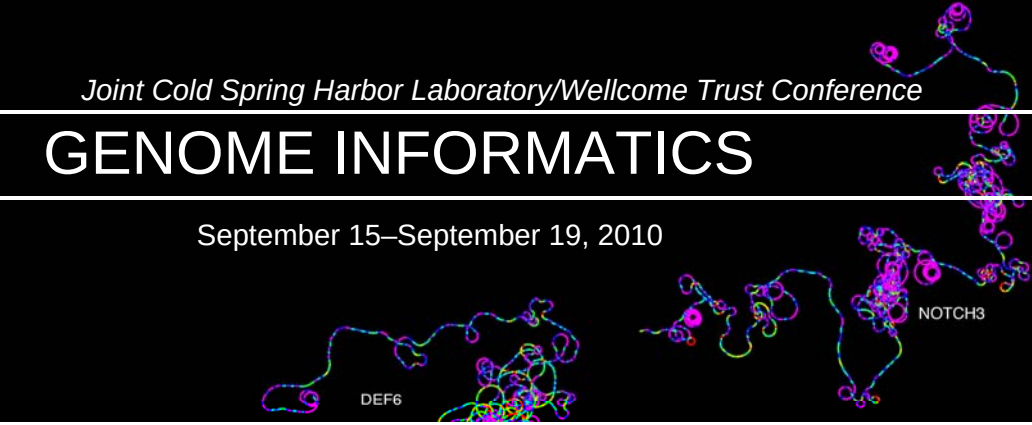
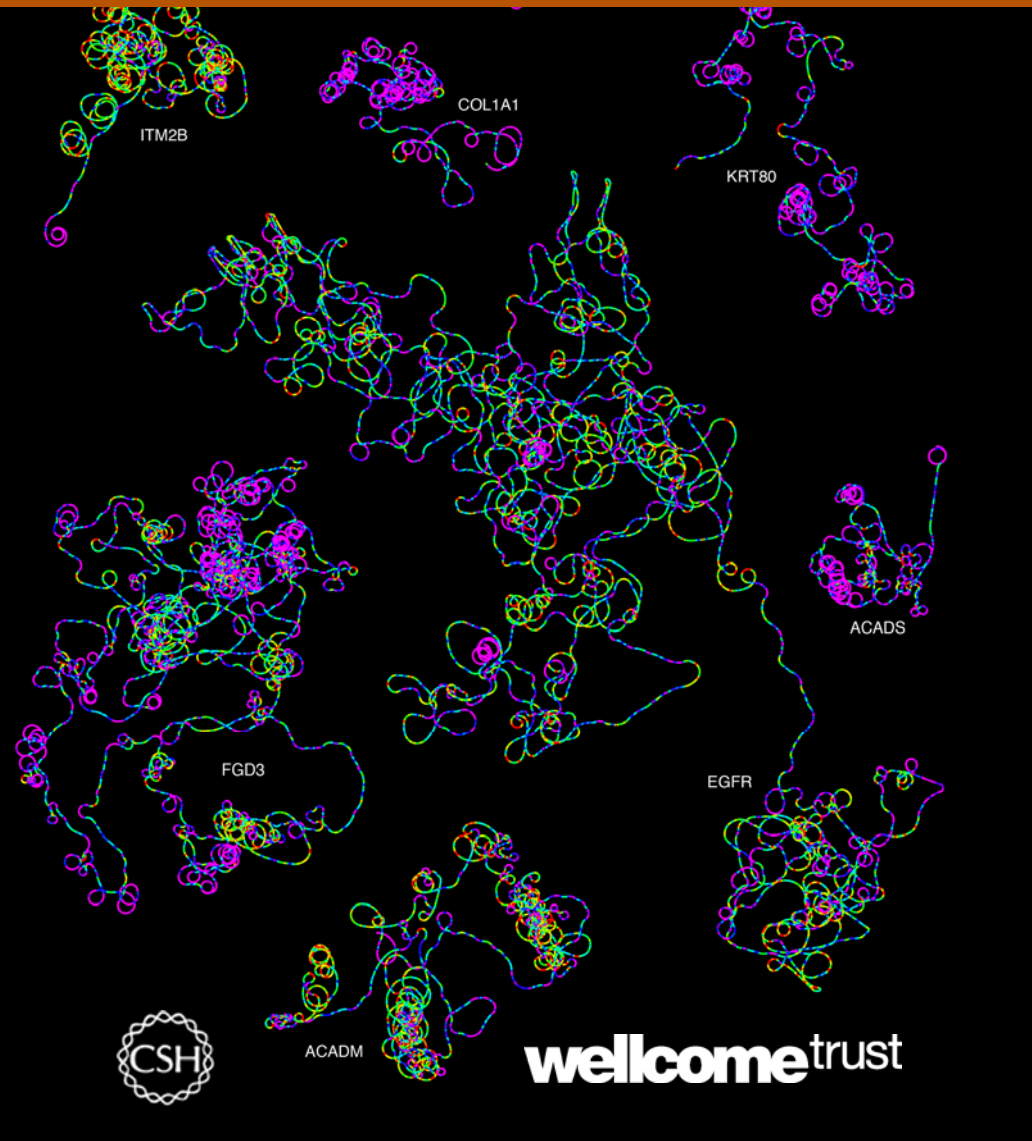
Joint Cold Spring Harbor Laboratory/Wellcome Trust Conference

GENOME INFORMATICS

September 15–September 19, 2010





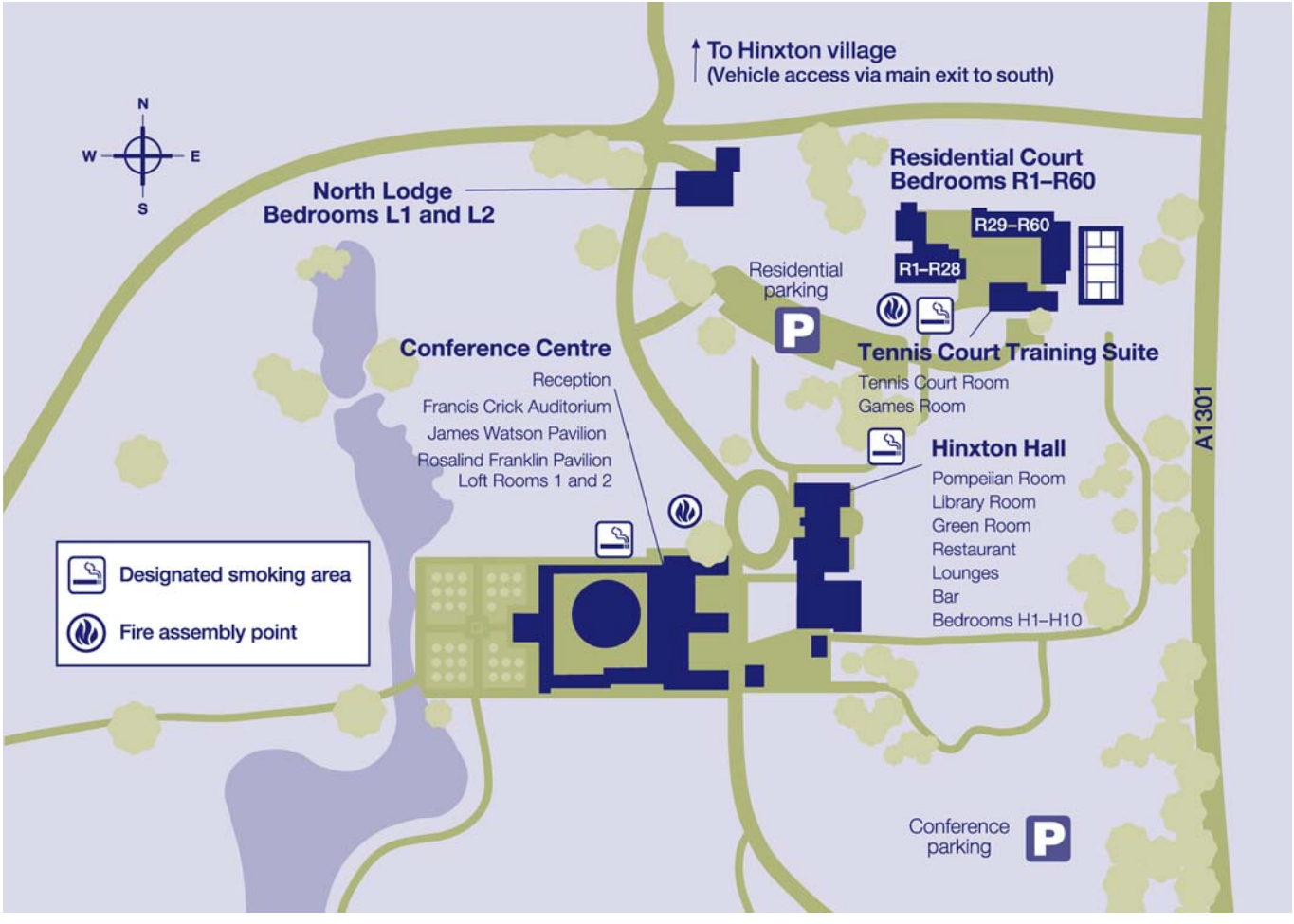
Joint Cold Spring Harbor Laboratory/Wellcome Trust Conference

GENOME INFORMATICS

September 15–September 19, 2010

Arranged by

Inanc Birol, *BC Cancer Agency, Canada*
Michele Clamp, *BioTeam, Inc.*
James Kent, *University of California, Santa Cruz, USA*


To Hinxton village
(Vehicle access via main exit to south)
N
W
E
S
North Lodge
Bedrooms L1 and L2
Residential Court
Bedrooms R1–R60
R29–R60
R1–R28
Residential
parking
P
Conference Centre
Reception
Francis Crick Auditorium
James Watson Pavilion
Rosalind Franklin Pavilion
Loft Rooms 1 and 2
Tennis Court Training Suite
Tennis Court Room
Games Room
Hinxton Hall
Pompeiian Room
Library Room
Green Room
Restaurant
Lounges
Bar
Bedrooms H1–H10
A1301
Designated smoking area
Fire assembly point
Conference
parking
P

SCHEDULE AT A GLANCE

Wednesday 15th September 2010

17.00-17.30	Registration – finger buffet dinner served from 17.30-19.30
19.30-20:50	Session 1: Epigenomics and Gene Regulation
20.50-21.10	Break
21.10-22.30	Session 1, continued

Thursday 16th September 2010

07.30-09.00	Breakfast
09.00-10.20	Session 2: Population and Statistical Genomics
10.20-10:40	Morning Coffee
10:40-12:00	Session 2, continued
12.00-14.00	Lunch
14.00-15.20	Session 3: Environmental and Medical Genomics
15.20-15.40	Break
15.40-17.00	Session 3, continued
17.00-19.00	Poster Session I and Drinks Reception
19.00-21.00	Dinner

Friday 17th September 2010

07.30-09.00	Breakfast
09.00-10.20	Session 4: Databases, Data Mining, Visualization and Curation
10.20-10.40	Morning Coffee
10.40-12.00	Session 4, continued
12.00-14.00	Lunch
14.00-16.00	Free afternoon
16.00-17.00	Keynote Speaker: Alex Bateman
17.00-19.00	Poster Session II and Drinks Reception
19.00-21.00	Dinner

Saturday 18th September 2010

07.30-09.00	Breakfast
09.00-10.20	Session 5: Sequencing Pipelines and Assembly
10.20-10.40	Morning Coffee
10.40-12.00	Session 5, continued
12.00-14.00	Lunch
14.00-15.20	Session 6: Comparative and Evolutionary Genomics
15.20-15.40	Break
15.40-17.00	Session 6, continued
17.00-19.00	Pre Dinner Drinks
19.00-21.00	Conference Dinner

Sunday 19th September 2010

07.30-09.00	Breakfast
09.00-10.20	Session 6: RNA Sequencing and Gene Prediction
10.20-10.40	Morning Coffee
10.40-12.00	Session 6, continued
12.00-14.00	Lunch
14.00	Return Transfers to Cambridge and LHR & STN Airports

General Information

Welcome to the Wellcome Trust Conference Centre and the **Genome Informatics** Conference.

Conference Badges
Please wear your name badge at all times to promote networking and to assist staff in identifying you.

If you have advised us of any dietary requirements, you will find a small coloured dot on your badge. Please make yourself known to the catering team and they will assist you with your meal request.

Scientific Session Protocol
Photography, audio or video recording of the scientific sessions is not permitted.

Internet Access
Wireless internet access is available throughout the campus. Please inquire at reception for a Wireless Connection token.

Presentations
If you are an invited speaker or your abstract has been selected for an oral presentation, please provide an electronic copy of you talk to a member of the AV team who will be based in the auditorium.

Poster Sessions
Posters will be displayed throughout the conference. Please post your materials in the Cloisters on arrival. The abstract page number indicates to the assigned poster board number

Social Events
Thursday, 16 September – A drinks reception will take place in the Conference Centre Cloisters from 17.00 during poster session I.

Friday, 17 September - A drinks reception will take place in the Conference Centre Cloisters from 17.00 during poster session II.

Saturday, 18 September - Pre dinner drinks will take place in the Conference Centre Cloisters from 17.00.

Conference Meals
All meals will be served in the Hall Restaurant. Please refer to the conference programme in this book as times will vary based on the daily scientific presentations.
All conference meals and social events are for registered delegates only.

For Wellcome Trust Conference Centre Guests
Check in
If you are staying on site at the Wellcome Trust you may check into your room from 2.00pm. If you plan to arrive late at night you can check into your room as the conference centre reception is open 24 hours. Please note there will be no lunch or dinner facilities available outside of the conference timetable; however, there is a local public house (The Red Lion), serving both lunch and evening meals, located just 2 minute walk from the campus in the village of Hinxton.

Breakfast
Your breakfast will be served in the Hall restaurant from 07.30 – 09.00 every morning.

Telephone
If you are staying on-site and would like to use the telephone in your room, you will need to contact the Reception desk (Ext. 5000) to have your phone line activated - we will require your credit card number and expiry date to do so.

Departures
You must vacate your room by 10.00 on the day of your departure. Please ask at reception for assistance with luggage storage in the Conference Centre.

Homerton College Guests
Check in
If you are staying at Homerton College in Cambridge, you are able to check into your room from 14.00

Breakfast
Your breakfast will be served at the Homerton College from 07.30-08.15

You must vacate your room by 09.00 on the day of your departure. A luggage store is available in the Conference Centre please ask at the reception.

Complimentary transfers have been arranged to and from the Conference Centre, please see below for times.

Transfers
Wednesday 15th September 2010
22.45 Conference Centre – Homerton College

Thursday 16th September 2010
08.15 Homerton College – Conference Centre
21.30 Conference Centre – Homerton College

Friday 17th September 2010
08.15 Homerton College – Conference Centre
21.30 Conference Centre – Homerton College

Saturday 18th September 2010
08.15 Homerton College – Conference Centre
21.30 Conference Centre – Homerton College

Sunday 19th September 2010
08.15 Homerton College – Conference Centre

Return Ground Transportation
Complimentary return transportation to Heathrow, Stansted Airport and the Cambridge Train Station and City Centre have been arranged for 14.00 on Sunday, 19 September. Please note: a sign-up sheet will be available at the registration desk. Places are limited so you are advised to book early.

Taxis
Please find a list of local taxi numbers should you require:
Panther – 01223 715715
Mid Anglia - Tel: 01223 836000
Phil's Taxi Services - Tel: 01223 521918
A&M Carriages (Airport Specialist) - Tel: 01223 513703

Messages and Miscellaneous
All messages will be posted on the registration desk in the Conference Centre Foyer.

A number of toiletry and stationery items are available for purchase at the conference centre reception. Cards for our self-service laundry are also available.

If you have any queries or comments, please do not hesitate to contact a member of staff who will be pleased to help you.

Joint Cold Spring Harbor Laboratory/Wellcome Trust conferences at Hinxton are supported in part with funding courtesy of The Wellcome Trust.

These abstracts should not be cited in bibliographies. Material contained herein should be treated as personal communication and should be cited as such only with consent of the author.

Printed on 100% recycled paper.

Front and Back Covers: Martin Krzywinski, British Columbia Cancer Agency, Vancouver, Canada.

Sequences of genes mentioned in the list of abstracts of the conference, rendered as paths. For details, see http://mkweb.bcgsc.ca/genomeinfo2010.

