observation suggest that ectopic recombination between *Galileo* or *BuT5* copies may have generated at least some of these inversions. However, the situation is much more complex for one inversion that has been accompanied by a gene duplication and subsequent microinversion of the duplicate gene. This inversion could have been caused by insertion of a hybrid *BuT5* element or by staggered breaks and reparation. We conclude that (i) TEs are involved in the generation of at least some fixed inversions, (ii) cytological coincidences are often confirmed at the molecular level and (iii) TEs are possibly responsible for some inversion breakpoint coincidences.

*Authors contributed equally.

GENOME-WIDE ANALYSIS OF *NEUROSPORA CRASSA* TRANSCRIPTS REGULATED BY THE NONSENSE-MEDIATED MRNA DECAY PATHWAY

Ying Zhang[1], Fei Yang[1], Mohammed Mohiuddin[2], Stephen K Hutchison[2], Lorri A Guccione[2], Chinnappa Kodira[2], Matthew S Sachs[1]

[1]Texas A&M University, Department of Biology, MS3258, College Station, TX, 77843, [2]Roche 454, 454 Research & Development, 20 Commercial Street, Branford, CT, 06405

Nonsense-mediated mRNA decay (NMD) is a surveillance pathway that rids cells of mRNAs that contain premature translation termination codons. It is active in all eukaryotes examined and the core factors are highly conserved. NMD pathways in higher eukaryotes can employ factors that are not present in the yeast *Saccharomyces cerevisiae*, such as components of the exon junction complex (EJC), which has a role in mRNA splicing. The genome of the model filamentous fungus *Neurospora crassa* contains core NMD components as well as EJC components, and, unlike *S. cerevisiae*, many of its mRNAs are spliced. We have established that knockouts of *N. crassa* genes for the NMD components UPF1 and UPF2 lead to the increased stability of specific mRNAs that are NMD substrates. We are using 454 whole transcriptome sequencing to perform studies of transcripts in *N. crassa* strains that are wild-type or deficient in NMD to evaluate at the genome-wide level the changes that occur when this surveillance pathway is eliminated. Here we present the results of our comparative analysis of the whole transcriptome data from wild type and knockout *Neurospora* strains and provide further evidence for the extent and complexity of NMD in regulating transcript metabolism.

SILK WEAVER: A WORKFLOW MANAGEMENT SYSTEM FOR NGS DATA ANALYSIS

Taro L Saito, Jun Yoshimura, Wei Qu, Shinichi Morishita

University of Tokyo, Department of Computational Biology, 5-1-5 Kashiwanoha, Kashiwa City, Chiba, 277-8562, Japan

In NGS data analysis, we have to organize variety of tasks into a workflow, including index construction of genomes, read alignment, detecting notable patterns from the results, etc. An ideal scenario would be that the workflows run perfectly without any modification. In reality, however, parameter values of the programs have to be tuned to achieve better alignment results. Sometimes we also need to switch alignment programs (e.g, BLAST, Bowtie, BWA, SOAP2, etc.) since each of them has pros and cons in their performance, tolerance to mismatches and the number of results to be produced. Refinement of workflows involves such try-and-error tests, and thus their management are strongly demanded. To ease the burden of designing workflows, we
developed a workflow management system, called Silk Weaver, which enables us to organize bioinformatics workflows and supports partial evaluation of the workflows, which can be used to test various program options. Silk Weaver also manages tuning options used to produce the results for ensuring the reproducibility of the data analysis. In addition, we extended the UTGB genome browser to visualize the execution of the workflows and the detailed read alignments including in/dels and mutations. Silk Weaver and UTGB are open-source projects, and their source code is freely available at http://utgenome.org/.

A GENOME-WIDE ANALYSIS OF POPULATION STRUCTURE IN SWEDEN

Elina Salmela[1,2], Tuuli Lappalainen[3], Päivi Lahermo[1], Jianjun Liu[4], Kamila Czene[5], Per Hall[5], Juha Kere[2,6]

[1]University of Helsinki, Institute for Molecular Medicine Finland FIMM, P.O. Box 20, Helsinki, FI-00014, Finland, [2]University of Helsinki, Department of Medical Genetics, P.O. Box 63, Helsinki, FI-00014, Finland, [3]University of Geneva Medical School, Department of Genetic Medicine and Development, CMU / Rue Michel-Servet 1, Geneva, 1211, Switzerland, [4] Genome Institute of Singapore, Human Genetics, 60 Biopolis Street, Singapore, 138672, Singapore, [5]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, P.O. Box 281, Stockholm, SE-171 77, Sweden, [6]Karolinska Institutet, Department of Biosciences and Nutrition, Novum, Huddinge, SE-141 57, Sweden

The genetic structure of human populations is a potential confounding factor in genome-wide association studies, and it can also yield information on population history. We have studied the population structure within Sweden using more than 410,000 autosomal single nucleotide polymorphisms (SNPs) genotyped in 754 Swedes on the Illumina HumanHap550 array. We have also compared the Swedes with over 3000 reference samples from other European populations based on almost 25,000 SNPs overlapping between the Illumina and Affymetrix 250K Sty arrays. We observed a general correspondence of genetic and geographic distances, which is consistent with previous findings from Europe. Within Sweden, different parts of the country showed varying amounts of influence from the neighboring populations of Finland and Germany, which may reflect historical immigration patterns. Furthermore, the northernmost part of Sweden, Norrland, clearly differed from the central and southern parts of the country and exhibited a loose north-south pattern, whereas the genetic substructure in the rest of the country was subtle. The distinctive genetic features of Norrland probably result mainly from isolation by distance and genetic drift caused by low population density. Overall, the population structure within Sweden appears less pronounced than that previously detected within the neighboring population of Finns. These results underline the potential of genome-wide data in analyzing populations that might otherwise appear relatively homogeneous, such as the Swedes.

CHD7 FUNCTIONS AS A REGULATOR OF BOTH NUCLEOPLASMIC AND NUCLEOLAR GENE EXPRESSION

Mike P Schnetz*, Gabriel E Zentner*, Peter C Scacheri

Case Western Reserve University, Genetics, Euclid Ave, Cleveland, OH, 44106

CHD7 is one of nine members of the chromodomain helicase DNA-binding domain family of ATP-dependent chromatin remodeling enzymes found in mammalian cells. De novo mutation of CHD7 is a major cause of CHARGE syndrome, a genetic condition characterized by multiple congenital anomalies. To gain insights to the function of CHD7, we used ChIP-Seq to map CHD7 sites in mouse ES cells, representing the earliest precursor to the cell types affected in CHARGE syndrome. We detected 10,483 sites on chromatin bound by CHD7 at high confidence. Most of the CHD7 sites show features of gene enhancer elements, i.e., they are located distal to transcription start sites, contain p300 and high levels of H3K4 mono-methylation, positioned within open chromatin, and enriched near genes that are specifically expressed in ES cells. We were surprised to find that CHD7 can modulate genes in either the positive or negative direction, given that enhancers are typically associated with gene activation. Next, we aligned CHD7 ChIP-seq reads to sequences corresponding to rDNA (not included in the genome assemblies), and detected multiple CHD7 sites on rDNA. The binding of CHD7 to rDNA was validated by standard ChIP and is consistent with immunofluorescence and western blot analysis of subcellular fractions, showing high levels of CHD7 in the nucleolus. ChIP-chop analyses demonstrate that CHD7 specifically associates with hypomethylated, active rDNA, suggesting a role as a positive regulator of rRNA transcription. Consistent with this hypothesis, siRNA-mediated depletion of CHD7 results in hypermethylation of the rDNA promoter and concomitant reduction of 45S pre-rRNA levels. Reduced cell proliferation and protein synthesis were also observed upon reduction of CHD7 expression. Furthermore, compared to wild-type, the levels of 45S pre-rRNA are reduced in both Chd7+/- and Chd7-/- mouse ES cells, as well as Chd7-/- whole mouse embryos and multiple tissues dissected from Chd7+/- embryos. Together, these results indicate that CHD7 dually functions as a transcriptional regulator in the nucleoplasm and the nucleolus. We propose that the multiple anomalies in CHARGE syndrome are due to the combined effects of dysregulated tissue-specific gene expression and reduced rRNA biogenesis. Lastly, the results serve as proof of principle for ChIP-seq analysis of rDNA-binding proteins.

ASSEMBLY AND EVOLUTIONARY ANALYSIS OF THE GORILLA GENOME

Aylwyn Scally

Sanger Institute, Informatics, Hinxton, Cambs, CB10 1SA, United Kingdom

Gorilla Genome Sequencing Consortium

Gorillas are our closest evolutionary cousins after chimpanzees, and in fact are sufficiently akin to us that ~17% of our genome is closer to gorilla than to chimpanzee. Thus the gorilla genome is essential to a fuller understanding and interpretation of our own evolution since divergence from the other great apes. We present the assembly and analysis of a whole genome sequence for gorilla, based on 2x capillary and 35x Illumina sequencing of a female Western Lowland gorilla (Gorilla gorilla gorilla), 'Kamilah'. The assembly is 3.0 Gbp in length, containing ~95% of Kamilah's genomic sequence, with typical base-pair contiguity (N50) of 11 kbp and scaffold contiguity of 913 kbp. Chromosomes comprising 96% of the assembly have been constructed from scaffolds using human homology and incorporating known large-scale human-gorilla rearrangements. A final assembly stage using BAC and fosmid end pairs to correct and confirm the chromosomal structure is nearly complete.

Gene annotation and orthology/paralogy analysis using the Ensembl-Compara pipeline has identified models for over 21,000 protein-coding genes, with orthologs to more than 90% of those in human. We have also sequenced gorilla transcriptome data and are using it to add further annotation and to investigate non-coding RNA and gene expression in comparison with human and other primates. We have generated a multiple alignment of the human, chimpanzee, gorilla, orangutan and macaque genomes in which 90% of the gorilla assembly is represented, and using a coalescent inference model we are able to significantly improve the estimates of timescales and population sizes involved in the great ape speciation. We also identify those regions genome-wide where humans are closest to gorilla, and are investigating the characteristics of such regions. In the orthology analysis, 26% of gene trees involving 1-1 orthologs had human/gorilla as the closest pair.

We have compared the sequence content of human, chimpanzee and gorilla genomes using reference assemblies and Illumina data from all three. Surprisingly, chimpanzee and gorilla share ~100 Mbp of sequence not found in human, of which 20% is non-repetitive and contains functional material. We have also obtained sequence data for other gorillas from both the Eastern and Western species, and are able to measure sequence diversity and copy-number variation, providing insights into recent evolution within the gorilla genus.

GENOME-WIDE ASSOCIATION UNCOVERS A NOVEL ANTIMALARIAL RESISTANCE GENE IN *P. FALCIPARUM*

Stephen F Schaffner*[1], Daria Van Tyne*[2], Daniel J Park*[1], Daniel E Neafsey*[1], Elaine Angelino*[3], Joseph Cortese[1], Kayla Barnes[2], David Rosen[2], Amanda Lukens[2], Rachel Daniels[1,5], Danny Milner[2], Charles Johnson[1], Ilya Shlyakhter[1,5], Shari Grossman[1,5], Daniel Yamins[5], Dyann F Wirth[1,2], Sarah K Volkman[1,2], Pardis C Sabeti[1,5]

[1]Broad Inst, Infect Disease, 7 Cambridge Ctr, Cambridge, MA, 02142, [2]Harvard Sch of Pub Health, Immun & Infect Disease, 665 Huntington Ave, Boston, MA, 02115, [3]Harvard, Engineering, Mass Hall, Cambridge, MA, 02138, [4] Harvard, Mol & Cell Biology, Mass Hall, Cambridge, MA, 02138, [5]Harvard, FAS Ctr for Sys Biol, 52 Oxford St, Cambridge, MA, 02138

Malaria's rapid adaption to new drugs has allowed it to remain a devastating infectious disease; understanding and tracking adaptation's genetic basis is critical to the success of therapeutic and intervention strategies. We developed a high-density genotyping array with 17,000 single nucleotide polymorphisms (SNPs) across the *P falciparum* genome (~1 SNP/kb), and applied it to 48 culture-adapted malaria parasites from 3 continents, characterizing population structure and identifying signatures of selection throughout the genome. We also created a platform for high throughput assessment of drug-sensitivity phenotypes and performed genome-wide association studies (GWAS) for chloroquine and halofantrine resistance. Besides detecting the known chloroquine resistance locus *pfcrt*, we discovered a novel halofantrine resistance locus, *PF10_0355*. Functional analysis showed that overexpression of *PF10_0355* reduces sensitivity to halofantrine, mefloquine and lumefantrine but not to unrelated antimalarials. Our results demonstrate the power of genome-wide approaches to identify resistance loci and point to *PF10_0355* as mediator of drug resistance in the parasite.
* Contributed equally.

DE NOVO ASSEMBLY OF LARGE GENOMES USING CLOUD COMPUTING.

Michael C Schatz, Dan Sommer, David Kelley, Mihai Pop

Univ. of Maryland, Center for Bioinformatics and Computational Biology, 3104B Biomolecular Sciences Building #296, College Park, MD, 20742

The first step towards analyzing a previously unsequenced organism is to assemble the genome by merging together the sequencing reads into progressively longer contig sequences. New assemblers such as Velvet, Euler-USR, and SOAPdenovo attempt to reconstruct the genome by constructing, simplifying, and traversing the de Bruijn graph of the reads. These assemblers have successfully assembled small genomes from short reads, but have had limited success scaling to larger mammalian-sized genomes, mainly because they require memory and compute resources that are unobtainable for most users.

Addressing this limitation, we are developing a new assembly program Contrail (http://contrail-bio.sf.net), which uses the Hadoop/MapReduce distributed computing framework to enable de novo assembly of large genomes. MapReduce was developed by Google to simplify their large data processing needs by scaling computation across many computers, and the open-source version called Hadoop (http://hadoop.apache.org) is becoming a de facto standard for large data analysis, especially in so called “cloud computing” environments where compute resources are rented on demand. For example, we have also successfully leveraged Hadoop and the Amazon Elastic Compute Cloud for Crossbow (http://bowtie-bio.sf.net/crossbow) to accelerate short read mapping and genotyping, allowing quick (< 4 hours), cheap (< $100), and accurate (> 99% accuracy) genotyping of an entire human genome from 38-fold short read coverage.

Similar to other leading short read assemblers, Contrail relies on the graph-theoretic framework of de Bruijn graphs. However, unlike these programs Contrail uses Hadoop to parallelize the assembly across many tens or hundreds of computers, effectively removing memory concerns and making assembly feasible for even the largest genomes. Preliminary results show contigs produced by Contrail are of similar size and quality to those generated by other leading assemblers when applied to small (bacterial) genomes, which scales far better to large genomes. We are also developing extensions to Contrail to efficiently compute a traditional overlap-graph based assembly of large genomes within Hadoop, a strategy that will be especially valuable as read lengths increase to 100bp and beyond.

COMPARATIVE POPULATION GENOMICS OF THE PLANT PATHOGENIC FUNGI MYCOSPHAERELLA GRAMINICOLA AND ITS WILD RELATIVE SPECIES

Eva H Stukenbrock[1], Troels T Hansen[1], Julien Y Dutheil[1], Thomas Bataillon[1], Ruiqiang Li[2], Marcello Zala[3], Bruce A McDonald[3], Wang Jun[2], Mikkel H Schierup[1]

[1]Aarhus University, Bioinformatics Research Center, CF Mollers Alle, Aarhus, 8000, Denmark, [2]BGI, Beijing Genomics Institute, Shenzhen, 518083, China, [3]ETH Zurich, Inst. Integrative Biology, Universitätsstr. 2, Zurich, 9082, Switzerland

The fungal pathogen Mycosphaerella graminicola is a recently evolved pathogen of wheat. We performed whole genome sequencing of 2 individuals M. graminicola and 10 individuals of two very closely related progenitors.
Paired end Illumina sequencing at 30-50X allowed de novo assembly of the 12 genomes with scaffold N50 of 100 kb. Alignment of these scaffolds to a reference genome of M. graminicola resulted in a ~30Mb multiple alignment covering >80% of the reference genome including 13 essential chromosomes of the pathogen and a variable number of small dispensable chromosomes. Pulse field gel electrophoresis revealed presence/absence polymorphism of these chromosomes predating the speciation as well as several large scale rearrangements.
Levels of polymorphism and divergence were largest on dispensable chromosomes. We assessed Ka/Ks ratios in 9725 aligned genes and identified 9 candidate genes (4 with signal peptides) likely involved in host specialization or speciation. Signal peptides were also highly enriched among the genes showing evidence for adaptive evolution using the McDonald-Kreitman test. Global analysis of MK tables showed that adaptive evolution has played an important role in shaping patterns of variation within and between species.
We applied a coalescent HMM to decipher the evolutionary history of M. graminicola and its close relatives. We found that speciation of M. graminicola was associated with a bottleneck ~10.000 years ago from a more variable ancestral species. Incomplete lineage sorting (ILS) affects 40% of nucleotides. The amount of ILS decreases with gene density and increases with the amount of recombination, both in accordance with a large role of natural selection in the evolution of these genomes.

LINKING ALLELE SPECIFIC EXPRESSION AND DNA METHYLATION IN THE H1 HUMAN EMBRYONIC STEM CELL GENOME

Robert J Schmitz[1], Matthew D Schultz[1,2], Ryan Lister[1], Mattia Pelizzola[1], Tanya Biorac[3], Delia Ye [3], Miroslav Dudas[3], Gavin D Meredith[3], Christopher C Adams[3], Joseph R Ecker[1]

[1]The Salk Institute of Biological Studies, Genomic Analysis Laboratory, 10010 North Torrey Pines Rd, La Jolla, CA, 92037, [2]Univeristy of California San Diego, Bioinformatics, 9500 Gilman Drive, La Jolla, CA, 920903, [3]Invitrogen, a division of Life Technologies Corporation, Genetic Systems Business Unit, 5791 Van Allen Way, Carlsbad, CA, 92008

Recent dramatic decreases in the cost of DNA sequencing have enabled the rapid production of human genomes, revealing vast genetic diversity in the genome sequence between individuals. In addition, the first genome-wide single-base map of DNA methylation in humans was recently reported for H1 human embryonic stem cells (hESC) and differentiated fibroblasts, revealing widespread differences in cell-type DNA methylomes. However, to date, no comprehensive characterization of the consequences of sequence diversity on the epigenome has been reported. Here we present the high coverage genome sequencing of the widely used H1 hESC line. We have used the SOLiD platform to sequence the H1 hESC line with short single-end reads in addition to sequencing short (1.5 kb) and long (5 kb) mate-pair libraries for a total of ~50X read coverage and >1000X clone coverage. The combination of these library types has allowed us to characterize extensive genetic variation revealing SNPs, small insertion/deletions, inversions and structural variants. Because of the high depth of sequence coverage in our dataset and the two-base encoding strategy used in SOLiD sequencing we have confidently identified heterozygous loci within the H1 hESC genome. These heterozygous loci have allowed us to identify allele specific RNA expression and DNA methylation. These results demonstrate the importance of matching the genome with the transcriptome(s) and the epigenome(s) from the same individual for studying the relationship between genetic and epigenetic variation.

CHANGING WITH THE TIMES: THE HUMAN REFERENCE GENOME

V Schneider[1], P Flicek[2], T Graves[3], T Hubbard[4], D Church[1]

[1]NCBI, Bethesda, MD, 20892, [2]EBI, Hinxton, Cambridge, CB10 1SD, United Kingdom, [3]The Genome Center at Washington University, St. Louis, MO, 63108, [4] The Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom

Major efforts to sequence and assemble the human genome culminated in 2004, resulting in a reference assembly of very high coverage and quality. Since that time, the reference genome has played critical and diverse roles in many aspects of biological research. Among the important findings facilitated by the reference genome was the discovery of an unanticipated degree of genetic variation between individuals, coupled with the realization that the extant assembly is insufficient in its representation of these often divergent and complex genomic regions. Providing alternative assemblies of such regions is a major focus of the Genome Reference Consortium (GRC), the group responsible for ongoing efforts to improve the reference genome (http://genomereference.org). In 2009, the GRC released the current version of the human reference genome, GRCh37, which for the first time puts alternate locus representations into a chromosome context. We will review these alternate assemblies and demonstrate their importance to personal genomics projects and improving phenotype to genotype correlations. The GRC also strives to correct sequencing errors and close the remaining gaps in the reference genome. While such assembly updates are crucial to researchers working in poorly-represented regions, frequent assembly releases disrupt the coordinate system relied upon by other researchers working at the whole-genome level. To balance the need for timely updates with the need for a relatively stable coordinate system, the GRC has developed a system of assembly patches. Patches consist of new or updated sequence contigs defined outside the coordinate system of the current assembly that can be released between full assembly updates. Patches will be incorporated into the reference assembly at the time of the next full update. We will present the first set of assembly patches for GRCh37 and review their relationship to the current and future reference genomes, as well as their biological importance. The work presented is a collaborative effort of the GRC member institutions: NCBI, EBI, The Genome Center at Washington University and The Wellcome Trust Sanger Institute.

FLASH SEQUENCING: STRUCTURE AND SEQUENCE INFORMATION FROM SINGLE MOLECULES

Timothy M Schramm[1,2], Konstantinos Potamousis[1], Steve Goldstein[1], Quinglin Pei[4], Shiguo Zhou[1], Michael Newton[4], David C Schwartz[1,2,3]

[1]University of Wisconsin-Madison, Laboratory for Molecular and Computational Genomics, 425 Henry Mall, Madison, WI, 53706, [2]University of Wisconsin-Madison, Chemistry, 1101 University Ave, Madison, WI, 53706, [3]University of Wisconsin-Madison, Genetics, 425 Henry Mall, Madison, WI, 53706, [4] University of Wisconsin-Madison, Biostatistics and Medical Informatics, 1300 University Ave, Madison, WI, 53706

The Flash Sequencing System allows for acquisition of DNA sequence information from individual molecules—without use of cycles or amplification. Single molecules of genomic DNA, from hundreds of kilobasepairs to megabasepairs in length, are site-specifically labeled with a fluorochrome-labeled nucleotide. At each site, the amount of fluorochrome label is quantitated when each individual molecule map is interrogated; thus, the quantity of the labeled base is determined. Sequence information from multiple sites spread across the context of large DNA molecules leads to the phasing of haplotype and elucidation of single nucleotide polymorphism as well as larger scale structural alteration. The information rich physical maps derived from the single molecules can be assembled to provide comprehensive analysis of human genomes.

T.M.S. is supported by an NHGRI training grant to the Genomic Sciences Training Program (5T32HG002760).

USE OF LARGE SCALE LINKAGE AND GENOME-WIDE ASSOCIATION STUDY RESULTS TO ESTIMATE THE UPPER BOUND FOR THE EFFECT SIZES OF LESS COMMON CAUSAL VARIANTS AND THE LIKELIHOOD THEY ARE RESPONSIBLE FOR COMMON GENOME-WIDE ASSOCIATION SIGNALS FOR TYPE 2 DIABETES

Laura J Scott[1], Weihua Guan[2], Michael Boehnke[1]
[1]University of Michigan, Biostatistics, 1420 Washington Hghts, Ann Arbor, MI, 48109, [2]University of Minnesota, Biostatistics, 420 Deleware St. SE, Minneapolis, MN, 55455

Rapid advances in next-generation sequencing technologies allow ever more complete surveys of targeted genomic regions and whole genomes and are beginning to generate a flood of genetic association studies of rarer variants. Assessment of results from existing genome-wide association studies (GWAS) and linkage studies for complex diseases, such as type 2 diabetes (T2D), can help provide estimates of the upper limits of allelic relative risks (RR) that might be expected for variants with risk allele frequency (RAF) < 1% and also address whether variants with RAF < 1% can account for the observed GWAS signals. A meta-analysis of 23 T2D linkage studies, containing an equivalent of ~6500 affected sib pairs, did not detect genome-wide significant linkage signals (Guan et al., 2008). Under a multiplicative model, this study would have had 80% power to detect causal variants with total RAF of 0.5% and RR = 6 or RAF of 1% and RR = 4. Taking these RRs as the plausible upper limits of RR for causal variants of frequency 0.5% - 1% we next asked if the associations we detected in our T2D GWAS could have been caused by variants consistent with the linkage results. The three previously known T2D loci, as well as the initial six T2D loci identified through GWAS, were identified in sample sizes of up to ~32,000 cases and controls. In a total sample of ~32,000 cases and controls, we would have had < 5% power at $p = 5 \times 10^{-8}$ to detect an association from a causal variant with RAF = 0.5% (1%) and RR < 6(4) for a GWAS SNP in D' = 1 with the causal variant with a RAF > 45% (>55%). Four of the initial nine T2D loci have RAF > 50% suggesting that causal variants with combined RAF < 1% are unlikely to underlie these signals. In a subsequent DIAGRAM meta-analysis, with an effective sample size of up to 64,000, two of six associated SNPs had RAF > .65 and would have had < 5% power to detect causal variants with RAF = 0.5% and RR < 6. In conclusion, a substantial portion of the T2D GWAS loci identified to date are unlikely to be explained by variants with RAF < 1%.

ENHANCING THE ANNOTATION OF GENOMES IN ENSEMBL

Stephen Searle, Bronwen Aken, Julio Banet, Susan Fairley, Felix Kokocinski, Magali Ruffier, Amy Tang, Jan Vogel, Simon White

Wellcome Trust Sanger Institute, Informatics, Hinxton, Cambridge, CB10 1SA, United Kingdom

A major aim of Ensembl is generating annotation on vertebrate genomes. We have previously developed systems for annotating protein coding and short non coding (ncRNA) genes. These methods have been applied to over 50 genomes. Several recent additions to the annotation system will be described including an improved pipeline for the merging of manual annotation into the automatically predicted set, a pipeline for predicting long ncRNA genes, and the systems we are developing for generating annotation from RNASeq data.

Ensembl had previously incorporated a subset of models from manual annotation from the HAVANA group. The latest merge process now incorporates all HAVANA annotation into the final set. The manual annotation contains more alternative isoforms than the Ensembl automatic annotation, and includes extra non coding gene types and more pseudogene annotation, but currently only covers 70% of the genome. Merging manual and automatic annotation provides a set which covers the complete genome while retaining the extra depth from the manual set. Although increasing the number of transcripts approximately 3 fold, there has been only a small change in the number of protein coding genes in the merged set (22320 compared to 22218).

We have developed a pipeline to predict long ncRNA genes, based on a method previously described by Guttman et al. The approach is based on the distinctive chromatin signature which marks actively transcribed genes located outside of known protein-coding loci. This is implemented in Ensembl utilising data from the Ensembl functional genomics, comparative genomics and annotation databases. When applied to human the pipeline identified 4502 long ncRNA genes, and is currently being applied to mouse.

We have developed a system for generating transcript models from RNASeq data and applied this to a large set of RNSeq data for Zebrafish, and are working on merging this with the standard Ensembl annotation. The method has also been entered in the RGASP RNASeq model building competition. Results for this will be released shortly.

HIGH RESOLUTION ANALYSIS OF CNV BREAKPOINTS REVEALS POTENTIALLY PREDISPOSING SEQUENCE MOTIFS AND VARIABLE MECHANISMS OF GENOMIC REARRANGEMENT.