PROGRAM

WEDNESDAY, September 15—7:30 PM

SESSION 1 SEQUENCING PIPELINES AND ASSEMBLY

Chairpersons: **Z. Torok,** Astrid Research, Debrecen, Hungary
I. Birol, BC Cancer Agency, Vancouver, Canada

THURSDAY, September 16—9:00 AM

SESSION 2 POPULATION AND STATISTICAL GENOMICS

Chairpersons: **R. Durbin,** Wellcome Trust Sanger Institute, Hinxton, UK
E. Margulies, NHGRI, National Institutes of Health, Bethesda, Maryland, USA

THURSDAY, September 16—2:00 PM

SESSION 3 ENVIRONMENTAL AND MEDICAL GENOMICS

Chairpersons: **T. Thorgeirsson,** University of California, Santa Cruz, USA
M. Clamp, BioTeam, Middleton, Massachusetts, USA

THURSDAY, September 16—5:00 PM

SESSION 4 POSTER SESSION I and DRINKS RECEPTION

FRIDAY, September 17—9:00 AM

SESSION 4 DATABASES, DATA MINING, VISUALIZATION AND CURATION

Chairpersons: **D. Dooling,** Washington University School of Medicine, St. Louis, Missouri, USA
M. Krzywinski, BC Cancer Agency, Vancouver, Canada

FRIDAY, September 17—4:00 PM

KEYNOTE SPEAKER

FRIDAY, September 17—5:00 PM

SESSION 5 POSTER SESSION II

SATURDAY, September 18—9:00 AM

SESSION 5 EPIGENOMICS AND GENE REGULATION

Chairpersons: **W. Noble,** University of Washington, Seattle, USA
Z. Weng, University of Massachusetts Medical School, Worcester, USA

SATURDAY, September 18—2:00 PM

SESSION 6 COMPARATIVE AND EVOLUTIONARY GENOMICS

Chairpersons: **S. Batzoglou,** Stanford University, California, USA
P. Flicek, European Bioinformatics Institute, Hinxton, UK

Retrotransposons in the orangutan (*Pongo*) lineage—A new evolutionary tale
Miriam K. Konkel, Jerilyn A. Walker, Brygg Ullmer, Leona G. Chemnick, Oliver A. Ryder, Robert Hubley, Arian F A. Smit, Mark A. Batzer.
Presenter affiliation: Louisiana State University, Baton Rouge, Louisiana. 154

Mining the allohexaploid wheat genome for useful sequence polymorphisms
Rachel Brenchley, Rosalinda D'Amore, Gary Barker, Keith Edwards, Michael Bevan, Anthony Hall, Neil Hall.
Presenter affiliation: University of Liverpool, Liverpool, United Kingdom. 155

Mapping the evolution of transcription factor binding
Dominic Schmidt, Michael Wilson, Benoit Ballester, Petra C. Schwalie, David Thybert, Klara Stefflova, Michelle Ward, Duncan Odom, Paul Flicek.
Presenter affiliation: European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom. 156

Analysis of the primate genomes using the 5-way EPO multiple alignments
Kathryn Beal, Stephen Fitzgerald, Paul Flicek, Javier Herrero.
Presenter affiliation: EBI, Hinxton, United Kingdom. 157

SnoWMAn—High throughput phylotyping, analysis and comparison of microbial communities
Gernot Stocker, Renè Snajder, Johannes Rainer, Slave Trajanoski, Gregor Gorkiewizc, Zlatko Trajanoski, Gerhard Thallinger.
Presenter affiliation: Innsbruck Medical University, Innsbruck, Austria; Graz University of Technology, Graz, Austria. 158

Comparative genomics of all 24 described species within the genus *Campylobacter*
William G. Miller, Craig T. Parker, Emma Yee, Robert E. Mandrell.
Presenter affiliation: USDA, Agricultural Research Service, Albany, California. 159

SATURDAY, September 18—7:00 PM

CONFERENCE DINNER

SESSION 7 SEQUENCING AND GENE PREDICTION

Chairpersons: **A. Mortazavi,** California Institute of Technology, Pasadena, USA
L. Pachter, University of California, Berkeley, USA

REFERENCE ASSEMBLY IN COLOUR SPACE

Balazs Gór [1], Anett Balla*[1], Edit Tukacs*[1], Miklós Laczik*[1], István Nagy*[2], Zsolt Török*[1,3]

[1]Astrid Research Inc., Debrecen, H-4029, Hungary, [2]Bay Zoltán Foundation for Applied Research, Institute for Plant Genomics, Human Biotechnology and Bioenergy, Szeged, H-6701, Hungary, [3]University of Debrecen, Medical and Health Science Center, Clinical Research Center, Debrecen, H-4010, Hungary

Researchers in the field of genomics have been facing new challenges ever since the next generation sequencing (NGS) technologies appeared. Handling the huge amount of outcoming data is a challenge on its own, but it is the complexity of data interpretation that is the primary problem. The unique characteristic of the Applied Biosystem's SOLiD NGS System is that its output consists of a dataset containing short reads with 35 to 50 nucleotides (also called tags). A tag is built of colour codes, in which two adjacent nucleotides define a single colour, subsequently, the nucleotide and the colour coded sequence is one to one correspondent. This unique, two base-encoding chemistry is the key in achieving >99.91% accuracy. Yet, most of the current reference assembly algorithms does not work in colour space, therefore the read errors occurring during sequencing cause difficulties in sequence alignment and identification of single or multi nucleotide polymorphisms (SNPs and MNPs, respectively) and insertions/deletions (indels).

Our algorithm aligns the raw output tags to the colour coded transformation of the reference sequence, and by doing so it detects polymorphisms and indels with the help of reads overlapping these structural modifications of the genome. In addition, the algorithm can also make use of reads containing several read errors in the assembly, thereby increasing the coverage. A further benefit of the algorithm is that the complementary translation of the reads is unnecessary owing to the assembly in colour space, which significantly shortens the run time.

We have used this algorithm to compare the genomes of 14 isolates of a skin commensal bacteria, for which a reference genome was available. Subsequent to the reference assembly we identified SNPs, MNPs, short indels and longer deletions that were not detected by other assembly software tools. Importantly, the majority of these structural variations could be validated by conventional molecular biology techniques, such as Sanger sequencing or PCR, emphasizing the power of the newly developed algorithm.

COLUMBUS: HYBRID *DE NOVO* AND MAPPED ASSEMBLY OF SHORT READ TRANSCRIPTOMIC OR GENOMIC DATA

Daniel R Zerbino[1], Wendy Lee[1], Chris Wilks[1], Mark Diekhans[1], Benedict J Paten[1], Zachary J Sanborn[1], Marcel H Schulz[2], David Haussler[1]

[1]University of California, Santa Cruz , Center for Biomolecular Science and Engineering, Santa Cruz, CA, 95060, [2] Max Planck Institute for Molecular Genetics, Department of Vertebrate Genomics, Berlin, 14195, Germany

Analysis methods to process short-read transcriptome data can generally be divided into two categories: on one hand, mapping based approaches rely on the reference genome sequence to cluster fragments into exons which are then assembled into transcripts. On the other, *de novo* assemblers use overlaps between the reads to construct contigs, which can then be aligned to a reference genome or to a database of known genes. Whereas mappers have the advantages of speed, specificity and robustness, they are unable to use the reads spanning even short novel sequences, such as rearrangement breakpoints or indels.

We developed *Columbus*, an extension to the *Velvet de novo* assembler, which uses mapping data when assembling the reads. Knowledge of the reference sequence and of the read placement allows Columbus to distinguish between identical repeat copies, and ignore spurious overlaps which were previously resolved by the full length mapping of the reads.

For example, used downstream of the *Cufflinks ab initio* transcriptome assembler, *Columbus* can extend and enrich the assemblies produced. It maintains the consistency of the reference-based assembly while integrating the previously discarded reads that are spanning indels or undetected exons.

We applied our pipeline to cancer transcriptome data, providing us with the ability to detect and directly observe alterations of the coding sequence such as fusion genes within a high-throughput framework. As a Genomic Data Analysis Center of the TCGA project, we are analyzing RNA-seq data for several hundred tumors from different cancer types. Such an approach can also be directly extended to comparative transcriptome assembly and to the reconstruction of structural variant breakpoints.

EFFICIENT OVERLAP-BASED ASSEMBLY METHODS AND APPLICATIONS TO SEQUENCE VARIATION DETECTION

Jared T Simpson, Richard Durbin

Wellcome Trust Sanger Institute, Informatics, Hinxton, CB10 1SA, United Kingdom

Previously, we developed a novel overlap-based assembly algorithm using the Burrows-Wheeler transform and the FM-index [ISMB, 2010]. Based on a compressed data structure, this method is both time and memory efficient and substantially lowers the computational requirements for overlap assembly, allowing overlap assembly of the very large data sets common to second generation sequencing experiments. To characterize the performance of our algorithm, we have performed an assembly of simulated, error-free, 100bp reads from the human genome at an average sequencing depth of 20X (57 Gbp of sequence data). The construction of the FM-index required 48GB of memory and 187 CPU-hours. To compute the overlaps used to construct the string graph required 50GB of memory and 245 CPU-hours. Finally, generating the assembled contigs from the string graph required 240GB of memory and 24 CPU-hours. The N50 value for the assembled contigs was 2.3Kbp.

We have extended our approach to real data sets, accounting for sequencing errors in the reads by allowing inexact overlaps between reads. Our method first performs error correction on the entire read set in parallel using the FM-index without requiring the construction of the entire assembly graph, yielding a space and time efficient correction algorithm. The assembly string graph is then constructed from the corrected reads and the graph is processed to remove erroneous vertices and edges arising from uncorrectable sequencing errors. We demonstrate the effectiveness of our error correction algorithm on real and simulated data. Further, we show that correcting reads using overlaps improves the quality of the resulting assembly in both the string graph and de Bruijn frameworks.

We will also discuss the application of our assembly method to reference-free variation detection in populations. We are exploring the joint assembly of multiple strains of *Saccharomyces cerevisiae* to capture the variation present between strains in the assembly graph. Our view is that assembly-based methods of variation detection will become increasingly important as sequencing technology advances and read lengths increase.

SEQUENCING AND ANALYSES OF THE HEXAPLOID WHEAT CHROMOSOME 3B

Frédéric Choulet[1], Etienne Paux[1], Philippe Leroy[1], Arnaud Couloux[2], Michael Alaux[3], Hadi Quesneville[3], Patrick Wincker[2], Catherine Feuillet[1]

[1]INRA, UMR 1095 GDEC, Clermont-Ferrand, 63100, France, [2]Genoscope, Institut de Génomique, CEA, Evry, 91057, France, [3]INRA, UR 1164 URGI, Versailles, 78026, France

Because of its large (17 Gb, 5x the human genome and 40x the one of rice), polyploid (3 homoeologous A-, B- and D-genomes within a same nucleus) and highly repetitive (>80% of DNA corresponding to transposable elements) genome, the development of wheat genomics has been lagging behind the one of the other major crops. Two years after the establishment of the first physical map of the biggest wheat chromosome, the 3B, which represents 1 Gb (Paux et al. Science 2008), its complete sequencing is now underway (ANR project 3BSEQ) by combining Roche 454 sequencing of pools of contiguous BACs and Whole Chromosome Shotgun sequencing by Solexa/Illumina. High throughput marker development and functional analyses based on RNASeq, tiling array and copy number variants detection are also planned in the framework of this project. In order to prepare for its complete sequencing and analysis, we performed a pilot project on 18 Mb of contiguous sequences which allowed us to improve our understanding of the wheat genome composition and evolution. Comparative and evolutionary analyses revealed a large amount of nonsyntenic genes interspersed into a conserved ancestral grass gene backbone, suggesting that the wheat gene content has been extensively rearranged probably through transposable element-mediated gene capture. Finally, bioinformatics tools and databases has been developed in order to manage automatic annotation and analyses of such a large amount of data.

A GENERAL METHOD FOR ASSEMBLING GENOMES FROM SHORT READS

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

Broad Institute, Genome Sequencing and Analysis Program, 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, 9 for which a reference sequence was available and 7 from completely new samples. These genomes range in size from 2 Mb to 2.6 Gb, 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 for the smaller genomes their quality exceeds the general quality level of draft assemblies that had been achieved using Sanger sequencing. For the largest (bushbaby), we obtained contigs and scaffolds having N50 sizes 18 kb and 3.9 Mb, respectively. We compared to a preexisting 3.5x Sanger assembly, and by so doing estimated the combined misjoin rate in the two assemblies: one per 10 Mb. In addition we will report on a half dozen vertebrate assemblies that are in progress using the same methodology.

CAPTURING BIOLOGICAL FUNCTION IN *DE NOVO* ASSEMBLIES OF CEREAL GENOMES

Shiran Pasternak[1], Jer-Ming Chia[1], Andrew Olson[1], Joshua Stein[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, 11724, [2]USDA-ARS, NAA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY, 14853

While monumental genome projects have traditionally sought to produce near-finished reference sequence as a tool for biological inquiry, it has become recently apparent that the 80/20 rule holds: 20% of the cost, time, and effort of such projects produces 80% of the necessary biological content, while 80% of the cost, time, and effort is then expended to elucidate the difficult remaining 20%. This is particularly true in large and complex cereal genomes such as maize and wheat, wherein most of the functional landscape relevant for translational research (bioenergy and agriculture) is in low-copy, relatively simple stretches of DNA. We offer a methodology, manifested in a large-scale analysis pipeline, that uses short next-generation sequencing reads to rapidly produce raw de novo assemblies and to characterize their gene content. We employ several computational (e.g., ab initio genefinding and repeat annotation) and comparative strategies (e.g., gene trees and whole-genome alignment) based on the nature of the assembly as well as the availability of data from related species. The pipeline is being used in our group for a variety of NSF-funded objectives, including identification of genespace in the maize genome previously not included in the reference (B73 RefGen_v2), comparative analysis of wild varieties of tomato and grape to their domesticated counterparts, elucidation of genic scaffolds in the wheat D-genome, and preliminary analysis of as-yet-sequenced cereal genomes.

WAVESEQ: A NOVEL ALGORITHM FOR CNV DETECTION FROM NEXT-GENERATION SEQUENCING DATA

Bojan Losic[1], Sujata Syam[1], Quang Trinh[1], Richard de Borja[1], Irina Kalatskaya[1], John McPherson[1], Lakshmi Muthuswamy[1,2,3]

[1]Ontario Institute for Cancer Research, Bioinformatics and Biocomputing, Toronto, M5G0A3, Canada, [2]Cold Spring Harbor Laboratory, Bioinformatics, Cold Spring Harbor, NY, 11724, [3]University of Toronto, Medical biophysics, Toronto, M5G0A3, Canada

It is well established that copy number variations (CNVs) are a major source of genomic variability between any two individuals. At present, the effective genomic resolution of CNVs that can be detected using high resolution microarray technologies is approximately 5kb. Advances in next-generation sequencing technologies enable us to identify structural variations at the single-nucleotide level.

In this work we describe a new algorithm to detect breakpoints of CNVs from next-generation sequencing data. Our method is unique in the sense that it carries out an inherently multi-scale signal processing analysis using translation-invariant discrete wavelet transforms, enabling us to identify both INDELs and CNVs at all genomic scales. We test the veracity of these event-calls using a probability measure based on a number of genomic and experimental variables, including the quality of bases, GC content, and k-mer frequency. Here we describe our findings on somatic copy number and INDEL variations based on whole genome sequencing (with an Illumina GAII sequencer) of two pancreatic cancer genomes.

We identify somatic and germ-lime mutations based on Primary, matched Normal, and Xenograft tissues. All three of them have an average coverage of 20X and approximately 80% of the genome is covered. Preliminary pathway analysis on the genes present within somatic mutations in both genomes reveals that the most significant cluster involves the Androgen receptor signaling pathway.

RNA-SEQ IN ENSEMBL

Simon J White, Bronwen Aken, John E Collins, Susan Fairley, Thibaut Hourlier, Magali Ruffer, Steve Searle, Derek Stemple, Amy Tang, Jan-Hinnerk Vogel, Amonida Zadissa, Tim Hubbard

Wellcome Trust Sanger Institute, Informatics, Cambridge, CB10 1SA, United Kingdom

New sequencing technologies are creating large transcriptome datasets for many species, which are of great value in helping to inform gene annotation. Here we present an overview of how we have used Illumina paired end transcriptome data for the de-novo construction of gene sets for multiple species. In particular, we have created an Ensembl gene set for zebrafish which incorporates novel coding and non-coding RNA-Seq genes as well as tissue specific splicing and expression data.

We have developed tools for handling RNA-Seq data in the Ensembl genome annotation pipeline. These have enabled us to produce novel datasets to assist in gene annotation that quantify read support for exons and introns, many of these are now employed by the HAVANA annotators. We have also produced complete de-novo gene sets, which have been assessed as part of the RGASP competition (publication in preparation).

By combining RNA-Seq technology with a traditional Ensembl gene build we have been able to produce a novel hybrid gene set for the zebrafish Zv9 assembly. The addition of RNA-Seq has significantly increased the number of genes built using zebrafish specific data when compared to what would have been attainable using cDNA and protein evidence alone. RNA-Seq has also been used to add value by creating tissue specific expression and splicing data. The intuitive Ensembl web display provides visualisation of exon and transcript level RPKM values along with tissue specific intron support. In addition, we have used a novel technique that employs RNA-Seq to determine the precise three prime ends of transcripts. Single and often multiple possible three prime ends can be viewed on the website.

The new zebrafish gene set provides a greater depth of automatic gene annotation than has previously been available and has demonstrated how RNA-Seq technology may be used to improve Ensembl gene annotation in other species in the future.

EFFICIENT STRATEGIES FOR POPULATION GENOME SEQUENCING

Richard Durbin

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom

In the pilot phase of the 1000 Genomes Project we have demonstrated that low coverage (2-4x) sequencing can efficiently find shared sequence variation with high accuracy, and high power for variants present in at least 5 copies or so in the sequenced sample. Based on this (a) the 1000 Genomes Project has started sequencing 2500 people in five groups of 500 each at 4x genome wide (and also deep in exomes), (b) in collaboration with clinical investigators we have launched the UK10K project which involves sequencing 4000 cohorts samples with rich phenotypes, (c) we are beginning to explore strategies to identify almost all variation segregating in the population, starting with defined population isolates. Each project has its own requirements and design issues. Interestingly, even to maximise the number of singletons found for fixed budget, a relatively low depth of 6-8x per sample is optimal because it allows more samples to be sequenced. Alongside presenting progress cross-sample calling strategies, population modelling and depth analysis, we are also exploring population sequence assembly to find variation in a way that is not dependent on the reference sequence.

DISSECTING COMMON AND RARE REGULATORY VARIATION IN HUMAN GENOMES USING RNA SEQUENCING

Stephen B Montgomery, Tuuli Lappalainen, Emmanouil T Dermitzakis

University of Geneva, Department of Genetic Medicine and Development , Geneva, 1214, Switzerland

Our understanding of common and complex disease is being enhanced by our ability to uncover the effects of genetic variation on cellular state. Specifically, by understanding which variants have an impact on the expression of genes, it is likely that we can also find those variants which inform important human conditions. Now, with the increasing availability of complete genomes we are confronted with the challenge of dissecting rare and common as well as small and large variants into a more complete model of association with the aim of pinpointing specific causal variants. To approach this, we analyzed 60 complete genomes from the 1000 Genomes Project with respect to gene expression assayed by RNA-Seq. We have assessed association of 6.5 million common genetic variants (6,554,051 SNPs and 561,844 indels) with the expression of 22,194 GENCODE annotated protein coding genes. We have compared the relative effects of SNPs versus indels on gene expression. We have investigated features of alternative splicing. We have furthermore developed novel strategies for assessing allele-specific expression and have integrated this information to further understand the impact of rare variants and to uncover new structural features of the transcriptome which may be underlying human phenotypic diversity. We will also report the degree to which coding heterozygotes are supported by RNA-sequencing data.

DETERMINING DETERMINANTS OF DIFFERENTIATION: A MULTIVARIATE ANALYSIS OF ERYTHROID EPIGENOMIC FEATURES.

Swathi A Kumar[1], Weisheng Wu[1], Ross C Hardison[1], Francesca Chiaromonte[1,2]

[1]The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, State College, PA, 16802, [2]The Pennsylvania State University, Department of Statistics, State College, PA, 16802

The generation of mature erythroid cells relies on the establishment of lineage-specific differentiation programs that are regulated by the transcription factors, co-activators, co-repressors and chromatin modifications. The full extent of this transcriptional control, however, has not yet been understood. We explore these questions in a mouse cell line model of erythroid differentiation. G1E is an ES-derived cell line with a genetic knockout of the gene Gata1. G1E cells closely mirror erythroid progenitors BFU-E and CFU-E. An estradiol-dependent rescued subline of G1E (G1E-ER4) mimics normal erythroid maturation on treatment with estradiol. This system provides a physiologically meaningful assay to evaluate genes that are induced or repressed by GATA1.

Using ChIP-Seq, we have mapped transcription factors GATA1, GATA2, TAL1 and CTCF, histone modifications H3K4me1, H3K4me3 and H3K27me3, and chromatin accessibility (DHS), genome-wide, in the G1E and G1E-ER4 cells. We analyse the relative impact of these features on the levels and changes in gene expression using multivariate regression and discriminant analysis techniques. We employ density-weighing methods based on the distribution of the level and change in gene expression, offering a higher predictive power than traditional models. Our models explain 71% of the variation in the level of gene expression and account for 39% of the change in gene expression. We identify combinations of epigenomic features that discriminate differentially expressed genes using Linear Discriminant Analysis (LDA). Our model classifies high versus low expression level genes with 96.4% accuracy (Sn = 0.95; Sp= 0.97) and induced versus repressed genes with 76% accuracy. We find that the levels of H3K4me3 and H3K4me1 strongly influence the level of gene expression while the regulation of genes is dependent on changes in transcription factor occupancy. Further, we employ feature extraction methods that utilize all peak signals (for transcription factors) and segmentation states (for histone modifications) to identify combinatorial binding of these epigenetic features.