Hung-Chun Yu[1], Chad Haldeman-Englert[2], Elizabeth A Geiger[1], Hongbo M Xie[3], Juan C Perin[3], Xiaowu Gai[3], Tamim H Shaikh[1]

[1]Univ. of Colorado, Pediatrics, Aurora, CO, 80045, [2]Wake Forest Univ. Sch. of Medicine, Pediatrics, Winston-Salem, NC, 27103, [3]The Children's Hosp. of Phila., BIC, Phila., PA, 19104

Copy number variations (CNVs) are an increasingly recognized cause of human disease and variation. We have analyzed rearrangement breakpoints in 100 pathogenic CNVs detected in a cohort with intellectual and developmental disabilities. 17/100 (17%) CNVs resulted from NAHR between highly homologous segmental duplications (SDs), a significant enrichment, considering that SDs only make up 5% of the genome. Almost all of these 17 CNVs were recurrent and the frequency was directly proportional to the size and sequence identity shared between the paralogous SDs mediating NAHR. The remaining 82 CNVs were mostly non-recurrent, singletons with a few instances of overlapping CNVs but with different breakpoints. We used custom-designed, tiling microarrays to refine the rearrangement breakpoints in a subset of these allowing the rapid cloning and sequencing of 42 breakpoints (21 CNVs). 25 breakpoints localized to repetitive DNA elements like Alus and LINES, but only three CNVs appeared to result from NAHR between repeats. The majority of the breakpoints appeared to result from Non-Homologous End Joining (NHEJ). In each of the CNVs analyzed, 1 -14 base pairs of microhomology was observed at the breakpoints further supporting the involvement of NHEJ. We identified a Poly(dA:dT) tract at an average distance of 85 bp from 41/42 breakpoints. This motif has been shown to be important in nucleosome organization, suggesting a correlation between chromosome breakpoints and chromatin structure. We further analyzed each of the breakpoint regions for sequence motifs that have previously been identified at or near rearrangement breakpoints. Although many such “hotspot” motifs were identified, none were significantly enriched when compared to control sequences. Further analysis of additional CNV breakpoints, both pathogenic and normal variants, will help determine the relative contribution of these variable rearrangement mechanisms in generation of CNVs in the human genome.

TANDEM REPEAT SEQUENCES AS CAUSATIVE CIS EQTLS FOR PROTEIN-CODING GENE EXPRESSION VARIATION: THE CASE OF *CSTB*

Andrew J Sharp[1], Christelle B Borel[1], Eugenia Migliavacca[1,2], Emmanouil T Dermitzakis[1], Maryline Gagnebin[1], Stylianos E Antonarakis[1]
[1]University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva, 1211, Switzerland, [2]Swiss Institute of Bioinformatics, Geneva, 1211, Switzerland

Individual variation of gene expression contributes to phenotypic variability. Thus the understanding of the genetic control of gene expression is important to decipher the etiology of genetic traits and disorders. Genetic analysis in cells and tissues from different individuals has revealed both cis and trans expression quantitative trait loci (eQTLs) for many genes; the effect of these cis-eQTLs is relatively strong since they have been discovered and replicated with relatively small sample sizes. Recent studies have revealed that the majority of the discovered eQTLs are tissue specific (Dimas et al. Science 2009), a finding compatible with the tissue specificity of the phenotypic characteristics of genetic disorders. The challenging question for each eQTL is to identify the functional variant that controls the gene expression variation.

We hypothesize that copy number variation of short sequence repeats in the human genome contributes to the gene expression variation of some genes. Our laboratory has previously identified that rare expansions of a dodecamer repeat (CGGGGCGGGGCG) in the promoter region of the *CSTB* gene on chromosome 21q lead to silencing of the gene, resulting in progressive myoclonus epilepsy (Lalioti et al Nature 1997). The majority of alleles in the human population contain either 2 or 3 copies of this dodecamer repeat. Since the large expansion to more than 200 repeat copies results in silencing of the *CSTB* gene, we hypothesized that the common 2 or 3 copy variation may be a causative cis-eQTL for *CSTB* expression variation.

We used PCR to genotype the repeat length and quantified *CSTB* expression by TaqMan qRT-PCR in 170 lymphoblast and fibroblast samples from the GenCord collection. Although there is considerable variation in expression of this gene in the normal population, we observed that in lymphoblasts repeat length is strongly correlated with *CSTB* expression ($p=3x10^{-11}$), with individuals homozygous for the 3-repeat allele showing ~2-fold higher expression than individuals homozygous for the 2-repeat allele. In fibroblasts a weaker effect was observed ($p=0.03$), showing that this effect is cell-type specific. Examination of both genotyped and imputed SNPs within 1Mb of *CSTB* revealed none that were significantly correlated with *CSTB* expression. Therefore, the dodecamer repeat represents the strongest cis eQTL for *CSTB* in lymphoblasts, explaining 23% of expression variation. We conclude that polymorphic tandem repeats likely represent the causative variation of a fraction of eQTLs in the genome.

COMPARING THE TRANSCRIPTIONAL CIRCUITS CONTROLLING HUMAN AND MOUSE HEMATOPOIESIS

Tal Shay*[1], Vladimir Jojic*[2], Noa Novershtern[1,3,4], The Immunological Genome Project Consortium - -[3], Benjamin L Ebert[1,5], John L Rinn[1,6], Daphne Koller [2], Aviv Regev [1,3,7]

[1]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02142, [2]Stanford University School of Medicine, Department of Computer Science, MSOB X319, 251, Stanford, CA, 94305, [3]Massachusetts Institute of Technology, Department of Biology, 77 Massachusetts Avenue, Cambridge, MA, 02140, [4] Hebrew University, School of Computer Science, Givat Ram Campus, Jerusalem, 91904, Israel, [5]Brigham and Women's Hospital, Harvard Medical School, Division of Hematology, 75 Francis Street, Boston, MA, 02115, [6]Beth Israel Deaconess Medical Center, Harvard Medical School, Department of Pathology, 1 Deaconess Road, Boston, MA, 02215, [7]Howard Hughes Medical Institute, 25 Francis Avenue, Cambridge, MA, 02138

Hematopoietic differentiation – the process of differentiation from a hematopoietic stem cell into all blood and immune cells – is a much-studied differentiation process in mammals. Despite decades of research, the transcriptional circuitry controlling this process is still only partially understood. The generation of two unprecedented compendia of gene expression measured in a range of multipotent and differentiated cell types across the human (DMAP) and mouse (www.immgen.org) hematopoietic lineages opened the way to systematically decipher the regulatory circuitry controlling this process. Here, we build on a novel extension of probabilistic graphical models to model gene expression using the topology of the differentiation tree to construct a unified regulatory model in the lineage (Jojic, Shay et al, in preparation). We applied this approach to both human and mouse compendia, identifying the key transcription factors that drive this process. We compare the resulting models between the two organisms to detect the conserved and divergent features of the regulatory code. Our analysis represents a first comprehensive model of gene regulation within a complete cell lineage, highlights similarity and differences between mouse – a common model in immunology – and humans, and can shed light on the role of dis-regulation of cell circuits in hematologic disorders and malignancies.

* Equal contribution

THE DESIGN OF WHOLE GENOME RESEQUENCING FOR ASSOCIATION STUDIES

Yufeng Shen[1,2], Itsik Pe'er[1,2]

[1]Columbia University, Center for Computational Biology and Bioinformatics, 1130 St Nicholas Avenue, New York, NY, 10032, [2]Columbia University, Computer Science, 1214 Amsterdam Avenue, New York, NY, 10027

Whole-genome resequencing (WGS) allows direct interrogating of rare variants, which will facilitate the study of human diseases. One of the emerging problems where rare variants are crucial is the missing heritability observed in genome-wide association studies (GWAS) of common diseases. However, the cost of sequencing is a limiting factor in designing large-scale case-control association studies. Here we describe theoretical considerations to maximize the power of detecting association under the constraint of cost. First we considered the scenario where the cost was proportional to the total amount of base pairs to be sequenced, i.e., proportional to the number of subjects times the average depth-coverage per subjects. In an association study, the power of detecting association increases when the number of subjects increases. Meanwhile, the power of correctly calling rare variants increases when the average depth-coverage increases. Therefore it is important to strike a balance between the number of subjects and the average depth-coverage per subject. We established a depth-coverage distribution model based on empirical data. Assuming publicly available population control data (such as that from the 1000 Genomes Project) will be used in a study, and the rare variants of interest are near absent in the control data, we demonstrated that the optimal power of detecting association was always achieved at medium depth-coverage (<15x) under realistic conditions. We then considered a second scenario where the number of subjects alone determined the cost and the power of variant detection is always close to 1. This is in line with whole exome sequencing, where a smallest unit (lanes or slides) of high-throughput sequencing platform yields large depth-coverage for a subject. Assuming it is necessary to sequence both cases and controls, the problem of study design is then what is the optimal ratio of cases and controls that maximizes the power of detecting association of rare variants. We deduced that if the rare variants of interest are near absent in controls, the optimal ratio of cases is 1/e.

METHYLATION DETECTION IN A MCF-7 CELL LINE USING ULTRA HIGH-THROUGHPUT BISULFITE-SEQUENCING WITH THE SOLID™ SYSTEM

Vrunda Sheth[1], Stephen F McLaughlin[1], Zheng Zhang[2], Christina Chung[2], Melissa A Barker[2], Victoria L Boyd[2], Heather E Peckham[1]

[1]Life Technologies, Genetic Systems, 500 Cummings Center, Beverly, MA, 01915, [2]Life Technologies, Genetic Systems, 850 Lincoln Center, Foster City, CA, 01915

Epigenetic modification plays an important regulatory role in diseases such as cancer. While bisulfite sequencing is the most powerful method to study DNA cytosine methylation,approximately 99% of cytosine bases are converted to uracil thus creating essentially a 3 base bisulfite-converted genome that has less signature than a 4 base genome and is more difficult to map against with short reads. Traditional approaches to map the bisulfite-converted tag to both a bisulfite-converted reference and an unconverted reference can lead to loss of information when mapping reads with a moderate number of methylated cytosines. For example a read that has 6 methylated and 6 unmethylated bases would not map to either reference when mapped allowing up to 5 mismatches. In order to enable accurate mapping to the highly redundant bisulfite-converted genome,we have developed a mate-pair scheme in which only one mate-pair tag is bisulfite converted while the other tag remains unconverted. The adapters and non-converted tag are protected via incorporation of 5-methyl-cytosine,which is resistant to bisulfite conversion.We used this protocol to make a mate-pair library from a human MCF-7 cell line. The non-converted sequence provides an “anchor” in the genome and facilitates the identification and the methylation status of the bisulfite-converted tag. This approach reduces mis-mapping since the anchor sequence regulates the mapping of the bisulfite-converted tag and allows the non-anchored tag to tolerate more mismatches and thus map more of the moderately methylated tags and open a window to these previously difficult to obtain regions. This novel technique has the potential to provide a routine and reliable method for hypothesis free genome-wide methylation detection. Studying the role of methylation on gene expression will provide useful information on the role of epigenetic mutations in human breast cancer.
The SOLiD instrument is For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

IN-DEPTH METABOLIC CHARACTERIZATION OF GENETIC LOCI UNDERLYING SERUM-LIPIDS

S-Y Shin[1], A-K Petersen[2], W Römisch-Margl [3], G Zhai[4], K Small[4], R Wang-Sattler [2], E Grundberg[1,4], J Ried[2], A Döring[2], H-E Wichmann[2,5,6], M Hrabé de Angelis [7,8], H-W Mewes[3,9], T Illig[2], TD Spector[4], J Adamski[7,8], K Suhre[3,10], C Gieger[2], N Soranzo[1,4]

[1]Wellcome Trust Sanger Institute, Human Genetics, Hinxton, CB10 1HH, United Kingdom, [2]Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany, [3]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany, [4] King's College London, DTR, London, SE1 7EH, United Kingdom, [5]Institute of Medical Informatics, Ludwig-Maximilians-Universität, Munich, , Germany, [6]Klinikum Grosshadern, Munich, Munich, Germany, [7]Institute of Experimental Genetics, Helmholtz Zentrum München, Neuherberg, Germany, [8]Institute of Experimental Genetics, Technische Universität München, Munich, Germany, [9]Technische Universität München, Department of Genome-oriented BioinformaticsMunich, Germany, [10]Ludwig-Maximilians-Universität, Faculty of Biology, Munich, Germany

Emerging metabolomics technologies enable monitoring of small metabolites from body fluids, and can help elucidate processes underlying inherited variation in established disease risk factors. To dissect the effect of recently published genetic variants influencing serum lipids and heart disease, we profiled 151 metabolites in 1,797 participants from the KORA population (Germany), replicating results in the TwinsUK cohort (n=1,236). We aim to identify cases where a genetic locus is associated with both a lipid and a metabolite concentration. We report here the initial results of this effort, as well as discussing methodological approaches to metabolite analyses. Among others, we identify associations of *GCKR* variants with different ratios of plasmalogens and phosphatidylcholines ($P = 3.2 \times 10^{-8}$).

USING BIOSCOPE™ SOFTWARE AND THE SOLID™ SYSTEM TO INVESTIGATE VARIATION IN THE HUMAN GENOME AT HIGH COVERAGE AND ACCURACY

Asim Siddiqui, Heather Peckham, Fiona Hyland, Aaron Kitzmiller, Jeff Ichikawa, Somalee Datta, Eric Tsung, Charles Scafe, Yutao Fu, Rajesh Gottimukkala, Caleb Kennedy, Stephen McLaughlin, Onur Sakarya, Paolo Vatta, Zheng Zhang, Gina Costa, Ellen Beasley

Life Technologies, Genetic Systems, 850 Lincoln Centre Drive, Foster City, CA, 94404

The advent of high throughput next generation sequencing enables experiments to study genome variations, transcription of coding and non-coding RNAs and epigenetic profiles. The ability to design effective experiments that make efficient use of samples and highly parallel sequence generation requires a clear understanding of the impact of system accuracy on the power to detect a variety of meaningful biological differences between or within samples. Standard expressions of sequence quality are important, but insufficient, to support rational experimental design.
In developing the SOLiD System we include system validation and the release of system data sets. With mapped data accuracy at greater than 99.9%, we continue to analyze reference samples so that we can continue to report concordance with a reference genome.
Using the SOLiD 3+ System, we generated 128 GBases of alignable sequence for the HuRef genome. We used the Bioscope Software to produce BAM files, calling SNPs, indels, inversions and CNVs. Validating against the 7x Sanger and SNP chips for this individual (Levy et al., 2007), we found 99.7% SNP concordance. Small indel sensitivity was 80-85% depending on the choice of mapping parameters.
Using the SOLiD 4 System, we have generated over 100 GBases of additional HuRef genome sequence data to date. We will present data that demonstrates the power of Bioscope and SOLiD to detect SNPs, indels and CNV in standard resequencing experiments and comparisons of the SOLiD 4 system with its predecessor.

ESTIMATION OF ANCESTRAL HUMAN DEMOGRAPHY FROM INDIVIDUAL GENOME SEQUENCES

Ilan Gronau[1], Brad Gulko[1], Charles G Danko[1], Melissa J Hubisz[1], Stephan C Schuster[2], Webb Miller[2], Vanessa M Hayes[3], Adam Siepel[1]

[1]Cornell U., Biol Stats & Comput Biol, 102 Weill Hall, Ithaca, NY, 14853, [2]Penn State U., Ctr Comp Genomics & Bioinformatics, 310 Wartik Lab, University Park, PA, 16802, [3]U. New South Wales, C25 Lowy Cancer Research Ctr, High St., New South Wales, 2031, Australia

Complete genome sequences are now available for individuals representing several distinct human populations. Interest in these sequences so far has focused on the technical feasibility of individual genome sequencing, the identification of single nucleotide and structural variations, and implications for personalized medicine. However, these data also represent a rich source of information about human evolution. Here we describe an effort to estimate key evolutionary parameters -- including the times at which major human population groups diverged, the effective sizes of these populations, and the rates of migration between them -- from the complete genome sequences of seven individuals, including two European Americans, a Yoruban, a Han Chinese, a Korean, a Bantu, and a Khoisan.

To extract information about ancestral populations from this data set, we rely on patterns of variation in genealogies across loci, using a coalescent-based method called MCMCcoal (Rannala and Yang, *Genetics*, 2003). We adapt this method to accommodate (unphased) diploid genomes and migration between populations. In addition, we realign all reads using a uniform pipeline, and employ a novel algorithm to integrate over uncertainty in both genotypes and haplotype phasings in the calculation of genealogy likelihoods. Finally, we apply a comprehensive set of filters to minimize biases from both sequencing/alignment error and natural selection.

Preliminary results indicate a split time of about 100 thousand years ago (kya) for the ancestral Khoisan population, a split time of ~50 kya for the Eurasian and African populations, and a split time of ~40 kya for the Europeans and East Asians. Estimates of Eurasian population sizes are sharply reduced in comparison with ancient and modern African sizes. We estimate an effective size of ~5000 for the population ancestral to all modern humans.

PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES USING BOTH SEQUENCE AND EXPRESSION INFORMATION FROM MULTIPLE SPECIES

Elizabeth Siewert, Katerina Kechris

University of Colorado Denver, Biostatistics & Informatics, 13001 E. 17th Place, B-119, Aurora, CO, 80045

Genome-wide detection of transcription factor binding sites (TFBSs) is difficult because they are short, degenerate sequences buried in unknown locations within regulatory regions of a gene. Earlier prediction methods incorporated genome-wide expression data and promoter sequences into a linear-model framework, regressing measures of expression onto counts of putative TFBSs in promoters for a single species. More recently, it has been shown that including genomic sequence data from multiple species improves the predictive ability of this regression model. We describe two extensions of this single-species, linear regression method using a multivariate modeling framework. The resulting algorithms extend the search space to both sequence *and* expression information from *all* available genes across *multiple* species. We also constrain our multivariate models to account for the phylogenetic relationships among the multiple species. We show that the multiple-species method results in an improvement in the prediction of TFBS over the single species method using several evaluation criteria. In summary, utilizing both sequence *and* expression data across species appears to be beneficial for genome-wide prediction of TFBSs.

SEQUENCE CAPTURE TECHNOLOGY FOR RE-SEQUENCING IN NON-HUMAN GENOMES

Snaevar Sigurdsson[1,2], Gerli Pielberg[2], Evan Mauceli[1], Claire Wade[1,3], Cord Drogemuller[4], Mia Olsson[2], Leeb Tosso[4], Matthew Webster[2], Kerstin Lindblad-Toh[1,2]

[1]Broad Institute, Sequencing, 7CC, Cambridge, MA, 02142, [2]Uppsala University, Department of Medical Biochemistry and Microbiology, BMC, Uppsala, SE-75237, Sweden, [3]University of Sydney, Faculty of Veterinary Sciences, NSW, Sidney, 2006, Australia, [4] University of Berne, Institute of Genetics, Vetsuisse Faculty, PO-Box 8466, Berne, CH-3001, Switzerland

The aim of the project was to develop targeted re-sequencing of candidate genomic regions to identify novel genomic variants using sequence capture microarrays in non-human species. Recent technology advances in DNA sequencing has revolutionized how genetic variants are discovered, allowing millions of nucleotides to be sequenced in one single run. For re-sequencing of megabase sized candidate genomic regions a single sequencing run of Illumina Genome analyzer would be more than enough. The major obstacle has been to efficiently select or amplify the region of interest from the rest of the genome prior to sequencing. For this purpose selective sequence-capture methods using hybridization on microarrays have been developed. On example is the custom-made microarray from Roche-Nimblegen that includes 385 K probes that can capture 1-5 Mb of genomic sequence. This technique has been developed for human and mouse samples, but has so far not been applied on other genomes.

We have modified the protocol to perform sequence capture on microarrays followed by Illumina sequencing in a variety of species including dogs, horses and cattle. Our protocol involves sheering to ~300bp, ligating Illumina adapters, sequence capture on Nimblegen microarray, adapter mediated PCR followed by sequencing. We typically map 60% of sequence back to the target, and with an average coverage of ~100x we expect to see >96% of the target covered by at least 5x.

The methodology has worked robustly for several applications including regions in the dog genome to detect SNPs, indels and CNVs associated with several canine diseases, the Leopard Complex spotting in the horse and arachnomelia in cattle. We conclude that the method works robustly across species.

IDENTIFICATION OF STRUCTURAL VARIATION IN NEXT-GENERATION SEQUENCE DATA BY MULTIPLE-SIGNAL INTEGRATION

Selim Önal[1], Luke C Peng[1,3], Anna Ritz[1], Hsin-Ta Wu[1], Suzanne S Sindi[2,3], Benjamin J Raphael[1,3]

[1]Brown University, Dept. of Comp. Sci., Providence, RI, 02912; [2]Brown University, Div. of Molecular and Cellular Biology, Providence, RI, 02912; [3]Brown University, Center for Computational Molecular Biology, Providence, RI, 02912

Projects such as 1000 Genomes and The Cancer Genome Atlas are employing next-generation sequencing technologies to resequence many individual genomes. Alignment of sequenced reads to the reference genome allows one to discover both single nucleotide variants and structural variants. Numerous methods have been developed to identify structural variation from resequencing data. The two primary signals used are read-depth (RD) coverage and paired-end mapping (PEM). Both rely on the correct alignment of reads to a reference genome; thus the power to detect variants with breakpoints in repetitive sequence is reduced. To overcome this, other methods incorporate reads with multiple possible mappings.

We introduce two probabilistic models for structural variation detection. One combines signals from both RD and PEM (GASV+CC), while a second also includes ambiguous reads (GASV-Prob). Each is built on our established Geometric Analysis of Structural Variants (GASV) program that precisely localizes the breakpoints of variants (Sindi et al, 09). To our knowledge, GASV+CC and GASV-Prob are the first methods that integrate RD and PEM (of all discordant pairs) into a single probabilistic model.

We compared GASV+CC and GASV-Prob with existing methods on simulated structural variants from the Venter genome and resequencing data from two individuals in the 1000 Genomes project (NA18507 and NA12878). For the individual genomes we used a set of validated structural variants from fosmid data (Kidd et al, 08) to estimate the true positive rate. On simulated deletions our probabilistic methods, GASV+CC and GASV-Prob, out performed competing methods using only PEM or RD. GASV+CC has the most balanced performance, making 50% fewer predictions at equal sensitivity of other methods. GASV-Prob showed the highest sensitivity, detecting 65% of deletions with 40% fewer predictions than other methods. On the 1000 Genomes data the relative performance of GASV-Prob and GASV+CC was maintained. GASV-Prob was the most sensitive method (>70%), while GASV+CC was the most specific. These results demonstrate combining multiple signals of structural variation into a single model achieves increased sensitivity and specificity of detection.

EXPEDITED BATCH PROCESSING AND ANALYSIS OF TRANSPOSON INSERTION SITES IN NON-MAMMALIAN VERTEBRATES

Jeremy D Smith[1,2], David A Ray[1]

[1]Mississippi State University, Department of Biochemistry and Molecular Biology, 402 Dorman Hall, Starkville, MS, 39762, [2]Mississippi State University, Sciences and Biotechnology Institute, 650 Stone Blvd, Starkville, MS, 39762

With advances in sequencing technology, greater and greater amounts of genomic data are becoming available every day. A large portion of these genomics sequences consists of transposable elements; frequently 50% or more in vertebrates. Transposable elements are known to act as drivers of genomic evolution and diversification and are important genetic markers. Each transposable element family may have thousands of copies within a given genome, and therefore it can take an exorbitant amount of time and effort to process data in a meaningful fashion. In order to combat this problem, we developed a set of modern bioinformatics techniques and programs to streamline the analysis. This includes a unique perl script which automates the process of taking BLAST, Repeatmasker and similar data and extracting the hit sequences from the genome. This script, called Process_hits uses an Object-oriented methodology to compile all hit locations from a given file for processing, organize this data into useable categories, and output it into multiple formats. It is capable of handling large amounts of transposon data in an efficient fashion, with each of the major sub-functions, hit processing, nucleotide sequence extraction, and hit-object methods, are contained within their own sub-modules to allow for greater expandability and as foundation for future program design.

TIGHT REGULATION OF LARGE-SCALE SOMATIC REARRANGEMENT IN A BASAL VERTEBRATE GENOME

Jeramiah J Smith[1], Evan E Eichler[2], Chris T Amemiya[1]

[1]Benaroya Research Institute, at Virginia Mason, 1201 9th Ave., Seattle, WA, 98101, [2]University of Washington, Department of Genome Sciences, 1705 NE Pacific St, Seattle, WA, 98195

Somatic genome rearrangement is often a cause and consequence of cancers or other "genomic disorders". However, a few metazoan and protist lineages are known to undergo tightly-regulated and large-scale somatic recombinations during the normal course of their development. The discovery of programmed genome rearrangement in lamprey (a vertebrate) fills an important gap in our understanding of dysregulated rearrangement of vertebrate genomes and the capacity for tight regulation of genome rearrangement other taxa. Several lines of evidence demonstrate that the lamprey undergoes a dramatic remodeling of its genome, resulting in the elimination of hundreds of millions of base pairs (~20% of the genome) from many somatic cell lineages during embryonic development. Embryological studies reveal that many of these rearrangements take place early in development, resulting in a situation wherein an individual's "germline" and "somatic" cell lineages differ substantially in genome structure and gene content. Computational, array CGH (comparative genomic hybridization), and 454 sequencing studies reveal that several distinct genomic regions are altered during this process and have identified specific rearrangement breakpoints that differentiate germline and somatic genomes. Genomic regions that are removed via programmed rearrangements include hundreds of genes, many of which are transcribed in adult and juvenile testes or during early embryonic development. A large fraction of these somatically-deleted genes have homologs that are known to contribute to genome stability or the specification/maintenance of pluripotent cell lineages.

It is worth noting that ostensibly similar rearrangements have also been observed in hagfish (another basal chordate lineage), suggesting that this dynamic genome biology can be traced to a point very near the common ancestor of all vertebrates. Understanding the mechanisms by which lamprey regulates such extensive remodeling of its genome will provide invaluable insight into factors that can promote stability and change in vertebrate genomes, as well as the consequences of reorganization in the context of "normal" vertebrate development and cell biology.

ANALYSIS OF METAGENOMIC HUMAN SPECIMENS AT THE WASHINGTON UNIVERSITY GENOME CENTER

Erica Sodergren[1], Hongyu Gao[1], Kathie Mihindukulasuriya[1], Yanjiao Zhou[1], Kristine Wylie[1], Tiffany Williams[1], Makedonka Mitreva[1], John Martin[1], Sahar Abubucker[1], Karthik Kota[1], Lynn Carmichael[1], Eric deMello[1], Josh Peck[1], WIlliam Shannon[2], Elena Deych[2], Jia Wang[2], George M Weinstock[1]

[1]Washington University School of Medicine, The Genome Center, 4444 Forest Park, St. Louis, MO, 63108, [2]Washington University School of Medicine, Dept. of Medicine, Division of General Medical Sciences, Biostatistical Consulting Center, 660 S. Euclid Ave, St. Louis, MO, 63110

The HMP has produced some of the largest metagenomic data sets thus far generated in the field (~50 subjects, ~900 specimens) and allow a number of questions to be approached in more detail than previously: reproducibility of sampling methods, Sanger vs 454 16S sequencing, stability of microbiomes between visits, variation in microbiomes between or within individuals for different body sites, variation between body sites, etc. These specimens have been subjected to 16S rRNA sequencing by both Sanger and 454 (using 3 windows within the 16S gene). Tools for analyzing these 16S community descriptions are being developed to answer these questions.

Shotgun data provides superior information compared to 16S, allowing species and gene content description. Using both Illumina and 454 will produce terabases of shotgun sequences, requiring the development of new approaches for metagenomic analysis. An integrated analysis system is being developed to use sequence overlaps and coverage, GC content, phylogenetic markers, similarity to known sequences and association to pathway networks. Species identification and abundance is being approached by comparison to ~1000 reference genomes.

TARGETED SEQUENCING OF INDEXED LIBRARIES USING POOLS OF BIOTINYLATED OLIGONUCLEOTIDE CAPTURE PROBES

Frank J Steemers[1], Kerri York[1], Wiehua Chang[1], Jean Lozach[1], Casey Turk[1], Jerry Kakol[1], Jennie M Le[1], Natasha Pignatelli[1], Mostafa Ronaghi[1], Niall Gormley[2], Johanna Whitacre[1], Melissa Shults[1], and Kevin L Gunderson[1]

[1]Illumina, Inc., R&D, 9885 Towne Centre Dr., San Diego, CA, 92121, [2]Illumina, Inc., R&D, Little Chesterford, Essex, CB10 1QY, United Kingdom

Targeted enrichment and indexed library sample preparation are two key technologies that, when combined, enable full utilization of next generation sequencing technologies. We have developed a targeted sequencing assay capable of enriching up to 10 Mb of genomic region (exonic or contiguous) in a single-tube “pull-down” assay. The enrichment protocol uses single and/or tiled biotinylated 80-100mer oligonucleotides to specifically target short DNA library elements of interest (150-450 bases). After annealing of the biotinylated oligo pool to standard Genome Analyzer (GA) DNA libraries, excess biotinylated oligos are removed, and the oligo-library duplex pulled down onto streptavidin beads. The bound library elements are stringently washed to remove non-specifically bound DNA. The library elements are eluted and subjected to amplification before clustering and sequencing on the GA. In order to match the throughput of the GA with the desired sample coverage (e.g., 50-80X) per lane, sample indexing in targeted sequencing is required. We demonstrate simultaneous, uniform enrichment of 12 different indexed sample libraries in a single pull-down across over 5500 different exons. Additionally we have demonstrated scale-up of enrichment complexity to over 55,000 exons from a single sample and demonstrated mock complexity to over 300,000 pull-downs. At this complexity, the human exome can be enriched in a single pull-down. In summary, our enrichment assay exhibits scalability, high specificity with 60-80% of reads on target, good coverage uniformity with greater than 90% of the targeted bases covered at 0.2X or greater of the mean coverage, minimal allelic bias, and concordance with Infinium® genotyping of over 99.8%. We show practical utility of our exon enrichment assay for both detection of cancer mutations and copy number changes in DNA from tumor cell lines. These technologies provide a flexible, automatable, highly-specific and relatively unbiased route towards rapid targeted sequencing of the genome.

ARE NUCLEOSOME POSITIONS *IN VIVO* PRIMARILY DETERMINED BY HISTONE-DNA SEQUENCE PREFERENCES?

Arnold Stein, Taichi E Takasuka, Clayton K Collings

Purdue University, Biological Sciences, 915 W. State St., West Lafayette, IN, 47907

Large-scale and genome-wide studies have concluded that approximately 80% of the yeast (*Saccharomyces cerevisiae*) genome is occupied by positioned nucleosomes. *In vivo* this nucleosome organization can result from a variety of mechanisms, including the intrinsic DNA sequence preferences for wrapping the DNA around the histone core. Recently, genome-wide studies were reported using massively parallel sequencing to directly compare *in vivo* and *in vitro* nucleosome positions. In one case, it was concluded that intrinsic DNA sequence preferences indeed have a dominant role in determining the in vivo nucleosome organization of the genome, consistent with a genomic code for nucleosome positioning. Some other studies disagreed with this view. We have attempted to clarify a fundamental question concerning the packaging of genomic DNA: To what extent are nucleosome positions in vivo determined by histone-DNA sequence preferences? We have analyzed data obtained from different laboratories in the same way, and have directly compared these data. We examine the difference between nucleosome occupancy and nucleosome positioning. We compute the precise effect of statistical nucleosome positioning for yeast chromatin. We question using read numbers from massively parallel sequencing to determine nucleosome occupancies. We also identify possible problems with some of the data analyses. Our findings suggest that DNA sequence preferences have only small effects on the positioning of individual nucleosomes throughout the genome *in vivo*. However, these small effects may still be significant in influencing chromatin higher-order structure.

LEVERAGING THE 1000 GENOMES PROJECT FOR NEXT-GENERATION MICROARRAYS

Michael A Eberle, Jennifer L Stone, Karine Viaud, Luana Galver, Chan Tsan, Ken Kuhn

Illumina Inc., DNA Analysis Products, 9885 Towne Centre Drive, San Diego, CA, 92121

High-throughput sequencing is expanding the catalogue of variation at an unprecedented rate, enabling a more comprehensive understanding of the underlying linkage disequilibrium (LD) patterns within and across populations. This information is required to optimally design the next generation of whole-genome genotyping (WGA) arrays. Until recently the primary source of information for developing WGA arrays was the International HapMap Project as it provided detailed information on frequencies and LD between almost four million SNPs in three distinct populations. Selecting an optimal subset of these SNPs allows the creation of arrays that interrogate over 90% of all common HapMap variants either directly or indirectly through LD with only 600-800k tagSNPs. However, the catalog of SNPs available through HapMap represents only a small, and potentially biased, subset of the total variation in the human population. With the advent of high-throughput sequencing, the 1000 Genomes Project has greatly increased the spectrum of known variants and provides an excellent resource for content to develop the next generation of microarrays for "rich" genome-wide association studies (GWAS). These microarrays will interrogate the entire genome, including rarer content down to ~1% MAF. We have evaluated the whole-genome coverage provided by chips whose content was based on the HapMap data. Using the phased genotype data from the 1000 Genomes Project, we calculate that less than 70% of the common (>5%) variants and less than 60% of all variants seen at least twice are tagged by even the most comprehensive arrays currently available. Using a "greedy" approach to selecting tag SNPs, we show that ~2 million intelligently selected markers can effectively cover the SNPs from the 1000 Genomes Project in all three HapMap populations, excluding singletons. Additionally, because the 1000 Genomes Project is sequencing to ~4x depth, many of the available genotype calls have been imputed and/or are based on only a limited number of reads leading to potential false positives and genotype errors. We will also present the results of validation experiments using hundreds of thousands of SNPs to assess false positive rate and categorize potential errors in these imputed genotype calls.

COPY NUMBER VARIATION AND GENE FAMILY DIVERSITY FROM 151 SEQUENCED HUMAN GENOMES

Peter Sudmant*[1], Jacob Kitzman*[1], Katie Campbell[1], Nick Sampas[2], Anya Tsalenko[2], Maika Malik[1], Francesca Antonacci[1], 1000 Genomes Consortium[1], Jay Shendure[1], Evan Eichler[1]

[1]University of Washington, Genome Sciences, Pacific, Seattle, WA, 98195, [2]Agilent Technologies, A Labs, Park, Palo Alto, CA, 94306

Copy number variants (CNVs) play significant roles in human biology, but existing array platforms struggle to accurately genotype multi-allelic CNVs within segmental duplications. As a result, these regions and the genes therein have generally been regarded as inaccessible. We develop methods to accurately estimate copy using NGS read depth and to assess content using 'singly unique nucleotides' (SUNs) that discriminate the copies. We report here the first large-scale, accurate genotyping of multi-allelic CNVs from the sequence of 151 human genomes. Using FISH, qPCR and aCGH, we show 86-100% concordance in predicting the copy number of segments >1.6kb that range from 0-56 copies. We show that 17% of previous CNV genotypes from SNP microarrays are in error and that 95% lie within these multicopy regions. We detect the most variable and most population stratified genes (n=383) in the human species and show that most of this diversity has not been previously characterized. To genotype specific duplicated segments, we identify 4.6 million SUN positions and show that 91% of these are informative in distinguishing specific copies. Using read depth at SUNs we estimate individual paralogs' copy, even for deeply nested duplications, and show that ~12% of SDs are diploid and nearly copy-invariant across genomes sampled. Conversely, we identify specific paralogs with high copy number variability and population stratification. We detect previously undetectable patterns of structural variation embedded within SDs, including paralog-specific deletions of duplicated genes and also find patterns suggestive of large-patch gene conversion events between tandem SDs at several loci. We demonstrate, for the first time, the ability to assay both copy and content of complex regions of the human genome, opening these regions to disease association studies and further population and evolutionary analysis.
(*) contributed equally

DIRECT ESTIMATION OF THE MICROSATELLITE MUTATION RATE

James X Sun[1,3], Agnar Helgason[2], Gisli Masson[2], Sunna Ebenesersdóttir[2], Nick Patterson[4], Augustine Kong[2], David E Reich[3,4], Kari Stefansson[2]

[1]MIT, Division of HST, 77 Mass Ave, E25-519, Cambridge, MA, 02139, [2]deCODE Genetics, Sturlugata 8, 101 Reykjavik, Iceland, [3]Harvard Medical School, Dept of Genetics, 77 Ave Louis Pasteur, #260, Boston, MA, 02115, [4] Broad Institute, Medical and Population Genetics, 7 Cambridge Center, Cambridge, MA, 02139

Mutations in the germline are the source of genetic variation. While there have been observations of germline mutations in humans at specific disease-causing loci, a large-scale genome-wide direct observation of mutations transmitted from parent to offspring has not been possible because the mutation rate is low.

To provide an accurate estimate of the human mutation rate at a common type of polymorphism, we focus on microsatellites, which are pervasive in the genome and mutate at a much higher rate than single nucleotide substitutions. We estimate the mutation rate using data from over 95,000 Icelanders at up to 2555 loci. We estimated mutation rates using two approaches: (1) using 19,134 trios and (2) using 2,406 extended families. We discovered 2,131 germline mutations in 6 million parent-offspring allele-transfers. Experimental re-genotyping of a subset of the mutations demonstrated a false-positive rate of <10%.

Our results show:
i. The overall mutation rate is 3.5×10^{-4}, which is 3-times lower than previous studies. Tetra-nucleotide loci almost exclusively mutate by 1-repeat unit, while the di-nucleotide loci show a spectrum. Based on the mutation rates and length distributions, we demonstrate that the stepwise-mutation-model is inadequate in estimating absolute coalescent times.
ii. Mutation rate is strongly correlated with microsatellite length.
iii. The paternal-to-maternal mutation rate ratio (α) is 3.1.
iv. The paternal mutation rate is significantly correlated to paternal age, but none on the maternal side. This suggests that the maternal rate may be relatively well described as proportional to the number of generations that elapse, while the male mutation rate may be better described as a quantity proportional to the number of years.

DEEP SEQUENCING ANALYSIS AND CHARACTERZIATION OF TRANSCRIPTIONAL START SITES

Yutaka Suzuki, Riu Yamashita, Kenta Nakai, Sumio Sugano

University of Tokyo, Department of Medical Genome Sciences, 5-1-5 Kashiwanoha, Kashiwa, 277 8562, Japan

Although recent studies have revealed the complete structure of the transcripts for the major part of the human genes, they are mostly virtual transcript models deduced from collective information from hundreds of different cell types. Detailed figure of the transcriptome of a given particular cell type still remains elusive. We recently developed a method to enable deep sequencing analysis of the transcriptional start sites (TSSs) by combining our oligo-capping method and Illumina GA sequencing technology. Briefly, the 5'- and 3'-adaptor sequences necessary for the Illumina GA sequencing were introduced as the 5'-end-oligo at the RNA ligation and as the random hexamer primer at the first strand cDNA synthesis, respectively.
By appling this TSS Seq technology, we characterized 140 million transcriptional start sites (TSSs) in 12 human cell types. Despite the large number of TSS clusters (TSCs), the TSCs with significant expression levels were rare; having strikingly characteristic sequence features and expression patterns from the other minor TSCs. Also, the significantly expressed TSCs had several distinctive biological features in their surrounding regions. Nucleosome-Seq analysis revealed highly-ordered nucleosome structures, and ChIP-Seq analysis detected strong RNA polymerase II binding signals. Evaluations of 16,080 previously sequenced and 846 newly shotgun-sequenced complete cDNA sequences revealed that the preferable transcripts for translation are more frequently associated with those TSCs. Furthermore, RNA Seq analysis of polysome-incorporated RNAs yielded direct evidence that those transcript products are actually used for protein translation. Similar discriminative features were also found between the intergenic TSCs of significant expression levels and those of very low expression levels. We demonstrate that integrative transcriptome analysis provides a powerful tool to discriminate TSCs having clear biological significance from the other possible noise level transcriptions.

POLYMORPHIC LTR RETROTRANSPOSONS CAN TERMINATE TRANSCRIPTS AT A DISTANCE, CAUSING MOUSE LINEAGE VARIATION

Jingfeng Li[1], Keiko Akagi[1], Yongjun Hu[2], Natalia Volfovsky[3], Robert M Stephens[3], David E Smith[2], David E Symer[1]

[1]Ohio State Univ., MVIMG, Columbus, OH, 43221, [2]U. Michigan, Pharm. Sci., Ann Arbor, MI, 48109, [3]NCI-Frederick, ABCC, Frederick, MD, 21702

Retrotransposons comprise nearly half of mammalian genomes. Widespread genomic variation between normal mouse strains has been driven by recent endogenous mobilization of these elements, resulting in tens of thousands of insertional polymorphisms. Here we describe the identification, using next generation sequencing, of previously uncharacterized LTR (long terminal repeat) retrotransposon integrants in various mouse lineages. Integrant loci in divergent, wild mouse lineages such as MOLF/EiJ, CAST/EiJ and SPRET/EiJ are almost entirely different from those in laboratory mouse lines such as the C57BL/6J reference strain, suggesting independent, lineage-specific retrotransposition by ancestral master elements. In addition, we linked altered transcript structures and expression levels to nearby polymorphic LTR retrotransposon integrants in a variety of strains, tissues and developmental stages. The vast majority of genomic LTR polymorphisms appear to have little or no effect on neighboring genes' transcripts. However, a small percentage of polymorphic LTR retrotransposons can initiate fusion transcripts, while others can terminate transcripts prematurely. We characterized expression of *Slc15a2* (also called *PEPT2*), whose structure and function have been altered very strikingly by an intronic, polymorphic LTR retrotransposon integrant that can terminate its transcription at a distance. Substantially increased premature polyadenylation of *Slc15a2* transcripts, triggered by the polymorphic LTR element, leads to dramatically reduced protein expression and significant functional consequences in affected mouse strains. Other candidate genes whose transcription may be terminated similarly by LTR polymorphisms are described. Thus endogenous mouse LTR retrotransposons can contribute to extensive genomic variation, leading in numerous cases to substantial alterations of nearby genes' regulation, structure and function. The resulting phenotypic diversity may provide selective advantages for their hosts.

A NEXT-GENERATION OF METHODS FOR CHARACTERIZING COMPLEX BALANCED REARRANGEMENTS CONTRIBUTING TO DEVELOPMENTAL DISORDERS

Michael E Talkowski[1,2], Bhavana Muddukrishna[1], Carl Ernst[1], Andrew Kirby[1,2], Toshiro Ohsumi[1], Mark Borowsky[1], Mark J Daly[1,2], Cynthia C Morton[1,2], James F Gusella[1,2]

[1]Massachusetts General Hospital/Brigham and Women's Hospital and Harvard Medical School, Center for Human Genetic Research, Departments of Molecular Biology, Genetics, Neurology, Obstetrics and Gynecology, 185 Cambridge St, Boston, MA, 02114, [2]Broad Institute, Program in Medical and Population Genetics, 7 Cambridge Center, Cambridge, MA, 02142

The contribution of balanced genomic rearrangements to common complex disorders remains unclear. Current genotyping methods preclude accurate detection of such changes and these events represent a complementary opportunity for determining causative pathways. We describe a series of experiments utilizing massively parallel paired-end (PE) sequencing to rapidly identify apparently balanced genomic rearrangements contributing to neurodevelopmental phenotypes. At the whole genome level we mapped balanced translocations, inversions, and other complex rearrangements using standard insert PE sequencing to 10X coverage on multiple lanes of an Illumina flow cell and large insert jumping libraries (3-8kb) on a single lane of a flow cell. In ongoing efforts to restrict genomic complexity, we used flow sorting to facilitate complete deep sequencing of derivative and normal chromosomes as well as a novel capture-seq approach to isolate chimeras and split reads in patients that have had rearrangement breakpoints narrowed to a finite region (less than 4 Mb). Our initial pilot experiment involves sequencing captured regions from ten cases pooled together on a single lane of an Illumina flow cell without the need for indexing. Notably, in addition to resolving genomic breakpoints to base pair resolution, our experiments have detected significant complexity at the breakpoint and from independent regions in the vicinity of breakpoints. Taken together, our experiments suggest that next-generation sequencing is a powerful tool to characterize these previously intractable sources of genomic variation and will enable detection of novel targets contributing to autism and other developmental disorders.

HIF-1α CHIP-SEQ ANALYSIS OF CANCER CELL LINE DLD-1

Kousuke Tanimoto[1], Katsuya Tsuchihara[2], Yutaka Suzuki[1], Sumio Sugano[1]

[1]the University of Tokyo, Department of Medical Genome Science, 5-1-5 Kashiwanoha, Kashiwa, 277-8562, Japan, [2]National Cancer Center Hospital East, Research Center for Innovative Oncology, 6-5-1 Kashiwanoha, Kashiwa, 277-8577, Japan

Cancer cells in solid tumors are frequently lacking oxygen (hypoxia) because aberrantly growing cancer cells cause shortage of blood flow. Tumor cells adapt themselves to such hypoxic condition by regulating cellular responses. The central regulator of these responses is the transcription factor hypoxia-inducible factor 1 (HIF-1). In this study, we performed HIF-1α (oxygen-sensitive subunit of HIF-1) ChIP-Seq analysis to obtain the overview of HIF-1 target genes by using human colorectal cancer cell line DLD-1. We collected 4,134,678 sequence tags from the ChIP sample and 6,964,466 tags from the whole cell extract (WCE) sample. We selected binding sites as the regions having more than 60 bp continuous 15-fold enrichment between the ChIP and WCE (whole cell extract). As a result, we identified 458 potential HIF-1α binding sites. Of these, HRE (Hypoxia Responsible Element) motif was observed in 95(21%) cases. In addition, we performed TSS (Transcription Start Site)-Seq analysis under hypoxia and normoxia to obtain the information about exact positions of transcription start sites. The TSS-tags were further clustered to generate TSS clusters (TSCs; Tsuchihara et.al. NAR 2009). In total, we generated 19,213,284 tags and identified 1,428,455 TSCs under hypoxia. Of these, 619 RefSeq gene TSCs were located from 10Kb upstream to 1Kb downstream of the HIF-1α binding sites which were identified by ChIP-Seq. Also, we identified 297 intergenic TSCs which were located within the same distance from the identified HIF-1α bindig sites. Combination of the ChIP-Seq and TSS-Seq analyses should be a powerful approach to identify HIF-1α binding sites.

EVALUATING THE EFFICACY OF CROSS-SPECIES MICROARRAY-BASED GENOMIC CAPTURE AND ITS APPLICATION TO TARGETED SEQUENCING IN A NONHUMAN PRIMATE MODEL FOR HIV/AIDS RESEARCH

K Mondal[1], J K Davis[1], V C Patel[1], A C Shetty[1], Z P Johnson[2], G Silvestri[3], M E Zwick[1], James W Thomas[1]

[1]Emory University School of Medicine, Human Genetics, 615 Michael St, Atlanta, GA, 30322, [2]Emory University, Yerkes National Primate Research Center, 954 North Gatewood Road, N.E., Atlanta, GA, 30322, [3]University of Pennsylvania School of Medicine, Pathology and Laboratory Medicine, 3400 Spruce St, Philadelphia, PA, 19104

Comparative studies in nonhuman primates provide a unique and powerful perspective on human biology, disease and evolution. However, most primates lack the basic genomic sequence resources that are required for modern genetic and comparative genomic studies. The goal of our research project is to test efficacy of cross-species microarray-based genomic selection (MGS) and to then apply this methodology toward targeted population-based sequencing in a nonhuman primate model of HIV/AIDS. To evaluate the general cross-species capability of MGS, we are testing the ability of a human MGS array to capture ~500 -1000 kb segments near the *CFTR* locus in a 4 nonhuman primates that represent a range of sequence divergence levels of 1-10% from human. Preliminary studies completed to date have shown that the human probes can effectively enrich for the orthologous target regions in chimpanzees and orangutans. Enrichment of the target regions in the more divergent rhesus and marmoset are, however, greatly reduced. The results of ongoing detailed studies correlating the probe-target species divergence levels with along with additional experiments quantifying the efficiency of cross-species MGS will be presented. An overview of the second phase of our project which will involve the sequencing ~100 genes that are candidates for the benign versus pathogenic nature of SIV infection in two closely related Old World monkeys, the sooty mangabey and the rhesus macaque, will also be presented. In summary, we are investigating the utility of MGS as a method for simultaneously producing comparative and population-based sequence data from targeted regions in multiple species, and are applying this technology toward the identification of genetic factors that suppress the progression to AIDS in a natural nonhuman primate host of SIV.

GLOBAL ANALYSIS OF RNA AND PROTEIN CHANGES IN RESPONSE TO OSMOTIC STRESS

Scott E Topper[1], M. Violet Lee[2], Joshua J Coon[2], Audrey P Gasch[1]

[1]University of Wisconsin-Madison, Department of Genetics, Madison, WI, 53706, [2]University of Wisconsin-Madison, Department of Chemistry, Madison, WI, 53706

Yeast respond to significant changes in their environment by activating massive alterations in their transcriptome. While different stresses stimulate transcriptional profiles that are unique in specific details, the broad strokes of the transcriptional responses are strikingly similar: about 300 genes (involved in all manner of cell defense and physiology) are reliably induced, and about 600 genes (mostly involved in RNA and protein synthesis) are reliably repressed. While this transcriptional response is very well characterized, the consequences for the proteome are still poorly understood.

To improve our ability to appropriately interpret transcriptional data, this project simultaneously characterizes global changes in RNA and protein of a single population of *Saccharomyces cereviciae*, in response to osmotic stress (NaCl). Using time-course microarrays and quantitative, time-course proteomics, this experiment describes the coordinated temporal dynamics of both types of molecules. The data identifies different relationships between RNA and protein levels for subsets of genes, identifies candidate genes for phenomena such as acquired stress resistance and cellular memory, and provides evidence for a group of genes that are primarily regulated post-transcriptionally.

CUFFLINKS: TRANSCRIPT ASSEMBLY, ABUNDANCE ESTIMATION, AND DIFFERENTIAL EXPRESSION WITH RNA-SEQ

Cole Trapnell[1,2], Brian A Williams[3], Geo Pertea[1], Ali Mortazavi[3], Gordon Kwan[3], Marijke J van Baren[4], Steven L Salzberg[1], Barbara J Wold[3], Lior Pachter[2]

[1]University of Maryland, College Park, Center for Bioinformatics and Computational Biology, 3115 Biomolecular Sciences Building #296, College Park, MD, 20742, [2]University of California, Berkeley, Department of Mathematics, Evans Hall, Berkeley, CA, 94720, [3]California Institute of Technology, Division of Biology, 130 Kerckhoff, Pasadena, CA, 91125, [4] Genome Sciences Center, Washington University, One Brookings Drive, St. Louis, WI, 63130

High-throughput transcriptome sequencing has the potential to report the sequence and abundance of every RNA in a cell or tissue. This goal has not been fully realized due to computational and experimental challenges. We introduce a new transcript assembly algorithm coupled with a statistical model for RNA-Seq experiments that produces estimates of transcript abundances. Cufflinks constructs a parsimonious set of transcripts that "explain" the reads observed in an RNA-Seq experiment. Cufflinks implements a constructive proof of Dilworth's Theorem by constructing a covering relation on the read alignments, and finding a minimum path cover on the directed acyclic graph for the relation. Cufflinks uses a statistical model of paired-end sequencing experiments to compute a unique maximum likelihood set of abundances for transcripts in a sample. The package includes a tool, "Cuffdiff", which tests for statistically significant differential expression of genes and individual transcripts. Cuffdiff also tests for changes in relative abundance of transcripts within biologically relevant groups, such as those transcripts that share a common start site (TSS). By tracking relative changes within TSS groups, Cuffdiff can infer not only changes in expression, but changes in transcriptional and post-transcriptional regulation. In a timeseries RNA-Seq experiment on differentiating myoblasts, Cufflinks assembled 13,689 and 3,724 known and novel transcripts, respectively, and detected differential splicing or promoter use in 330 genes (FDR < 5%). Cufflinks and its companion tools are available open source at http://cufflinks.cbcb.umd.edu

NEXT GENERATION WHOLE EXOME SEQUENCING IN FAMILIAL CANCER.

Lisa R Trevino[1], David A Wheeler[1], Kyle Chang[1], Donna M Muzny[1], Jeffrey G Reid[1], Richard A Gibbs[1,2], Sharon E Plon[1,2,3,4]

[1]Baylor College of Medicine, Human Genome Sequencing Center, 1 Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, Department of Molecular and Human Genetics, 1 Baylor Plaza, Houston, TX, 77030, [3]Baylor College of Medicine, Department of Pediatrics, 1 Baylor Plaza, Houston, TX, 77030, [4] Texas Children's Hospital, Cancer Center, 6621 Fannin St, Houston, TX, 77030

We are investigating genetic variation contributing to childhood cancer by sequencing probands with a family history suggestive of susceptibility to cancer. In each case the proband had cancer diagnosed in childhood and met one or more of these criteria, i)a second malignancy before the age of 30, ii)multiple members of a single family with a diagnosis of cancer before the age of 30,and iii)a developmental anomaly or significant developmental delay. An initial study using Sanger sequencing of 45 candidate cancer associated genes in genomic DNA from 48 families led to the discovery of pathogenic mutations in 14% of probands. To broaden the study,we are sequencing with whole exome capture and next generation sequencing methodologies to identify novel, cancer susceptibility loci. Our first kindred to undergo this analysis is a family predisposed to childhood lymphocytic leukemia/lymphoma with 4 affected individuals. The pattern of inheritance suggests autosomal dominant inheritance with incomplete penetrance. Genomic DNA was extracted from normal tissue of 3 affected and 1 non-transmitting parent with subsequent sequencing of exons from the CCDS transcript sets (17,000 genes comprising 23 Mb of target). In total we obtained greater than 10X coverage of 80% of all targets via the SOLiD system of next-generation sequencing. The intersect of all three affected individuals results in 11,906 mutation sites,8% of which are novel sites. Of the novel variant sites,66% (n=629) were splice-site,missense or nonsense. We eliminated variants identified in the non-transmitting parent, reducing to 180 the novel non-silent variants in affected probands. A systematic bioinformatic analysis is determining the severity of missense changes. We are validating these novel variants from this first level analysis in other existing families with lymphoid malignancies.

COMPARING GENOMIC SEQUENCE OF SELECT LARGE STRETCHES OF INBRED RAT STRAINS USING THREE DIFFERENT SEQUENCING PLATFORMS IN TANDEM

Michael Tschannen[1], Elizabeth Worthey[1], Kathrin Saar[2], Marek Tutaj[1], Oliver Hummel[2], Giannino Patone[2], Wei Chen[2], Howard Jacob[1], Norbert Hubner[2]

[1]Medical College of Wisconsin, Human and Molecular Genetics Center, 8701 Watertown Plank Rd, Milwaukee, WI, 53226, [2]Max-Delbruck-Center, Molecular Medicine, Robert-Rössle-Str. 10, Berlin, D13092, Germany

The most commonly used inbred laboratory rat strains have originated from limited founder populations which have been exclusively derived from Rattus norvegicus and the inherited genetic variation plays a crucial role in correlating genotype to phenotype. Here we report the re-sequencing of two commonly used inbred rat strains using paired-end (PE) technology on both Illumina and AB SOLiD platforms: FHH/Mcwi and SS/JrMcwi rats represent the parental founder strains of the consomic rat panel. These consomics serve as renewable animal resources and provide a valuable tool to generate and validate mapping data. For the PE sequencing on the Illumina platform, we used 2 different libraries with insert sizes of 150 bp and 1750 bp, respectively. For mate-pair sequencing on the SOLiD platform, we used a library of approx. 1600bp insert size. Both strains were sequenced to a minimum of 10-fold coverage. For subsequent analyses of the Illumina raw data we used the Illumina-Software-Package Pipeline 1.4.0, CASAVA and GenomeStudio 2009. Overall we produced approx. 60Gb post filtering with the Illumina platform. We also generated 1.3Gb of genomic sequence from 200 BACs for the SS, FHH and BN (reference strain originally sequenced) by capillary and 454 sequencing for select QTL regions.

We compare the sequencing results with respect to alignment, assembly and accuracy for the different sequencing platforms in the QTLs of interest. We will also report SNPs, structural variants and copy number variants between the three rat strains. For example, in the FHH/Mcwi rat we identified some 2.6 million SNPs compared to the BN reference sequence, of which 899,000 are novel. Our data will guide the identification of the molecular variants underlying quantitative physiological traits that have been characterized in the FHH/SS/BN consomic panel and provide insight into how best to generate and assemble genomic regions of interest.