The nature of our cell lines and high performance of our models make our data uniquely suited to study the covariation between epigenomic features and their role in differentially expressed erythoid genes. Additionally, these data are a valuable resource for further studying gene expression and regulatory mechanisms.

HIGH THROUGHPUT GENETIC MAPPING USING THREAD MAPPER

Jitender Cheema[1], Noel Ellis[2], Jo Dicks[1]

[1]John Innes Centre, Computational and Systems Biology, Norwich, NR4 7UH, United Kingdom, [2]John Innes Centre, Crop Genetics, Norwich, NR4 7UH, United Kingdom

The accurate construction of a genetic linkage map is a vital component in the development and exploitation of a genome sequence. The linkage map helps to guide a genome assembly via anchor markers, and mapped markers then enable varietal improvement through marker assisted selection in many plant and animal species. We have recently introduced a new, visual approach to genetic map construction. Our software tool, THREaD Mapper Studio[1], employs a series of machine learning techniques to construct global genetic maps in three dimensions . By combining proven techniques from graph partitioning and advances in manifold embedding and learning, we have introduced fresh ideas that have changed the way that biologists view, literally, their genetic maps. Furthermore, the THREaD Mapper tool allows the user to interact with the visual map and to refine the final result using additional information such as supercontig assignment of the markers.

Following the success of our initial approach, used for example in the recent international *Brachypodium distachyon* mapping project[2], we are developing our algorithms further for high throughput map estimation. In particular, we are examining high throughput mapping datasets with up to tens of thousands of Single Nucleotide Polymorphism (SNP) markers. Here, we present our recent results in high throughput analysis and comparisons of our approach with alternative algorithms in a key case study. In conclusion, our new algorithms place THREaD Mapper firmly in the toolkit for analysis and exploitation of novel genome sequences.

1) Cheema J, Ellis THN and Dicks J (2010) THREaD MAPPER Studio: a novel, visual web server for the estimation of genetic linkage maps. *Nucleic Acids Research* doi: 10.1093/nar/gkq430

2) Garvin DF, McKenzie N, Vogel JP, Mockler TC, Blankenheim ZJ, Wright J, Cheema JJS, Dicks J, Huo N, Hayden DM, Gu Y, Tobias C, Chang JH, Chu A, Trick M, Michael TP, Bevan MW, and Snape JW (2009) An SSR-based Genetic Linkage Map of the Model Grass *Brachypodium distachyon*. *Genome* 53(1):1-13.

ANALYSES OF IDENTICAL TWINS' GENOMES REVEAL SOURCES OF FALSE-POSITIVE VARIANT DETECTION

Elliott H Margulies[1], Subramanian S Ajay[1], Stephen C J Parker[1], Hatice Ozel Abaan[1], Rachel L Goldfeder[1], Nancy F Hansen[2], Karin Fuentes Fajardo[3], Thomas C Markello[3], William A Gahl[3], James C Mullikin[2]

[1]Genome Informatics Section, Genome Technology Branch, National Human Genome Research Institute, Bethesda, MD, 20892, [2]Comparative Genomics Unit, Genome Technology Branch, National Human Genome Research Institute, Bethesda, MD, 20892, [3]NIH Undiagnosed Diseases Program, National Institutes of Health, Bethesda, MD, 20892

We have sequenced the genomes of monozygotic twins concordant for a unique neurological phenotype. Each of the twins was sequenced on a single flowcell giving 185Gb and 162Gb of sequence from 2x100 base reads from a single HiSeq2000 run. This resulted in 55X and 50X depth of coverage from bwa-aligned and de-duplicated reads for each twin's genome. From this dataset, we used a Bayesian genotype caller (called MPG, for Most Probable Genotype) to confidently call genotypes across 97.6% of the hg18 reference genome in both twins with an initial concordance rate of 99.9995% (13,297 discordant genotypes). Inspection of a random set of discordantly genotyped positions revealed that a majority occurred in regions with poorly aligning reads. We subsequently filtered reads with a low alignment score (representing these poorly aligning reads) and re-called genotypes using only reads with high alignment scores. This filter reduced the total number of genotype calls that could be made by 4.4% (down to 93.2% of the hg18 reference sequence) but reduced the number of detected differences between the twins by 80% (2,719 discordant genotypes). Additional analyses revealed that over half of the remaining variants occurred at or within 5bp of an identified small insertion/deletion (indel; representing 2.4% of the hg18 reference sequence) in one of the twin's genomes, suggesting that these too might be false positives due to incorrect alignments of short reads across indels. At this point, 1,188 discordant genotypes between the twins remain from analyzing 90.8% of the hg18 reference sequence. We are currently evaluating these differences to determine what proportion is likely to be real, and what proportion represents additional false positives. From a clinical perspective, it remains to be determined whether whole genome sequencing can be successfully used to identify a disease-causing variant, or whether the intrinsic background false positive rate will overwhelm our current analytic tools. From a basic biological perspective, this research will help determine the true genetic level of "identicalness" between monozygotic twins.

COLLECTING GENE SEQUENCE VARIANTS IN ALL MENDELIAN DISEASE GENES FROM RESEQUENCED PERSONAL GENOMES IN LSDBS.

Peter E Taschner, Ivo F Fokkema, Jacopo Celli, Johan T den Dunnen

Leiden University Medical Center, Human Genetics, Leiden, 2300 RC, Netherlands

Since the late eighties, starting with monogenic disorders, we have successfully linked sequence variants in specific genes to specific diseases. Now we approach the time that we can explore not a single gene but the entire genome for sequence variants and start to link these to complete individual phenotypes, disease-related, healthy or even prognostic. For the interpretation of all these variants, powerful tools are needed to efficiently sift through all variants found, link them to existing knowledge and prioritizing those that need further attention, especially those related to an individuals health. To support DNA diagnostics of monogenic diseases, gene-specific collections have been generated listing all variants identified world-wide, so called gene variant databases (Locus-Specific DataBases, LSDBs). Although generally accepted as essential in supporting an accurate and fast clinical diagnosis based on the latest findings, effective collection of these variants has been shown to be difficult. Many reasons for this exist, one of them being the lack of a central web-based gene-based repository in a format familiar to clinical geneticists. In the setting of the EU-funded Gen2Phen project we have now established a gene variant database (LSDB) for all genes accepting all gene variant data available, including those obtained from complete exome or genome resequencing (see www.LOVD.nl). However, to make this resource most useful, we need the help of guardians for these database, who will curate incoming information and thereby ensure data quality. Therefore, we invite clinicians and researchers working on genetic disorders world-wide to become the guardian of their gene(s) of interest. As an incentive for the submission of new variants, tools linked to LOVD enable automatic conversion of chromosomal positions to gene-related positions used by LSDBs as well as links to available variant information. This approach would use the database as a reference to allow identification of variants occurring with low frequency, which are likely to be involved in genetic disorders and should be prioritized for further investigation.

Funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 200754 - the GEN2PHEN project.

A MATTER OF LIFE AND DEATH: HOW MICROSATELLITES EMERGE AND DISAPPEAR FROM THE HUMAN GENOME.

Yogeshwar Kelkar, Francesca Chiaromonte, Kateryna Makova

Penn State University, Center for Medical Genomics, University Park, PA, 16802

Microsatellites - tandem repeats of short motifs - are abundant in human genome, and have high mutation rates. Their instability is implicated in 40 known genetic diseases; however, a systematic study of the molecular processes involved in microsatellite emergence and disappearance in the genome has been lacking. Microsatellite loci are hypothesized to follow a life cycle, wherein microsatellite are born (the birth phase), expand leading to the adulthood phase, until interruptions or large deletions cause their degradation and death (the death phase). Here we carried out a large-scale comparative genomic study to identify microsatellite births/deaths in three primate species, determined the causal mutational processes, and found the local genomic features that may drive them. Using publicly available genome alignments, we identified a genome-wide set of de novo microsatellite birth/death events in human, chimpanzee, and orangutan, using macaque and marmoset as outgroups. Using maximum parsimony, we determined the causal mutational steps (substitutions, insertions and deletions) for the majority of the identified birth/death events. We mapped the events to transposable elements and genes in the human genome, and discerned the effect of these environments on microsatellite birth/death propensities. To identify the drivers of birth/death events among many regional genomic features, we implemented a multiple regression approach. We came to the following conclusions. First, although substitutions were the predominant cause for births/deaths of small microsatellites (under 12 bp), insertions and deletions were increasingly important for births/deaths of larger ones. Second, LINE and younger SINE elements showed higher and lower propensity for birth/death events, respectively. There was a dearth of births/deaths in the proximity of genic regions, indicating that such events have long-term functional consequences. Supporting these observations, multiple regressions indicated that regions surrounding birth/death loci are marked by high substitution rates, skewed nucleotide composition, and high and low density of LINE and SINE elements, respectively. Our study also allowed us to estimate the probability of a region to give birth to a microsatellite, as well as of a microsatellite to pass away, in functional regions. This has important applications for predicting the susceptibility of certain genomic regions to bear novel disease-causing microsatellites.

UNBIASED GENOME-WIDE DETECTION OF SIGNATURES OF SELECTION

Pamela Russell[1], Evan Mauceli[1], Federica Di Palma[1], Kerstin Lindblad-Toh[1,2], Manfred Grabherr[1]

[1]Broad Institute of MIT and Harvard, Genome Sequencing and Analysis Program, Cambridge, MA, 02139, [2]Uppsala University, Dept. of Medical Biochemistry and Microbiology, Uppsala, 751 23, Sweden

An increasing number of population studies aim to detect signatures of selection. Methods designed for this purpose tend to share certain computational drawbacks. Sensitivity is compromised as they search only for user-specified patterns in the data. On top of this, the desired signal is obscured by noisy data and misclassification of samples.

Here we present an unsupervised method to detect recurring local patterns in a population dataset. Patterns are identified automatically; the method does not require initial hypotheses or division into groups.

Genomic regions are associated with phylogenies as representations of local evolutionary history. A self-organizing map (SOM) identifies genome-wide patterns and determines when to add a new phylogeny: the background phylogeny is identified first, followed by decreasingly prevalent patterns. A Hidden Markov Model (HMM), robust to noise in the dataset, optimizes local assignment of phylogenies to regions. Several iterations of the SOM and HMM give a high-resolution mapping between regions and underlying evolutionary relationships.

The method has been applied to detect loci underlying selection, conservation and speciation. In a low-coverage resequencing dataset in stickleback, it has successfully detected signatures of adaptation to freshwater environments. It has also been used to locate speciation islands among diverging mosquito strains, and local patterns of selection and conservation in mammalian whole-genome multiple alignments.

GENETICS OF NICOTINE DEPENDENCE AND SMOKING-RELATED DISEASES

Thorgeir E Thorgeirsson

University of California Santa Cruz, Jack Baskin School of Engineering, Santa Cruz, CA, 95064

Genome-wide association (GWA) studies have identified solid associations of a common sequence variant on chromosome 15q25 with nicotine dependence and smoking behavior[1-3]. The same variant is associated with risk of several smoking-related diseases, including lung cancer[3-5], peripheral arterial disease[3], and chronic obstructive pulmonary disease[6], and it has been implicated in addiction to cocaine with a reversed direction of association[7].

Since the original findings, a number of studies testing for association with the main variant at chrs 15q25 variant have become available, but there is still some debate regarding interpretation of the results. The key questions are: Does the variant confer risk of smoking-related diseases primarily through its effect on smoking behavior, or is there also a more direct effect on disease vulnerability? Is the association specific to nicotine dependence, or does the variant confer risk of addiction to other substances as well?

Recently three large GWA studies of smoking behavior were conducted[8-10], using a reciprocal replication model with the combined discovery and replication sample sizes of over 140,000 subjects for smoking initiation and over 85,000 for the cigarettes per day (CPD) phenotype. The chromosome 15q25 region was the strongest finding in all three studies, and additional independent signals within the region were discovered. Furthermore, several novel variants associating with CPD and smoking initiation were unraveled. At least some of the novel variants associating with CPD appear to associate with lung cancer as well[10].

Understanding these associations will provide insight into the gene-environment interactions involved in addiction and leading to other serious diseases, and into how addiction risk variants can be expected to impact diagnosis and treatment of addiction in the near future.

1) Saccone, S.F. et al. *Hum Mol Genet* **16**, 36-49 (2007). 2) Berrettini, W. et al. *Mol Psychiatry* **13**, 368-73 (2008). 3) Thorgeirsson, T.E. et al. *Nature* **452**, 638-42 (2008). 4) Amos, C.I. et al. *Nat Genet* **40**, 616-22 (2008). 5) Hung, R.J. et al. *Nature* **452**, 633-7 (2008). 6) Pillai, S.G. et al. *PLoS Genet* **5**, e1000421 (2009). 7) Grucza, R.A. et al. *Biol Psychiatry* **64**, 922-9 (2008). 8) *Nat Genet* **42**, 441-7 (2010). 9) Liu, J.Z. et al. *Nat Genet* **42**, 436-40 (2010). 10) Thorgeirsson, T.E. et al. *Nat Genet* **42**, 448-53 (2010).

GENOMIC ARCHITECTURE OF TWO REPEAT POLYMORPHISMS IN THE HUMAN SLC6A3 GENE

Elena Shumay[1], Elisabeth Mulligan[1], Joanna S Fowler[1], Nora D Volkow[2]

[1]Brookhaven National Laboratory, Medical Department, Upton, NY, 11973, [2]National Institute on Drug Abuse, National Institute of Health, NIDA , Bethesda, MD, 20892

• Our group previously reported that the two VNTRs (one – in 3'UTR and another – in intron8) of the DAT1 gene are associated with the DAT density in human midbrain. Considering a possibility that the functional effect of the repeat polymorphism stems from sequence variations within the repeated increments rather than from the variance in the length of the polymorphic region, we sequenced a number of primary samples of individual genomic DNA aiming to reveal the genetic architecture of these polymorphic regions in fine details and to assess the possible mechanisms by which the sequence characteristics of the repeats can contribute to vastly diverse DAT phenotype in human.
• About 500 primary samples of genomic DNA were genotyped to establish prevalence of rare alleles in population, and then about 20 randomly selected samples with major alleles (homo-and heterozygous genotypes) and about 10 samples per each detected rare variance were sequenced with a single-base resolution.
• The analysis revealed the MAF (minor allele frequency) of the variants traditionally regarded as "rare" exceed 5% in our population sample. Also, while genetic architecture of the common alleles of the 3'UTR VNTR was established previously, we discovered that high degree of conservation observed in common alleles, does not hold in rare variants, where we observed an increased complexity of the sequence composition of the repeated increments. We also ascertained that these variations can potentially interfere with the cellular machinery that controls the DAT1 expression. At the same time, the fact that the homozygous carriers of the 3'UTR rare alleles do not have phenotypes poses that this variance is well-tolerated. In contrast, rare alleles of the intron8 VNTR were ultimately detected as an allele of a heterozygous genotype, thus pointing on their potentially deleterious impact. We showed that the intron8 VNTR resided in an immediate proximity to the hot spot in human genome and, thus, it is likely that the sequence features of the rare alleles might confer increase in regional instability. Furthermore, according to our preliminary analysis, the DAT1 rare alleles are more frequent in individuals that are using elicit drugs, this finding prompted further investigations which are ongoing.

HOW ACCURATE ARE POLYMORPHISM ESTIMATES FROM NGS DATA? AN EMPIRICAL APPROACH.

Benjamin Dickins, Anton Nekrutenko

The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, PA, 16802

Because they rely on parallel sequencing, new high-throughput methods permit the analysis of mixed DNA samples. Recent papers (including our own) have exploited this to analyze changing patterns of polymorphism in evolving populations of microorganisms (Dickins and Nekrutenko, 2009; Barrick and Lenski, 2009). This application is very exciting, but accurate polymorphism estimation is difficult because we do not understand the nature and impact of sequencing errors. One rational approach, adopted in these studies, is to estimate errors for a given set of presumably polymorphic samples by sequencing DNA that is minimally processed or clonal, from the same genome. But the illusion of predictability can prevail if experimenters attend to particular genomic sites after the fact or because estimation procedures, leveraging real data, make assumptions about error profiles. To address this we have formalized an empirical approach by Illumina-sequencing of a small plasmid genome and cross validating the results with dideoxy sequencing of multiple subclones. We were surprised to discover that some sites exhibited high levels of apparent polymorphism at prodigious coverage, but were invariant in subclones. We sought explanations for these rare events and, crucially, tested their tendency to recur by repeating the entire procedure. These results and our approach have important implications for polymorphism detection, a major application of next-generation sequencing.

HIGH THROUGHPUT SEQUENCING STRATEGIES FOR FAMILIES AFFECTED WITH A SPECTRUM OF TWO MENTAL DISORDERS

Simon L Girard[1], Jean-Baptiste Rivière[1], Julie Gauthier[1], Ron Lafrenière[1], Isabelle Bachand[2], Paul Lespérance[1], Yves Dion[1], Geneviève Tellier[2], François Richer[3], Sylvain Chouinard[1], Patrick Dion[1], Guy Rouleau[1,2]

[1]Center of Excellence in Neuromics, Medecine, Montréal, H2L 2W5, Canada, [2]Hopital Ste-Justine, Medecine, Montreal, H3T 1C5, Canada, [3]Université du Québec à Montréal, Psychiatry, Montréal, H3C 3P8, Canada

The search for genes involved in mental disorders has always been a difficult quest, partially due to the large spectrum of symptoms. Although Genome Wide Association (GWA) studies have been a successful approach for many diseases, much is still to be found. With the emergence of Next Generation Sequencing (NGS) technologies, it is expected that a new part of the genetic etiology of mental disorders will be revealed. We used NGS on families affected with a combination of two mental disorders in order to find new variation potentially linked to the disease pathogenesis.

Obsessive compulsive disorder (OCD) is a complex and poorly treated disorder of the brain. Converging evidence suggests that OCD symptoms are the results of malfunctioning of the synapses that play a crucial role in the transmission of signal throughout the brain and the body. Other evidences also suggest that genetics play an important role in OCD. However, research projects in the genetic of OCD have so far failed to identify genes causing or predisposing to the disease. A similar story is observed for Tourette Syndrome (TS), a disorder mainly characterized by the presence of tics. It is well established that TS can be cause by both genetic and environmental factors. However, so far, most of the genetic studies conducted on TS families have been inconclusive.