THE TRANSCRIPTOMES OF TWO HERITABLE CELL TYPES HELP ILLUMINATE THE CIRCUIT GOVERNING THEIR DIFFERENTIATION

Brian B Tuch[1,2], Quinn M Mitrovich[1], Oliver R Homann[1], Aaron D Hernday[1], Francisco M De La Vega[2], Alexander D Johnson[1,3]

[1]University of California, San Francisco, Dept. of Microbiology and Immunology, 600 16th St., San Francisco, CA, 94143, [2]Life Technologies, Research and Development, 850 Lincoln Center Dr., Foster City, CA, 94404, [3]University of California, San Francisco, Dept. of Biochemistry and Biophysics, 600 16th St., San Francisco, CA, 94143

The differentiation of cells into distinct cell types is frequently driven by epigenetic mechanisms governed by regulators of transcription. White and opaque cells of the fungal pathogen *Candida albicans* are two such heritable cell types, each thought to be adapted to unique niches within their human host. While the transcriptional circuit of a master regulator of the switch between white and opaque cell types, White opaque regulator 1 (Wor1), was previously described, several mysteries about this circuit remain. For example, it was unclear why so few genes bound by Wor1 are differentially expressed between the two cell types. Here we performed strand-specific massively-parallel ligation sequencing of RNA from *C. albicans* white and opaque cells. Combining the resulting data from both cell types, we first substantially re-annotated the *C. albicans* transcriptome, finding 1422 novel coding and non-coding transcriptionally active regions (nTARs). We then compared the new annotation to genomic regions bound by Wor1, finding that the revised transcriptional landscape substantially alters our understanding of the circuit controlling white-opaque switching. Helping to explain the improved concordance between binding and differential expression, we find that 36% of bound and differentially expressed transcripts were previously undiscovered, including transcripts antisense to ORF-encoding transcripts and others that are entirely isolated. Interestingly, 17 of the novel opaque-specific transcripts cluster in three genomic locations and encode short ORFs with homologs found only in the two species known to undergo white-opaque switching. Taken together, the results presented here greatly enhance our understanding of the circuit that differentiates these two cell types, a circuit bearing many similarities to the circuit that specifies the pluripotency of embryonic stem cells.

CONSTRUCTION OF A REAL-TIME DISEASE WEATHER MAP.

Stephen W Turner, Eric E Schadt

Pacific Biosciences, 1505 Adams Dr., Menlo Park, CA, 94025

Continued emergence of highly pathogenic viral agents (e.g., West Nile Virus) that infect large numbers of individuals across broad geographic regions has driven the need to discover and more fully characterize the genomes of viruses present in the environment over time. While different viruses fluxing through the environment vary with respect to overall rates of infection and the severity of health and financial complications they induce, understanding the real-time geographic distributions of these pathogens would lead to significant public health benefit by enabling government agencies to better discover and contain infectious disease, providing advanced notification to at-risk populations; and providing advanced warning to healthcare providers and drug manufacturers of the increased risk so they may mobilize appropriate resources. With the advent of second generation sequencing technologies, it is possible to sequence DNA and RNA isolated from environmental swabs taken in areas at high risk of infection to identify the pathogens responsible. However, increasing cost-per-run and time-to-result create challenges for these applications. Unlike the major second generation sequence by synthesis technologies, single molecule real time (SMRT™) sequencing exploits the high catalytic rates and processivity of DNA polymerase to radically increase the rate of synthesis (1–3 bases/second) and read length (from tens to thousands of bases/read). It is now possible to not only rapidly sequence the genomes of many viruses simultaneously at high fold coverage to identify the entire complement of viruses present in a DNA sample, but also to assess variants in the genomes found. SMRT Sequencing technology now makes it possible to sequence pathogens of public health concern from environmental samples at high coverage for under $99 in 15 minutes. We demonstrate the feasibility of SMRT sequencing for use in widespread disease monitoring through a pilot study. We collected and sequenced environmental swabs from a selection of sources including airports, rapid transit systems, fitness centers, and sewage treatment plants over a period of time. Samples were enriched for likely pathogens by amplification with primers selected from conserved regions of viral pathogens. The results and analysis will be presented with emphasis on the feasibility of a disease weather map based on molecular analysis of environmental exposures to disease. Please wash your hands before attending this talk.

ORIGINS AND EVOLUTION OF SULFADOXINE RESISTANCE IN HUMAN MALARIA PARASITE, *PLASMODIUM FALCIPARUM*

Sumiti Vinayak[1,2], Md Tauqeer Alam[1], Kanungnit Congpuong[3], Chansuda Wongsrichanalai[4], Laurence Slutsker[1], Ananias A Escalante[5], John W Barnwell[1], Venkatachalam Udhayakumar[1]

[1]Centers for Disease Control and Prevention, Malaria Branch, Division of Parasitic Diseases, 4770 Buford Highway, Atlanta, GA, 30341, [2]Atlanta Research and Education Foundation, VA Medical Center, 1670 Clairmont Road, Decatur, GA, 30033, [3]Ministry of Public Health, Laboratory Reference Center, Bureau of Vector Borne Diseases, Nonthaburi, 11000, Thailand, [4] Independent Scholar, 130, Sub Street, Bangkok, 10500, Thailand, [5]Arizona State University, School of Life Sciences, PO Box 874501, Tempe, AZ, 85287

Gene flow has played an important role in the intercontinental spread of drug resistant *Plasmodium falciparum* malaria. Using microsatellite data flanking *pfcrt* (determinant for chloroquine resistance) and *dhfr* (determinant for pyrimethamine resistance), it has been shown that resistance to chloroquine and pyrimethamine originated at least four times globally, with one origin in the Thailand-Cambodia region of Southeast Asia. It was further shown that chloroquine and pyrimethamine resistant alleles spread from Southeast Asia to Africa. However, the global origins and evolution of sulfadoxine resistance in *P. falciparum* are not known. In this study, we characterized *dhps* genotypes (determinant for sulfadoxine resistance) and its flanking microsatellites in a large number of singly-infected *P. falciparum* isolates from Thailand-Cambodia region to: (i) reveal the origin(s) of sulfadoxine resistance; and (ii) acquire evidence of selection operating on this gene. We also wanted to know if sulfadoxine-resistant alleles in Southeast Asia have any relationship with African and South American *dhps* alleles. Our results show that resistant *dhps* alleles, especially those with 3 mutations, are fixed in the Thailand-Cambodia region, with strong evidence of selection, consistent with the extensive use of sulfadoxine-pyrimethamine in the past in this region. The most common resistant alleles were AGEAA, SGEGA, and SGNGA (amino acids at codons 436, 437, 540, 581 and 613; mutated codons are underlined). We provide evidence for at least two independent origins for the triple mutants in Thailand-Cambodia region, one for both, SGEGA/SGNGA and another for AGEAA, with neither sharing any evolutionary relationships with the alleles prevalent in Africa and South America. Thus, unlike CQ and pyrimethamine-resistance, sulfadoxine-resistance in Southeast Asia has more than one origin and, there is no evidence of intercontinental spread of sulfadoxine-resistant *dhps* alleles from Southeast Asia to Africa.

STRAND-SPECIFIC RNA SEQUENCING OF HEPG2 CELLS IDENTIFIES GENES THAT ARE DIFFERENTIALLY EXPRESSED, ALTERNATIVELY SPLICED AND ALLELICALLY IMBALANCED IN RESPONSE TO TGF-BETA

Stefan Enroth[1], Ola Wallerman[2], Brian Tuch[3], Catalin Barbacioru[3], Madhu Bysani[2], Robin Andersson[1], Stefan Thermén[4], Aristidis Moustakas[4], Carl-Henrik Heldin[4], Niclas Eriksson[5], Sarah Stanley[3], Jian Gu[6], Scott Kuersten[6], Melissa Barker[3], Jan Komorowski[1,7], Kevin McKernan[8], Francisco M De La Vega[3], Claes Wadelius[2]

[1]Uppsala University, Linnaeus Centre for Bioinformatics, Uppsala, SE-75185, Sweden, [2]Uppsala University, Dept of Genetics and Pathology, Uppsala, SE-75185, Sweden, [3]Life Technologies, Foster City, CA, 94404, [4]Ludwig Institute for Cancer Research, Uppsala, SE-75185, Sweden, [5]Uppsala University, Uppsala Clinical Research Center, Uppsala, SE-75185, Sweden, [6]Life Technologies, Austin, TX, 78712, [7]Warsaw University, Interdisciplinary Centre for Mathematical and Computer Modeling, Warsaw, 00-927, Poland, [8]Life Technologies, Beverly, MA, 01915

Transforming growth factor-beta (TGF-beta) controls many complex behaviors of normal and transformed cells. In early stage adenomas TGF-beta acts as a tumor suppressor, whereas in advanced carcinomas it promotes tumor cell invasiveness and metastasis. HepG2 cells provide a good model for advanced tumor cells. RNA sequencing (RNA-Seq) can provide a detailed view of the dynamic transcriptional landscape of cells experiencing environmental perturbation and was therefore applied here to study TGF-beta response. We serum starved HepG2 cells for 24 hours and then treated for 1 hour with 2 ng/ml TGF-beta1 or with vehicle control. Total RNA was isolated, depleted of rRNA, fragmented and then used to create strand-specific cDNA libraries with the SOLiD™ Small RNA Expression Kit. Approximately one billion 50 bp reads were sequenced with the SOLiD™ System, over half of which could be uniquely aligned to the human genome. The aligned transcriptomes cover roughly 4% of each strand of the genome and more than half of the reads align within known transcribed regions. The expression level of each gene was defined and hundreds of genes were determined to be differentially expressed between the two conditions. The 50 bp length of the sequenced fragments permitted us to accurately align reads to splice junctions. In all, ~5 million reads aligned to ~90,000 known and putative splice junctions in both the control and stimulated conditions, which allowed us to determine the effects of TGF-beta on alternative splicing in fine detail. By genotyping we identified over 300,000 heterozygous SNP positions, which we employed to examine allele-specific gene expression. We observed many differences in the balance of expressed alleles between TGF-beta treated and untreated cells, which could be explained by linked SNPs in cis-regulatory elements. We are now comparing these results to ChIP-Seq data gathered under the same conditions to understand the relationship between allele-specific binding of transcriptional regulators and allele-specific expression of nearby genes. Furthermore, we have determined positions of nucleosome and critical histone modifications to see how they are related to the transcriptional output.

GENOTYPING STRUCTURAL VARIANTS FROM NEW SEQUENCING TECHNOLOGY DATA

Klaudia Walter[1], Lorenz Wernisch[2], Le Si Quang[1], Richard Durbin[1], Matthew E Hurles[1], and the Structural Variation Group of the 1000 Genomes Consortium[1]

[1]Sanger Institute, Wellcome Trust, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom, [2]MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 0SR, United Kingdom

Analyses of sequencing data enable the discovery of a greater diversity of structural variants (SV) than is possible on microarrays, including novel insertions and inversions, as well as deletions and duplications. The presence of an SV can reveal itself through one or more of the following signatures: anomalous spacing and/or orientation of read-pairs, discrepant read-depth and split-reads. With sufficient sequencing coverage, the precise breakpoints of the SV can be reliably assembled. However, in order to assess the functional impact of SVs using genetics, it is necessary not only to discover SVs but also to genotype individuals for those SVs. The most efficient design is to spread the available sequencing resources across as many individuals as possible while generating sufficient data in each individual to allow genotypes to be robustly estimated.

We have developed a probabilistic genotyping model that makes efficient use of information in the data by integrating read-pair, read-depth and split-read information to estimate genotype likelihoods for SVs. This approach is, in principle, applicable to all types of structural variation. These likelihoods can be further refined by incorporating SNP genotype information on the same individuals and imputing haplotypes. To demonstrate our approach we have focussed on genotyping large deletions (> 50bp) from the pilot phase of the 1000 Genomes Project where 180 samples were sequenced at low-coverage (2-4X). We compared the resultant SV genotypes to published array-based genotypes for the same variants in the same samples and show high concordance indicating these genotypes are of high quality. Our genotyping approach, in combination with SV discovery and breakpoint assembly, promises to enable genetic study of structural variation across the full range of sizes and classes of SVs.

EXPLORING THE DIGITAL "TREE OF LIFE" BY DECODING THE GENOMES OF 1000 PLANTS AND ANIMALS

Xun Xu, Jun Wang

BGI, Shenzhen, Main Building, Beishan Industrial Zone,Yantian, Shenzhen, 518083, China

A determined genome sequence is the basis that will facilitate studies in the species and related creatures. In addition, comparison in between genomes could reveal interesting relationship and biological stories, which help understanding the evolution of life. Previous studies are generally restricted by funding or technical feasibility, yet recent advances in sequencing technology and bioinformatics analytical tools have unblocked the difficulties and demonstrated the possibility to decipher new genomes at an affordable cost. With the effort of collaborators, we have initiated a large-scale project that aims to determine the genome sequences of 1,000 plants and animals. Over 20 genomes were already deciphered, with scores in production. Here we report the analysis results of recently completed the genome assemblies of polar bear, penguin as well as soybean and potato that are important for agriculture. By *de novo* assembly of new generation sequencing reads, the assemblies achieved high quality and displayed interesting biological points. The project is rapidly in progress and we believe it will significantly benefit the scientific community as an abundant public repository in future.

INTERPRETATION OF ASSOCIATION SIGNALS AND IDENTIFICATION OF CAUSAL VARIANTS FROM GENOME-WIDE ASSOCIATION STUDIES

Kai Wang[1], Samuel P Dickson[2], Catherine A Stolle[3], Ian D Krantz[4], David B Goldstein[2], Hakon Hakonarson[1,4]

[1]Children's Hospital of Philadelphia, Center for Applied Genomics, 3615 Civic Center Blvd, Philadelphia, PA, 19104, [2]Duke University, Center for Human Genome Variation, 450 Research Drive, Durham, NC, 27708, [3]Children's Hospital of Philadelphia, Department of Pathology and Laboratory Medicine, 3615 Civic Center Blvd, Philadelphia, PA, 19104, [4]University of Pennsylvania, Department of Pediatrics, 3615 Civic Center Blvd, Philadelphia, PA, 19104

Genome-wide association studies (GWAS) have been successful in identifying disease susceptibility loci, but it remains a challenge to pinpoint the causal variants in subsequent fine-mapping studies. A conventional fine-mapping effort starts by sequencing dozens of randomly selected samples at susceptibility loci to discover candidate variants, which are then placed on custom arrays or used in imputation algorithms to find the causal variants. We propose that one or several rare or low-frequency causal variants can hitchhike the same common tag SNP, so causal variants may not be easily unveiled by conventional fine-mapping efforts. Here, we first demonstrate that the true effect size and proportion of variance explained by a collection of rare causal variants can be severely underestimated by a common tag SNP, thereby accounting for some of the "missing heritability" in GWAS. We next describe a case-selection approach based on phasing long-range haplotypes and sequencing cases predicted to harbor causal variants. We compared this approach with conventional strategies on a simulated dataset, and demonstrated its advantages when multiple causal variants are present. We also evaluated this approach in a GWAS on hearing loss, where the most common causal variant has a minor allele frequency (MAF) of 1.3% in the general population and 8.2% in 329 cases. With our case-selection approach, it is present in 88% of the 32 selected cases (MAF=66%), so sequencing a subset of these cases can readily reveal the causal allele. Our results suggest that thinking beyond common variants is essential to interpret GWAS signals and identify causal variants

RAP: RNA-SEQ DATA ANALYSIS PACKAGE

Liguo Wang, Yuanxin Xi, Wei Li

Division of Biostatistics, Dan L. Duncan Cancer Center, Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

Whole transcriptome shotgun sequencing (RNA-seq) provides massive and invaluable information about the functional elements transcribed from the genome, such as (allele specific) gene expression, alternative splicing, novel transcribed regions, aberrant transcripts such as gene fusions, etc. However, one of the biggest challenges is to bridge the knowledge gap between our ability to generate these massive RNA-seq data and to interpret them. For instance, people have to deal with long transcript bias, large variation between replicates when profiling gene expression, and people also have to consider gene families, pseudo genes, and other paralog elements when using pair-end RNA-seq data to detect gene fusions. Here to overcome these obstacles and make use of the “gold mine” of RNA-seq data, we developed a RNA-seq analysis package (RAP). RAP is implemented in Python and Perl; it also incorporates several public available tools including samTools, BedTools and various R packages.

IDENTIFICATION OF RARE DNA VARIANTS IN MITOCHONDRIAL DISORDERS WITH IMPROVED ARRAY-BASED SEQUENCING

Wenyi Wang[1,2], Peidong Shen[1], Sreedevi Thyagarajian[1], Curtis Palm[1], Rita Horvath[3], Thomas Klopstock[4], Lynn Pique[5], Iris Schrijver[5,6], Ronald W Davis[1], Michael Mindrinos[1], Terence P Speed[2], Curt Scharfe[1]

[1]Stanford U, SGTC, 855 California Ave, Palo Alto, CA, 94304, [2]UC Berkeley, Statistics, 367 Evans, Berkeley, CA, 94720, [3]Newcastle U, Inst for Aging and Health, Framlington Place, Newcastle upon Tyne, NE24HH, United Kingdom, [4] Ludwig Maximilians U, Neurology, Bavariaring 19, Munich, 80539, Germany, [5]Stanford U, Pathology, 300 Pasteur Dr, Stanford, CA, 94305, [6]Stanford U, Pediatrics, 300 Pasteur Dr, Stanford, CA, 94305

Rare functional variants in nuclear genes can lead to secondary mtDNA defects that present with clinically similar or different disease phenotypes. Here we analyzed thirty-nine candidate genes for disorders of mtDNA maintenance in forty medical cases and controls (4.3Mb) using a customized Affymetrix resequencing array. To identify DNA variants in each case with a minimal number of false discoveries, we performed single and multi-array data analysis that accounts for technical variables and the possibility of both low and high frequency genomic variation. Our Sequence Robust Multi-Array analysis (SRMA) in comparison to traditional Sanger sequencing achieved a false discovery rate of 2% (false negative rate 5%). In addition to mutations confirmed in positive control cases in the mitochondrial DNA polymerase gamma (POLG), we detected novel functional variants in multiple cases in the mitochondrial topoisomerase 1 gene (TOP1MT), the human homolog of E.coli mutY mismatch repair gene (MUTYH), and the apurinic-apyrimidinic endonuclease 2 (APEX2). Several cases carried rare heterozygous variants in multiple genes, which might indicate synergistic genetic effects in similar diseases. In summary, our optimized statistical methods increase the accuracy of identifying rare DNA variants and reduce the resequencing costs. This is particularly useful when studying sporadic cases and small families in which linkage cannot be established.

DYSREGULATION OF GENE EXPRESSION AND ALLELIC IMBALANCE IN MAMMALIAN INTERSPECIFIC HYBRIDS

Xu Wang, Don Miller, Doug Antczak, Andrew Clark

Cornell University, Ithaca, NY, 14853

Interspecific hybrids provide an excellent opportunity to study the effects of a profound but well-defined and replicable genome-wide perturbation. The successful function of interspecific hybrids is a testament to the robustness of many biological processes, and is surprising given the profound dysregulation at the transcriptional level (at least in Drosophila). Here we focus on classical mammalian interspecific hybrids resulting from reciprocal crosses between horse and donkey. The mule is the product of a male donkey mated to a female horse, while a hinny is produced by the mating of a stallion horse with a female donkey. The wide disparities in function of mules and hinnies despite their identical genomes poses an interesting challenge for functional genomics.

To obtain a quantitative genome-wide picture of the degree of dysregulation in these mammal hybrids, we did RNA-seq of invasive trophoblast cells (chorionic girdle) from horse, donkey and their reciprocal F1 hybrid progeny using Illumina technology. This tissue was chosen because it represents a focal point of interaction between the mother and fetus in the developing placenta. A total of 11.37 Gb of sequences were obtained from the four types of animal, and 70% of them were uniquely mapped to the horse RefSeq database, and they cover more than 12000 RefSeq genes in the two parents and reciprocal F1 hybrids. Total expression level for every gene in the trophoblast transcriptome was quantified by the normalized number of mapped reads. The expression profile of both mule and hinny placenta strongly resemble the horse (correlation coefficient 0.906 and 0.967) more than the donkey (0.810 and 0.857). Whole genome Agilent expression microarray confirmed this finding. In addition, we found many instances of genes whose expression level in mule and hinny was outside the range of the parental horse and donkey.

With the novel SNPs we discovered between horse and donkey, we have been able for the first time to quantify the allelic imbalance in these F1 hybrid mammals. Our data provide a detailed quantitative assessment of the dysregulation of gene expression in interspecific hybrids, and identify many instances of cis-acting SNPs that drive the species-specific imbalance. Whole biological pathways generally express alleles from both species, but significant skews in horse vs. donkey allelic expression are rampant.

RECURRING HUMAN LEUKEMIA MUTATIONS DISCOVERED BY SEQUENCING A MOUSE ACUTE PROMYELOCYTIC LEUKEMIA (APL) GENOME

Lukas D. Wartman, MD[1], David E. Larson, PhD[2], Li Ding, PhD[2], Ken Chen, PhD[2], Zhifu Xiang, MD, PhD[1], John S. Welch, MD, PhD[1], Patrick Cahan, PhD[1], Jacqueline E Payton, MD, PhD[3], Michael D. McLellan, BS[2], Heather Schmidt, BS[2], Ling Lin, MS[2], Robert S. Fulton, MS[2], Rachel M. Abbott, BS[2], Lisa Cook, AA[2], Sean D. McGrath, MS[2], Xian Fan, MS[2], Adam F. Dukes, BA[2], Tamara L. Lamprecht, BS[1], Michael H. Tomasson, MD[1], Elaine R. Mardis, PhD[2], Richard K. Wilson, PhD[2] and Timothy J. Ley, MD[1]
[1]Depts. of Internal Medicine and Genetics, Div. of Oncology, Stem Cell Biology Section, Washington University School of Medicine, Siteman Cancer Center, St. Louis, MO, [2] Genome Center, Washington University School of Medicine, St. Louis, MO, [3]Dept. of Pathology and Immunology, Washington University, St Louis, MO.

Acute promyelocytic leukemia (APL, FAB M3 AML) is a subtype of AML characterized by the t(15;17)(q22;q11.2) translocation that creates a fusion oncogene, PML-RARA. We previously targeted a human PML-RARA cDNA to the 5' untranslated region of the mouse cathepsin G gene (mCG-PR) in a 129/SvJ derived ES line. F1 129/SvJ x C57Bl/6 mice were subsequently backcrossed onto the B6/Taconic background for 10 generations. About 60% of the mCG-PR mice in the B6 background develop a disease that closely resembles APL after a latent period of 7-18 months, suggesting that additional progression mutations are required for APL development. To identify these progression events, we sequenced a cytogenetically normal mouse APL genome on the Illumina platform. We created 2 Illumina paired-end libraries (insert sizes of 300-350 bp and 550-600 bp) and generated 59.64 billion base pairs of sequence with 3 full sequencing runs; mapped reads generated 15.6x haploid coverage. We detected 87,778 heterozygous Single Nucleotide Variants (SNVs) and 23,439 homozygous SNVs compared to the mouse C57Bl6/J reference sequence. Of the predicted heterozygous SNVs, 695 were non-synonymous (missense, nonsense, or altering a canonical splice site). Next, the 129/SvJ genome was sequenced using the same approach (13.997x haploid coverage). After filtering out the SNVs present in the 129/SvJ genome, 392 putative somatic SNVs remained. 361 were not analyzed further because they occurred in contiguous blocks, suggesting that they were inherited. Of the remaining 31, 23 were false positives. 9 were tumor-specific, and by deep readcount analysis, present in virtually all tumor cells. Six of the 9 mutations were non-synonymous, and were screened for recurrence in 89 additional murine APL tumor samples. Validated mutations in the Jarid2 (L915I) and Capns2 (N149S) genes occurred only in the proband. Four mutations were found in additional samples; 3 of these were derived from a common ancestor, and are therefore inherited, not somatic. The other recurring mutation was in the pseudokinase domain of JAK1 (V657F), and was identified in 6 additional mice that were not closely related to the proband. This mutation is orthologous to the known activating mutation V617F in human JAK2, *and is identical to a recently described JAK1 pseudokinase domain mutation (V658F) found in human APL* and T-ALL samples (EG Jeong et al, Clin Can Res 14: 3716, 2008). We proved that this mutation cooperates with PML-RARA by expressing human *Jak1* cDNAs (via retroviral transduction) in the bone marrow cells of young WT vs. mCG-PR mice. 9/10 mice transplanted with mCG-PR bone marrow transduced with the V658F JAK1 retrovirus developed a fatal myeloproliferative disease with a mean latency of 34 days (range 28-52 days), whereas the JAK1 WT or control viruses did not cause a phenotype; the mutant JAK1 virus did not cause disease in WT marrow cells. Finally, structural variation analysis of the genome from this cytogenetically-normal tumor revealed a somatic 150kb deletion involving the *Kdm6A* gene. *Kdm6A/Utx*, a histone H3K27 demethylase, has recently been shown to be deleted in 2 human AML cell lines (G van Haaften et al, Nat Gen 41: 5, 2009). 3/14 addition mouse APL samples screened also had somatic deletions of *Kdm6A/Utx* that reduced its expression. In summary, unbiased whole genome sequencing of a mouse APL genome has identified recurring mutations in *Jak1* and *Kdm6A/Utx* in both human and mouse AML samples.

FUNCTIONAL CONSEQUENCES OF BIDIRECTIONAL PROMOTERS

Wu Wei, Zhenyu Xu, Julien Gagneur, Lars Steinmetz

EMBL Heidelberg, Genome Biology Unit, Meyerhofstraße 1, Heidelberg, 69117, Germany

Genome-wide pervasive transcription has been reported in many eukaryotic organisms, revealing a highly interleaved transcriptome organization that involves hundreds of non-coding RNAs. We have provided a complete transcriptome architecture including SUTs (Stable Unannotated Transcripts) and CUTs (Cryptic Unstable Transcripts) in yeast by profiling the transcriptome in multiple conditions, a mutant of the exosome machinery and different strain backgrounds. We have shown that most of non-coding RNAs (SUTs and CUTs) initiate from nucleosome-free regions (NFRs) associated with the promoters of other transcripts. We are now investigating the functional consequences of bidirectional promoters. When a non-coding RNA shares a bidirectional promoter with an ORF transcript and overlaps another ORF on the opposite strand, the possibility exists for spreading of regulatory signals from one ORF to the next. Our data provides preliminary evidence to suggest that non-coding RNAs may provide a mechanism for local spreading of regulatory signals.

A TRANSCRIPTOME-WIDE SURVEY OF PARENT-OF-ORIGIN EFFECTS IN HUMAN CELL LINES

Jens R Wendland[1], Johannes Schumacher[1,2], Bertram Muller-Myhsok[3], Francis J McMahon[1]

[1]National Institute of Mental Health, Mood and Anxiety Disorders Unit, 35 Convent Dr., Bethesda, MD, 20892, [2]University of Bonn, Institute of Human Genetics, Sigmund-Freud-Str. 25, Bonn, 53127, Germany, [3]Max Planck Institute of Psychiatry, Statistical Genetics Research Group, Kraepelinstr. 2, Munich, 80804, Germany

Parent-of-origin effects (POE) refer to a genetic phenomenon where maternally and paternally-derived alleles differentially contribute to phenotypic variance. While the etiologic relevance for POE to several human disorders is well-established, less is known about POE in the context of systematic gene expression regulation. To this end, we conducted a transcriptome-wide survey of POE in human cell lines.

We analyzed 865 transcripts and 968 autosomal linkage markers in lymphoblastoid cell lines derived from 15 CEPH families. Using MERLIN, we calculated identity-by-descent scores separately for maternal and paternal alleles and regressed these on transcript levels. This POE linkage test requires the simultaneous presence of imprinting effects and allele-specific effects on gene expression levels. We used a logarithm of the odds score of 3.3 (two-sided $p<9.8e-5$) as threshold for genome-wide statistical significance per transcript. At this level, we expect 5% (N=43) of our 865 transcripts to reach this threshold by chance alone.

We observed N=151 transcripts whose expression levels were significantly linked to at least one maternally-derived locus, and N=213 transcripts linked to at least one paternal locus. In contrast, when using a joint model not accounting for POE, N=158 transcripts reached genome-wide significance. Thus, POE are common and approximately equally distributed between maternal and paternal inheritance models. Two transcripts showed particularly strong linkage to maternal loci: *CDK12* on 17q12 (best marker rs1033348 on 8q21.3, p=7.5e-9) and *KDM3A* on 2p11.2 (D11S533, 11q13.5, p=2.32e-8). Further, we identified several hotspot regions in the genome enriched for POE transcripts as well as a transcription factor binding motif (PBX1) that was highly enriched in the promoters of POE transcripts. Our results suggest that POE are a widespread, significant regulatory mechanism of gene expression. Additional studies in untransformed cell lines or tissues and biological studies now appear warranted.

UNTANGLING HYBRID SEQUENCING READS

Harris A Jaffee[1], Rafael A Irizarry[1], Sarah J Wheelan[2,1]

[1]The Johns Hopkins Bloomberg School of Public Health, Biostatistics, 615 N. Wolfe Street, Baltimore, MD, 21205, [2]The Johns Hopkins University School of Medicine, Oncology Biostatistics and Bioinformatics, 550 N. Broadway, Baltimore, MD, 21205

Some high-throughput sequencing experiments are designed find boundaries between two sequence elements. The sequencing results are thus expected to contain "hybrid" reads which will be derived partly from the reference sequences and partly from another type of sequence, and the sequence boundaries are unknown in advance.

A few different protocols produce such hybrid reads:
— to find all retroviral or transposon insertions in a genome, one would look for sequencing reads that have a genomic segment fused to the inserted sequence.

— sequences, such as miRNAs, are too short to span the entire length of a read and thus contain adaptor sequences on one or both ends. If the adaptor placement were always identical, it could be trimmed quite easily, but the miRNAs are of varying lengths, so the adaptor sequence can take up a variable amount of each read.

— excessive (accidental) shearing causes fragments to be shorter than expected so that the sequencing reads run into the adaptors.

In each of these cases, a short read aligner will not be able to align the hybrid sequences, as the index structures for these programs are typically optimized to align a known length of the read instead of dynamically deciding where boundaries could be. One solution is to first align all sequencing reads to the genome that was sequenced, and then take everything that did not align and examine those reads more closely for the presence of extraneous sequences. Depending on how many reads did not align and the number and types of errors allowed, this could be a lengthy process.

We have developed a straightforward algorithm, using modifications to the Biostrings functions in R, that trims sequences from either or both ends of a sequencing read, allowing for a preset number of errors and indels on either side, based on a set of potentially contaminating sequences provided by the user. Comparisons made on several datasets suggest that this is both faster and more sensitive than the two-stage approach that we and others have used to date.

WHOLE EXOME SEQUENCING IN HEPATOCELLULAR CARCINOMA

David A Wheeler[1,2], Marie-Claude Gingras[1], Donna M Muzny[1,2], Ronnald T Cotton[3], Jacfranz J Guiteau[3], John A Goss[3], Lara M Bull[1], Betty L Slagle[4], Richard A Gibbs[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030, [2]Baylor College of Medicine, Molecular and Human Genetics, One Baylor Plaza, Houston, TX, 77030, [3]Baylor College of Medicine, Michael DeBakey Department of Surgery, One Baylor Plaza, Houston, TX, 77030, [4] Baylor College of Medicine, Department of Virology and Microbiology, One Baylor Plaza, Houston, TX, 77030

Hepatocellular carcinoma (HCC) is a global healthcare phenomenon, with 1,000,000 new cases diagnosed annually and local incidences as high as 120 cases per 100,000. It is the fifth most common solid tumor worldwide, and fourth leading cause of cancer-related death. While HCC is seen in the United States much less frequently than in developing countries, incidence has risen sharply with increased rates of Hepatitis C viral infection. Surgical resection or liver transplantation, the only potentially curative modalities, is an option in only a small fraction of patients. HCC is also of special interest due to the known roles of hepatitis B and C viruses in malignancy. To date we have collected 52 HBV, 35 HCV and 6 non-viral HCC patients. Initial sequencing of 28 candidate cancer genes known to be associated with HCC revealed significantly different mutation profiles in HBV and HCV-associated cancers. TP53 and WNT pathway genes are more highly mutated in HBV than in HCV, in parallel with published results in studies with limited sets of candidate genes. Using targeted DNA capture, based on the Consensus Coding Sequence, we are sequencing whole exomes in these same patients. This will afford an unprecedented view of the mutational landscape obtained by sequencing in >17,000 genes (23 Mb) in HCC and provide new insights into pathways leading to cancers of viral etiology.

RE-SEQUENCING OF CANDIDATE REGIONS TO FIND MUTATIONS FOR A CANINE SLE-RELATED DISEASE COMPLEX

Maria Wilbe[1], Katarina Truvé [1], Michael C Zody [2], Gerli Pielberg [3], Päivi Jokinen [4], Hannes Lohi [4], Helene Hansson-Hamlin[5], Göran Andersson [1], Kerstin Lindblad-Toh [2,3]

[1]Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, P.O. Box 597, Uppsala, S-75124, Sweden, [2]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Department of vertebrate genome biology, 7 Cambridge Center, Cambridge, MA, 02142, [3]Uppsala University, Department of Medical Biochemistry and Microbiology, P.O. Box 582, Uppsala, S-752 37, Sweden, [4] Univeristy of Helsinki and Folkhälsan Research Center, Department of Veterinary Biosciences and Dept of Medical Genetics, P.O. Box 63, Helsinki, 00014, Finland, [5]Swedish University of Agricultural Sciences, Department of Clinical Sciences, Box 7054, Uppsala, SE-75007, Sweden

We have studied a systemic lupus erythematosus (SLE)-related disease complex in the canine breed Nova Scotia duck tolling retriever that comprises two different types. The first, which mostly resembles human SLE-related diseases, involves chronic musculoskeletal signs and pain from joints. The dogs often display antinuclear antibodies. The other variant is steroid-responsive meningitis-arteritis.
Genome-wide association mapping identified five loci for the disease complex (Wilbe et al. Nature Genet 2010). Three of our associated regions contain genes involved in T-cell activation through the nuclear factor of activated T cells (NF-AT) pathway (PPP3CA, HOMER2, DAPP1 and PTPN3), which might be a novel SLE-pathway. We also identified BANK1 as a strong candidate gene.
To identify candidate mutations, re-sequencing of candidate regions was performed. 10 dogs were selected for hybridization of 3.2 Mb of sequence. Percentage on target reads ranged between 79-98% and the average coverage was 67x. SNP, indel and copy number variant calling was performed and ~ 3,000 SNPs were identified. Candidate mutations will be identified by searching in coding and non-coding conserved regions and then further validated in more dogs.
Finding new causative variants will give better diagnostic methods for dogs and enhance our understanding of SLE for both canine and humans.

FOSILLS: FOSMID LIBRARIES FOR PAIRED END ILLUMINA SEQUENCING

Louise J Williams, Na Li, Diana G Tabbaa, Aaron Berlin, Terrance P Shea, Sarah Young, Chad Nusbaum, Andreas Gnirke

The Broad Institute, Sequencing, 320 Charles St, Cambridge, MA, 02141

Libraries of large-insert clones are key requisites of traditional Sanger-based whole-genome shotgun sequencing projects. For example, end sequencing of Fosmid libraries with forward and reverse primers generates read pairs spanning ~40 kb in the genome that provide long-range contiguity of sequence assemblies.

To generate next-generation ~40 kb read-pairs, we modified the pFOS1 cloning vector such that the cloning site is flanked by Illumina sequencing primer sequences and two Nb.BbvC1 nicking endonuclease sites. Using this vector we constructed *Fos*mid libraries that can be paired-end sequenced by *Ill*umina (*Fosill*). Primary Fosill libraries were amplified by overnight liquid culture and maxi-prepped. To bring the ends of the cloned fragments together on short sequenceable PCR amplicons, we nicked the Nb.BbvC1 sites, translated the nicks a few hundred bp into the cloned inserts and cleaved there by digestion with nuclease S1. After re-circularization and inverse PCR out of the vector using Illumina PE enrichment primers we sequenced the thus converted Fosill libraries by Illumina.

Paired end sequencing reads for *E. coli* and *R. sphaeroides* had the gap-size distribution expected for Fosmid end sequences and a low rate of chimerism (~2%). The general approach of converting plasmid libraries into "jumping" libraries is an alternative to constructing mate-pair libraries by *in vitro* circularization of size-selected genomic DNA fragments. Since *in vitro* jumping libraries are not amplified until the very end, any prior loss of DNA reduces the complexity of the library. In contrast, for Fosill cloning, large DNA fragments are packaged and size-selected in bacteriophage lambda, circularized and amplified *in vivo*. Once amplified, any subsequent loss of DNA has little effect on library complexity.

Our Fosill approach for generating read pairs spanning ~40 kb will be useful for *de novo* next-generation sequencing of complex genomes as well as for detecting chromosomal structural rearrangements such as translocations or inversions.

LIFE HISTORY TRAITS AFFECT THE MAGNITUDE OF MALE MUTATION BIAS ACROSS 32 MAMMALIAN GENOMES

Melissa A Wilson Sayres[1,2,3], Chris Venditti[4], Francesca Chiaromonte[2,3,5], Mark Pagel[4,6], Kateryna D Makova[1,2,3]

[1]The Pennsylvania State University, Department of Biology, University Park, PA, 16802, [2]The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, PA, 16802, [3]The Pennsylvania State University, Integrative Biosciences Program, University Park, PA, 16802, [4] University of Reading, School of Biological Sciences, Reading, RG6 6BX, United Kingdom, [5]The Pennsylvania State University, Department of Statistics, University Park, PA, 16802, [6]Santa Fe Institute, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM, 87501

Male mutation bias theory predicts that the mutation rate in males is often higher than in females because male gametes, sperm, undergo significantly more rounds of replication than female gametes, eggs. Although observed in mammals, birds, fish, and even plants, estimates of the magnitude of male mutation bias vary substantially across species. There are two explanations for this. First, many genomic factors (e.g. repetitive elements, GC content, recombination rate) influence mutation rates regionally across the genome and, when not accounted for, can skew estimates of male mutation bias. Second, variations observed across species may be influenced by differences in life history traits, specifically metabolic rate, sexual selection, and generation time. Male mutation bias is expected to be influenced by metabolic rate because sperm not only live in a more reactive oxygen species rich environment, they are also more susceptible to mutations through oxidative stress than eggs. Male mutation bias might become elevated with stronger post-copulatory sexual selection; males in species where sperm from multiple males compete to fertilize eggs produce more sperm, potentially at the expense of a higher mutation rate, than males in species without competition. Additionally, species with shorter generation times might experience less male mutation bias because their sperm undergo fewer rounds of replication before conception.

Few studies have examined the factors that influence variation in the magnitude of male mutation bias across multiple species using more than a subset of any genome. Utilizing the 32 eutherian mammal genome sequences we are able to investigate male mutation bias on a genome-wide scale across mammalian taxa with diverse life history traits. We ask which life history traits affect the magnitude of male mutation bias observed in mammals. To answer this question we collected literature on life history traits for all 32 mammals, filtered whole-genome alignments of factors known to influence substitution rates regionally, and computed global and context-dependent substitution rates. Then, after accounting for phylogenetic dependence, we developed a model to describe how variations in life history traits affect variations in the magnitude of male mutation bias. We found that representatives of three major life history traits (metabolic rate, generation time, and sexual selection) all affect the magnitude of male mutation bias, explaining at least 70% of the variation observed across these diverse mammals. Our results corroborate and expand upon previous research, and suggest the significant influence of life history traits on genome evolution.

DIFFERENTIAL PATTERNS OF OPEN CHROMATIN SUGGEST ALTERNATE MODES OF GENE REGULATION

Deborah R Winter[1,2], Lingyun Song[2], Zhancheng Zhang[2], Alan P Boyle[2], Elizabeth A Rach[1,2], Uwe Ohler[2,3], Gregory E Crawford[2], Terrence S Furey[2,3]

[1]Duke University, Computational Biology & Bioinformatics Graduate Program, 304 Research Drive, Durham, NC, 27708, [2]Duke University, Institute for Genome Sciences & Policy, 101 Science Drive, Durham, NC, 27708, [3]Duke University, Biostatistics & Bioinformatics, 101 Science Drive, Durham, NC, 27708

Much of the human genome exists in the nucleus as DNA wrapped around histones to form nucleosomes which are further compacted into higher-order structure. However, in any given cell line, certain stretches are more loosely arranged and accessible to binding factors. DNase Hypersensitivity (DHS) can be exploited to reveal these “open chromatin” and identify functional elements that are actively involved in transcriptional regulation – such as promoters, enhancers, insulators, suppressors, and locus control regions. From the results of our genome-wide assay, known as DNase-seq, we are able to study the locations of open sites with respect to the nearest gene. Moreover, using microarray data, we can compare the prevalence of DHS sites across different cell-lines with their expression values. We observe some genes where the DHS signal at the promoter increases with expression as expected; while other promoters are open in all cell-lines regardless of expression. In these cases, control of transcription appears to be more complex, suggesting the participation of additional binding factors and/or distal elements. We further analyzed these regions in the context of transcription start site (TSS) usage. These alternate gene architectures may reflect different methods of transcriptional regulation.

OLIGONUCLEOTIDE MICROARRAYS ON ONE SQUARE MILLIMETER GLASS CHIPS: DEVELOPMENT AND APPLICATIONS OF THE 'MILLICHIP'

Jamison Wolfer, Kurt Heinrich, DongGee Hong, Melissa LeBlanc, Michael Sussman

University of Wisconsin, Biotechnology, 425 Henry Mall, Madison, WI, 53706

Development of the millichip provides researchers with a straightforward, economical method for monitoring the transcription of the entire genome of simple organisms under varying sets of conditions. The millichip is a 1 mm^2 glass chip containing up to 4489 unique 70mers on its surface with the ability to be hybridized with 5 μL solutions. These millichips are fabricated using the maskless array synthesizer developed in the Sussman lab and later commercialized by NimbleGen Systems, Inc. An array of covalently attached oligonucleotides is synthesized *in situ* on a pre-scored glass microscope slide. The surface of the glass on which the oligonucleotides are attached is then subdivided along the pre-scored grid into 62 separate chips, each ~1 mm^2 and 1 mm thick. These millichips can then be placed in microfuge tubes and used for hybridization to fluorescently labeled RNA. The fluorescence intensity of the 16 μm sized fluorescent features is then determined by measurement under a fluorescent microscope and quantified with software developed in our lab. The advantages of this approach include: (1) Each millichip costs ~1% of the typical DNA chip (~$10) so that many more experiments can be performed. Each millichip currently contains 4489 70mers, but additional improvements to the technology will increase this number dramatically. (2) The hybridization is done in the bottom of a microfuge tube, rather than in larger volume flatbed chambers, minimizing the amount of reagents. (3) The entire procedure can be performed at the bench top, rather than at a core facility. The goal of this work is to create a method for performing whole genome transcriptome analysis at the price and ease of an agarose gel, to allow scientists to more readily explore the full spectrum of changes in RNA expression that occur in biological systems. The millichip is being utilized to compare transcription levels of protein kinases in *Arabidopsis thaliana* grown under varying conditions. The ability to perform numerous chip experiments at a low cost allows for the comparison of hundreds of combinations of growth conditions, whereas previously only one or two changes have been evaluated in a single experiment.

ATLAS-LINK: SCAFFOLDING DRAFT GENOME ASSEMBLIES USING NEXT-GEN MATE PAIR DATA

Jixin Deng, Huaiyang Jiang, Yue Liu, Xiang Qin, Jiaxin Qu, Xing-Zhi Song, Kim C Worley, Richard A Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030

The recent advances in sequencing technologies have allowed large amounts of additional sequence data to be generated for relatively little cost to improve existing genome assemblies. Extracting information from these data for good genome assembly requires meticulous and effective methodologies and tools. We present a tool, Atlas-Link, that can produce a de novo scaffold structure, or upgrade an existing scaffold structure of a draft assembly using mate pair sequencing data.

The Atlas-Link software implements a greedy algorithm and uses graph theory to link and orient assembled existing contigs quickly and accurately using mapped mate pair information and an AGP file describing the contig locations in the existing assembly. Atlas-Link can also be applied to an ACE file from an assembler. The program has versatile and flexible features that provide users freedom to tune to different data. The algorithms associated with Atlas-Link are designed for high sequence coverage. But, based on simulated and benchmark data analyses, Atlas-Link achieves highly accurate and continuous scaffold structures representing the actual genome from both high and low coverage sequencing data. This new module of the Atlas assembly suite has been applied and significantly improved the previously released draft genome assemblies of the Tammar wallaby, the sea urchin and the honeybee using SOLiD and/or 454 sequencing data.

COPY NUMBER VARIATION DETECTION FROM 1000 GENOMES PROJECT EXON CAPTURE SEQUENCING DATA

Jiantao Wu, Chip Stewart, Gabor T Marth 1000 Genomes Project,

Boston College, Biology Department, 140 Commonwealth Ave., Chestnut Hill, MA, 02467

High-throughput next-generation sequencing technology enables detection of DNA copy number variations (CNVs) with higher accuracy than microarray based approaches. The 1000 Genomes Poject Pilot 3 capture sequencing data targeted more than eight thousand exon regions of the human genome, in nearly 697 individual samples from 7 populations. We developed computational and statistical methods for CNV detection that can deal effectively with the high-level of inherent sequence coverage fluctuation usually seen in capture sequencing data.

As the primary measured quantity, our method utilizes the number of reads mapped to a given capture target region within a given individual. The expected read count is modeled as a linear function of the overall read coverage of that sample, multiplied by a target-specific proportionality constant that accounts for target probe affinity and possible target specific PCR biases. CNV candidates are detected as low p-value outliers to the null model fit of the expected read count distribution. Sensitivity and selectivity were estimated using “simulated” CNV events (e.g. heterozygous or homozygous deletions, or amplifications) mixed in with the measured data.

We first applied our methods for 107 samples in a dataset produced by the Wellcome Trust Sanger Institute with paired-end Illumina sequencing. We identified a total of 164 exonic heterozygous-deletion events, of which 55.5% overlaps with CNV events in the Database Of Genomic Variants (DGV). Experimental validation of these events, as well as the analysis of the remaining ~600 samples, sequenced in three other data producing centers within the 1000 Genome Project, are underway.

Currently, multiple large medical sequencing projects generate whole-exome capture data from thousands of individual samples. Our study demonstrates that such datasets are a useful substrate for detecting copy number changes in the coding regions of genes. Our methods, trained on the 1000 Genomes Project pilot data, form a turn-key pipeline that can be readily applied for CNV detection in capture sequencing projects.

A MODULAR PIPELINE FOR DETECTING GENETIC VARIATIONS FROM NEXT-GENERATION SEQUENCING DATA AT NCBI

Chunlin Xiao[1], Tom Blackwell[2], Alistair Ward[3], Anatoly Mnev[1], Paul Anderson[2], Michael Stromberg[3], Chip Stewart[3], Richa Agarwala[1], Mike DiCuccio[1], Goncalo Abecasis[2], Gabor Marth[3], Stephen Sherry[1]

[1]National Institute of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, MD, 20892, [2]University of Michigan at Ann Arbor, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, MI, 48109, [3]Boston College, Biology Department, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467

Next-generation sequencing (NGS) technologies have revolutionized genome sequencing by coupling extremely high throughput with low cost, thereby providing researchers with unprecedented opportunities to address many important biomedical problems efficiently. Large-scale resequencing projects, e.g. 1000 Genomes, TCGA, and TSP, have been initiated to extend our knowledge of single nucleotide polymorphisms (SNPs), short insertions/deletions (INDELs) and structural variations (SVs) and relate these variants to human diseases. The amount of NGS data submitted to public repositories such as the Short Read Archives (SRA) at the NIH National Center of Biotechnology Information (NCBI) is growing exponentially and submissions represent a wide array of technology platforms and sequence collection strategies. To process and analyze these data for variation detection in a uniform manner is a challenge requiring a standard modular pipeline. In collaboration with investigators at Boston College and the University of Michigan, NCBI is developing a framework Variation Discovery and Annotation Pipeline (Gpipe). The pipeline generates quality input sequence data from the Short Read Archives, checks sample identities, aligns the read data with the human reference genome sequences, refines the mapping of placed reads (duplicate removal and base quality recalibration etc), and calls SNPs, INDELs, and SVs according to data availability and project-specific policies. A centrally implemented pipeline streamlines the data processing workflow for the data generated by next-generation sequencing technologies.

DNA METHYLATION CONSERVATION IN MAMMALIAN BRAIN

Yurong Xin[1], Yongchao Ge[2], Anne O'Donnell[1], Benjamin Chanrion[1], Maria Milekic[1], Andrew J Dwork[1], Victoria Arango[1], J. John Mann[1], Fatemeh Haghighi[1]

[1]Columbia University, Department of Psychiatry, 1051 Riverside Drive, New York, NY, 10032, [2]Mount Sinai School of Medicine, Department of Neurology, One Gustave L Levy Place, New York, NY, 10029

DNA methylation plays an important role in genome organization and function. However, little is known about the evolutionary conservation of DNA methylation patterns in mammalian genomes. Here we investigate whole-genome DNA methylation profiles in the central nervous system, and present the first large-scale comparative DNA methylation study in human and mouse genomes. The methylation data were generated using a novel method, methylation mapping analysis by paired-end sequencing (Methyl-MAPS). Genomic DNA from postmortem tissue was fractionated into methylated and unmethylated compartments and subsequent sequencing performed via the ultra high-throughput SOLiD sequencing platform.

In this large-scale DNA methylation profiling study, we have mapped DNA methylation states of over 36% and 89% of CpG sites in human and mouse brains. The human brain metholome data covers 10,262,160 CpGs across 10 normal brain samples from cerebral cortex. The mouse data represents 5 mouse brain specimens from the 129S6/SvEv inbred strain. These data revealed that DNA methylation patterns are not conserved in evolutionary conserved regions of the human and mouse genomes ($r=0.25$). Intriguingly, the extent of human-mouse DNA methylation is not driven by sequence conservation, rather it is driven by CpG dinucleotide density. DNA methylation conservation increases with increasing CpG density, showing >80% correlation for regions containing >5 CpG dinucleotides in 100bp window. These highly conserved regions are enriched in gene promoters and first exons, with conserved methylation patterns representing both unmethylated and methylated states. In addition, these data show distinct methylation signatures that are characteristic of specific genomic features (e.g, promoters and internal exons) that are highly conserved within human and mouse genomes. This study underscores the impact of DNA methylation in mammalian genome evolution.

DNA METHYLATION PROFILING OF NORMAL HUMAN CEREBRAL CORTEX

Yurong Xin, Fatemeh Haghighi

Columbia University, Dept of Psychiatry, NY, NY, 10032

DNA methylation may play an important role in the etiology of neuropsychiatric disorders, perhaps as equally important as genetics and the environment. However, In order to better understand both the wild type genomic DNA methylation patterns and aberrant methylation events that occur in disease states, we first examined DNA methylation profiles within the normal human brain. We have developed a cost-effective, unbiased, whole-genome methylation profiling technique that can assay the methylation state of more than 80% of the CpG sites in the human genome. This method, methylation mapping analysis by paired-end sequencing (Methyl-MAPS) couples advances in next generation sequencing with enzymatic fractionation of DNA by methylation state. In this large-scale study we have mapped the methylation state of 36% of CpG sites in the human cerebral cortex of 10 normal non-psychiatric subjects (including 6 prefrontal and 4auditory cortical samples). We focused on the prefrontal cortex (PFC) due to converging evidence from neuroimaging and functional studies implicating this region in both depression and schizophrenia. Secondarily, we also examined the auditory cortex, because schizophrenia disorder includes defects in sensory perception and processing. With these data we are for the first time able to explore DNA methylation profiles within two distinct brain regions with differing neurodevelopmental trajectories; the evolutionarily conserved auditory temporal cortex developing early as compared to the prefrontal cortex which undergoes maturation well into early adulthood. Our data reveal that DNA methylation is significantly more conserved in the auditory cortex then the PFC ($P<10-15$). Despite this significant difference, DNA methylation signatures in the cortex are highly conserved, with >25% of the total CpG sites in cortex showing less than 20% difference in methylation state across the 10 samples examined. Cross-species analysis of DNA methylation conservation between human and mouse brains show that DNA methylation is not correlated with sequence conservation. Instead, increase in cross-species DNA methylation conservation is correlated with increasing CpG density. Genomic regions with significant human-mouse DNA methylation conservation (correlation >80%) typically have greater than 5 CpG dinucleotide in 100bp window. Although enriched in gene promoters, these regions also cover gene bodies, as well as repeat sequences that represent both methylated and unmethylated states. These data provide insight in studies of neuropsychiatric disorders, in identifying genomic regions that are developmentally and evolutionarily conserved that when aberrantly methylated may confer increase risk for disease.

THE DNA METHYLOME OF HUMAN PERIPHERAL BLOOD MONONUCLEAR CELLS

Yingrui Li, Geng Tian, Ning Li, Xiuqing Zhang, Jun Wang, Huanming Yang

BGI-Shenzhen, Dept. of Sequencing, Beishan Road, Shenzhen, 518083, China

DNA methylation plays a vital role in genome dynamics and, in the human genome, occurs predominantly at cytosine guanine dinucleotide (CpG) sites1. The unique part of haploid human genome2 analysed here contains around 20 million CpG sites (methylome) where DNA methylation can vary, affecting many biological processes in health and disease3. Using whole-genome bisulfite sequencing4-6, we report the essentially complete (92.62%) methylome of human peripheral blood mononuclear cells (PBMC) which constitute an important source for clinical blood tests world-wide. We find the majority of CpG sites (68.4% at false positive rate of 0.46%) and only <0.2% of non-CpG sites to be methylated, demonstrating that non-CpG cytosine methylation is negligible in human PBMC. Analysis of the PBMC methylome revealed a rich landscape of epigenomic data for 20 distinct features including regulatory, protein-coding, RNA gene coding, non-coding and repeat sequences. Alu element mobility, for instance, was found to negatively correlate with their methylation levels, emphasizing the critical role of DNA methylation in genome stability. Integration of our methylome data with the previously determined genome sequence2 of the same Asian individual analysed here, enabled a first assessment of allele-specific methylation (ASM) differences between the two haploid methylomes of any individual. Using a conservative cut-off ($p < 0.001$), we identified 599 haploid differentially methylated regions (hDMRs) covering 287 genes. Of these, 76 genes had hDMRs within 2kb of their transcriptional start sites of which >80% displayed allele-specific expression (ASE) after random testing using TA clone sequencing of the same PBMC sample. These data show, that ASM is a recurrent phenomenon and highly correlated with ASE, suggesting that imprinting may be more common than previously thought. Our study not only provides a comprehensive resource for future epigenomic research but also demonstrates a paradigm for large-scale epigenomics studies through new sequencing technology.

DETECTING BREAKPOINTS OF LARGE DELETIONS AND MEDIUM SIZED INSERTIONS FROM PAIR-END SHORT READS IN 1000 GENOMES PROJECT AND CANCER GENOME PROJECT

Kai Ye[1], Erin Pleasance[2], Klaudia Walter[2], Matthew Hurles[2], Zemin Ning[2]

[1]Leiden University Medical Center, Medical statistics and Bioinformatics, Einthovenweg 20, Leiden, 2300RC, Netherlands, [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

There is a strong demand in the genomic community to develop effective algorithms to identify INDELs and structural variants from multiple samples, in order to investigate disease-related variants and genetic survey of large populations. We developed a so-called Pindel [1] method to identify breakpoints of large deletions (1bp-50kb) and medium sized insertion (1-60bp) at base level precision from paired-end short reads. From those one-end mapped paired-end reads, our Pindel program uses the mapped read to determine the anchor point on the reference genome and the direction of unmapped read. Knowing the anchor point, the direction to search for the unmapped read and the user defined Maximum Deletion Size, a sub-region in the reference genome can be located, where Pindel will break the unmapped reads into 2 (deletion) or 3 (short insertion) fragments and map the two terminal fragments separately.