We performed exome captures on 5 big families with multiple patients affected with both Tourette Syndrome (TS) and Obsessive Compulsive Disorder (OCD). Using ABI SOLiD 3+ ©, we sequenced the exome of 30 person with an average coverage of 30x. Using a combination of segregation analysis and annotation filters, we were able to identify multiple SNPs and indels that might be linked to either of the diseases. Different genetic models have been used, including heterozygosity with incomplete penetrance and multifactorial inheritance. All potentially interesting variants have been validated experimentally. The pooling of candidate genes from each pedigree has led us to establish a list of good candidates for TS and OCD that are currently investigated in a larger cohort.

COMPARATIVE ANALYSIS OF *SALMONELLA* PATHOGENICITY ISLANDS, PLASMIDS AND MOBILE ELEMENTS BY MASSIVELY PARALLEL SEQUENCING

Craig A Cummings*[1], Andrea I Moreno Switt*[2], Gregory R Govoni*[1], Matthew L Raineri[2], Henk C den Bakker[2], Joseph E Peters[3], Lovorka Degoricija[1], Elena Bolchacova[1], Manohar R Furtado[1], Martin Wiedmann[2]

[1]Applied Biosystems, a part of Life Technologies Corporation, Foster City, CA, 94044, [2]Cornell University, Department of Food Science, Ithaca, NY, 14853, [3]Cornell University, Department of Microbiology, Ithaca, NY, 14853

The bacterial pathogen, *Salmonella enterica*, infects millions of humans, killing thousands each year. This organism is phenotypically heterogeneous, comprising more than 2,500 serotypes, with differences in host range, virulence, and antimicrobial resistance. Underlying this diversity is a flexible species genome encompassing a variety of genomic islands, mobile elements, and plasmids. Although sequencing has been applied to explore some of the genomic diversity of this species, most investigations have focused on a small subset of *S. enterica* serotypes that are commonly isolated from humans. In order to assess the pan-genome of *S. enterica* in a less biased fashion, we used comparative whole genome sequencing with the SOLiD™ system to characterize pathogenicity islands (SPIs), mobile elements, and plasmids in 16 rare human disease-associated *S. enterica* serotypes. Following de novo assembly, each genome was analyzed for the presence of all 21 reported SPIs as well as four *S.* Typhimurium and *S.* Typhi prophages. Many SPIs were universally present among this collection (SPIs 1-6, 9, 11-13), but others were not found in this subset of strains, and seem to be serotype-specific. The prophage were present in 13% to 25% of the strains. Detailed analysis of SPI and prophage gene content revealed that ten SPIs had no missing ORFs, while only one prophage, Fels-2, was substantially intact in other strains. Three serotypes were found to encode a type IVB pilus operon in a putative mobile element that is a component of SPI-7. Some known virulence genes had elevated mean pairwise distances (Dn). *sipD* was most divergent with a Dn of 0.0543, versus 0.005 for typical core genes. Phylogenetic analysis of virulence factor sequences suggested that these strains can be grouped into four main clusters. The Uganda and Give strains which were isolated from cow and human, respectively, share the same serogroup (E1), but belong to different clusters based on the virulence factor analysis, suggesting possible mechanisms for discerning *Salmonella* strain host specificity. Large plasmids were also identified, subsequent to de novo assembly, as scaffolds that did not match any known *Salmonella* chromosomes. These were found to represent IncI1, IncH12, and IncW family plasmids that encode putative virulence and resistance genes.

*Authors contributed equally

ANALYSIS OF HERV-K (HML-2) ENV RNAS REVEALS THE MOBILIZATION OF HERV-K IN THE PLASMA OF HIV-1 PATIENTS AND UNCOVERS THE NOVEL CENTROMERIC HERV-K111

Rafael A Contreras-Galindo[1], Mark H Kaplan[1], Angie C Contreras-Galindo[1], Scott D Gitlin[1], Yasuhiro Yamamura[2], David M Markovitz[1]

[1]University of Michigan, Internal Medicine, Ann Arbor, MI, 48109, [2]Ponce School of Medicine, AIDS Research Program, Ponce, PR, 00716

We previously reported finding the RNA of a type K human endogenous retrovirus, HERV-K (HML-2), at high titers in the plasma of HIV-1-infected patients. However, whether these viruses are truly capable of being mobilized in modern humans remains an important question. We report the presence of HERV-K viral particles in the plasma of HIV-1-infected patients. The full-length RNA from the env gene of HERV-K (HML-2) was amplified and sequenced from the plasma of six HIV-1-infected patients collected over a period of one to three years, in order to reconstruct the genetic evolution of these viruses. We found a high frequency of recombinant env RNA sequences, accumulation of synonymous rather than non-synonymous mutations, and conserved N-glycosylation sites, indicating that some HERV-K (HML-2) viruses might have been mobilized by reinfection. Using "reverse genomics", a novel HERV-K (HML-2) provirus, termed K111, was found within the human genome. Surprisingly, this previously undiscovered provirus was found in the DNA of all healthy and HIV-1-infected people tested, but appears to be transcriptionally active exclusively during HIV-1 infection. Indeed, infection of lymphocytes with HIV-1 leads to marked expression of HERV-K 111 (K111). K111 is "polyproviral" (present in more than five different proviral forms in each individual), and coexists with a K111 soloLTR. K111 sequences are integrated into tandemly repetitive D22Z3 sequences, which have been found uniquely in the centromere of chromosome 22. In silico sequence analysis revealed that the K111 copy number has likely been expanded during evolution by homologous recombination. These findings suggest that HERV-K (HML-2) is unexpectedly mobilized in HIV-1-infected patients by multiple mechanisms. Furthermore, the discovery of the first centromeric and "polyproviral" endogenous retrovirus thus far described appears to have been made possible only by HIV infection, which allows for active transcription of K111 sequences that are normally silenced in the centromere of chromosome 22.

LOSS-OF-FUNCTION MUTATIONS IN A NATURAL ISOLATE OF CAENORHABDITIS ELEGANS

Ismael A Vergara[1], Maja Tarailo-Graovac[1], Jun Wang[1], Rong She[2], Ke Wang[2], Nansheng Chen[1]

[1]Simon Fraser University, Molecular Biology and Biochemistry, Burnaby, V5A 1S6, Canada, [2]Simon Fraser University, Computer Science, Burnaby, V5A 1S6, Canada

A growing number of genomic variations (GVs) have been identified in humans based on sequenced genomes of individuals. Many instances of GVs are loss-of-function mutations and have been associated with genetic diseases, including many types of cancer and mental retardations. However, how these GVs cause a disease condition is not well understood, due to two major challenges. First, due to the large size of human genes and intergenic regions, as well as the large amount of GVs, the correlation between GVs and genes is not well defined. More importantly, it is difficult to assess the functional impact of GVs and their contribution to the pathogenesis of disease conditions in humans. In our laboratory, we use the nematode Caenorhabditis elegans, which has a small and compact genome and is a well-established model organism for studying human biology, as a platform to associate GVs with phenotypes. Now we have sequenced the genome of the Hawaiian wild isolate (CB4856) of C. elegans using 454 sequencing technology and developed a new software named variationBlast to accurately detect GVs at the base-pair resolution. Using variationBlast and the N2 strain as reference, we have identified 194,945 putative SNPs, 17,142 putative deletions (with 1,553 of them larger than 100 bp), and 16,296 insertions (some of them larger than 1,000 bp). Correlation of these GVs with annotated protein-coding genes in C. elegans revealed that many of these GVs disrupt protein-coding genes by creating frameshifts, premature stop codons, or even by removing entire genes. Interestingly, we have identified candidate loss-of-function mutations in 141 different essential genes in C. elegans. Of these, nine genes are orthologs of human disease genes, according to OMIM annotation. The biological impact of these mutations is being analyzed by generating these mutations in the N2 background. This study is supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada.

SOMATIC SNIPER: A BAYESIAN PROBABILITY MODEL FOR MUTATION DETECTION

Christopher C Harris, David E Larson, Ken Chen, Daniel C Koboldt, Li Ding, Elaine R Mardis, Richard K Wilson

Washington University, The Genome Center, Saint Louis, MO, 63108

Whole genome re-sequencing of matched tumor/normal pairs is currently being used as an unbiased screen for cancer associated mutations. Sensitive identification of tumor specific mutations, also known as somatic mutations, is essential for these screens to succeed. The challenge of such an analysis is separating normal data variation in the datasets from genuine genomic differences while combating relatively high sequencing and mapping error. Here we present our somatic single nucleotide variation (SNV) detection tool, Somatic Sniper. It includes a Bayesian probability model that calculates the likelihood that the genotypes in the tumor and normal are different as well as several downstream filters designed to remove known variations and reduce errors caused by common whole genome re-sequencing artifacts such as paralog alignments, strand bias, and loss of heterozygosity. We achieve an estimated sensitivity of 98% on externally generated real world data and an estimated specificity of 84%

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

Subramanian S Ajay, Stephen C Parker, Hatice Ozel Abaan, Jamie K Teer, Praveen F Cherukuri, Nancy F Hansen, Pedro Cruz, Karin Fuentes Fajardo, Thomas C Markello, William A Gahl, James C Mullikin, Elliott H Margulies

National Institutes of Health, National Human Genome Research Institute, 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 to achieve 78X coverage from two flowcells on the HiSeq 2000 platform and 34X coverage from two flowcells on the GAIIx platform, giving us a total combined coverage of more than 110X (or 97X when duplicate molecules are removed). We observed that G+C bias is reduced in the HiSeq 2000 dataset, resulting in more uniform coverage and better representation of CDSs. Concordance rates of Single nucleotide variants (SNVs) in each of the datasets were equally high (>= 99.95%), compared to calls made on an array-based technology. However, when normalized to equal depth of coverage, we were able to call genotypes at a greater number of positions in the HiSeq 2000 dataset, compared to the GAIIx dataset (96.7% vs. 93.5%).

We used these combined datasets to address what coverage is required to attain different levels of comprehensiveness and accuracy, allowing us to make an informative decision about the depth of coverage needed for future whole-genome sequencing endeavors. Specifically, we generated 5X, 10X, 15X… 95X subsets of the data and processed them through our variant calling pipeline. At 30X genome average depth of coverage, 96.26% of the genome has 10X or greater Q20 base-wise coverage; at 60X there is 99.09% coverage and at 95X there is 99.50% coverage. Interestingly, we note that SNV calling concordance drops slightly with increasing depth of coverage. We also note a correlation between novel SNVs detected at higher depths of coverage and their overlap with a greater proportion of known segmental duplications, possibly causing this slight decrease in concordance.

These results will help more objectively address how much sequencing is needed to obtain a certain level of comprehensiveness.

PYICOS: A FLEXIBLE TOOL FOR ANALYZING PROTEIN-DNA AND PROTEIN-RNA INTERACTIONS WITH MAPPED READS FROM DEEP SEQUENCING

Sonja Althammer, Juan Ramón González-Vallinas, Eduardo Eyras

Computational Genomics, Universitat Pompeu Fabra, Barcelona, 08003, Spain

Background
Deep DNA and RNA sequencing has become indispensable for the discovery of protein binding sites. The result of these techniques are millions of short sequences that have to be mapped back to a reference genome or transcriptome in order to obtain the locations of the binding sites of interest. The next challenging task is to find significantly enriched regions from the mapped reads, which does not have a unique solution: Depending on the type of protein and its binding to the sequence we might want to proceed in different ways.

Results
There have been several methods (1,2) introduced that do a so called "peak calling" in order to detect enriched regions from deep-sequencing data that most likely interact with the protein of interest. We present here a new tool, Pyicos, which unlike other published methods, is very flexible and provides transparency by allowing the user to perform the operations of choice independently. Our method thereby helps to control the analysis of different kinds of data. Pyicos is already being used in several collaborations that our group is carrying out, such as the analysis of ChIP-Seq, CLIP-Seq and Mnase-seq data coming from Solexa/Illumina sequencing machines. In order to show that Pyicos keeps up with other methods (1,2) in terms of performance and efficiency as well as specificity and sensitivity we performed a benchmarking on the ChIP-Seq data-set of neuron-restrictive silencer factor (NRSF).

Conclusions
With our approach we offer a step-wise analysis of mapped reads which we consider to be more appropriate than the monolithic software packages that encourage bioinformaticians to use them as “black boxes”. Considering that it is very important to be able to analyze different kinds of data in different ways, we give the user the possibility to explore all the individual operations within Pyicos, as independent functions. Apart from this, we also provide protocols for the typical steps that are done for different types of data, such as punctuated ChIP-Seq data.

References:
1) Model-based Analysis of ChIP-Seq (MACS), Zhang et al., Genome Biology 2008
2) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology , Fejes et al., Bioinformatics. 2008

ANALYSIS PIPELINES FOR THE 1000 GENOMES AND UK10K PROJECTS

Sendu Bala, Petr Danacek, Jim Stalker, Thomas Keane, Richard Durbin

Wellcome Trust Sanger Institute, Vertebrate Resequencing, Cambridge, CB10 1SA, United Kingdom

The 1000 Genomes Project is sequencing the genomes of over 1000 individuals from populations around the world. The UK10K Project aims to sequence the genomes of over 10000 individuals from the UK. Both projects utilize "next-generation" sequencing platforms to generate both low-coverage whole-genome sequence and higher-coverage exome sequence. Their goals include finding low frequency genetic variants, with the 1000 Genomes Project able to find most accessible variants with a frequency over 1% and UK10K aiming to improve on this to 0.1%.

To reach these goals, after quality checking the sequencing data, sequencing reads in fastq files are mapped to the human reference, quality values are re-calibrated, likely PCR duplicates are marked, data are merged to produce single files with all the alignments for a particular individual and sequencing platform, and finally multiple variant callers are used on the resulting bam files. While tools exist to handle each of these individual steps, an informatics challenge presents itself when facing the sheer quantity of data associated with projects of this scale. For the 1000 Genomes Project we currently track over 40,000 fastq files and have generated more than 19TB of compressed alignment data; and these figures continue to grow.

We developed a custom tracking database and associated API, along with a generic pipeline system capable of running arbitrary sequences of actions on sets of directories simultaneously across a compute cluster. It can query and write to the database to keep track of which parts of which pipelines need to be run, and which have completed. It also has a system in place to reattempt failed parts of pipelines multiple times before failing those parts permanently, requiring user intervention. A series of pipelines were implemented using the system, one for each of the major steps outlined above.

BIOINFORMATICS-DRIVEN IDENTIFICATION OF CANDIDATE GENES IN WHICH VARIATIONS ASSOCIATE WITH NON-ALCOHOLIC FATTY LIVER DISEASE (NAFLD)-RELATED METABOLIC PHENOTYPES

Karina Banasik[1,3], Johanne M Justesen[1], Thomas S Jensen[2], Søren Brunak[2], Oluf Pedersen[1,3], Torben Hansen[1]

[1]Hagedorn Research Institute, Diabetes Genetics, Gentofte, 2820, Denmark, [2]Technical University of Denmark, Center for Biological Sequence Analysis, Lyngby, 2800, Denmark, [3]University of Copenhagen, Biomedical Sciences, Copenhagen N, 2200, Denmark

Objective Candidate genes for non-alcoholic fatty liver disease (NAFLD) identified by a bioinformatics approach were examined for variant associations to quantitative traits of NAFLD-related phenotypes.

Methods By integrating public database text mining, trans-organism protein-protein interaction transferal, and information on liver protein expression a protein-protein interaction network was constructed and from this a smaller isolated interactome was identified. Five genes from this interactome were selected for variant analysis. Twenty-one tagSNPs (HapMap, CEU) which captured all variation in these genes ($r^2<0.8$) were genotyped in 10,196 Danes, and analyzed for association with NAFLD-related quantitative traits (waist circumference, fasting levels of plasma glucose, and serum fasting levels of insulin and triglycerides), and for association with type 2 diabetes (T2D), obesity, and WHO-defined metabolic syndrome (MetS).

Results A total of 273 genes were included in the protein-protein interaction analysis and *EHHADH*, *ECHS1*, *HADHA*, *HADHB*, and *ACADL* were selected as likely candidates. Variation in several of the genes was associated with quantitative metabolic traits, e.g., the minor C-allele rs7093778 in *ECHS1* showed an association with increased fasting plasma glucose (per allele effect = 0.03mmol/l (0.01;0.05), P=0.002). Also, the case-control study showed associations between variation in the five genes and T2D, obesity, and MetS, respectively.

Conclusions Using a bioinformatics approach we identified five candidate genes in which variation was associated with NAFLD-related phenotypes. Given the explorative nature of our studies these findings need replication.

PUBMED2ENSEMBL: A RESOURCE FOR LINKING THE BIOMEDICAL LITERATURE TO GENES AND GENOMES.

Joachim Baran[1], Martin Gerner[1], Maximilian Haussler[1], Goran Nenadic[2], Casey M Bergman[1]

[1]University of Manchester, Faculty of Life Sciences, Manchester, M13 9PT, United Kingdom, [2]University of Manchester, School of Computer Science, Manchester, M1 7DN, United Kingdom

Advances in DNA sequencing technology over the last 20 years have drastically increased the rate of production of genomic sequence data, which in turn has directly accelerated the rate of biological discovery and publication. Despite the fact that genome sequence data and publications are two of the most heavily relied-upon sources of information for many biologists, very little effort has been made to systematically integrate genome sequence data directly with the biological literature. For a limited number of model organisms (e.g. yeast, *Drosophila*, mouse) dedicated teams manually curate publications about genes, however many thousands of articles refer to species with no such dedicated staff and are therefore never mapped to genes or genomic regions. Furthermore, no resource currently provides the capability of automatically performing queries across both the biological literature and genomic data. To overcome the lack of integration between genomic data and biological literature, we have developed pubmed2ensembl (http://www.pubmed2ensembl.org), an extension to the BioMart system that links over 2,000,000 articles in Pubmed to Ensembl genes for 50 species. We exploit several sources of curated (e.g. Entrez Gene) and automatically generated (e.g. gene name recognition in MEDLINE records, BLAST of EMBL records) gene-publication information, allowing users to filter and combine different data sources to suit their needs for information extraction and biological discovery. In addition to extending the Ensembl BioMart database to include information on publications, we also implemented a novel BioMart interface that allows text-based search queries on PubMed abstracts to be performed in conjunction with queries on genomic features. By allowing biologists to find the relevant literature on specific regions of the genome or set of functionally related genes more easily, pubmed2ensembl offers a much-needed genome informatics inspired solution to accessing the ever-increasing biomedical literature.

CUSTOM-TRACK GENOMIC BROWSER PLUGIN FOR THE BIOGPS GENE PORTAL SYSTEM

Serge Batalov

Genomics Institute of the Novartis Research Foundation, (GNF), San Diego, CA, 92121

BioGPS [1] is an easily extensible and customizable gene portal. Utilizing a simple HTML-based plugin interface, BioGPS enables users to easily aggregate data on a gene by gene basis from more than 200 external sources, and to personalize their gene report using BioGPS layouts. By registering new plugins, the entire community is empowered to increase the breadth and depth of accessible gene annotation and allow external developers to take advantage of the BioGPS user base and core searching functionality.