In this study we developed a novel procedure based on Pindel algorithm to process multiple samples. We add tags to the reads to indicate their sources. Then we run Pindel using the entire pool of reads as the input. We modified our Pindel program to report the sample sources of the supporting reads for each identified indel event. With such information we are able to discern which samples have what indels. We also adapted the algorithm to include detection of deletions with small insertions of non-templated sequences at the breakpoint, and to report a confidence score that is monotonically related to false discovery rate.

We demonstrated our method with the low coverage data of human chr1 from 170 individuals in the 1000 genomes project and with the high coverage sequence data of COLO-829 cancer cell lines from both normal and tumor tissues of the same individual [2].

[1] Ye, K. et al. Bioinformatics 2009 25(21):2865-2871
[2] Pleasance, E. et al. Nature 2010, 463, 191-196

HIGH LEVEL OF AUTOSOMAL NUCLEOTIDE AND HAPLOTYPE DIVERSITY IN SOUTH INDIA POPULATIONS

Jinchuan Xing[1], Ya Hu[2], W S Watkins[1], Chad D Huff[1], Richard A Gibbs[2], Lynn B Jorde[1], Fuli Yu[2]

[1]University of Utah, Department of Human Genetics, Eccles Institute of Human Genetics, Salt Lake City, UT, 84112, [2]Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, Houston, TX, 77030

Genetic studies of populations from the Indian subcontinent are of great interest for many reasons, including India's large population size, its complex demographic history, and its unique social structure. Despite recent large-scale efforts in discovering human genetic variation (e.g., ENCODE, HapMap, and PopRes), India's vast reservoir of genetic diversity remains largely unexplored. To address this issue and to provide a supplement to the ENCODE resequencing project, we resequenced one of the 100Kb ENCODE regions in 94 South Indian samples collected from five populations – representing four castes and one tribal population – from the state of Andhra Pradesh in South India. By examining the unbiased distribution of common and rare (MAF<5%) variants in these non-HapMap populations, we sought to assess the additional information that can be gained by sampling more diverse populations, especially in geographic regions with little or no current coverage.

By comparing the five Indian populations with eight HapMap III populations that are resequenced in the same region, we found that more than 15% of the total SNPs are Indian-specific. Nine percent of the total were found only in the five South Indian populations. Several Indian population samples have nucleotide and haplotype diversity as high as HapMap African populations (nucleotide diversity estimates of 0.089 in middle-caste Yadava, 0.085 in lower-caste Mala/Madiga, and 0.083 in tribal Irula, compared to 0.082 and 0.083 in YRI and LWK). Unlike HapMap populations, whose sequence diversity decreases in proportion to their distances from Africa, some Indian populations (Yadava, Mala, and Irula) have significantly elevated sequence diversity relative to their geographic distance from Africa. Our study highlights the need to further study Indian genetic variation.

CORRELATING TRAITS OF GENE ESSENTIALITY, DUPLICABILITY AND FUNCTIONALITY WITH SELECTION TRENDS ACROSS VERTEBRATES, ARTHROPODS, AND FUNGI.

Robert M Waterhouse[1,2], Evgeny M Zdobnov[1,2,3], Evgenia V Kriventseva[1,2]

[1]University of Geneva Medical School, Genetic Medicine and Development, rue Michel-Servet 1, Geneva, 1211, Switzerland, [2]Swiss Institute of Bioinformatics, Computational Evolutionary Genomics, rue Michel-Servet 1, Geneva, 1211, Switzerland, [3]Imperial College London, Molecular Biosciences, South Kensington Campus, London, SW7 2AZ, United Kingdom

The recent availability of a growing number of sequenced eukaryotic genomes allowed us to examine gene genealogies along the most sampled lineages of vertebrates, arthropods, and fungi in unprecedented detail. We classified about 86% of 1.36M protein-coding genes from 95 annotated genomes into hierarchical orthologous groups, with over 90% of these groups having Gene Ontology (GO) or InterPro functional annotation. Using these data, we explored the links between gene essentiality, functionality, and accumulated numbers of homologs with constrained by selection ortholog sequence divergence, gene copy-number, and phyletic distribution.
We find a clear enrichment of essentiality among universal orthologs, with surprising consistency of the universal fraction across vertebrates, arthropods, and fungi. Essential genes consistently exhibit greater conservation of sequence identity, and are enriched among the majority of biological processes in animals, but not in yeast. Single-copy orthologs are characterized by markedly higher sequence identities, and the strength of purifying selection also vary substantially among different functional gene classes. The othology data are available from OrthoDB at http://cegg.unige.ch/orthodb.

IDENTIFICATION AND ANALYSIS OF UNITARY PSEUDOGENES: HISTORIC AND CONTEMPORARY GENE LOSSES IN HUMANS AND OTHER PRIMATES

Zhengdong D Zhang[1], Adam Frankish[2], Toby Hunt[2], Jennifer Harrow[2], Mark Gerstein[1,3]

[1]Yale University, Department of Molecular Biophysics and Biochemistry, 266 Whitney Ave, New Haven, CT, 06520, [2]Wellcome Trust Sanger Institute, HAVANA project, Wellcome Trust Genome Campus, Hinxton, CB10 1HH, United Kingdom, [3]Yale University, Department of Computer Science, 51 Prospect Street, New Haven, CT, 06520

Unitary pseudogenes are class of unprocessed pseudogenes without functioning (i.e. genic) counterparts in the genome. Numerically, they constitute only a small fraction of tens of thousands of annotated pseudogenes in the human genome. However, as they represent distinct functional losses over time, they shed particular light on the unique features of humans in primate evolution. Here, we develop a pipeline to detect human unitary pseudogenes through analyzing the global inventory of orthologs between the human genome and its mammalian relatives. We apply conservative cutoffs meant to filter out false positives and focus on gene losses along the human lineage after the divergence from rodents ~75 million years ago. In total, we identify 76 unitary pseudogenes, including previously annotated ones, such as ψGULO and ψUOX, and many novel ones. By comparing each of these to its functioning ortholog in other mammals, we can approximately date the creation of each unitary pseudogene (i.e. the gene 'death date') and show that for our group of 76, the functional genes appear to be disabled at a fairly uniform rate throughout primate evolution—and not all at once, correlated, for instance, with the 'Alu burst'. Furthermore, we identify 11 'polymorphic' pseudogenes that have both nonfunctional and functional alleles currently segregating in the human population. Comparing them with their orthologs in other primates, we find that two of them are in fact pseudogenes in non-human primates suggesting that they actually represent cases of a gene that is in the process of being resurrected in the human lineage.

DISTINCT FACTORS CONTROL HISTONE VARIANT H3.3 LOCALIZATION AT SPECIFIC GENOMIC REGIONS

Aaron Goldberg[1], Laura Banaszynski[1], Kyung-Min Noh[1], Peter Lewis[1], Deyou Zheng[2], David Allis[1]

[1]Rockefeller University, Laboratory of Chromatin Biology, 1230 York Ave Box 78, New York, NY, 10065, [2]Albert Einstein College of Medicine, Departments of Neurology, Genetics, Neuroscience, 1410 Pelham Parkway South, Bronx, NY, 10461

The incorporation of histone H3 variants has been implicated in the epigenetic memory of cellular state. Using genome editing with zinc finger nucleases to tag endogenous H3.3, we report genome-wide profiles of H3 variants in mammalian embryonic stem (ES) cells and neuronal precursor cells. Genome-wide patterns of H3.3 are dependent on amino acid sequence, and change with cellular differentiation at developmentally regulated loci. The H3.3 chaperone Hira is required for H3.3 enrichment at active and repressed genes. Strikingly, Hira is not essential for localization of H3.3 at telomeres and many transcription factor binding sites. Immunoaffinity purification and mass spectrometry reveal that the proteins Atrx and Daxx associate with H3.3 in a Hira-independent manner. Atrx is required for Hira-independent localization of H3.3 at telomeres, and for the repression of telomeric RNA. Our data demonstrate that multiple and distinct factors are responsible for H3.3 localization at specific genomic locations in mammalian cells.

1000 GENOMES PROJECT – DATA FLOW AND QUALITY ASSURANCE

Xiangqun Zheng-Bradley[1], Laura Clark[1], Richard Smith[1], Chunlin Xiao[2], Martin Shumway[2], Steve Sherry[2], Paul Flicek[1], 1000 Genomes Project DCC[1,2]

[1]European Bioinformatics Institute, Vertebrate Genomics, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, United Kingdom, [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD, 20894

The main goal of the 1000 genomes project is to establish a comprehensive and detailed catalogue of human genome variations; which in turn will empower association studies to identify disease-causing genes. The pilot studies for the 1000 genomes project are complete and the main project of sequencing 1000 additional whole genomes is underway. This is generating unprecedented amount of data everyday. A joint Data Coordination Center (DCC) between EBI and NCBI is responsible for maintaining the data.

The DCC manages the data flow for the project, which starts from getting sequence data from the sequencing centers via Sequence Read Archive (SRA) and ends at placing both the raw data and analysis results on publicly visible ftp sites. DCC data QA pipelines are integrated into the workflow to ensure data integrity and consistency. After retrieving the sequence data from the SRA, the DCC checks the data to make sure the reads are of high quality (Fastq QA) and are labelled correctly as appropriate samples and individuals (Sample QA), and then make the Fastq files available on the ftp sites. Data processing teams at the Sanger institute and the Translational Genomics Research institute subsequently align the high quality Fastq files to the reference genome and recalibrate base qualities. The alignments are returned to DCC as BAM files, which have to pass a BAM QA process before they can be released on the ftp sites. Analysis group uses the alignments to make variant calls, which are also distributed to public in Variant Call Format (VCF) through the DCC ftp sites. The variant calls can also be browsed in an Ensembl style browser.

The ftp sites and browser site are:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/
http://browser.1000genomes.org/

MOSAIC RECOMBINATION IN GENE FAMILIES: GENOME STRUCTURE CHANGE AND HOST-PARASITE COEVOLUTION

Martine M Zilversmit[1,2,3], Ella K Chase[2], Natalia Tichshenko[1], Diego Czul[1], Karen P Day*[3], Gil McVean*[2], Philip Awadalla*[1]

[1]University of Montreal, CHU Sainte-Justine centre de recherche, 3175 Ch. de la Cote-Ste-Catherine, Montreal, CA, H3T 1C5, [2]University of Oxford, Statistics Department, 1 South Parks Road, Oxford, OX13TG, United Kingdom, [3]NYU Langone Med. Center, Medical Parasitology, 341 East 25th Street, New York, NY, 10010

*These authors contributed equally

Reconstructing the evolutionary history at genomic loci where there is a high frequency of recombination can be difficult because the exchange of discrete blocks of genetic information (mosaic recombination) can obscure a clear signal of a bifurcating, tree-like history. This problem may be particularly pronounced when examining gene families because there can be non-allelic homologous recombination among gene copies (paralogs), in addition to allelic recombination. Using a new statistical method we are able to detect recent and ancestral recombination in gene families and among the paralogs. We apply this method to two gene families with unusually high levels of diversity, each with thousands of alleles: the *variant antigen* (*var*) genes that code for cell-surface proteins in the human malaria parasite *P. falciparum*, and the Human Leukocyte Antigen (HLA) Class I loci, which code for proteins that present pathogen antigens to the immune system. Here we show how mosaic recombination is generating diversity at these loci over evolutionary time, and how detecting blocks of exchanged genetic material may resolve outstanding issues in the evolution of these gene families. In the case of the hyper-variable *var* genes, we show how this family has a surprising old structure, with recombination blocks dating from before speciation, whereas many HLA alleles that may appear to be ancient are in fact recently evolved mosaics of old sequences recombined with newer ones.

GENETIC VARIATION IN ANIMAL BEHAVIOR: GENES, NEURONS, AND MAYBE SOME PRINCIPLES

Cori Bargmann[1], Andres Bendesky[1], Patrick McGrath[1], Evan Macosko[1], Matt Rockman[2], Leonid Kruglyak[3]

[1]Rockefeller University, HHMI, New York, NY, 10065, [2]New York University, Department of Biology, New York, NY, 10003, [3]Princeton University, HHMI, Princeton, NJ, 08544

Behavior arises from the interplay between genes, the environment, and experience, as interpreted and integrated by the nervous system. Interesting behavioral differences between individuals can be heritable, but attempts to map "behavioral genes" typically result in the identification of many QTLs with small effects, and only a few of these genes have been followed to the molecular level. We are studying the genetic architecture of foraging and social behaviors in the nematode worm Caenorhabditis elegans, a simple model animal with interesting but tractable behaviors. Our experimental approach combines neuroscience, molecular genetics, and an analysis of recombinant inbred advanced intercross lines (RIAILs) generated from a cross between two wild-type C. elegans strains. We find that multiple QTLs with epistatic interactions affect single behaviors. These genes intersect by regulating a common neuronal circuit. Comparing these results with results in flies and mice suggests that G protein-coupled receptor signaling may be a common target for natural genetic variation in very different systems.

THE ISEEM PROJECT: PHYLOGENETIC APPROACHES TO MICROBIAL METAGENOMICS

Thomas J Sharpton[1], Samantha Riesenfeld[1], Joshua Ladau[1], Steven W Kembel[2], Jessica L Green[2], Jonathan A Eisen[3], Katherine S Pollard[1]

[1]Gladstone Institutes, University of California San Francisco, 1650 Owens St., San Francisco, CA, 94158-2261, [2]University of Oregon, Center for Ecology & Evolutionary Biology, 5289 University of Oregon, Eugene, OR, 97403-5289, [3]University of California Davis, Genome Center, 451 Health Sciences Drive, Davis, CA, 95616-8816

Metagenomics - shotgun sequencing a microbial community directly from its environment - enables studies of the unseen world and its impacts on humanity and the planet. Our goal is to develop phylogenetic methods to study microbial diversity using metagenomic data. However, it is not clear *a priori* that meaningful trees relating short, fragmentary metagenomic sequence reads can be constructed.
We approach this problem with two insights. First, we focus on gene trees in which the leaves are individual metagenomic reads, rather than species trees. Second, we leverage "reference" gene sequences from fully sequenced microbial genomes and probabilistic profiles of their evolution to classify metagenomic reads, construct alignments, and guide phylogeny construction (from which reference sequences can be pruned). We constructed a metagenomic simulation pipeline and evaluated how well this method resolves phylogenies of metagenomic reads compared to using full-length rRNA or protein sequences. Performance is reasonable as long as the number and phylogenetic breadth of reference sequences is fairly large. We therefore generated trees for a collection of taxonomically informative genes and used these in several novel ways.
First, we developed a method to define operational taxonomic units (OTUs), a microbial surrogate for species. Typical OTU methods cluster sequences using pair-wise sequence identity. To enable clustering of non-overlapping metagenomic reads, we utilize branch lengths in our SSU-rRNA tree to derive a phylogenetic distance measure. We applied this method to human distal gut, human saliva, and global open ocean metagenomes. Deeply sequenced metagenomic libraries recover most OTUs found using traditional PCR methods, plus additional OTUs including previously unknown bacterial taxa. Our second method estimates geometric characteristics of OTU ranges by comparing sets of OTUs found at sites separated by different distances (i.e. distance-decay). We found great variability in geographic ranges, potentially reflecting different dispersal processes and environmental requirements.

We also developed taxonomy-free methods to quantify bacterial diversity within and between metagenomic samples. Our statistics are based on mean branch-lengths separating pairs of reads in protein-coding gene phylogenies. Applying this method to an ocean depth gradient, we found high community diversity at intermediate depths compared to shallow and very deep samples, a pattern that is not revealed using OTUs based on pair-wise sequence identity.

Phylogenetic approaches to metagenomics shed light on the rare biosphere and the mechanisms structuring microbial communities.

QUANTIFYING PROPERTIES OF REGULATORY MUTATION IN *SACCHAROMYCES CEREVISIAE*

Jonathan D Gruber[1,2], Patricia J Wittkopp[1,2]

[1]University of Michigan, Ecology and Evolutionary Biology, 830 N University Ave, Ann Arbor, MI, 48109-1048, [2]University of Michigan, Molecular, Cellular, and Developmental Biology, 830 N University Ave, Ann Arbor, MI, 48109-1048

Gene expression varies within and between species and is an important source of phenotypic differences. The potential for evolutionary changes in gene expression depend on the properties of regulatory mutations, but no quantitative description of the characteristics of these mutations exists. For a single focal gene's expression, we empirically determine the regulatory mutation rate (*i.e.*, the proportion of new mutations that affect its expression), the distribution of mutational effects, the prevalence of dominance, and the relative frequency of *cis*- and *trans*-acting mutants. To address these issues, we collected ~250 novel genotypes of a *Saccharomyces cerevisiae* strain expressing Yellow Fluorescent Protein (YFP) regulated by the promoter sequence of the *TDH3* gene. Briefly, EMS mutagenesis was used to elevate the mutation rate, and 1728 candidate mutants were isolated by Fluorescence Assisted Cell Sorting (FACS). In a secondary screen, a clonal population of each of these candidate mutant genotypes was cultured and analyzed by FACS to estimate the mean and variance of YFP fluorescence for each novel genotype. Nonsynonymous YFP coding sequence changes or YFP gene duplications were found in 43 candidates; we noted that these genotypes have the most-extreme phenotypes but otherwise excluded them from further analysis. Using the remaining mutants, our data suggest that mutations causing statistically significant changes in YFP expression should occur spontaneously at rate of approximately 6×10^{-7} /genome/generation. Mutations with small effects on expression are more common than those with large effects, but intriguingly we found that the mutations that increase expression is were approximately twice as common as those that decrease expression. Mutant genotypes were also crossed to a wildtype strain that enabled us to assess dominance and distinguish between *cis*- and *trans*-action. We found that 85.3% of mutant alleles were recessive to the wildtype allele in a diploid background. Of the 28 genotypes harboring a dominant regulatory mutation, we observed that 9 affected YFP expression in an allele-specific manner (*i.e.*, in *cis*), while the rest show *trans*-acting effects.

NEARLY IDENTICAL GENOMES WITH COMPLEX CONDITIONAL ESSENTIAL PHENOTYPES

Robin D Dowell*[1,5], Owen Ryan*[4], Gerald R Fink[2,3], Charles Boone[4], David K Gifford[1,2,3]

[1]MIT, CSAIL, 32 Vassar Street, Cambridge, MA, 02139, [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA, 02142, [3]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02142, [4] University of Toronto, Terrence Donnelly Center for Cellular & Biomolecular Research, Toronto, M5S 3E1, Canada, [5]University of Colorado, 347 UCB, Boulder, CO, 80309

To address the genotype to phenotype problem we developed a relevant but simple comparative model system for the budding yeast, Saccharomyces cerevisiae. Our system enables a comprehensive assessment of the genetic mechanisms that lead to different phenotypes for the same mutation in two different genetic backgrounds. We focused on a strain called Σ1278b because it mates and forms viable meiotic progeny with the reference strain, S288C, but it is separated from S288C by the same level of genetic variation typical of two human individuals.

As a first step we sequenced and assembled the 12 Mb Σ1278b genome. By computational and tiling array analysis, we annotated a total of 6923 open reading frames (ORFs) in Σ1278b, of which 6848 have orthologs within S288c. Differences between the strains were largely caused by small indels or SNPs, with an average SNP density of 3.2 per kilobase, which is similar to the divergence between two humans.

We deleted all ~5500 genes within Σ1278b to enable a systematic comparison of its deletion mutant phenotypes to those of S288c. In particular, we identified genes that are required uniquely for viability in either Σ1278b or S288c, a phenotype termed "conditional essentiality". We expected such conditional essential genes to be rare because the genomes of Σ1278b and S288c are nearly identical.

Although 894 genes were essential in both S288c and Σ1278b, 44 genes were essential only in Σ1278b and 13 genes were essential only in S288c. Our analysis showed that conditional essentiality is almost always a consequence of complex genetic interactions involving multiple modifiers.

The ability to identify these conditional essential phenotypes in yeast provides the framework to unravel the fundamental principles of genetic networks based upon natural variation, including those that underlie human disease.

A TRANSCRIPTOME OF THE MIGRATING POSTEMBRYONIC *C. ELEGANS* LINKER CELL

Erich M Schwarz, Mihoko Kato, Paul W Sternberg

California Institute of Technology, HHMI, Division of Biology, 156-29, Pasadena, CA, 91125

Transcriptional profiling in the nematode *Caenorhabditis elegans* has so far been performed on large-scale harvests of stably differentiated embryonic or postembryonic cells, yet key events in development rely on individual cells that migrate and invade tissues dynamically. Male gonadal development in *C. elegans* requires that the linker cell trace an arced pathway through the posterior body; during this, the linker cell both migrates and changes shape in a highly reproducible way, and this program partially requires the nuclear hormone receptor NHR-67 (Kato and Sternberg [2009], Development, *136*, 3907-3915). We have thus adapted laser microsurgery, patch-clamp pipettes, and 3'-tailed RT-PCR to enable RNAseq of individually dissected migrating linker cells from wild-type L3- and L4-stage male larvae, and from *nhr-67*(RNAi) L4 larvae. We have detected expression of 7,439 genes in wild-type L3 or L4 linker cells, versus 13,136 genes in bulk RNA from wild-type larvae (out of 20,252 genes in the *C. elegans* genome). 853 genes (~11% of 7,439) are robustly expressed in L3 or L4 linker cells (with an RPKM of at least 1), but have at least 20-fold lower expression in bulk larvae; this subset of linker-cell-specific genes includes *nhr-67*, along with over 40 other transcription factors such as the bZIP *atf-5*, three Hox genes, three HLH genes, and the heterochronic gene *lin-29*. We are currently assaying particularly interesting genes for function by GFP expression patterns and RNAi phenotypes. This analysis should provide the first transcriptional portrait of a migrating cell in *C. elegans*.

VARIATION, SEX AND SOCIAL COOPERATION: MOLECULAR POPULATION GENOMICS OF THE SOCIAL AMOEBA DICTYOSTELIUM DISCOIDEUM.

Jonathan Flowers*[1], Si Li*[1], Angela Stathos*[1], Gerda Saxer[2], David Queller[2], Joan Strassmann[2], Michael Purugganan[1]

[1]New York University, Center for Genomics and Systems Biology, New York, NY, 10003, [2]Rice University, Ecology and Evolutionary Biology, Houston, TX, 77005

Dictyostelium discoideum is a eukaryotic microbial model system for multicellular development, cell-cell signaling and social behaviour. Key models of social evolution require an understanding of genetic relatedness between individuals across the genome or possibly at specific genes, but the nature of variation within D. discoideum is largely unknown. We conducted the first molecular population genomic study in this species, examining the levels and patterns of nucleotide variation in this social microbial species as well as phenotyping wild strains for levels of kin discrimination, social dominance and mating type. We find surprisingly low levels of nucleotide variation in D. discoideum across these strains, with a mean $\pi = 0.0008$, and there is no strong population stratification within North American strains. Contrary to expectations of kin selection theory on the evolution of social cooperation, we do not find any clear relationship between nucleotide divergence and levels of social dominance and kin discrimination in D. discoideum. We do, however, observe a significant negative correlation of kin discrimination between strains with geographic distance. Finally, despite the fact that sex has been rarely observed in this species, we document a rapid distance decay of linkage disequilibrium between SNPs, the presence of recombinant genotypes among natural strains, and high levels of the population recombination parameter ρ. Our results indicate that greater levels of self/non-self recognition may arise within D. discoideum populations, possibly to mitigate against the evolution of cheating behaviours in the spatially unstructured populations of this species. Our SNP data also indicates that recombination is widespread within D. discoideum, and that sex as a form of social interaction is likely to be an important aspect of this model organism's life cycle. These results suggest that genetic studies in this species should be possible, and would potentially expand the molecular toolkit available for this model organism.

GENOMES, TRANSCRIPTOMES, METHYLOMES AND SMRNAOMES OF ARABIDOPSIS ACCESSIONS

Ronan O'Malley[1,2], Ryan Lister[1,2], Robert Schmitz[1,2], Jarrod Chapman[3], Issac Ho[3], Jason Affourtit[4], Zhoutao Chen[4], Brian Desany[4], Srinivasan Maithreyan[4], James Knight[4], Daniel Rokshar[3], Michael Egholm[4], Tim Harkins[4], Joseph Ecker[1,2]

[1]The Salk Institute for Biological Studies, Plant Biology Laboratory, 10010 N. Torrey Pines Road, La Jolla, CA, 92037, [2]The Salk Institute for Biological Studies, Genomic Analysis Laboratory, 10010 N. Torrey Pines Road, La Jolla, CA, 92037, [3]U.S. DoE, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, [4] 454 Life Sciences, A Roche Company, 15 Commercial Street, Banford, CT, 06405

Phenotypic variation results from a combination of genetic and epigenetic differences. With the advent of deep sequencing, it is possible to examine the consequences of genetic polymorphisms genome-wide. *Arabidopsis thaliana* provides an excellent model for such a study as it exists in the wild as a collection of naturally occurring inbred populations with significant phenotypic differences. Using 454 mate-pair and XLR reads, we generated a de novo assembly of a Cape Verde Island (Cvi-0) accession. Comparison to the Columbia (Col-0) reference identifies SNPs, indel, and structural variations constituting a 3% sequence difference between these accessions. Using the Cvi-0 assembly as a reference, we have mapped the Cvi-0 transcriptome, methylome and smRNAome. A comparison of methylomes between Cvi-0, Col-0 and the Landsberg Erecta (Ler-0) accession reveals that Cvi-0 lacks half the methylation of the other accessions due to a decrease in the methylated cytosines at CG dinucleotides. The highest conservation of CG methylation between the accessions occurs in a region upstream of the transcription start site suggesting a possible role for retention of mCG at specific regulatory regions. While there is significant conservation of smRNA loci between the accessions, hundreds of accession–specific epialleles have been identified. Additionally, the relationship of structural, genetic and epigenetic variation is also presented.

A FINE-SCALE GENETIC MAP OF THE CHIMPANZEE GENOME FROM SEQUENCE VARIATION DATA

Oliver Venn[1], Adi Fledel-Alon[2], Adam Auton[1], Cord Melton[2], Susanne Pfeifer[3], Ryan Hernandez[2], Rory Bowden[1,3], Zamin Iqbal[1], Simon Myers[1,3], Peter Donnelly*[1,3], Molly Przeworski*[2], Gilean McVean*[1,3]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom, [2]University of Chicago, Dept of Human Genetics, 920 East 58th Street, Chicago, IL, 60637, [3]University of Oxford, Dept of Statistics, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom

Statistical analysis of sequence variation data has enabled the construction of a fine-scale genome-wide map of recombination rates in humans. These maps revealed a landscape dominated by recombination hotspots and led to the discovery of a 13-bp motif associated with the activity of 40% of human hotspots. Recently, we showed that the rapidly evolving zinc-finger protein PRDM9 binds this motif and that the motif is being lost from the human genome through a self-destructive drive. These observations, and the fact that humans and chimpanzees show low, if any, sharing of hotspots, suggest a scenario in which the evolution of PRDM9 and the hotspot motif are coupled. However, many questions remain about whether there exist classes of hotspot that are conserved and whether recombination rates are conserved at broader scales between species. To address these questions we have sequenced 10 *Pan troglodytes verus* to 6-10X coverage using paired end sequencing with 50-bp reads. This project increases the amount of chimpanzee variation data by nearly 10-fold and provides a comparative map of genetic variation between humans and their closest living relatives. From these data we intend to construct a high-resolution genome-wide recombination map using coalescent methods that incorporate genotype uncertainty. This will enable four innovations: (1) locating chimpanzee hotspots and observing their properties, (2) determining whether hotspot classes predicted by PRDM9 are enriched in chimpanzee, (3) identifying shared hotspot classes between chimpanzee and human, (4) exploring the scales at which recombination factors operate in chimpanzee.
*Equal contributors

DE NOVO ASSEMBLY AND EVOLUTIONARY ANALYSES OF LIVER-EXPRESSED GENES IN 16 MAMMAL SPECIES

John C Marioni[1], George H Perry[1], Pall Melsted[1], Ying Wang[1], Katelyn Michelini[2], Matthew Stephens[1,3], Jonathan K Pritchard[1,2], Yoav Gilad[1]

[1]University of Chicago, Department of Human Genetics, 920 E. 58th Street, Chicago, IL, 60637, [2]University of Chicago, Howard Hughes Medical Institute, 920 E. 58th Street, Chicago, IL, 60637, [3]University of Chicago, Department of Statistics, 5734 S. University Avenue, Chicago, IL, 60637

Changes in gene regulation have been proposed as critical in the evolution of phenotypic diversity of primates. However, the lack of high-quality reference genomes for most species and the limited number of independently derived transcripts for non-human primates has made it difficult to study gene regulation across multiple primates. To overcome these problems, we used massively-parallel sequencing to interrogate mRNA samples extracted from the livers of 16 species (12 primates including human, and 4 non-primate out groups; 4 samples per species). Of the 12 primate species, 8 do not have currently available reference genome sequences (vervet, galago, slow loris, and 5 lemur species), which means that we had to assemble the transcriptomes de novo for these species. This study design results in nucleotide sequence, quantitative expression, and gene structure data from thousands of genes, providing insight into gene regulation and sequence evolution across a broad spectrum of primate species.

Our analysis has revealed sets of genes whose expression level patterns are consistent with the action of natural selection along individual or ancestral primate lineages. We also investigated the relationship between alternative splicing within species and complete exon gains and losses between species. Interestingly, we also identified a large number of genes, highly conserved at the sequence level, that are expressed in the livers of some taxa, but that are unexpressed in others. Further, we have identified genes whose function and expression pattern we hypothesize may underlie specific adaptive processes. For example, PRKAA2, which inactivates enzymes that regulate the synthesis of fatty acids and is targeted by drugs that treat diabetes, is lowly expressed in primates but highly expressed in non-primate out groups.

Introducing HiSeq™ 2000

Redefining the trajectory of sequencing.

What if you could:

- Sequence a normal and a cancer human genome at 30x coverage?
- Perform gene expression profiling on 200 samples?
- Sequence a genome on one flow cell and its epigenome and transcriptome on the other flow cell?

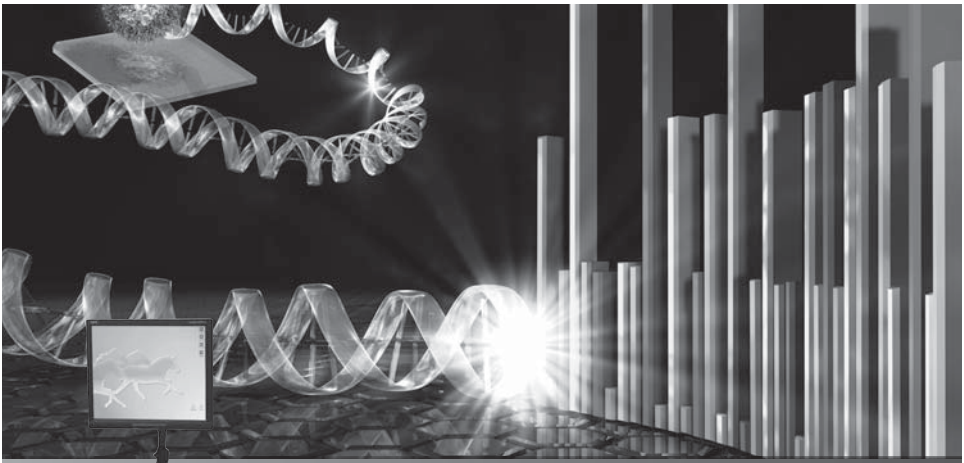
Each in a single run?

Now you can with HiSeq 2000. It's a new standard in output, user experience, and cost-effectiveness.

Sequence on a scale never before possible.

Learn more at
www.illumina.com/HiSeq2000

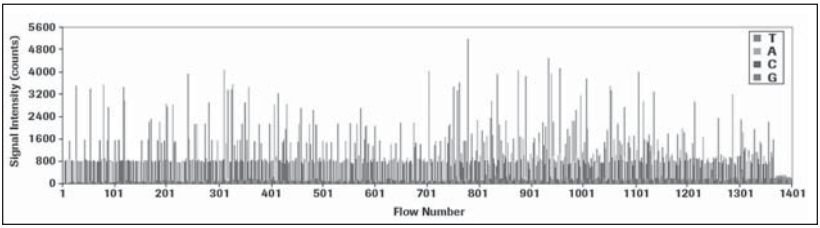
illumina®



www.454.com

Genome Sequencer FLX System

~~Some~~ Less assembly required



DNA Sequencing Flowgram: Each bar within the flowgram represents a discrete nucleotide (A, T, C, or G) and the height of the bar corresponds to the number of nucleotides detected. The flowgram above represents a 1008-base-pair sequencing read from *E. coli* K-12.

Make assembly easier by using the **Genome Sequencer FLX System** – featuring the longest read length available in next-generation sequencing (up to 1,000 bp) and a powerful suite of analysis tools.

- **GS *De Novo* Assembler**
  Generate a full consed directory structure.
- **GS Reference Mapper**
  Map 1 Gb of sequencing data in 1 hour.
- **GS Amplicon Variant Analyzer**
  Create reports and graph results of variations.

Learn more at *www.454.com*

For life science research only.
Not for use in diagnostic procedures.

454 SEQUENCING

Roche Diagnostics Corporation
Roche Applied Science
Indianapolis, Indiana

Roche

Participant List

Dr. David Adams
Wellcome Trust Sanger Institute
da1@sanger.ac.uk

Dr. Véronique Adoue
McGill Univ. & Genome Quebec Innovation Centre
veronique.adoue@mail.mcgill.ca

Dr. Subramanian Ajay
National Institutes of Health
ajayss@mail.nih.gov

Dr. Md Tauqeer Alam
Centers for Disease Control & Prevention (CDC)
hsf1@cdc.gov

Dr. Cornelis Albers
University of Cambridge
caa43@cam.ac.uk

Dr. Frank Albert
Max Planck Institute for Evolutionary Anthropology
falbert@eva.mpg.de

Dr. Ido Amit
The Broad institute
iamit@broad.mit.edu

Mr. Guruprasad Ananda
Penn State University
guru@psu.edu

Ms. Andrea Anderson
GenomeWeb Daily News
anderson@genomeweb.com

Dr. Leif Andersson
Uppsala University
leif.andersson@imbim.uu.se

Dr. Bjorn Andersson
Karolinska Institutet
bjorn.andersson@ki.se

Prof. Stylianos Antonarakis
University of Geneva, Medical School
Stylianos.Antonarakis@unige.ch

Mr. Timo Verneri Anttila
Wellcome Trust Sanger Institute
verneri.anttila@gmail.com

Prof. Alan Archibald
Roslin Institute
alan.archibald@bbsrc.ac.uk

Mr. Chris Armour
NuGEN Technologies, Inc.
carmour@nugeninc.com

Dr. Peter Arndt
Max Planck Institute for Molecular Genetics
arndt@molgen.mpg.de

Dr. Adam Auton
Oxford University
auton@well.ox.ac.uk

Dr. Philip Awadalla
University of Montreal
philip.awadalla@umontreal.ca

Dr. Jeanette Axelsson
Swedish University of Agriculture
Jeanette.Axelsson@hgen.slu.se

Dr. Orli Bahcall
Nature Genetics
o.bahcall@natureny.com

Dr. Suganthi Balasubramanian
Yale University
suganthi.bala@yale.edu

Dr. Michaela Banck
Mayo Clinic Rochester
banck.michaela@mayo.edu

Dr. Cori Bargmann
The Rockefeller University
cori@mail.rockefeller.edu

Mr. Derek Barnett
Boston College
barnetde@bc.edu

Dr. Jeffrey Barrett
Wellcome Trust Sanger Institute
barrett@sanger.ac.uk

Dr. Mark Batzer
Louisiana State University
mbatzer@lsu.edu

Dr. Cindy Bell
Genome Canada
cbell@genomecanada.ca

Dr. David Bentley
Illumina, Inc
dbentley@illumina.com

Dr. Eugene Berezikov
Hubrecht Laboratory
e.berezikov@hubrecht.eu

Dr. Andreas Beutler
Mayo Clinic
beutler.andreas@mayo.edu

Prof. Peter Beyerlein
University of Applied Sciences Wildau
peter.beyerlein@tfh-wildau.de

Dr. Andy Bhattacharjee
Agilent
andy_bhattacharjee@agilent.com

Dr. Henry Bigelow
The Broad Institute
petersen@broadinstitute.org

Dr. Minou Bina
Purdue University
Bina@Purdue.edu

Dr. Ewan Birney
EBI/EMBL
shelley@ebi.ac.uk

Dr. Bruce Birren
Broad Institute of MIT and Harvard
bwb@broadinstitute.org

Dr. Thomas Blackwell
University of Michigan
tblackw@umich.edu

Dr. Martin Blaser
New York University, Langone Medical Center
martin.blaser@med.nyu.edu

Dr. Helmut Blöcker
HZI - Helmholtz Centre for Infection Research
bloecker@helmholtz-hzi.de

Dr. Vivien Bonazzi
National Institutes of Health
bonazziv@mail.nih.gov

Ms. Mary Kate Bonner
University of Wisconsin-Madison
mbonner@wisc.edu

Dr. Gerard Bouffard
NIH/NHGRI
bouffard@mail.nih.gov

Prof. Michael Boutros
German Cancer Research Center
m.boutros@dkfz.de

Dr. Adam Boyko
Stanford University School of Medicine
aboyko@stanford.edu

Prof. Allan Bradley
Wellcome Trust Sanger Institute
abradley@sanger.ac.uk

Dr. Lisa Brooks
NIH
brooksl@mail.nih.gov

Mr. Clive Brown
Oxford Nanopore Technologies Ltd
clive.brown@nanoporetech.com

Dr. Christopher Brown
University of Chicago
caseybrown@uchicago.edu

Dr. Alayne Brunner
Veterans Affairs Palo Alto Health Care System
alayne.brunner@stanford.edu

Ms. Katarzyna Bryc
Cornell University
kb282@cornell.edu

Ms. Marija Buljan
Wellcome Trust Sanger Institute
mb9@sanger.ac.uk

Dr. Lara Bull
Baylor College of Medicine
bull@bcm.edu

Prof. Ralf Bundschuh
The Ohio State University
bundschuh@mps.ohio-state.edu

Mr. Hernan Burbano
Max Planck Institute for Evolutionary Anthropology
hernan_burbano@eva.mpg.de

Mr. Joshua Burton
The Broad Institute
petersen@broadinstitute.org

Ms. Michele Busby
Boston College
busbym@bc.edu

Dr. Harmen Bussemaker
Columbia University
hjb2004@columbia.edu

Prof. Carlos Bustamante
Stanford University
cdbustam@stanford.edu

Dr. Mario Caccamo
The Genome Analysis Centre
mario.caccamo@bbsrc.ac.uk

Dr. Benjamin Callahan
Stanford University
benjc@stanford.edu

Dr. John Capra
Gladstone Institutes, UCSF
tony.capra@gladstone.ucsf.edu

Dr. Lucia Carbone
Children's Hospital of Oakland Research Institute
lcarbone@chori.org

Dr. Piero Carninci
RIKEN
carninci@riken.jp

Mr. Richard Carter
Oxford Nanopore Technologies Ltd
sidra.moazzam@nanoporetech.com

Dr. John Carulli
Biogen Idec
john.carulli@biogenidec.com

Dr. Ferran Casals
Université de Montréal
ferran.casals.lopez@umontreal.ca

Dr. Tanita Casci
Nature Publishing Group
t.casci@nature.com

Dr. Todd Castoe
University of Colorado School of Medicine
todd.castoe@ucdenver.edu

Dr. Susan Celniker
Lawrence Berkeley National Laboratory
celniker@bdgp.lbl.gov

Dr. Wei Chen
MDC Max-Delbrück-Center for Moleculare Medicin
wei.chen@mdc-berlin.de

Dr. Jack Chen
Simon Fraser University
chenn@sfu.ca

Dr. Rui Chen
Baylor College of Medicine
ruichen@bcm.tmc.edu

Dr. Ken Chen
Washington University School of Medicine
kchen@watson.wustl.edu

Dr. Chao Cheng
Yale University
chengchao12@gmail.com

Dr. Derek Chiang
University of North Carolina
chiang@med.unc.edu

Dr. Jeremy Chien
Mayo Clinic College of Medicine
Chien.Jeremy@mayo.edu

Mr. Brian Chin
MIT
brianc@mit.edu

Dr. Stephanie Chissoe
GlaxoSmithKline
stephanie.l.chissoe@gsk.com

Dr. Deanna Church
DHHS/NIH/NLM/NCBI
church@ncbi.nlm.nih.gov

Dr. Kevin Clancy
Invitrogen; Life technologies
kevin.clancy@lifetech.com

Dr. Matthew Clark
Wellcome Trust Sanger Institute
mc3@sanger.ac.uk

Dr. Andrew Clark
Cornell University
ac347@cornell.edu

Ms. Laura Clarke
EMBL-EBI
laura@ebi.ac.uk

Dr. Sandra Clifton
Washington University School of Medicine
sclifton@watson.wustl.edu

Dr. Cristian Coarfa
Baylor College of Medicine
coarfa@bcm.edu

Dr. Barak Cohen
Washington University School of Medicine
cohen@genetics.wustl.edu

Dr. Gregory Cooper
University of Washington
coopergm@uw.edu

Dr. Chris Cotsapas
Massachusetts General Hospital
chrisc@chgr.mgh.harvard.edu

Dr. Alex Coventry
Clark Lab, Cornell University
coventry@cornell.edu

Dr. Mark Cowley
Garvan Institute of Medical Research
m.cowley@garvan.org.au

Dr. David Craig
Translational Genomics Research Institute - TGEN
dcraig@tgen.org

Prof. Gregory Crawford
Duke University
greg.crawford@duke.edu

Ms. Ciara Curtin
Genome Technology magazine
ccurtin@genomeweb.com

Dr. Christina Curtis
University of Cambridge
cc529@cam.ac.uk

Mr. Bryce Daines
Baylor College of Medicine
daines@bcm.tmc.edu

Dr. Charles Danko
Cornell University
dankoc@gmail.com

Mr. Lawrence David
Massachusetts Institute of Technology
ldavid@mit.edu

Dr. Aaron Day-Williams
Wellcome Trust Sanger Institute
adw@sanger.ac.uk

Dr. Pieter De Jong
Children's Hospital & Research Cntr Oakland CHRCO
pdejong@chori.org

Mr. Jason de Koning
University of Colorado Denver, School of Medicine
Jason.DeKoning@UCDenver.edu

Dr. Francisco De La Vega
Life Technologies
francisco.delavega@lifetech.com

Dr. Ricardo del Rosario
Genome Institute of Singapore
delrosariorc@gis.a-star.edu.sg

Dr. Panagiotis Deloukas
Wellcome Trust Sanger Institute
panos@sanger.ac.uk

Dr. Mark DePristo
Broad Institute
depristo@broad.mit.edu

Dr. Emmanouil Dermitzakis
University of Geneva
emmanouil.dermitzakis@unige.ch

Dr. Brian Desany
454 Life Sciences
brian.desany@roche.com

Dr. Scott Devine
Unversity of Maryland School of Medicine
sdevine@som.umaryland.edu

Dr. Federica Di Palma
Broad Institute of MIT and Harvard
fdipalma@broad.mit.edu

Dr. Sabine Dietmann
Mayo Clinic
dietmann.sabine@mayo.edu

Dr. Laura Dillon
NIH/NHGRI
lliefer@mail.nih.gov

Dr. Li Ding
Washinton University School of Medicine
lding@watson.wustl.edu

Prof. Peter Donnelly
Wellcome Trust Centre for Human Genetics
directorpa@well.ox.ac.uk

Mr. Kory Douglas
Texas A&M University
kodouglas@tamu.edu

Dr. Radoje Drmanac
Complete Genomics Inc.
rdrmanac@completegenomics.com

Dr. Marie Pierre Dube
Montreal Heart Institute
marie-pierre.dube@statgen.org

Dr. Ian Dunham
EBI
dunham@ebi.ac.uk

Dr. Richard Durbin
Wellcome Trust Sanger institute
rd@sanger.ac.uk

Dr. Laurent Duret
Laboratoire de Biometrie et Biolgie Evolutive
duret@biomserv.univ-lyon1.fr

Dr. Joseph Ecker
The Salk Institute for Biological Studies
ecker@salk.edu

Dr. John Edwards
Washington University School of Medicine
jedwards@dom.wustl.edu

Dr. Michael Egholm
454 Life Sciences
michael.egholm@roche.com

Dr. James Engert
McGill University
jamie.engert@mcgill.ca

Dr. Xavier Estivill
Center for Genomic Regulation
xavier.estivill@crg.cat

Prof. Peggy Farnham
University of California-Davis
pjfarnham@ucdavis.edu

Mr. Andrew Farrell
Boston College
farrelac@bc.edu

Ms. Gloria Fawcett
Baylor College of Medicine
glfawcet@bcm.edu

Dr. Elise Feingold
National Human Genome Research Institute/NIH
feingole@mail.nih.gov

Dr. Adam Felsenfeld
National Human Genome Research Institute/NIH
felsenfa@mail.nih.gov

Mr. Lars Feuerbach
Max Planck Institut für Informatik
lfbach@mpi-inf.mpg.de

Dr. Lars Feuk
Uppsala University
lars.feuk@genpat.uu.se

Dr. Paul Flicek
European Bioinformatics Institute
flicek@ebi.ac.uk

Dr. Sneh Lata FNU
Cold Spring Harbor Laboratory
slata@cshl.edu

Dr. Kelly Frazer
UCSD
kafrazer@ucsd.edu

Ms. Karin Fredrikson
Roche Applied Science
karin.fredrikson@roche.com

Mr. Gordon Freeman
University of Wisconsin - Madison
gfreeman@wisc.edu

Dr. Nelson Freimer
University of California, Los Angeles
nfreimer@mednet.ucla.edu

Dr. Yan FU
Monsanto Company
yan.fu@monsanto.com

Dr. Asao Fujiyama
Research Organaization of Information and Systems
afujiyam@nii.ac.jp

Mr. Robert Fulton
Washington University School of Medicine
bfulton@watson.wustl.edu

Prof. Yin Wan Wendy Fung
Chinese University of Hong Kong
wendy.fung@cuhk.edu.hk

Dr. Rebecca Furlong
Genome Medicine
rebecca.furlong@genomemedicine.com

Dr. Daniel Gaffney
University of Chicago
dgaffney@uchicago.edu

Dr. Julien Gagneur
European Molecular Biology Laboratory
gagneur@embl.de

Dr. Caroline Gallant
Uppsala University
caroline.gallant@genpat.uu.se

Dr. Manuel Garber
The Broad Institute of MIT and Harvard
mgarber@broad.mit.edu

Ms. Nandita Garud
Stanford
ngarud@stanford.edu

Dr. Clare Garvey
Genome Biology
clare.garvey@genomebiology.com

Dr. Michel Georges
University of Liège
michel.georges@ulg.ac.be

Dr. Mark Gerstein
Yale University
Mark.Gerstein@yale.edu

Dr. Mohammad Ghahramani Seno
Hospital for Sick Children
mgseno@yahoo.com

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.tmc.edu

Prof. David Gifford
MIT
mahony@mit.edu

Dr. Michael Gilchrist
National Institute for Medical Research
m.gilchrist@nimr.mrc.ac.uk

Dr. Andreas Gnirke
Broad Institute of MIT and Harvard
gnirke@broadinstitute.org

Dr. Loyal Goff
MIT
lgoff@csail.mit.edu

Dr. Peter Good
National Institutes of Health
goodp@mail.nih.gov

Mr. David Goode
Stanford University
dgoode@stanford.edu

Dr. Bettie Graham
National Institutes of Health
bettie_graham@nih.gov

Dr. Brenton Graveley
University of Connecticut Health Center
graveley@neuron.uchc.edu

Dr. Eric Green
National Human Genome Research Institute/ NIH
egreen@nhgri.nih.gov

Dr. Richard Green
Max Planck Institute for Evolutionary Anthropology
green@eva.mpg.de

Dr. Ilan Gronau
Cornell University
ig67@cornell.edu

Ms. Shari Grossman
Harvard University and the Broad Institute
grossm@fas.harvard.edu

Dr. Jonathan Gruber
University of Michigan
gruberjd@umich.edu

Dr. Elin Grundberg
King's College London
eg5@sanger.ac.uk

Dr. Yongtao Guan
University of Chicago
ytguan@gmail.com

Dr. Roderic Guigo
Centre de Regulacio Genomica (CRG)
roderic.guigo@crg.es

Dr. Kiranmai Gumireddy
The Wistar Institute
kgumireddy@wistar.org

Dr. Kevin Gunderson
Illumina, Inc.
kgunde@illumina.com

Dr. Chris Gunter
HudsonAlpha Institute for Biotechnology
cgunter@hudsonalpha.org

Dr. Mitchell Guttman
The Broad Institute of MIT and Harvard
mguttman@mit.edu

Dr. Mark Guyer
National Human Genome Research Institute/ NIH
guyerm@exchange.nih.gov

Mr. Kevin Ha
McGill University
kevin.ha@mail.mcgill.ca

Dr. Brian Haas
Broad Institute of MIT and Harvard
bhaas@broad.mit.edu

Mr. Lukas Habegger
Yale University
lukas.habegger@yale.edu

Dr. Fatemeh Haghighi
Columbia University
fgh3@columbia.edu

Dr. Kevin Hall
Illumina
KHall@illumina.com

Dr. Ira Hall
University of Virginia
irahall@virginia.edu

Dr. Molly Hammell
University of Massachusetts Medical School
molly.hammell@umassmed.edu

Dr. Jian Han
HudsonAlpha Institute for Biotechnology
jhan@hudsonalpha.org

Dr. Nancy Hansen
NHGRI/National Institutes of Health
nhansen@mail.nih.gov

Dr. Oscar Harari
Washington University in St. Louis, Med School
harario@psychiatry.wustl.edu

Prof. Ross Hardison
Penn State University
rch8@psu.edu

Dr. Olivier Harismendy
University of California San Diego
oharismendy@ucsd.edu

Dr. R. Alan Harris
Baylor College of Medicine
rharris1@bcm.edu

Dr. Jennifer Harrow
Wellcome Trust Sanger Institute
jla1@sanger.ac.uk

Dr. Ronald Hart
Rutgers University
rhart@rci.rutgers.edu

Mr. Chris Hartl
The Broad Institute
petersen@broadinstitute.org

Dr. Shin-ichi Hashimoto
The University of Tokyo
hashimot@m.u-tokyo.ac.jp

Prof. Masahira Hattori
The University of Tokyo
hattori@k.u-tokyo.ac.jp

Dr. Jacqueline Heard
Monsanto Company
jacqueline.e.heard@Monsanto.com

Ms. Eliana Hechter
University of Oxford
hechter@stats.ox.ac.uk

Ms. Monica Heger
In Sequence
mheger@genomeweb.com

Dr. Ryan Hernandez
UCSF
ryan.hernandez@ucsf.edu

Ms. Jaqueline Hess
EMBL
jacky@ebi.ac.uk

Dr. Peter Heutink
VUMC
p.heutink@vumc.nl

Dr. Axel Hillmer
Agency of Science Technology & Research
hillmer@gis.a-star.edu.sg

Dr. Emily Hodges
Cold Spring Harbor Laboratory
hodges@cshl.edu

Dr. Margret Hoehe
Max Planck Institute for Molecular Genetics
hoehe@molgen.mpg.de

Mr. Lewis Hong
Stanford University
lhongz@stanford.edu

Dr. Roger Hoskins
Lawrence Berkeley National Laboratory
Hoskins@fruitfly.org

Dr. Bryan Howie
University of Chicago
howie.bryan@gmail.com

Mr. Hao Hu
University of Utah
yunathestar@gmail.com

Dr. Qihong Huang
The Wistar Institute
qhuang@wistar.org

Dr. Timothy Hubbard
Wellcome Trust Sanger Institute
th@sanger.ac.uk

Prof. Norbert Hubner
Max Delbruck Centrum for Molecular Medicine
nhuebner@mdc-berlin.de

Dr. Thomas Hübsch
MPI Molecular Biology
huebsch@molgen.mpg.de

Dr. Thomas Hudson
Ontario Instutite for Cancer Research
tom.hudson@oicr.on.ca

Dr. Peter Humburg
University of Oxford
peter.humburg@well.ox.ac.uk

Mr. Sean Humphray
Illumina Cambridge
shumphray@illumina.com

Dr. Tim Hunkapiller
Discovery Biosciences
tim@discoverybio.com

Ms. Julie Hussin
University of Montreal
julie.hussin@umontreal.ca

Dr. Fiona Hyland
Life Technologies
fiona.hyland@lifetech.com

Dr. Oleg Iartchouk
PCPGM
oiartchouk@partners.org

Dr. Youssef Idaghdour
Université de Montréal
youcarolina@gmail.com

Mr. Amit Indap
Boston College
indapa@gmail.com

Ms. Aino Inkeri Järvelin
EMBL
melanie.rauscher@embl.de

Dr. Zamin Iqbal
Oxford University
zam@well.ox.ac.uk

Mr. Sergii Ivakhno
Cancer Research UK Cambridge Research Institute
si245@cam.ac.uk

Dr. Howard Jacob
Medical College of Wisconsin
jacob@mcw.edu

Dr. David Jaffe
Broad Institute of MIT and Harvard
jaffe@broad.mit.edu

Dr. Anna Jasinska
UCLA Center for Neurobehavioral Genetics
ajasinska@mednet.ucla.edu

Ms. Hui Jiang
BGI-shenzhen
jianghui@genomics.org.cn

Dr. Anna Johansson
Uppsala University
anna.johansson@genpat.uu.se

Dr. Vladimir Jojic
Stanford University
vjojic@gmail.com

Dr. Yann Joly
McGill University
yann.joly@mail.mcgill.ca

Mr. Luke Jostins
Wellcome Trust Sanger Institute
LJ4@SANGER.AC.UK

Dr. Scott Kahn
Illumina
skahn@illumina.com

Mr. Sotaro Kanematsu
the University of Tokyo
kk077637@ims.u-tokyo.ac.jp

Mr. Rahul Kanwar
Mayo Clinic
kanwar.rahul@mayo.edu

Dr. Elinor Karlsson
Harvard University & The Broad Institute
elinor@broad.mit.edu

Dr. Julia Karow
GenomeWeb
jkarow@genomeweb.com

Dr. Arek Kasprzyk
Ontario Institute for Cancer Research
arek.kasprzyk@oicr.on.ca

Dr. Mamoru Kato
CSHL
mkato@cshl.edu

Dr. Kazuto Kato
Kyoto University
kato@zinbun.kyoto-u.ac.jp

Dr. Jane Kaye
University of Oxford
jane.kaye@ethox.ox.ac.uk

Dr. Thomas Keane
Wellcome Trust Sanger Institute
thomaskeane@gmail.com

Dr. Katerina Kechris
University of Colorado Denver
katerina.kechris@ucdenver.edu

Dr. Jonathan Keebler
NC State University, University of Montreal
jkeebler42@gmail.com

Dr. Alon Keinan
Cornell University
ak735@cornell.edu

Prof. Manolis Kellis
MIT
manoli@mit.edu

Mr. Pouya Kheradpour
Massachusetts Institute of Technology
pouyak@mit.edu

Dr. Ekta Khurana
Yale University
ekta.khurana@yale.edu

Mr. Ari Kiirikki
Knome, Inc.
akiirikki@knome.com

Ms. Helena Kilpinen
Wellcome Trust Sanger Institute
HK2@SANGER.AC.UK

Mr. Martin Kircher
Max-Planck-Institute for Evolutionary Anthropology
martin.kircher@eva.mpg.de