Much of the information for a gene is better understood within its genomic context. To bridge the gene- and genome-centric information, we aimed to create a template for a genomic viewer plugin that would 1) allow to visualize and configure user's data side-by-side with publically available genome annotation tracks in a genomic interval, 2) be easy to install and maintain, 3) allow for non-standard, free-style visual custom tracks.

After evaluating a dozen existing genomic browsers, we found the well-established UCSC Genome Browser [2] the easiest to embed and extend. We have implemented a simple proxy for viewing the UCSC web pages augmented with free-content custom tracks that are served from the user's local web server. This solution may appeal to small labs, as it does not require the UCSC GB database installation and maintenance. Furthermore, this solution is easily refactored for the up-coming UCSC distributed (remote) track platform [3].

[1] Wu C, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 2009;10(11):R130. http://biogps.gnf.org

[2] Rhead B, et al. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 2010 January; 38: D613–D619.

[3] http://users.soe.ucsc.edu/~kent/presentations/distributed2010.pdf

REFEX: REFERENCE EXPRESSION DATASET FOR PRACTICAL USE OF GENE EXPRESSION DATA

Hidemasa Bono[1], Hiromasa Ono[1], Kousaku Okubo[1,2], Toshihisa Takagi[1,2]

[1]Research Organization for Information and Systems, Database Center for Life Science(DBCLS), Bunkyo-ku, 113-0032, Japan, [2]Research Organization for Information and Systems, National Institute of Genetics, Mishima, 411-8540, Japan

In Database Center for Life Science (DBCLS), we have been tackling the problem how to organize various types of gene expression data and huge amount of transcript sequences.

For the former part, we made the integrated interface to browse gene expression data by iAFLP (introduced amplified fragment length polymorphism), microarray (GeneChip) and expressed sequence tag (EST) counts for human and mouse. Gene expression data calculated from CAGE (Cap Analysis Gene Expression) by counting CAGE tags around transcription start sites, which is developed by RIKEN group under Genome Network Project in Japan, is now tentatively added for comparison. In order to tackle the latter issue, we developed and maintained the analysis pipeline (called Buncher) for transcript sequence information from the next generation sequencers as gene expression data (RNA-seq) as a natural extension of EST-based gene expression database (BodyMap).

All these data is integrated as Reference Expression dataset (RefEx) for comparative analyses of gene expression data. Web interface for RefEx contains the form in which users can search by gene names, various types of IDs, chromosomal regions in genetic maps, and keywords. Relative values of gene expression are mapped to the 3D body image as well as the graphical histograms for those are available for different types of measurement methods.

We will present current status of the project and utility of the output data.

POST-PROCESSING STATISTICAL ANALYSIS OF QUANTITATIVE PROTEOMIC DATA

Tyler S Bray[1], Juliesta E Sylvester[2], Stephen J Kron[2]

[1]The University of Chicago, Computer Science, Chicago, IL, 60637, [2]The University of Chicago, Molecular Genetics and Cell Biology, Chicago, IL, 60637

The use of proteomics to identify characteristic biomarkers of disease is critically hindered by our current ability to analyze and interpret proteomic data in a biological context. Computational approaches based on compilations of observed molecular interactions have been developed as references from which probable networks of protein activity can be constructed. Even so, the amount of time required to generate biologically reasonable hypotheses from proteomic data can be a significant challenge for data analysis. Biological interpretations based on protein interaction networks are not useful without a standardized method of data analysis that determines quantitative significance at a global scale and meaningfully organizes data using graphical aids for the visualization of patterns. Toward this end, we developed a software suite that uses population-based statistics and heuristics to guide researchers to proteins and function-based keyword categories of the greatest biological interest. The program was developed as a post-processing platform that calculates the relative abundance of proteins and keyword categories from quantitative proteomic data, determines significance based on changes in the sampled population, and generates heat maps according to the correlation between measurements.

A TOOL FOR SIGNIFICANCE TESTING AND FUNCTIONAL CLASSIFICATION OF QUANTITATIVE PROTEOMIC DATA

Tyler S Bray[1], Juliesta E Sylvester[2], Stephen J Kron[2]

[1]The University of Chicago, Computer Science, Chicago, IL, 60637, [2]The University of Chicago, Biochemistry and Molecular Biology, Chicago, IL, 60637

Proteomic research aims to identify characteristic biomarkers of disease; however, current efforts are challenged by the amount of time required for data analysis, the lack of standardization among analytical methods, and the resulting difficulty in identifying the most relevant changes in protein activities. With the aim of addressing these problems, we developed Correlator, a software application that uses population-based statistics and heuristics to guide researchers to proteins and function-based keyword terms that show a significant change in abundance. The program was used to identify proteins and keywords that responded to lipopolysaccharide in neutrophil-differentiated HL60 cells. Data were acquired by tandem mass spectrometry and technical replicates with reverse 18O labeling were used to validate quantitative analyses. Six optional filters were incorporated to control the quality of data, thereby reducing measurement redundancy and outliers affecting population statistics. An alternative metric to compare relative peptide abundance was evaluated to increase normality, decrease standard deviations, and improve correlations among experiments. Results are presented as a heat map and organized by cluster analysis to demonstrate functional groupings among proteins and keywords. The program was developed as an independent post-processing platform that runs on all common operating systems, thereby providing a useful tool with which to measure significant changes in the proteome.

HIGH-THROUGHPUT SEQUENCING-BASED DNA METHYLOMICS

Guillermo Carbajosa[1], Lisa Nanty[1], Michelle Holland[1], Sarah Finer[1], Thomas A Down[2], Vardhman K Rakyan[1]

[1]BICMS, Queen Mary University, Diabetes, London, E1 2AT, United Kingdom, [2]Gurdon Institute, Univesrity of Cambridge , Transcription Informatics, Cambridge, CB2 1QN, United Kingdom

The overall goals of our lab are to understand how epigenetic mechanisms influence complex phenotypes and diseases in multi-cellular organisms. We are particularly interested in exploring the notion of the 'epiallele' – loci at which the epigenetic state varies as a result of genetic and/or environmental influences, and are pursuing several complementary lines of investigation that integrate molecular genetics, functional genomics, bioinformatics, model organisms, and human cohorts. Recently, we have focused on high-throughput sequencing-based DNA methylation analyses. DNA methylation is an indispensible epigenetic modification in many different eukaryotic organisms. We have either developed or adapted three different approaches: MeDIP-Seq (Methylated DNA Immunoprecipitation Sequencing), RRB-Seq (Reduced Representation Bisulfite Sequencing) and BS-Seq (Bisulfite sequencing). I will present results from the use of BS-Seq and RRB-seq to investigate the links between phenotypic plasticity and DNA methylation in the Honey Bee, and MeDIP-seq in mouse and human to study inter-individual epigenetic variation. In particular, I will focus on the computational biology and statistical aspects of the analysis of these DNA methylomic datasets.

CRAWL (CHADO RESTFUL ACCESS WEB-SERVICE LAYER) - A PROGRAMMATIC INTERFACE FOR QUERYING PATHOGEN GENOMICS DATA

Giles Velarde, Tim Carver, Matt Berriman, Jacqueline McQuillan

Wellcome Trust Sanger Institute, Pathogen Genomics Group, Cambridge, CB10 1SA, United Kingdom

Chado is a standard relational database schema used in many genomics resources. At the WTSI Pathogen Genomics group, we maintain the GeneDB pathogen database, using Chado, storing genomic annotation data for many of the organisms investigated here. To help deal with rapidly evolving requirements, a lightweight query framework has been developed for Chado. Written as a library, it is currently deployable both as a command line utility and a web-services application server. Unlike many contemporary approaches, which use object relational mappings and search engines, the bulk of the query logic is developed in targeted SQL statements, which are written and optimized for the queries required. This approach has allowed effective rapid prototyping of the framework, with several concrete use-cases continuously driving its design. For example, the command line utility has been used to automate data extraction from our pathogen
database for use in scriptable analysis tasks. The services are used to drive AJAX-based application development, allowing interactive Javascript UIs to be designed. When deployed on GeneDB, the services can be used by collaborating database resources (e.g., EupathDB). The result is a comprehensive, yet readily extensible, query infrastructure able to perform biologically relevant queries and applicable to many problems.

ARTEMIS AND ACT : BROWSING GENOMES AND VISUALISATION OF NEXT GENERATION DATA.

Tim Carver, Giles Velarde, Matt Berriman, Julian Parkhill, Jacqueline McQuillan

Wellcome Trust Sanger Institute, Pathogen Genomics Group, Cambridge, CB10 1SA, United Kingdom

The advent of next generation data has posed challenges for storing, distributing and visualising the vast quantity of data and information they contain. Artemis has traditionally been used as a genome browser and annotation tool and the Artemis Comparison Tool (ACT) is used to compare sequences and to highlight regions of similarity and differences. New functionality for both these tools is presented, including the ability to view next generation data in the context of the sequence, annotation and variation data. For example a window (BamView) has been incorporated to view sequence reads and other data can be imported by reading them in as a user plot and displayed as graphs or as a heat map. In this way it provides the annotator with extra levels of information that can inform them about structural annotation. Additionally a new javascript version of the Artemis tool is presented. This has the advantage of delivering genome annotation straight to the community via their browser.

CONSERVATION AND DIVERGENCE OF HIGHER ORDER CHROMATIN STRUCTURE DURING VERTEBRATE EVOLUTION

Emily V Chambers, Wendy A Bickmore, Colin A Semple

MRC, Human Genetics Unit, Edinburgh, EH4 2XU, United Kingdom

Chromatin structure is based on the interaction between DNA and proteins and encompasses several hierarchical layers of genome organisation from nucleosome arrays to inter-chromosomal interactions. Each layer is subject to differing epigenetic modifications and it is the relationships between the modifications at all levels of chromatin structure that create an 'epigenomic landscape'. The epigenome creates a bridge between genotype and phenotype, regulating the way the genome is expressed in different cell types, developmental stages and disease states, including cancer.

Although there are many different combinations of modifications, two distinct forms of higher order chromatin have emerged, open, active chromatin and compact, silent chromatin. These forms have a list of potentially opposing or differing properties which include levels of expression and accessibility, spatial positioning, replication timing, histone marks and evolutionary rate. However, it has been difficult to distinguish cause from effect within these groups of co-segregating properties.

I have made genome-wide comparisons using data relating to different aspects of higher order chromatin to investigate conservation and divergence across cell types and species. Whilst it is important to establish the degree of conservation between differing species, the mechanisms for divergence present another intriguing area of research. Here I present data demonstrating widespread conservation of structure between mammals but also clear examples of divergence. Such insights have implications for our understanding of higher order chromatin structure and genome evolution.

MOLECULAR COMBING PHYSICAL MAPS: IMPROVING ASSEMBLIES AND STUDYING GENOMIC LANDSCAPE/VARIABILITY.

Kevin Cheeseman[1,2], Grace Yao[1], Aaron Bensimon[1], Emmanuel Conseiller[1], Serge Casaregola[3], Pierre Renault[2], Maurizio Ceppi[1]

[1]Genomic Vision, Diagnostics Division, Paris, 75014, France, [2]Institut National de la Recherche Agronomique, Micalis, Jouy-en-Josas, 78352, France, [3]Institut National de la Recherche Agronomique, CIRM-levures, Thiverval-Grignon, 78850, France

The arrival of new sequencing technologies made sequencing easily accessible for small laboratories. However, the revolution brought by next-generation sequencing technologies has not come without drawbacks. The reads lengths, even if constantly increasing, are small, and the dedicated assemblers generate more fragmented assemblies, mostly due to difficulties in the resolution of repeats and sequencing biases. As a result, albeit considerably reduced sequencing costs and increased throughput, assembly validation and finishing steps remain labour-intensive and costly, and thus are often left aside sequencing projects, with contig- or scaffold-state sequences being the end product.

Molecular Combing represents an alternative and complementary technology, able to improve the quality of assemblies. Molecular Combing is a technology allowing to linearly and homogeneously stretch hundreds to thousands of genomes under the form of lone DNA fibres on a single microscopy coverslip. This technology enables the accurate positioning and measurement of genomic entities on the DNA molecule.
By providing a mean to easily position, order and orient all contigs from a *de-novo* sequencing project on a whole genome Combing-based physical map, the effort required to obtain a high quality genome sequence is considerably reduced. A proof of concept on several genomes with different levels of complexity, and belonging to the genus of bacteria, yeasts and fungi is currently under development. The generation of high resolution, Molecular Combing-based physical maps do not only provide a mean to improve assemblies, but also to study genomic structure and variability between strains or species. This methodology should help reduce experiments and time required to obtain high level assemblies or finished genomes.

HUMAN PIRNAS ARE UNDER POSITIVE SELECTION AND REPRESS TRANSPOSABLE ELEMENTS

Sergio Lukic, David Gould, Kevin Chen

Rutgers University, Genetics, Piscataway, NJ, 08854, [2]Rutgers University, Genetics, Piscataway, NJ, 08854, [3]Rutgers University, Genetics, Piscataway, NJ, 08854

Piwi-interacting RNAs (piRNAs) are a recently discovered class of 24-30nt noncoding RNAs whose best understood function is the repression of transposable elements (TEs) in animal germlines. In humans, sequences derived from TEs comprise ~45% of the genome and there are several active TE families, including LINE-1 and Alu elements, which are a significant source of de novo mutations and inter-individual variability. In the "ping-pong model", piRNAs are thought to alternatively cleave sense and antisense TE transcripts in a positive feedback loop. Since piRNAs are poorly conserved between closely related species, we used a population genomics approach to study piRNA function and evolution. We developed a novel statistical test of natural selection called DEEGEP (Density Estimation by Expansions of Gegenbauer Polynomials) that considers multi-population allele frequency data. We used this test to show that piRNA sequences are under significant selective constraint within humans, even though they have diverged between human and chimpanzee. Next we used copy number variation data to show that there is positive selection for increased piRNA copy number in humans, consistent with a coevolutionary arms race between piRNAs and TEs. To explicitly demonstrate the function of piRNAs as repressors of TEs in humans, we mapped the piRNA sequences to human TE sequences and found strong correlations between the age of each LINE-1 and Alu subfamily and the number of piRNAs mapping to the subfamily. Overall, our results elucidate the function and evolution of piRNAs in humans and highlight the utility of population genomics analysis for studying this rapidly evolving system.

THE ITERATIVE GRAPH ROUTING ASSEMBLER

Lei Chen, Ken Chen, George M Weinstock

Washington University at St Louis, Genome Sequencing Center, St Louis, MO, 63108

The Iterative Graph Routing Assembler (TIGRA) is a De Bruijn graph based assembler, which first generates a De Bruijn graph from genomic reads and then assembles the target genome by navigating and manipulating the graph. The nodes in De Bruijn Graph represent sequences of length K while edges represent K-1 sequence overlaps. For most De Bruijn graph based assemblers to generate good assembly, a series of assembly runs with different K values are done and best result picked, manually usually, from one of them. An advantage of TIGRA is that it can use multiple K values together. This leads to more contiguous contigs by TIGRA. Another advantage of TIGRA is that it records all the Kmer (sequence of length K) frequencies throughout the assembly process. This enables TIGRA to give a rough quality estimation for the assembled bases. TIGRA also comes with a viewer to visualize the connections between the assembled contigs or scaffolds.

Right now, cancer genome sequencing and other human re-sequencing projects put the focus at the variations among human genomes. Since the human reference genome is available, the genomic variations can be studied by first mapping the reads to the reference and calling potential variations based on mapping abnormality, then a local assembly can be carried out to confirm the calls and recover the variations to base-pair resolution. To better assemble the variations locally from diploid human genome data, TIGRA was modified to recover variations from heterozygous regions, and to be able to assemble the distinct alleles with common flanking regions when it's necessary to map them back to the reference.

A HIGH-RESOLUTION VARIATION MAP IN THE MAIZE: SIGNATURES OF RECOMBINATION, SELECTION AND DOMESTICATION

Jer-Ming Chia[1], Aaron Chuah[1], Robert Elshire[2], Qi Sun[2], Sherry Flint-Garcia[3], John Doebley[4], Jeffrey Ross-Ibarra[5], Michael McMullen[3,6], Edward Buckler[2,6], Doreen Ware[1,6]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, 11724, [2]Cornell University, Institute for Genomic Diversity, Ithaca, NY, 14853, [3]University of Missouri, Division of Plant Sciences, Columbia, MO, 65211, [4]University of Wisconsin, Laboratory of Genetics, Madison, WI, 53706, [5]University of California, Department of Plant Sciences, Davis, CA, 95616, [6]United States Department of Agriculture , Agricultural Research Service , Washington DC, DC, 20250

Maize is a highly diverse species that exhibits much phenotypic variation, and high-resolution descriptions of the genetic variation that underlie these have been the focus of recent studies. We have identified millions of segregating SNPs and small indels in a panel of diverse maize lines by generating extremely deep sequencing coverage of the genome. Included in this panel were the parental lines of the maize nested association mapping (NAM) population as well as teosinte inbred-lines. By identifying consecutive windows of contrasting read counts between the B73 reference and the other parental lines, we are also able to determine segments of structural and copy-number variations (CNVs) that segregate. The integrated SNP, indel and CNV variation map presented here will facilitate association mapping of complex traits in the maize, and in doing so accelerate breeding efforts aimed at agricultural sustainability.

We will discuss data from this variation map, focusing on signatures of selection and domestication, and also patterns of recombination across the genome.

CLASSIFYING AND VISUALIZING GENOME WIDE ASSOCIATION STUDY DATA IN THE ARABIDOPSIS 2010 PROJECT

Aaron Chuah[1], Jer-Ming Chia[1], Yu S Huang[2], Genevieve DeClerck[3], Athikkattuvalasu S Karthikeyan[3], Magnus Nordborg[2]

[1]Ware Lab, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, [2]Molecular and Computational Biology, University of Southern California, Los Angeles, CA, 90089, [3]Department of Plant Breeding & Genetics, Cornell University, Ithaca, NY, 14853, [4]U.S. Department of Agriculture (USDA), Agricultural Research Service (USDA-ARS), Ithaca, NY, 14853

To date, Genome Wide Association Studies have been performed to uncover evidence of linkage between genotypic variations and observed phenotypic differences. Typically, the number of SNPs on which the data is measured is vast (hundreds of thousands of SNPs per study) and each study may comprise tens to hundreds of phenotypes. In this poster we discuss efforts made to store and meaningfully present these associations, allowing the user to select and group phenotypes via an ontology-based controlled-vocabulary of phenotype descriptions. We also present two approaches in visualizing GWAS data for 107 published phenotypes from the Arabidopsis 2010 project, harnessing the established Ensembl Variation platform, and independently, the Google Web Toolkit which is capable of efficiently rendering and storing big data in a cloud computing environment.