Dr. Chinnappa Kodira
454 Roche
Chinnappa.kodira@roche.com

Dr. Isaac Samuel Kohane
Harvard Medical School
isaac_kohane@hms.harvard.edu

Dr. Kensuke Kojima
Kyowa Hakko Kirin Co., Ltd
kensuke.kojima@kyowa-kirin.co.jp

Dr. Lesheng Kong
University of Oxford
lesheng.kong@dpag.ox.ac.uk

Dr. Miriam Konkel
Louisiana State University
konkel@lsu.edu

Dr. Melissa Kramer
Cold Spring Harbor Laboratory
delabast@cshl.edu

Dr. Leonid Kruglyak
Princeton University
leonid@genomics.princeton.edu

Mr. Deniz Kural
Boston College
denizkural@gmail.com

Mr. Ahmet Kurdoglu
The Translational Genomes Research Institute
akurdoglu@tgen.org

Dr. Tony Kwan
McGill University
tony.kwan@mail.mcgill.ca

Dr. Eric Lander
The Broad Institute of MIT & Harvard
lander@broad.mit.edu

Mr. Benjamin Langmead
Johns Hopkins Bloomberg School of Public Health
blangmea@jhsph.edu

Dr. Tuuli Lappalainen
University of Geneva Medical School
tuuli.lappalainen@helsinki.fi

Dr. Denis Larkin
University of Illinois at Urbana-Champaign
dlarkin@uiuc.edu

Dr. Michael Lassig
University of Cologne
lassig@thp.uni-koeln.de

Dr. Patrick Law Tik Wan
Core Facilities
patricklaw@cuhk.edu.hk

Dr. Michael Lawrence
Broad Institute
lawrence@broadinstitute.org

Dr. Wan-Ping Lee
Boston College
leeamv@bc.edu

Dr. Young-Ae Lee
Charité
yolee@mdc-berlin.de

Dr. Alison Lee
Institute of Molecular and Cell Biology
alee@imcb.a-star.edu.sg

Dr. Boris Lenhard
University of Bergen
boris.lenhard@bccs.uib.no

Dr. Wen Fung Leong
Boston College
Leongwe@bc.edu

Mr. Louis Letourneau
Genome Quebec
louis.letourneau@mail.mcgill.ca

Prof. Harris Lewin
Univerisy of Illinois at Urbana Champaign
h-lewin@igb.uiuc.edu

Dr. Mingyao Li
University of Pennsylvania
mingyao@mail.med.upenn.edu

Mr. Yingrui Li
Beijing Genomics Institue, Shenzhen
wangweiwei@genomics.org.cn

Dr. Jun Li
University of Michigan
junzli@umich.edu

Prof. Yun Li
University of North Carolina
yunpersonal@gmail.com

Ms. Jian Li
Baylor College of Medicine
jianl@bcm.edu

Dr. Svetlana Limborska
Institute of Molecular Genetics, RAS
limbor@img.ras.ru

Dr. Yin Lin
UCSD
yclin@ucsd.edu

Dr. Gabriella Lindgren
Swedish University of Agricultural Sciences
gabriella.lindgren@hgen.slu.se

Dr. Ryan Lister
The Salk Institute for Biological Studies
lister@salk.edu

Dr. George Liu
USDA-ARS
George.Liu@ars.usda.gov

Dr. George Livi
GlaxoSmithKline
george.p.livi@gsk.com

Dr. Devin Locke
Washington University School of Medicine
dlocke@watson.wustl.edu

Mr. Benjamin Logsdon
Cornell University
bal47@cornell.edu

Dr. Ari Löytynoja
EMBL-European Bioinformatics Institute
ari@ebi.ac.uk

Dr. Xuemei Lu
Beijing Institute of Genomics
xuemeilu@gmail.com

Dr. Gerton Lunter
University of Oxford
gerton.lunter@dpag.ox.ac.uk

Dr. Robert Lyle
Oslo University Hospital
Robert.Lyle@medisin.uio.no

Dr. Daniel MacArthur
Wellcome Trust Sanger Institute
dm8@sanger.ac.uk

Dr. Aaron Mackey
University of Virginia
amackey@virginia.edu

Dr. Quino Maduro
NIH
gracechu@mail.nih.gov

Dr. Jared Maguire
The Broad Institute of MIT and Harvard
jmaguire@broad.mit.edu

Dr. Paul Magwene
Duke University
paul.magwene@duke.edu

Dr. Ana-Teresa Maia
University of Cambridge
ana-teresa.maia@cancer.org.uk

Dr. Thomas Mailund
Aarhus University
mailund@birc.au.dk

Dr. Jacek Majewski
McGill University
jacek.majewski@mcgill.ca

Dr. Robert Majovski
Genome Research
majovski@cshl.edu

Dr. Kateryna Makova
Penn State University
kdm16@psu.edu

Dr. Joel Malek
Weill Cornell Medical College in Doha
jom2042@qatar-med.cornell.edu

Dr. Teri Manolio
NIH
manoliot@mail.nih.gov

Dr. Elaine Mardis
Washington University School of Medicine
emardis@wustl.edu

Dr. Elliott Margulies
National Institutes of Health
elliott@nhgri.nih.gov

Dr. John Marioni
University of Chicago
marioni@uchicago.edu

Dr. Tomas Marques-Bonet
University of Washington
tmarques@uw.edu

Prof. Gabor Marth
Boston College
marth@bc.edu

Dr. Andrew McCallion
Johns Hopkins University School of Medicine
andy@jhmi.edu

Dr. William McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Dr. Jean McEwen
National Institutes of Health
jean.mcewen@nih.hhs.gov

Mr. Cory McLean
Stanford University
cmclean@stanford.edu

Dr. Joel McManus
University of Connecticut Health Center
jmcmanus@uchc.edu

Dr. Megan McNerney
University of Chicago
megan.mcnerney@uchospitals.edu

Dr. John McPherson
Ontario Institute for Cancer Research
john.mcpherson@oicr.on.ca

Dr. Daniël Melters
University of California, Davis
dpmelters@ucdavis.edu

Dr. Kerstin Meyer
Cancer Research UK
Kerstin.Meyer@cancer.org.uk

Dr. Jason Mezey
Cornell University
jgm45@cornell.edu

Mr. Christopher Miller
Baylor College of Medicine
camiller@bcm.edu

Dr. Ryan Mills
Harvard Medical School - BWH
remills@partners.org

Dr. Aleksandar Milosavljevic
Baylor College of Medicine
amilosav@bcm.edu

Ms. Kay Minn
Mayo Clinic College of Medicine
minn.kay@mayo.edu

Dr. Robi Mitra
Washington University
rmitra@genetics.wustl.edu

Dr. Makedonka Mitreva
Washington University
mmitreva@watson.wustl.edu

Dr. Zora Modrusan
Genentech
modrusan.zora@gene.com

Dr. Alexandre Montpetit
Genome Quebec
alexandre.montpetit@mail.mcgill.ca

Dr. Barry Moore
University of Utah
barry.moore@genetics.utah.edu

Dr. Michael Morgan
Genome Canada
mmorgan@genomecanada.ca

Prof. Shinichi Morishita
University of Tokyo
moris@cb.k.u-tokyo.ac.jp

Dr. Katherine Morley
Wellcome Trust Sanger Institute
km5@sanger.ac.uk

Prof. Leonid Moroz
University of Florida
moroz@whitney.ufl.edu

Dr. Ali Mortazavi
California Institute of Technology
alim@caltech.edu

Ms. Xinmeng Mu
Yale University
xinmengmu@gmail.com

Dr. Jonathan Mudge
The Wellcome Trust Sanger Institute
jm12@sanger.ac.uk

Dr. Loris Mularoni
Johns Hopkins University School of Medicine
lmularo1@jhmi.edu

Dr. Jim Mullikin
NHGRI/NIH
mullikin@mail.nih.gov

Dr. Kasper Munch
Aarhus University
kaspermunch@birc.au.dk

Dr. Donna Muzny
Baylor College of Medicine
donnam@bcm.edu

Ms. Rachel Myers
North Carolina State Unveristy
ramyers@ncsu.edu

Dr. Richard Myers
HudsonAlpha Institute for Biotechnology
rmyers@hudsonalpha.org

Mr. Nicholas Navin
Cold Spring Harbor Laboratory
navin@cshl.edu

Dr. Anton Nekrutenko
Penn State University
anton@bx.psu.edu

Ms. Alexandra Nica
University of Geneva Medical School
an2@sanger.ac.uk

Dr. Rasmus Nielsen
University of California, Berkeley
rasmus@binf.ku.dk

Dr. Koji Numata
RIKEN, BRC
numata@rtc.riken.jp

Dr. Chad Nusbaum
Broad Institute of MIT and Harvard
chad@broad.mit.edu

Dr. Carole Ober
University of Chicago
c-ober@genetics.uchicago.edu

Dr. Ronan O'Malley
The Salk Insititute for Biological Studies
omalley@salk.edu

Dr. Larsson Omberg
Cornell University
lgo2@cornell.edu

Dr. Brian O'Roak
University of Washington
oroak@uw.edu

Dr. Kenshiro Oshima
University of Tokyo
oshima@cb.k.u-tokyo.ac.jp

Mr. Omead Ostadan
Illumina, Inc.
oostadan@illumina.com

Mr. Francis Ouellette
Ontario Centre for Cancer Research
Francis@oicr.on.ca

Dr. Bradley Ozenberger
National Institutes of Health
bozenberger@mail.nih.gov

Dr. Svante Paabo
Max-Planck-Institute
paabo@eva.mpg.de

Prof. Lior Pachter
UC Berkeley
lpachter@math.berkeley.edu

Ms. Heidi Pagan
Mississippi State University
heidijtp@gmail.com

Prof. Aarno Palotie
Wellcome Trust Sanger Institute
ap8@sanger.ac.uk

Mr. Chungoo Park
The Penn State University
cxp440@psu.edu

Dr. Steve Parker
NIH
parkerst@mail.nih.gov

Dr. Brian Parker
University of Copenhagen
bparker@binf.ku.dk

Dr. Jennifer Parla
Cold Spring Harbor Laboratory
parla@cshl.edu

Mr. Leopold Parts
Wellcome Trust Sanger Institute
lp2@sanger.ac.uk

Dr. Tomi Pastinen
McGill U. and Genome QC Inn. Ctr.
tomi.pastinen@mcgill.ca

Mr. Dirk Paul
Wellcome Trust Sanger Institute
DP5@SANGER.AC.UK

Dr. Flo Pauli
HudsonAlpha Institute for Biotechnology
fpauli@hudsonalpha.org

Dr. William Pavan
NHGRI/NIH
bpavan@nhgri.nih.gov

Dr. Celia Payen
University of Washington
payen@u.washington.edu

Dr. Jakob Pedersen
University of Copenhagen
jsp@binf.ku.dk

Dr. Elizabeth Pennisi
Science Magazine
epennisi@aaas.org

Dr. Mihaela Pertea
University of Maryland
mpertea@umd.edu

Dr. Jane Peterson
National Human Genome Research Institute/NIH
petersoj@mail.nih.gov

Dr. Joseph Petrosino
Baylor College of Medicine
jpetrosi@bcm.edu

Dr. Dmitri Petrov
Stanford University
dpetrov@stanford.edu

Mr. Doug Phanstiel
University of Wisconsin, Madison
phanstiel@wisc.edu

Mr. Joseph Pickrell
University of Chicago
pickrell@uchicago.edu

Dr. Olli Pietilainen
Institute for Molecular Medicine Finland FIMM
olli.pietilainen@thl.fi

Mr. Roy Platt
Mississippi State University
neal.platt@gmail.com

Dr. Katherine Pollard
University of California, Davis
kspollard@ucdavis.edu

Dr. David Pollock
UC Denver School of Medicine
David.Pollock@ucdenver.edu

Prof. Christopher Ponting
UK Medical Research Council, University of Oxford
chris.ponting@dpag.ox.ac.uk

Mr. Ryan Poplin
The Broad Institute
petersen@broadinstitute.org

Dr. Rudy Pozzatti
National Institutes of Health
pozzattr@mail.nih.gov

Dr. Aparna Prasad
The Hospital for Sick Children
aprasad@sickkids.ca

Dr. Jonathan Pritchard
University of Chicago
pritch@uchicago.edu

Dr. Ludmila Prokunina-Olsson
NCI, NIH
Prokuninal@mail.nih.gov

Mr. Kay Pruefer
Max Planck Institute for Evolutionary Anthropology
pruefer@eva.mpg.de

Dr. Michael Purugganan
New York University
mp132@nyu.edu

Ms. Wei Qu
University of Tokyo
quwei@cb.k.u-tokyo.ac.jp

Dr. Raquel Rabionet
Center for Genomic Regulation (CRG)
kelly.rabionet@crg.es

Dr. Raja Ragupathy
Agriculture and Agri-Food Canada
rajaragupathy@gmail.com

Dr. Koustubh Ranade
Genentech
ranade.koustubh@gene.com

Dr. Ben Raphael
Brown University
braphael@brown.edu

Dr. Gunnar Ratsch
Friedrich Miescher Laboratory, Max Planck Society
Gunnar.Raetsch@tuebingen.mpg.de

Dr. Muthuswamy Raveendran
Baylor College of Medicine
raveendr@bcm.edu

Dr. David Ray
Mississippi State University
dray@bch.msstate.edu

Dr. Chris Raymond
NuGEN Technologies, Inc.
craymond@nugeninc.com

Dr. Timothy Reddy
HudsonAlpha Institute for Biotechnology
treddy@hudsonalpha.org

Dr. Martin Reese
Omicia Inc.
mreese@omicia.com

Dr. David Reich
Harvard Medical School
reich@genetics.med.harvard.edu

Dr. Jeffrey Reid
Baylor College of Medicine
jgreid@bcm.edu

Dr. Stephen Richards
Baylor College of Medicine
stephenr@bcm.edu

Dr. Samantha Riesenfeld
Gladstone Institutes & UCSF
samantha.riesenfeld@gladstone.ucsf.edu

Dr. John Rinn
Broad Institute of MIT and Harvard Medical School
jrinn@bidmc.harvard.edu

Dr. Patrizia Rizzu
VU university Medical Center
p.rizzu@vumc.nl

Dr. Juan Rodriguez-Flores
University of California at San Diego
juan@ucsd.edu

Ms. Yu-Hui Rogers
The J. Craig Venter Science Foundation
yrogers@jcvi.org

Prof. Jeffrey Rogers
Baylor College of Medicine
jr13@bcm.tmc.edu

Dr. Jane Rogers
The Genome Analysis Centre
jane.rogers@bbsrc.ac.uk

Mr. Michael Rosen
Stanford University
mjrosen@stanford.edu

Dr. Jeffrey Rosenfeld
Zucker Hillside Hospital
jrosenfeld@nshs.edu

Dr. Jonathan Rothberg
Ion Torrent
lstevens@iontorrent.com

Dr. Steve Rozen
Duke-NUS Graduate Medical School Singapore
steve.rozen@duke-nus.edu.sg

Dr. Joel Rozowsky
Yale University
joel.rozowsky@yale.edu

Dr. Yijun Ruan
Genome Institute of Singapore
ruanyj@gis.a-star.edu.sg

Dr. Alfredo Ruiz
Universitat Autònoma de Barcelona
Alfredo.Ruiz@uab.cat

Dr. Carsten Russ
Broad Institute of MIT and Harvard
petersen@broadinstitute.org

Dr. Pardis Sabeti
Harvard University
psabeti@oeb.harvard.edu

Dr. Matthew Sachs
Texas A&M University
msachs@mail.bio.tamu.edu

Dr. Taro Saito
University of Tokyo
leo@cb.k.u-tokyo.ac.jp

Ms. Elina Salmela
University of Helsinki
elina.t.salmela@helsinki.fi

Prof. Steven Salzberg
University of Maryland
salzberg@umiacs.umd.edu

Dr. Paul Samollow
Texas A&M Universtiy
psamollow@cvm.tamu.edu

Dr. Joel Savard
Cell Press (Trends in Genetics)
jsavard@cell.com

Dr. Peter Scacheri
Case Western Reserve University
pxs183@case.edu

Dr. Aylwyn Scally
Wellcome Trust Sanger Institute
as6@sanger.ac.uk

Dr. Stephen Schaffner
Broad Institute
sfs@broadinstitute.org

Mr. Michael Schatz
University of Maryland
mschatz@umiacs.umd.edu

Prof. Mikkel Schierup
Aarhus University
mheide@birc.au.dk

Mr. Dominic Schmidt
University of Cambridge
dominic.schmidt@cancer.org.uk

Dr. Bob Schmitz
The Salk Institute for Biological Studies
rschmitz@salk.edu

Dr. Valerie Schneider
NLM/NCBI
schneiva@ncbi.nlm.nih.gov

Mr. Timothy Schramm
University of Wisconsin - Madison
schramm@chem.wisc.edu

Prof. David Schwartz
University of Wisconsin - Madison
dcschwartz@wisc.edu

Dr. Erich Schwarz
California Institute of Technology
emsch@its.caltech.edu

Dr. Laura Scott
University of Michigan
ljst@umich.edu

Dr. Will Seabrook
National Intsitutes of Health

Dr. Stephen Searle
The Wellcome Trust Sanger Inst.
searle@sanger.ac.uk

Dr. Darren Segale
Illumina
dsegale@illumina.com

Dr. David Serre
Cleveland Clinic Foundation
serred@ccf.org

Dr. Tamim Shaikh
University of Colorado
tamim.shaikh@ucdenver.edu

Mr. B. Jesse Shapiro
Massachusetts Institute of Technology
jesse1@mit.edu

Dr. Andrew Sharp
Mount Sinai School of Medicine
andrew.sharp@mssm.edu

Ms. Tal Shay
Broad Institute/MIT
talshay@broadinstitute.org

Dr. Yufeng Shen
Columbia University
yshen@c2b2.columbia.edu

Dr. Jay Shendure
University of Washington
shendure@u.washington.edu

Ms. Vrunda Sheth
Life Technologies
vrunda.sheth@lifetech.com

Dr. Robert Shields
Public Library of Science
rshields@plos.org

Ms. So Youn Shin
Wellcome Trust Sanger Institute
ss22@sanger.ac.uk

Dr. Ilya Shlyakhter
Harvard University
ilya_shl@alum.mit.edu

Dr. Asim Siddiqui
Life Technologies
asim.siddiqui@lifetech.com

Dr. Arend Sidow
Stanford University School of Medicine
arend@stanford.edu

Prof. Adam Siepel
Cornell University
acs4@cornell.edu

Ms. Elizabeth Siewert
University of Colorado Denver
siewertb@gmail.com

Dr. Snaevar Sigurdsson
The Broad Institute
petersen@broadinstitute.org

Mr. Shripad Sinari
The Translational Genomes Research Institute
ssinari@tgen.org

Dr. Suzanne Sindi
Brown University
Suzanne_Sindi@Brown.edu

Dr. Magdalena Skipper
Nature
m.skipper@nature.com

Dr. Kerrin Small
Kings College London
kerrin.small@kcl.ac.uk

Mr. Jeremy Smith
Mississippi State University
auverus@gmail.com

Dr. Jeramiah Smith
Benaroya Research Institute
jsmith@benaroyaresearch.org

Dr. Michael Snyder
Stanford University School of Medicine
michael.snyder@yale.edu; mpsnyder@stanford.edu

Dr. Erica Sodergren
Washington University Medical School
esodergr@watson.wustl.edu

Dr. Nicole Soranzo
Wellcome Trust Sanger Institute
ns6@sanger.ac.uk

Prof. Tim Spector
King's College London
tim.spector@kcl.ac.uk

Dr. John Stamatoyannopoulos
University of Washington
jstam@uw.edu

Dr. Frank Steemers
Illumina, Inc
fsteemers@illumina.com

Dr. Arnold Stein
Purdue University
steina@purdue.edu

Dr. Derek Stemple
Wellcome Trust Sanger Institute
ds4@sanger.ac.uk

Dr. Chip Stewart
Boston College
chip.stewart@bc.edu

Dr. Jennifer Stone
Illumina
jstone@illumina.com

Dr. Eric Stone
North Carolina State University
eric_stone@ncsu.edu

Mr. Peter Sudmant
University of Washington
psudmant@uw.edu

Dr. Eun-Kyung (Anita) Suk
Max Planck Institute for Molecular Genetics
suk@molgen.mpg.de

Mr. James Sun
Harvard Medical School / MIT
xinsun@mit.edu

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Dr. Yutaka Suzuki
University of Tokyo
ysuzuki@k.u-tokyo.ac.jp

Dr. David Symer
Ohio State University Comprehensive Cancer Center
david.symer@osumc.edu

Dr. Michael Talkowski
Harvard Medical School / MGH
talkowski@chgr.mgh.harvard.edu

Mr. Kousuke Tanimoto
The University of Tokyo
kk077619@ims.u-tokyo.ac.jp

Dr. Alice Tay
Institute of Molecular and Cell Biology
mcbalice@imcb.a-star.edu.sg

Dr. Todd Taylor
RIKEN Advanced Science Institute
taylor@riken.jp

Dr. James Taylor
Emory University
james@jamestaylor.org

Dr. James Thomas
Emory University
jthomas@genetics.emory.edu

Dr. Daryl Thomas
DNAnexus
daryljthomas@gmail.com

Mr. Scott Topper
University of Wisconsin
stopper@wisc.edu

Mr. Cole Trapnell
University of Maryland
cole@cs.umd.edu

Dr. Lisa Trevino
Baylor College of Medicine
lt2@bcm.edu

Mr. Michael Tschannen
Medical College of Wisconsin
mtschann@mcw.edu

Dr. Brian Tuch
Applied Biosystems
brian.tuch@appliedbiosystems.com

Prof. Tom Tullius
Boston University
tullius@chem.bu.edu

Dr. Steve Turner
Pacific Biosciences, Inc.
sturner@pacificbiosciences.com

Dr. Anton Valouev
Stanford University
valouev@stanford.edu

Dr. Venky Venkatesh
Monsanto
t.v.venkatesh@monsanto.com

Prof. B. Venkatesh
Institute of Molecular and Cell Biology
mcbbv@imcb.a-star.edu.sg

Mr. Oliver Venn
Oxford University
oliver.venn@queens.ox.ac.uk

Dr. Albert Vilella
EMBL-EBI
ksmith@ebi.ac.uk

Dr. Sumiti Vinayak
Centers for Disease Control & Prevention (CDC)
gvk2@cdc.gov

Dr. Axel Visel
Lawrence Berkeley National Laboratory
avisel@lbl.gov

Dr. Claes Wadelius
Uppsala University
Claes.Wadelius@genpat.uu.se

Dr. Jeffrey Wall
UCSF
wallj@humgen.ucsf.edu

Dr. Klaudia Walter
Wellcome Trust Sanger Institute
kw8@sanger.ac.uk

Dr. Jinhua Wang
NYU Cancer Institute
jinhua.wang@nyumc.org

Dr. Lu Wang
NIH/NHGRI
wanglu@mail.nih.gov

Dr. JUN WANG
Beijing Genomics Institute at Shenzhen
wangj@genomics.org.cn

Ms. Jing Wang
Yale University
cindy.wj0821@yahoo.com

Mr. Xu Wang
Cornell University
xw54@cornell.edu

Dr. Ting Wang
Washington University
twang@genetics.wustl.edu

Dr. Kai Wang
Children's Hospital of Philadelphia
kaichop@gmail.com

Dr. Wenyi Wang
Stanford University
wenyiw@stanford.edu

Dr. Liguo Wang
Baylor College of Medicine
liguow@bcm.edu

Dr. Alistair Ward
Boston College
AlistairNWard@gmail.com

Dr. Doreen Ware
CSHL USDA-ARS
ware@cshl.edu

Mr. Lukas Wartman
Washington University School of Medicine
wartmanl@WUSTL.EDU

Dr. Daniel Weaver
Bee Power, LP
dbeeweaver@gmail.com

Dr. Wu Wei
EMBL
wuwei@embl.de

Dr. George Weinstock
Washington University School of Medicine
geowei@mac.com

Dr. Jens Wendland
National Institutes of Health
wendlandj@mail.nih.gov

Prof. Lorenz Wernisch
Medical Research Council
lorenz.wernisch@mrc-bsu.cam.ac.uk

Ms. Kris Wetterstrand
National Human Genome Research Institute/NIH
wettersk@mail.nih.gov

Dr. Sarah Wheelan
The Johns Hopkins University School of Medicine
swheelan@jhmi.edu

Dr. David Wheeler
Baylor College of Medicine
wheeler@bcm.tmc.edu

Mr. Nava Whiteford
Oxford Nanopore Technologies Ltd
sidra.moazzam@nanoporetech.com

Ms. Maria Wilbe
Swedish University of Agricultural Sciences (SLU)
Maria.Wilbe@hgen.slu.se

Dr. Louise Williams
Broad Institute of MIT and Harvard
williams@broad.mit.edu

Dr. Richard Wilson
Washington University
rwilson@wustl.edu

Dr. James Wilson
University of Mississippi Medical Center
james.wilson1@med.va.gov

Ms. Melissa Wilson Sayres
The Pennsylvania State University
maw397@psu.edu

Ms. Deborah Winter
Duke University
deborah.winter@duke.edu

Dr. Susan Wolf
University of Minnesota Law School
swolf@umn.edu

Dr. Jamison Wolfer
UW Madison - GSTP
wolfer@wisc.edu

Dr. Kim Worley
Baylor College of Medicine
kworley@bcm.edu

Dr. Jennifer Wortman
University of Maryland School of Medicine
jwortman@som.umaryland.edu

Mr. Jiantao Wu
Boston College
wuvw@bc.edu

Dr. Chunlin Xiao
NIH
xiao2@mail.nih.gov

Dr. Yurong Xin
Columbia University
xinyuro@pi.cpmc.columbia.edu

Dr. Zhenyu Xuan
University of Texas at Dallas
zhenyu.xuan@utdallas.edu

Mr. Xu Xun
Beijing Genomics Institute of Shenzhen
xuxun@genomics.org.cn

Prof. Huanming Yang
yanghm@genomics.org.cn

Mr. Jianchao Yao
Cold Spring Harbor Laboratory
jyao@cshl.edu

Ms. Moran Yassour
The Hebrew University
moran@cs.huji.ac.il

Dr. Kai Ye
Leiden University Medical Center
k.ye@lumc.nl

Ms. GEOK YEO
The Chinese University of Hong Kong
yenyeung@cuhk.edu.hk

Dr. Yong Yin
Monsanto Company
yong.yin@monsanto.com

Dr. Fuli Yu
Baylor College of Medicine
fyu@bcm.edu

Dr. Bingbing Yuan
Whitehead Institute for biological research
byuan@wi.mit.edu

Dr. Laura Zahn
AAAS
lzahn@aaas.org

Prof. Evgeny Zdobnov
University of Geneva
Evgeny.Zdobnov@unige.ch

Dr. Xinmin Zhang
Roche NimbleGen
xinmin.zhang@roche.com

Dr. Xiuqing Zhang
BGI-Shenzhen
zhangxq@genomics.org.cn

Dr. Zhengdong Zhang
Yale University
zhengdong.zhang@yale.edu

Dr. Deyou Zheng
Albert Einstein College of Medicine
deyou.zheng@einstein.yu.edu

Dr. Holly Zheng-Bradley
European Bioinformatics Institute
zheng@ebi.ac.uk

Dr. Degui Zhi
University of Alabama, Birmingham
zhi.degui@gmail.com

Dr. Martine Zilversmit
Université de Montréal
zmartine@gmail.com

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers
BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers
Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
Helpful tips - Obtain PIN from Meetings & Courses Office to enter Library after hours. See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail only
Lower level: Word processing and printing.
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar
Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, west wing, lower level
PIN#: **Press 64350 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($8.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)
Executive Limo	631-696-8000 (**1047**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322