This work was supported by the following grants:
NSF 0703908 Gramene: A Platform for Comparative Plant Genomics
NSF 0723510 Collaborative Research: An Arabidopsis Polymorphism Database

RELATING UNDERREPRESENTED GENOMIC DNA PATTERNS AND TIRNAS: THE RULE BEHIND THE OBSERVATION AND BEYOND

Miklos Cserzo, Gabor Turu, Peter Varnai, Laszlo Hunyady

Semmelweis University, Department of Physiology, Budapest, H - 1094, Hungary

One of the central problems of post-genomic biology is the understanding of regulatory network of genes. Traditionally the problem is approached from the protein-DNA interaction perspective. In recent years various types of noncoding RNAs appeared on the scene as new potent players of the game. The exact role of these molecules in gene expression control is mostly unknown at present, while their importance is generally recognized.

The Human and Mouse genomes have been screened with a statistical model for sequence patterns underrepresented in these genomes, and a subset of motifs, named 'spanions', has been identified. These motifs are arranged in clusters at close proximity of distinct genetic landmarks. The findings are in agreement with the known C/G bias of promoter regions while access much more sequential information than the simple composition based model.

In the Human genome the recently reported transcription initiation RNAs (tiRNAs) are typically transcribed from these spanion clusters according to the presented results. Apparently, the model access the common statistical feature of this new and mostly uncharacterized non-coding RNA class and, in this way, supports the experimental observations with theoretical background.

The presented results seem to support the emerging model of the RNA-driven eukaryotic gene expression control. Beyond that, the model detects spanion clusters at genetic positions where no tiRNA counterpart was considered and reported. The GO-term analysis of genes with high concentration of 'spanion' clusters in their promoter proximal region indicates involvement in gene regulatory processes. The results of the analysis suggest that the gene regulatory potential of the small non-coding RNAs is grossly underestimated at
present.

THE VARIANT CALL FORMAT AND VCFTOOLS

Petr Danecek[1], Adam Auton[2], Gonçalo Abecasis[3], Kees Albers[1], Eric Banks[4], Mark A DePristo[4], Bob Handsaker[4], Gerton Lunter[5], Gabor Marth[6], Steve Sherry[7], Gilean McVean[8], Richard Durbin[1]

[1]Wellcome Trust, Sanger Institute, Cambridge, CB10 1SA, United Kingdom, [2]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, United Kingdom, [3]University of Michigan, Center for Statistical Genetics, Ann Arbor, MI, 48109-2029, [4]Broad Institute, Genome Sequencing and Analysis, Cambridge, MA, 02142, [5]University of Oxford, Department of Physiology Anatomy & Genetics , Oxford , OX1 3QX, United Kingdom, [6]Boston College, Department of Biology, Chestnut Hill, MA, 02467 , [7]National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD, 20894, [8]University of Oxford, Department of Statistics , Oxford , OX1 3TG , United Kingdom

One of the main uses of next-generation sequencing is to discover variation amongst large populations of related samples. Recently the format for storing next-generation read alignments has been standardised by the SAM/BAM file format specification. This has significantly improved the interoperability of next-generation tools for alignment, visualisation, and variant calling. We propose the Variant Call Format (VCF) as a standarised format for storing the most prevalent types of sequence variation, including SNPs, indels and larger structural variants, together with rich annotations. VCF is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project, and has also been adopted by other projects such as UK10K, dbSNP, or the NHLBI Exome Project. VCFtools is a software suite that implements various utilities for processing VCF files, including validation, merging and comparing, and also provides a general Perl API. The VCF specification and VCF tools are available from http://vcftools.sourceforge.net.

THE RATIONALISER: AN ORGANISM-SPECIFIC APPROACH TO CURATING CONTROLLED VOCABULARIES IN CHADO DATABASES

Nishadi De Silva*, Adrian R Tivey*, Robin Houston, Matthew Berriman, Jacqueline A McQuillan

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

Established ontologies like the Gene Ontology (GO) play a pivotal role in genome analysis. They standardise and disambiguate the language used in gene annotations, and allow interoperability between genome databases. Database schemas such as Chado that are widely adopted by the genome informatics community are founded on the use of such standardised ontologies. However, these ontologies do not yet provide all the vocabularies that annotators need.

When there are no standard vocabularies, the practice has been to use free text. For instance, additional information about genes could be entered as notes in free-form text boxes. It became clear, over time, that these free-form notes contained common terms that were repeatedly used by annotators, and thus they started to resemble basic controlled vocabularies (CV). These CVs can also be stored in Chado databases. For instance, gene product names in the GeneDB database (www.genedb.org) at the Wellcome Trust Sanger Institute are maintained in a CV. However, due to misspellings and different naming conventions in the unformatted free text, these CVs contain many inconsistencies. For instance, one annotation can read ‘putative kinase’ and another ‘kinase, putative’ when they both ought to be the same. Curators are keen to fix such problems to improve the quality of the data. However, short of altering the database using Structured Query Language (SQL), identifying and correcting these problematic terms in all the annotations is non-trivial. Users can use tools such as Artemis to change conflicting terms. However, this would need to be done for every gene and, can become tedious and error-prone.

The Rationaliser is a tool that allows annotators to do this easily and, effectively, retrospectively create controlled vocabularies from free text lists. It can be configured to work on any CV in a database that implements the Chado schema. The user can browse through the terms and correct them for all the relevant annotations with the click of a button. Users can search terms and check any evidence codes to further aid their decisions. Moreover, organism-specific users can opt to make changes within the scope of their selected organisms; thus not conflicting with someone else's work. The Rationaliser can also suggest possible fixes for a chosen term and, since it runs via Java Web Start, requires no prior software installation. This tool has undergone several cycles of improvements, and is a work in progress.

* The authors contributed equally

HIGH THROUGHPUT RNA-SEQ DATA REVEALS THOUSANDS OF NOVEL SPLICING EVENTS.

Alexander Dobin, Carrie A Davis, Felix J Schlesinger, Chris Zaleski, Philippe Batut, Thomas R Gingeras

Cold Spring Harbor Laboratory, Functional Genomics, Cold Spring Harbor, NY, 11724

High throughput RNA-seq technology provides an opportunity for genome-wide studies of the gene isoforms. Reconstruction of transcripts from RNA-seq data presents a unique challenge because it comprises hundreds of millions of relatively short sequence fragments of the original RNA molecules which are derived from discontinuous regions of the genome. To address this problem we have developed the Spliced Transcript Alignment and Reconstruction (**STAR**) tool, which allows *de novo* detection of splice junctions in RNA-seq data. STAR aligns paired-end reads with several mismatches, identifies multiple splice junctions per read, detects poly-A tails, and trims out low quality tails at a speed of 60 Million 76nt reads per hour for the human RNA-seq data. As part of the **ENCODE** project, over **18 billions** reads from poly-A+/A- cytosolic and nuclear long RNA of 14 human cell lines were sequenced with Illumina 2x76 stranded paired end protocol. STAR was used to analyze this dataset and to identify a variety of previously unreported splicing events. We detected tens of thousands of **un-annotated canonical junctions** expanding significantly the collection of known gene isoforms. Many of these isoforms are expressed differentially in studied cell lines, and are present at different levels in separate cell compartments (cytosol vs. nucleus). In addition, hundreds of **chimeric RNA transcripts**, including canonical splicing events between exons of the neighboring genes, as well as inter-chromosomal, inter-strand and very long range junctions were also detected. Highly expressed, **previously unknown non-canonical** (not GT/AG intron motif) junctions were also observed in this dataset. The most abundant intron motifs for these junctions are single nucleotide variations from the canonical GT/AG motifs. In addition, a large number of reads were found to be spliced just a few bases away from the canonical junctions. While a majority of these junctions are likely to be caused by the **spliceosome errors**, we observed a slight enrichment for the junctions that conserve the open reading frame, which makes them capable of giving rise to proteins with a few inserted or deleted amino acids. To computationally validate our findings, we used exhaustive alignment algorithms for a small subset of reads, and also demonstrated the enhanced evolutionary conservation of the identified junctions. We will also present our strategies for experimental validation of the different types of the novel junctions.

ANALYIS PIPELINE FOR EXOME SEQUENCING DATA

Sebastian H Eck, Elisabeth Graf, Anna Benet-Pagès, Thomas Meitinger, Tim M Strom

Helmholtz Zentrum München, Institute of Human Genetics, Munich, 85764, Germany

Enrichment techniques for targeted sequencing of coding regions are currently applied to identify rare variants. We developed a pipeline to analyze exome sequencing data. The pipeline is a collection of Perl scripts which start with the sequence files generated by the Illumina software. It calculates quality metrics and performs read alignment to the reference sequence, variant calling, variant annotation and selection of candidate variants according to the genetic model. Subsequently variants are stored in a database.

Alignment and variant calling is performed with BWA and SAMtools. Subsequently, two main tasks are performed. First, quality metrics are calculated which include base quality, % mapped reads, % duplicates, % reads overlapping the target regions and read depth an a single base level. As a second task, variants are further filtered and annotated. Annotation includes presence in dbSNP, type of mutation and - if applicable - amino acid change. In addition, the frequency of the variants in our exome samples is determined. These information is then used, in conjunction with optional information such as inheritance model, affected siblings or a linkage region, to identify putative causative variants.

The pipeline has a modular composition and subsets of components may be run, for example to repeat key analysis steps with different parameters. For convenient manual inspection of the results, bed- and html-files usable with the UCSC Genome Browser are provided.

We applied the analysis pipeline to approximately 40 exomes. From an average of 6.5-7 GB of aligned sequence per exome, the pipeline calls ~16,000 coding variants. Approximately 7,500 of these are non-synonymous variants of which ~700 along with ~30 splice site variants and ~60 indels are not present in dbSNP (version 130). Depending on the number of affected individuals and the underlying inheritance model, we were able to confine this list to 1-10 putatively disease causing variants.

MAPPING AND ANALYSIS OF DNASE 1 HYPERSENSITIVITY IN MOUSE ESCS

Andre J Faure[1], Daniel Sobral[1], Yoichiro Shibata[2], Nathan Johnson[1], Damian Keefe[1], Steven Wilder[1], Ian Dunham[1], Paul Tesar[3], David Adams[4], Greg Crawford[2], Paul Flicek[1]

[1]European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom, [2]Duke University, Institute for Genome Sciences and Policy (IGSP), Durham, NC, 27708, [3]Case Western Reserve University, Department of Genetics, Cleveland, OH, 44106, [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1HH, United Kingdom

Large-scale remodelling of the chromatin landscape in embryonic stem cells (ESCs) accompanies changes in their transcriptional programme during differentiation. The interplay between these two factors suggests that characterising the chromatin state of ESCs is key to understanding such properties as their pluripotency and capacity for self-renewal.

We employed high-throughput sequencing to identify 136,115 DNase 1 hypersensitive sites (DHSs) within mouse ESCs. This map of open chromatin is expected to represent the vast majority of important regulatory regions in these cells. When compared to DHSs present in diverse human cell lines from the ENCODE project, we found that roughly half of those within alignable regions are unique to mouse ESCs. DHSs shared in at least one other human cell line occur mostly at the 5' end of genes, whereas more than half of unique DHSs occur further than 20kb away from the nearest annotated transcript. Genes proximal to DHSs shared across all cell lines are enriched for "house-keeping" functions whereas those near unique DHSs are enriched for terms related to embryogenesis.

By examining read profiles at genomic features we also reveal interesting properties of nuclear lamina-associated domains (LADs) and highlight differences between gene categories, including those that are/are not up-regulated in the pluripotent state. Focussing on gene-distal DHSs, we find that shared sites are enriched for promoter-associated histone modifications, many of which possess bivalent marks (H3K4me3/H3K27me3), whereas unique sites are more likely to be associated with LADs. Our results from *de novo* motif discovery further implicate CTCF in the maintenance of these repressive domains and reveal that a number of known developmental transcription factor motifs are overrepresented in gene-distal DHSs. Lastly, we provide evidence to suggest that these motifs represent the sites of bound factors by measuring a "footprint effect", or local depletion in DNase 1 cleavage events.

Our results from the analysis of this map of chromatin accessibility in mouse ESCs demonstrate the utility of this dataset and we expect that it will be a valuable resource for future research.

UNRAVELING THE COMMON FUNCTIONAL THEMES OF THE ENIGMATIC SIGMA FACTOR 54

Christof Francke[1,5], Tom Groot Kormelink[1,2,5], Yanick Hagemeijer[5], Lex Overmars[1,5], Vincent Sluijter[5], Roy Moezelaar[3], Roland J Siezen[1,4,5]

[1]TI Food and Nutrition, Kluyver Centre for Genomics of Industrial Fermentation and the Netherlands Bioinformatics Centre, Wageningen, 6700AN, Netherlands, [2]Wageningen University and Research Centre, Department of Microbiology, Wageningen, 6703HB, Netherlands, [3]Wageningen University and Research Centre, Food and Biobased Research and TI Food and Nutrition, Wageningen, 6700AA, Netherlands, [4]NIZO food research, bioinformatics, Ede, 6710BA, Netherlands, [5]Radboud University Nijmegen Medical Center, Center for Molecular and Biomolecular Informatics, Nijmegen, 6500HB, Netherlands

Transcription in bacteria necessitates the binding of a sigma factor at the promoter to recruit RNA polymerase. Bacteria often contain various sigma factors: the 'household' sigma factor 70 (sigma G) is responsible for ordinary gene expression, whereas several systemic responses are mediated by specialized sigma factors, like sigma S and sigma B for the response to stress in *Escherichia coli* and *Bacillus subtilis*, respectively. There is however one sigma factor that does not fit the general pattern. The enigmatic sigma factor 54 (sigma L or RpoN) is the sole representative of a different sigma factor protein family, recognizes a -24, -10 promoter sequence and, in contrast to the other sigma factors, requires the action of an enhancer protein to deliver the free energy to open the DNA.

Sigma 54 is found in bacteria of most phyla and has been linked to seemingly unrelated processes like motility, nitrogen assimilation, osmotolerance, virulence and biofilm formation. We have analyzed the presence and absence of sigma 54 and its enhancers in all sequenced genomes using a generic combined strategy based on Blast and similar motif searches. The strategy allowed improving the function annotation of far more than half of the enhancers. The similar motif search approach was then used to identify putative promoters and therewith putative species specific sigma 54 regulons. Finally, we have employed the concept of context bias to unravel surprising cross phyla common functional themes for sigma 54 mediated transcription regulation.

MANUAL CURATION IN UNIPROT KNOWLEDGEBASE: ENSURING COMPREHENSIVE AND ACCURATE REPRESENTATION OF PROTEIN DATA.

Michael J Gardner[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom, [2]Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, 1211, Switzerland, [3]Protein Information Resource, Georgetown University, Washington, DC, 20057-1414

The explosion in the availability of genomic and proteomic data over the last decade has generated both unrivalled opportunities and unexpected problems for biologists. One of the key challenges lies in the accurate collation and representation of biological knowledge in a manner that allows easy access on both an individual and programmatic basis. The UniProt Knowledgebase (UniProtKB) aims to provide the scientific community with a consistent and authoritative resource for protein sequence and functional information. Central to these efforts is the process of manual curation of entries. Manual curation involves the extraction and appraisal of experimental results from scientific literature and the use of a range of analysis programmes to ensure that sequences and sequence features are correctly reported. Consequently, manual curation plays a vital role in providing users with a complete overview of available data while ensuring its accuracy, reliability and accessibility. High-quality manual curation also provides the accurate and standardised data necessary for development of automated methods which enable the automatic annotation of uncharacterised proteins. Ongoing efforts in manual curation continue to include additional functional information as new data become available. All data are freely available from www.uniprot.org.

INTEGRATED GENOMICS BASED ABERRANT DIFFERENTIAL METHYLATION PROFILES IN NSCLC

Srinka Ghosh[1], Thomas Holcomb[2], Kimberly Walter[2], Thomas Januario[2], Robert Yauch[2], Lukas Amler[2], Robert Soriano[1], Zora Mordusan[1], Somasekar Seshagiri[1], David Shames[2]

[1]Genentech, Research, South San Francisco, CA, 94080, [2]Genentech, Development, Oncology Diagnostics, South San Francisco, CA, 94080

Lung cancer, constitutes two major histological subtypes: small cell lung cancer(SCLC), and non-small cell lung cancer(NSCLC). Research has shown that important biological differences exist within NSCLC that might be evident at disease onset. Specifically, NSCLC can be classified as epithelial-like(EL) and mesenchymal-like(ML) tumors. The transition from an epithelial to a mesenchymal state (EMT) is thought to correlate with increased malignancy; also there's evidence of an association of EMT biomarker status with EGFR tyrosine kinase inhibitor(TKI) sensitivity. Interrogating the molecular basis of such differences will significantly enhance our understanding of tumorigenesis and subsequent therapeutics development in NSCLC. To this end we've undertaken an integrated genomics approach, employing both gene expression profiling and methyl-DNA immunoprecipitation(meDIP). In the upcoming phase of the study, meDIP-seq will be used to acquire a whole-genome map of the methylation status of the CpG-rich sequences.

In the initial phase, 12 NSCLC cell-lines(6 each, EL and ML) without EGFR mutations were used. Samples were treated with the DNA methylation inhibitor 5-aza-2'-deoxycytidine(5-azadC) and gene induction was confirmed using qPCR. Gene expression(Affymetrix HGU133P) studies were performed on the DMSO(mock) and 5-azadC treated cell-lines. To ensure detection of differentially methylated regions(DMRs) in alternatively spliced and/or methylated variants of genes(eg. RASSF1A, p16), meDIP(Nimblegen Promoter2.0) was performed in parallel. An empirical Bayes based analysis was used to identify statistically significant associations of gene expression and the EL/ML phenotypes. The more significant changes were observed in the EL cell-lines, shown to be more sensitive to the EGFR-TKI, erlotinib. While there is overlap between the two array platforms, there are subsets of genes that are unique to each. This could be partially explained by the variation in the genomic regions interrogated by the two. Furthermore, the meDIP data required lowess normalization. The fitting used a subset of probes not enriched for CpG sequences. However, we confirmed by direct sodium bisulfite sequencing, that both approaches identify loci that are differentially methylated between the ML and EL cell-lines. The DMRs will be further characterized via meDIP-seq and validated using the infinium methylation assay(Illumina). The goal is to utilize these DMRs as potential anchors, in a wide panel of NSCLC tumors and cell-lines for biomarker characterization.

DECONVOLUTION OF BAF60-BASED SWI/SNF COMPLEXES IN MUSCLE STEM CELLS.

Lorenzo Giordani[1], Sonia Albini[1], Sonia Forcales[1], Pier Lorenzo Puri[1,2]

[1]Sanford-Burnham Institute for Medical Research, Cancer Epigenetics, La Jolla, CA, 92037, [2]Dulbecco Telethon Institute (DTI), IRCCS Fondazione Santa Lucia and European Brain Research Institute, Epigenetic pharmacology for regenerative medicine, Roma, 00143, Italy

The SWI/SNF chromatin remodeling complex plays a crucial role in chromatin remodeling and epigenetic regulation of gene expression. SWI/SNF activity is regulated in a spatio-temporal manner and this fine modulation is achieved through the combinatory assembly of the different subunits. The complex contains approximately 10 protein components and has two mutually exclusive catalytic subunits, Brm or Brg1. Formation of different SWI/SNF sub-complexes is emerging as a regulatory mechanism to control the transcription of distinct sub-sets of genes involved in different cellular functions.
We have shown that the structural sub-unit BAF60c is important for the signal-dependent recruitment of the SWI/SNF complex to muscle genes, via BAF60c phosphorylation by the p38 pathway (Simone et al. 2004 Nat Gen; Forcales et al. submitted). Our data support the formation of a BAF60c-based SWI/SNF complex dedicated to the activation of muscle gene expression in muscle progenitor cells, in response to cues that activate the p38 pathway.
There are three BAF60 variants, coded by different genes - BAF60a, b and c - which are mutually exclusive within the SWI/SNF complex and appear to be functionally distinct. Thus, we decided to characterize the gene network regulated by each BAF60 subunit in association with Brg1 or Brm, during myogenic differentiation. To this purpose, we used RNAi-mediated knockdown of each BAF60 variant and Brg1 or Brm in skeletal myoblasts and performed a gene microarray, which revealed the existence of specific sub-sets of genes regulated by each “minimal” BAF60-Brg1/Brm sub-complex. Our data indicate that BAF60 variants are key regulators of SWI/SNF complex composition, response to signal and promoter-specific activity. We are currently exploiting this system to further deconvolute the role of BAF60 variants, in association to Brg1 or Brm-based SWI/SNF complexes, in regulating gene expression during skeletal myogenesis, using Chip-based genome wide technique and mass spectrometry analysis.

INFERENCE OF SHARED ANCESTRAL HAPLOTYPES IN POPULATION ISOLATES

Dominik Glodzik[1], Paul McKeigue[2], Ruth McQuillan[2], Alan Wright[1], Harry Campbell[2], James F Wilson[2]

[1]MRC, Human Genetics Unit , Edinburgh, EH4 2XU, United Kingdom, [2]University of Edinburgh, School of Public Health, Edinburgh, EH8 9AG, United Kingdom

In an isolated population, because of recent common ancestry, it is possible to infer haplotypes from genome-wide SNP genotypes. Kong et al (2008, 2009) have described methods for this, based on screening of pairs of unrelated individuals to detect identity-by-descent (IBD) sharing, followed by reconstruction of the shared haplotypes and use of genealogical data to stitch these haplotypes together. A limitation of this method is that the genome is divided into chunks of fixed length (6 to 10 cM) and that sharing is enforced to start and end at the boundaries between these chunks.

We have developed a program ANCHAP to infer shared ancestral haplotypes in population isolates that overcomes these limitations. ANCHAP does not subdivide the genome into regions of fixed length, but uses a sliding window to infer IBD sharing across the entire genome, allowing break-points to occur at any position. This is more sensitive, but subsequent reconstruction of shared haplotypes becomes more challenging. Reconstruction is based on alignment of genotype sequences sharing a haplotype with the proband, so that the sequences are classified into to groups: one containing the proband's maternal and one the paternal haplotype. The alignment-based method does not require genealogical data, and can thus be used to stitch together IBD segments inherited from a common ancestor in the remote past. A robust implementation in R has been provided, which allows visualisation of haplotype alignments and inconsistencies.

The method was applied to a cohort of 749 individuals from the Orkney islands, who were genotyped with 300k Illumina SNP arrays. The mean number of IBD sequences for one proband at one locus was 6.61, of which 5.27 could be aligned consistently. On average, 87% of the genome was phased, with mean 1.86 continuous regions of phased haplotypes per chromosome. At an average locus, the most common haplotype occurred 32 times in the cohort. 9% of unique haplotypes occurred at least twice, and 0.8% more than ten times. The accuracy and consistency of phasing without genealogical data was confirmed by examining the inferred haplotypes of parent-offspring pairs.

The method has several applications, using either the inferred haplotypes, or the IBD sharing structure discovered. One application is to test traits for association with ancestral haplotypes rather than tag SNP alleles: this may make it possible to detect large effects of untyped alleles that are rare in the general population but have drifted to high frequency in the isolate. Another is to optimize the design of resequencing studies so that a minimal subset of individuals can be resequenced and the sequences of others inferred on the basis of their shared ancestral haplotypes.

DECIPHERING THE MOLECULAR TRAJECTORY IN DARJEELING TEA UNDER BIOTIC AND ABIOTIC STRESS

Bornali Gohain[1], Sangeeta Borchetia[1], Tirthankar Bandyopadhyay[1], Priyadarshini Bhorali[1], Raju Bharalee[1], Sushmita Gupta[1], Sourabh k Das[1], Neeraj Agarwal[1], Parveen Ahmed[1], Prasenjit Bhagawati[1], Neelakshi Bhattacharyya[1], Chiranjana Borah[1], M.C Kalita[2], Sudripta Das[1]

[1]Tea Research Association, Biotechnology, Jorhat, 785008, India, [2]Gauhati University, Biotechnology, Gauhati, 786014, India

Darjeeling teas are the highest grown teas in the world (40~ feet) and preferred for its flavour, aroma and quality. Apart from the genetic makeup of the plant, earlier reports suggest that insect infestation, particularly greenflies and thrips triggers the aroma and flavor formation in Darjeeling tea. This work determined the pattern of expression of genes from tea, mainly enzymes involved in the production of major classes of flavor metabolites (terpenoids) during insect infestation, at different stages of tea manufacturing and after mechanical injury. The selection of genes responsible for regulating flavour and aroma was based on a Subtractive Suppression Hybridization of B157 (tea clone with thrip infestation) with a control tea clone (without infestation) providing us around 500 ESTs encoding for transcripts viz. major monoterpenes encoding enzymes and precursors involved in hydrolysis of glycosides related to aroma and flavor, transcription factors, carbohydrate metabolizing enzymes. cDNA libraries from Camellia sinensis var.sinensis clone B157, clone AV2, clone SUNGMA following insect infestation and also during different stages of tea manufacturing viz. withered, 2nd hand rolled and mechanically injured were prepared. The expression profiles of the genes during infestation and manufacturing stages was evaluated by semi quantitative PCR and quantitative real-time PCR. We found the expression of genes viz. leucine zipper, ntd, nced, geraniol synthase, Rsa, Tsp, Amy, Farnesyl transferase, catalase, o methyl transferase, linalool synthase, peroxidases, elicitor responsive proteins, Linamarase, nerolidol linalool synthase 2, 12-Oxophytodieonate reductase, glucosidase, MYB transcription factor, alcohol dehydrogenase was regulated considerably due to insect attack, hand rolled and wounding. This regulation of gene expression can be extrapolated with increase in volatiles which is responsible for enhancing the quality of Darjeeling tea specially the strength and aroma of the teas. We hope to model these responses in relation to Darjeeling tea flavour and aroma, and thus understanding the molecular changes that occur during tea manufacturing.

DASH: DRAFT GENOME ANNOTATION BY STRENGTH OF HOMOLOGY

Allison Griggs, Clint Howarth, Matthew Pearson, Qiandong Zeng, Brian Haas

Broad Institute, Genome Sequencing and Analysis, Cambridge, MA, 02141

With the advent of next generation sequencing technologies, the bottleneck of genome data release has shifted to genome annotation, specifically the volume of manual curation required to generate high quality gene sets. The need for an automated genome annotation pipeline has been addressed by many groups, but the resulting gene sets from their processes do not have accuracy levels comparable to those that are manually curated. The automated creation of high-quality gene sets will become increasingly important as the number of new genome sequences for pathogens and other reference bacteria will soon be orders of magnitude higher.

Here, we present a new method for the generation of a structural genome annotation that can be applied to both prokaryotes and eukaryotes on an individual or comparative basis. We start with a candidate gene set of various ab initio gene predictions and BLAST-based open reading frames (ORFs). Each candidate gene is scored using BLAST evidence and PFAM domains. Then, accounting for positional constraints as well as tRNA and rRNA conflicts, we determine the highest scoring gene structure(s) at each locus.

Gene sets from our method, when compared to finished reference genomes, yield significant improvements over other pipelines. In performance benchmark test on E.coli K12, we correctly identified 97% of all genes, while the next best pipeline identified 88%. These improvements are a result of several factors. First, we do not rely on a solitary ab initio gene set; no single gene predictor identifies all of the genes in any reference genome. By combining several predictions, we obtain much higher coverage of the reference gene sets. A combination of Glimmer3, Metagene and Genemark predict 7% more of E.coli genes than Glimmer3 alone. Second, by using BLAST-based ORFs, we are able to identify putative pseudogenes that are not properly identified by ab initio gene predictors. Finally, by filtering gene structures by length and evidence score, our false positive rate is half that of the next best pipeline.

DASH provides for rapid and accurate gene structure annotation, closely approximating the product of manual annotation. The evidence for a DASH gene set can be generated in 12-24 hours, from which a gene set can be generated in under 10 minutes. DASH gene sets have been used to annotate more than 75 prokaryotic and 10 eukaryotic genomes over the course of 3 months, a rate that can scale with the demands of high-throughput genome sequencing.

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.

BACORTH: COMPUTING THE BACTERIAL ORTHOLOGUES

Mihail R Halachev, Nicholas J Loman, Mark J Pallen

University of Birmingham, School of Biosciences, Birmingham, B15 2TT, United Kingdom

Our goal is to compute orthologues between all complete bacterial genomes. Traditional approaches typically rely on all-against-all BLAST searching. With more than 1,000 completed bacterial genomes available in GenBank, we estimate computational effort for such an approach to take ≥ 5 years using a standard server (4 2.3GHz CPUs, 16 GB RAM).

We propose a system for ongoing orthologue annotation, which combines accuracy with speed. We also take into account the changing nature of genome annotations. In our approach, named BacOrth, we analyse 1,062 complete genomes from the NCBI's RefSeq collection (Release 39) containing 3,494,808 coding sequences (CDS). At the species level, we perform six-frame whole-genome pair-wise alignments in amino-acid space using the PROMER tool [http://mummer.sourceforge.net]. Orthologues are detected by PROMER computation of the similarity between CDSs in the aligned regions and selecting reciprocal best hits. This stage took 239 hours. Proceeding in the same brute-force manner for higher taxonomic levels would result in an exponential computational cost.

Instead, to reduce the number of alignments, we adopted a pragmatic scheme that trades accuracy for speed. For each species within a genus, we compute the pan-genome, a non-redundant set of CDSs from all genomes from the species - it is usually significantly smaller than the set of all CDSs from the species. Next, the orthologues between each pair of species pan-genomes are computed (using PROMER) and are mapped to each pair of genomes from this pair of species. The computation at higher taxonomic levels proceeds in similar manner, although to account for increasingly divergent nature of the genomes at family level and above, the CDS alignment is performed with the more sensitive (but slower) OrthoMCL tool [www.orthomcl.org]. Currently, we have finished the orthologue computation at species, genus, family, order and class levels, and found more than 70 million orthologous pairs, exploiting 33 days of computer time. Upon completion of the remaining levels, the orthologue and paralogue information will be made available as part of our xBASE database [http://xbase.bham.ac.uk].

There are numerous applications of the BacOrth database. For example, an analysis of the nine *H. pylori* genomes suggests that ~ 70 novel CDS are to be expected when sequencing the next genome and ~ 6 CDS for the 1,000th *H. pylori* genome. A comparison of the 32 *Streptococcus* genomes reveals eight CDSs with known function that occur in all 13 *S. pyogenes* genomes, but do not occur in the 11 *S. pneumoniae* or the 5 *S. suis* genomes. A phylogenetic study based on the number of orthologues found in the 36 *Escherichia*, 8 *Shigella* and 19 *Salmonella* genomes confirms that the genus *Shigella* is nested within the species *E. coli*, while *Escherichia* and *Salmonella* form 2 distinct and separate clusters.

A NOVEL NETWORK PROFILING OF GENE EXPRESSIONS IN HUMAN ADIPOSE TISSUE REVEALS AN IMPORTANT REGULATOR OF ADIPOCYTE HYPERTROPHY AND FUNCTION

Kazuo Hara[1], Momoko Horikoshi[1], Teppei Shimamura[2], Seiya Imoto[2], Satoru Miyano[2], Takashi Kadowaki[1]

[1]University of Tokyo, Department of Metabolic Diseases, Tokyo, 113-8655, Japan, [2]University of Tokyo, Laboratory of DNA Information Analysis, Tokyo, 108-8639, Japan

Adiponectin is a protein synthesized exclusively in adipose tissue and there are robust correlations of its levels with adipocyte hypertrophy, insulin resistance and risk of type 2 diabetes. However, the question concerning a transcriptional regulatory mechanism induced by adiponectin and its effect on adipocyte hypertrophy and function remains unsolved. To reconstruct adiponectin-dependent transcriptional regulatory networks, we performed network profiling analysis on gene expression data derived from human adipose tissue.

Network profiling is designed to build modulator-dependent transcriptional regulatory networks from gene expression data. A modulator is defined as a molecule affecting a regulation system between transcription factors and their target genes. A valuable feature of network profiling is that most of the traditional approaches only focused on the construction of a "static" network whose structure does not change under a process, while network profiling infers a sequence of networks related with a user-defined modulator and enables us to identify the variation of these network structures with a change of the modulator. Network profiling uses two inputs: (1) a gene expression matrix whose columns indicate patients and rows indicate genes and (2) its corresponding values of a user-defined continuous process. The key assumption of network profiling is that the network structure varies smoothly across the process, that is, the regulations between genes can be described as a smooth function of the process by a varying-coefficient structural equation model. Samples of adipose tissue were obtained from 60 Japanese subjects who underwent plastic surgery. Expectedly, in the constructed networks, *PPARG* and *CEBPA* have regulated the expression level of genes involved in adipocyte hypertrophy according to the alteration of adiponectin gene expression. We observed that *CDKN2C* had outstanding regulation effects on expression levels of genes related to mass of white adipose tissue and lipid metabolism. CDKN2C is a member of cyclin-dependent kinase inhibitors and reported to be remarkably induced during differentiation of 3T3-L1 to adipocyte. The present result indicates that CDKN2C is a putative master regulator of adipocyte differentiation and hypertrophy induced by the expression of adiponectin in human. The present study suggests that network profiling analysis is useful to explore the pathogenesis of human diseases using human tissues.

CLINICAL SEGMENTATION – THE ESSENTIAL KEY TO ANNOTATION.

Fritz E Hauser

Project SEGMENTA, Clinical Segmentation, Bad Lippspringe, D - 33167, Germany

“There’s no real creativity going on in the mammalian genome” (1) as long as one neglects the clinical features determined by them. Honorable attempts like CHANG (2) show that annotation will only be successful, when the principal of clinical segmentation “Patterns to Genes” is being observed.
Clinical segmentation (3) originating from dermatologic pattern formation and viscero-cutan reflexes (4) is found in most diseases and helps to read this creativity in humans and mammalians in order to develop new medical therapies (5).
Examples in infections, tumors, i.e. breast cancer, and metastases (6) will be shown and how diseases keep on this road map of segmentation. When aiming towards therapies for melanoma studying this clinical map therefore is likely to be more efficient than current techniques (7).
The applications of clinical segmentation in many diseases of the Project SEGMENTA, an upcoming research firm of Medical Scientific Knowledge Enterprise, is essential in order to achieve reliable returns of investments in sciences, biotechnology, and pharmaceutical industries basing on THE NEW MEDICAL ENTITIES, almost as “easy” as in drosophila: Patient + Pattern = Discovery + Product.

(1) [http://www.broadinstitute.org/news/102] introducing: Clamp M et al. Distinguishing protein-coding and noncoding genes in the human genome. Proc. Natl. Acad. Sci. USA. DOI: 10.1073/pnas.0709013104.
(2) Chang HY: Anatomic Demarcation of Cells: Genes to Patterns. In the special "Spatial Cell Biology - Location, Location, Location". Science, 27 November 2009, [http://www.sciencemag.org/content/vol326/issue5957/index.dtl].
(3) Hauser FE: Clinical segmentation as a major promotor of gene therapy applications. Nature Biotechnology symposium "Gene Therapy: Delivering the Medicines of the 21st Century." Washington, DC, USA, Nov 7-9, 1999.
(4) Hauser W: Lokalisationsprobleme bei Hautkrankheiten. In: Korting GW (1980) Dermatologie in Klinik und Praxis. Bd. I. Allgemeine Dermatologie. Stuttgart, New York 1980, 8.60-8.93.
(5) Hauser FE: Clinical Segmentation & New Medical Entities: Innovation in developing future Blockbusters. Lecture at NOVARTIS Headquarters, Basle, CH, October 24, 2008.
(6) Hauser FE: Clinical Segmentation and Signal Transduction in Cancer. Miami 2005 nature biotechnology winter symposia, Miami, Florida, USA, February 5-9, 2005. [http://www.med.miami.edu/mnbws/documents/HauserSR05.pdf].
(7) Falchi M, et al: Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. Nature Genetics, published online: 5 July 2009 | doi:10.1038/ng.410.

VITAMIN D RECEPTOR BINDING IN DISEASE AND EVOLUTION

Sreeram V Ramagopalan*[1,2], Andreas Heger*[3], Antonia J Berlanga[1,2], Narelle J Mageri[1], Matthew R Lincoln[1,2], Lahiru Handunnetthi[1,2], Sarah-Michelle J Orton[1,2], Adam E Handel[1,2], Corey T Watson[4], Julia M Morahan[1,2], Gavin Giovanni[5], Chris P Ponting[3], George C Ebers[1,2], Julian C Knight[1]

[1]Wellcome Trust, Centre for Human Genetics, Oxford, OX3 7BN, United Kingdom, [2]University of Oxford, Department of Clinical Neurology, Oxford, OX3 9DU, United Kingdom, [3]University of Oxford, MRC Functional Genomics Unit, Oxford, OX1 3QX, United Kingdom, [4]Simon Fraser University, Department of Biological Sciences, Burnaby, V5A 1S6, Canada, [5]Queen Mary University of London, Blizard Institute of Cell and Molecular Science, London, E1 2AT, United Kingdom

*authors contributed equally

Initially thought to play a restricted role in calcium homeostasis, the pleiotropic actions of vitamin D in biology and their clinical significance are only now becoming apparent. However the mode of action of vitamin D, through its cognate nuclear vitamin D receptor (VDR), and its contribution to diverse disorders, remain poorly understood.
Here we present the results of a ChiP-Seq analysis of the VDR after calcitriol stimulation in two lymphoblastoid cell lines. We find 2,776 genomic positions occupied by the VDR. Binding occurs predominantly outside promoter regions and is correlated with signatures of open chromatin. Motif analysis indicates that the majority of binding is due to the VDR receptor and its canonical motif. Binding is associated with regions previously implicated by genome-wide association studies in auto-immune diseases including multiple sclerosis, type 1 diabetes and rheumatoid arthritis. These result suggest that the VDR is strongly involved in the regulation of the immune system. VDR intervals are significantly concentrated among regions identified as having been subject to selective sweeps in individuals of Asian or European (but not African) descent (>1.4-fold; $P < 0.05$). The reasons behind these associations are unclear, but one suggestion is that evolutionary pressure has maintained vitamin D binding in some regions of the genome as humans migrated out of Africa.

TOWARD A TURNKEY SOLUTION FOR GENOME ANNOTATION AND DOWNSTREAM ANALYSIS

Carson Holt, Hadiul Islam, Mark Yandell

University of Utah & School of Medicine, Human Genetics, Salt Lake City, UT, 84114

It is generally accepted that within the next few years it will be possible to quickly sequence even human sized genomes for as little as $1,000, making whole genome sequencing routine even for small laboratories. Unfortunately, advances in genome annotation have not kept pace, and annotation is now the major bottleneck for many genome projects, especially those with limited bio-informatics expertise. These challenges extend beyond merely annotating the genome, as annotations must also be subjected to diverse downstream analyses, the complexities of which confound many smaller genome projects. Having developed a simple tool for automated genome annotation, MAKER[1], we have recently begun developing software to meet the downstream analysis and collaboration management needs of smaller genome projects.

Here we report our progress in creating a web-based, distributed, multi-user environment that permits automated genome annotation and downstream analysis via a simple web-browser. Using the web-based environment users can perform tasks such as identifying protein domains, putative gene functions, and gene ontology terms and add those data directly to the genome annotations. All results are automatically integrated into GMOD and Sequence Ontology tools creating a unified web-based resource for genome project collaboration which allows for easy distribution of data, global analysis of annotation features via SOBA[2], concurrent viewing of annotations via GBrowse[3] and JBrowse[4], and remote manual editing of annotations in the Apollo browser[5], thus greatly increasing the efficiency of genome project collaboration by enabling researchers in different locations to curate and analyze a shared genome data-set remotely. This tool has been used successfully in four recent MAKER driven annotation projects: *Pogonomyrmex barbatus* (Smith et al, submitted PNAS), *Linepithema humile* (Smith et al, submitted PNAS), *Atta cephalotes* (Currie et al, manuscript in preparation),and *Fusarium circunatum*, the first eukaryotic genome sequenced in Africa (Wingfield et al, manuscript in preparation).

1. Cantarel et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188-196 (2008).
2. Moore et al. SOBA: sequence ontology bioinformatics analysis. Nucl. Acids Res. 38, W161-164.
3. Stein et al. The Generic Genome Browser: A Building Block for a Model Organism System Database. Genome Res. 12, 1599-1610 (2002).
4. Skinner et al. JBrowse: A next-generation genome browser. Genome Res. 19, 1630-1638 (2009).
5. Lewis et al. Apollo: a sequence annotation editor. Genome Biology. 3, research0082.1-0082.14 (2002).

THE EBI METAGENOMICS PORTAL.

Christopher I Hunter, Sarah Hunter

EMBL-EBI, Interpro, Cambridge, CB102AY, United Kingdom

The EBI metagenomics portal (www.ebi.ac.uk/metagenomics) has been born of the need to collate and integrate outstanding EBI resources currently used by metagenomics researchers into a centralized and user friendly portal.

The EBI resources of UniProt, InterPro, Ensembl Genomes and IntAct are all used for analysis by metagenomic researchers, but in an ad hoc manner. We intend to provide a user friendly interface to these services allowing protein prediction, function analysis, comparison to complete reference genomes and metabolic pathway analysis, all coupled together with the submission of nucleotide sequence data to the European Nucleotide Archives (EMBL-Bank and SRA). Additionally we intend to offer tools for assembly, phylogeny, bioactive molecule discovery (ChEMBL) and eventually links to other 'omics resources such as PRIDE and ArrayExpress.

Here we present the current analysis pipeline, showing ~25% of reads per (Roche 454) dataset are found to be informative by comparison to known sequences. We also present future additions planned for the pipeline and an example of the analysis summary. This is all work in progress and we are very keen to canvas potential users for their views and opinions on how this resource should evolve in order to reach our goal of becoming the European focal point of metagenomic data archiving and analysis.

A RESOURCE FOR THE RATIONAL SELECTION OF DRUG TARGET PROTEINS AND LEADS FOR THE MALARIA PARASITE, *PLASMODIUM FALCIPARUM*

Christiaan J Odendaal, Claudia M Harrison, Michal S Szolkiewicz, Fourie Joubert

University of Pretoria, Bioinformatics and Computational Biology Unit, Pretoria, 0001, South Africa

The selection of drug targets and lead compounds in malaria has been mostly based on serendipitous discoveries and legacy compounds. The emergence of widespread drug resistance, even against current drugs is making the effective selection of new drug targets together with lead compounds essential. Currently available systems for selecting drug targets in malaria include PlasmoDB, the TDR Targets database and the Tropical Disease Kernel, but there is no system available that offers data mining of parasite proteins in a host-pathogen comparative context, together with ligand information and chemical properties.

The Discovery project is aimed at providing a publicly available informatics resource where comprehensive information on the parasite and host proteins are stored, together with the results from relevant 3rd-party investigations as well as results from our own high-throughput analysis. The comprehensive data included in the resource is aimed as wide as possible, including protein, gene-ontology, orthology, metabolic, structural, expression, interactome and chemoinformatics information. This is combined with a data-mining interface for researchers to perform the selection of putative drug target protein and lead compounds according to their specific highly-flexible criteria.

Protein information includes data from the human, mosquito and the various malaria genome projects. Chemical information is from PDB, KEGG and DrugBank. Information includes basic annotations, motifs, domains, binding sites, structural features, orthology information, ontology terms, protein-protein interactions, protein-ligand interactions, pathogen-host interactions and comparative genomics information. Chemical information includes protein interactions and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties. The researcher accessing the resource is then able to perform advanced searching and filtering of proteins and chemical compounds according to possible interactions and the different types of properties described in the database. The resource is currently at a stage where basic and advanced searches may successfully be performed across the proteins and compounds from all organisms, showing possible protein-ligand interactions together with the all the related protein and chemical properties. Additional work currently being performed includes the development of more accurate statistical scoring methods for the predictions, a literature mining component as well as the inclusion of additional chemical data sources.

IMPROVED IDENTIFICATION OF SEQUENCE VARIATION UPON PREDICTION OF MULTI-TAGS ALIGNMENTS BY IN SILICO RE-SEQUENCING IN YEAST.

Claire Jubin[1], Sophie Loeillet[1], Patricia Legoix-Né[4], Alexandre Serero[1], Emmanuel Barillot[2,3], Alain Nicolas[1]

[1]Institut Curie, UMR3244 CNRS, Université Pierre et Marie Curie, Research Department, Paris, 75248, France, [2]Institut Curie, INSERM U900, Research Department, Paris, 75248, France, [3]Ecole des Mines ParisTech, Research Department, Fontainebleau, 77305 , France, [4]Institut Curie, Translational Research Department, Paris, 75248, France

The analysis of high throughput sequencing tags is a challenging task in part because the reads are short (50nt) and at this size, genome contains multi-aligning regions. As a consequence, to identify sequence variation by aligning reads against a reference genome, multi-tag alignments create noise and the accurate assembly suitable for polymorphisms identification is not straightforward. Efficient algorithms are implemented in programs to identify SNP, indels, etc… but the results remains noisy because of initial mapping biases.

Thus, to improve our capability to detect SNPs and structural variants in the laboratory projects, using *Saccharomyces cerevisiae*, we have developed an in silico re-sequencing approach. The principle is to generate in silico NGS tags (varying the read length from 25bp, to 100bp without errors) and simulate reads mapping allowing mismatches (up to 6) as in real data alignment. This method allowed us to annotate the S288c reference genome (extracted from the Saccharomyces Genome Database) for potential multi-alignments, with or without mismatches, and thus define nucleotide positions and chromosomal regions as U-regions (unique) or R-regions (repeated).

This method of multi-tags alignment annotation, adapted to filter NGS data, will be presented. As well examples showing how it helps us to process SNP detection in various yeast strains and improve the de novo assembly of these strains.

THE MINOR C-ALLELE OF RS2014355 IN *ACADS* IS ASSOCIATED WITH REDUCED INSULIN RELEASE AFTER AN ORAL GLUCOSE LOAD

Malene Hornbak[1], Karina Banasik[1,2], Johanne M Justesen[1], Thorkild I Sørensen[2,4], Oluf Pedersen[1,2], Torben Hansen[1,3]

[1]Hagedorn Research Institute, Diabetes Genetics, Gentofte, 2820, Denmark, [2]University of Copenhagen, Faculty of Health Sciences, Copenhagen, 2200, Denmark, [3]University of Southern Denmark, Faculty of Health Sciences, Odense, 5000, Denmark, [4]Copenhagen University Hospital, Institute of Preventive Medicine, Center for Health and Society, Copenhagen K, 1357, Denmark

Background: A genome-wide association study (GWAS) using metabolite concentrations as proxies for enzymatic activity, suggested that two variants: rs2014355 in the gene encoding short-chain acyl-coenzyme A dehydrogenase (*ACADS*) and rs11161510 in the gene encoding medium-chain acyl-coenzyme A dehydrogenase (*ACADM*) impair fatty acid β-oxidation. Chronic exposure to fatty acids due to an impaired β-oxidation may down-regulate the glucose-stimulated insulin release and result in an increased risk of developing type 2 diabetes (T2D). We aimed to investigate whether the two variants associate with altered insulin release following an oral glucose load or with T2D.

Methods: The variants were genotyped using KASPar® PCR SNP genotyping system and investigated for associations with serum insulin levels following an oral glucose tolerance test (OGTT) in a population-based sample of 6,162 middle-aged individuals. The case-control analysis of T2D comprised 10,196 Danish individuals ascertained from the population-based Inter99 cohort (*n*=6,162), additional population-based participants (*n*=730) recruited by the Steno Diabetes Center (SDC); and T2D patients from the Addition Denmark screening study cohort (*n*=1,609), and SDC (*n*=1,695).

Results: In glucose tolerant individuals the minor C-allele of rs2014355 of *ACADS* associated with reduced measures of serum insulin 30 min following an oral glucose load (per allele effect (β)=-3.8% (-6.3%;-1.3%), *P*=0.003), incremental area under the insulin curve (β=-3.6% (-6.3%;-0.9%), *P*=0.009), insulinogenic index (β=-3.5% (-6.3%;-0.8%), *P*=0.01), and acute insulin response (β=-2.2% (-4.2%;0.2%), *P*=0.03). The C-allele was not associated with T2D in the case-control analysis (OR 1.07, 95% CI 0.96-1.18, *P*=0.21). rs11161510 of *ACADM* did not associate with any indices of glucose-stimulated insulin release or with T2D.

Conclusions: The minor C-allele of rs2014355 of *ACADS* was associated with reduced measures of glucose-stimulated insulin release during an OGTT, a finding which in part may be mediated through an impaired β-oxidation of fatty acids.

YASRAT: YET ANOTHER SHORT READ ALIGNMENT TOOL

Masahiro Kasahara

The University of Tokyo, Department of Computational Biology, Graduate School of Frontier Science, Kashiwa, 277-8583, Japan

Next-generation sequencers are improving and now HiSeq2000 is expected to yield 200Gb/run or more. Aligning obtained reads against reference genomes is often the first step for various analyses including whole-transcriptome sequencing, whole genome resequencing, copy-number variation analysis, structural variation detection and whole genome bisulfite sequencing. Fast aligners such as ELAND or bwa are often used for this purpose. However, the performance of the existing aligners deteriorates quickly as more mismatches and indels (insertions and deletions) are allowed in alignment. To align reads with more sequencing errors against more distant genomes, sensitive and not-so-slow alignment algorithms are demanded.

To this end we developed a new scalable and sensitive alignment algorithm, YASRAT. The primary goal of the algorithm is to provide a faster tool to align billions of short reads against a reference genome even when a higher discrepancy rate (>3%) between the reads and the reference genome is anticipated due to sequencing errors or polymorphisms. We implemented a new spaced seeding algorithm to gain specificity at the expense of memory requirement. However, a machine with big memory (e.g., 256GB) is expensive; we decided to distribute data structures over machines with relatively small memory, reducing memory requirement per machine (more precisely, per CPU core). For this purpose, YASRAT is built on top of Message Passing Interface, which is a standard parallel programming framework that runs on virtually any commodity clusters. Other features include partial read mapping (for structural variation detection and spliced alignment), “mappability” calculation of reference genomes, source code availability (licensed under GPL).

INTERNATIONAL CANCER GENOME CONSORTIUM DATA PORTAL

Junjun Zhang[1], Syed Haider[2], Anthony Cros[1], Saravanamuttu Gnaneshan[1], Jonathan Guberman[1], Jack Hsu[1], Yong Liang[1], Jianxin Wang[1], Christina Yung[1], Arek Kasprzyk[1]

[1]Ontario Institute for Cancer Research, Informatics & Bio-computing, Toronto, M5G 0A3, Canada, [2]University of Cambridge, Computer Laboratory, Cambridge, CB3 0FD, United Kingdom

The International Cancer Genome Consortium (ICGC) (http://www.icgc.org) orchestrates a multi-national effort to catalogue genomic abnormalities in 50 different tumour types and subtypes. For each type, 500 pairs of matched tumour and normal tissues will be studied using multiple technological platforms. ICGC data, which will be generated independently by each of its 50 member institutions, will be linked to cancer etiology, drug response and patient survival. The size of the data, its complexity and the fact that the individual data sources are distributed around the world make the data management task a considerable challenge.

To meet this challenge, the ICGC has adopted BioMart (http://www.biomart.org), an open source data federation system. Each ICGC member manages and processes their site-specific data and stores it in a local BioMart database. BioMart software provides an integrated view of each member's database and makes them available to the public through the ICGC data portal (http://dcc.icgc.org).

The first version of the portal includes four different data types: simple mutations, copy number mutations, structural rearrangements, and gene expression derived from pancreatic, liver, skin, lung, colorectal and breast cancer. In addition to the data generated by the ICGC, several external data sets, such as Ensembl Gene, KEGG Pathway, and Pancreatic Expression Database have also been federated to provide support for different type of queries. These range from gene oriented such as 'find all non-synonymous coding mutations identified in PIK3R1 for all cancers' to queries that integrate several different datasets, for instance: 'find all members of the Toll-receptor pathway having deletions in stage III breast cancer'. As the ICGC data increases and the portal's functionality is further improved it is expected that this will become an increasingly important resource for cancer researchers with diverse user requirements.

NEXT-GENERATION SEQUENCE DATA ANALYSIS PIPELINE AND NGS-BASED GENE PREDICTION

Yoshihiro Kawahara[1], Hironobu Wakimoto[2], Hiroaki Sakai[1], Takashi Matsumoto[1], Takeshi Itoh[1]

[1]National Institute of Agrobiological Sciences, Division of Genome and Biodiversity Research, Ibaraki, 305-8602, Japan, [2]Hitachi GP, Ltd., , Tokyo, 135-8633, Japan

Emerging techniques of massive parallel sequencing have dramatically reduced the time and cost of DNA sequencing, but have increased importance of bioinformatics processes. Here we present our analysis pipeline for short reads of illumina Genome Analyzer, which can be used for various purposes: detections of SNPs and structural variations, transcriptome analysis by mRNA-seq, methylation profiling, and analysis of DNA-protein interaction by ChIP-Seq. After preprocessing of short reads (trimming of low-quality bases and adapter sequences), and mapping to a reference sequence by BWA, our pipeline estimates expression levels on the RPKM (reads per kilobase of exon per million mapped sequence reads) basis for each transcript, and also predicts novel gene structures by Cufflinks. We applied this pipeline to find differentially expressed genes under different environmental conditions. We also developed an NGS-based gene prediction tool that utilizes coverage information of mRNA-seq reads on a reference sequences and creates exon-intron structures on the basis of the Hidden Markov Model (HMM) defined by known genes. To increase accuracy of splice site predictions, prior knowledge of positions of splicing junctions can be included in HMM. To assess accuracy of gene prediction, we applied it to publicly available *Homo sapiens* and *Oryza sativa* mRNA-seq data, and compared the predicted gene structures with those predicted by other NGS-based and ab initio programs. Furthermore, to examine whether an HMM of related species can be used to predict gene structures of another species, we applied the tool to predict gene structures of *Brachypodium distachyon* using Brachypodium mRNA-seq data and the HMM of *Oryza sativa* known gene structures.

EPIGENOMIC AND RNA STRUCTURAL CORRELATIVES OF POLYADENYLATION SITE USAGE

Mugdha Khaladkar[1], Mark Smyda[2], Sridhar Hannenhalli[1]

[1]University of Pennsylvania, Penn Center for Bioinformatics, Department of Genetics, Philadelphia, PA, 19104, [2]University of Pennsylvania, School of Engineering and Applied Science, Philadelphia, PA, 19104

Gene expression is regulated critically at the level of mRNA polyadenylation. However, the factors responsible for the usage and strength of polyadenylation site, especially in the face of alternatives, are not entirely known. In contrast to the analysis of sequence elements responsible for polyadenylation site usage, the epigenomic and mRNA structural determinants have not been systematically explored. Higher usage of poly(A) sites has previously been shown to associate with a greater nucleosome affinity downstream of the poly(A) site. In addition we found that the highly used poly(A) sites are also associated with a more favorable mRNA structure upstream of them. Moreover, a stronger nucleosome affinity downstream of the poly(A) site correlated with a more stable structure immediately upstream. Thus there seems to be an interplay between chromatin and RNA structure in determining the polyadenylation site usage. We explored this relationship further and found that a greater nucleosome occupancy downstream of poly(A) site also correlated with a greater accumulation of PolII at the poly(A) site. These findings taken together suggest a mechanism whereby a more compacted chromatin downstream of the poly(A) site promotes PolII pausing thus facilitating the folding of mRNA in a structure favorable for polyadenylation. In addition, upon examining the chromatin signature of the region flanking the poly(A) sites, we found distinct patterns of various histone modifications as well as CpG methylation. A support vector machine classifier, based solely on epigenetic patterns, was able to classify true poly(A) sites with ~82% accuracy. Furthermore, the epigenetic information in the polyadenylation region differed significantly between the genes having a single poly(A) site versus genes with multiple poly(A) sites. Classification between single and multiple poly(A) sites based solely on chromatin signature resulted in an accuracy of ~72%. Overall, our results reveal hitherto unknown epigenomic and mRNA structural correlatives for polyadenylation and also suggests causative link between chromatin structure and mRNA structure favorable for polyadenylation site usage.

GENOME FEATURES UNDERLYING TRAIT-ASSOCIATED SNPS

Alida S Kindt, Pau Navarro, Colin A Semple, Chris S Haley

MRC HGU, Institute of Genetics and Molecular Medicine, Edinburgh, EH4 2XU, United Kingdom

Genome Wide Association Studies (GWAS) have been made possible by the development of high-throughput genotyping platforms for single nucleotide polymorphisms (SNPs). Over 2,000 trait-associated SNPs have so far been reported with new ones being identified and published on a weekly basis (see http://www.genome.gov/gwastudies/). Each report identifies SNPs associated with a particular phenotypic trait at the genome-wide level of significance. However, identifying biologically plausible candidate loci for associated SNPs is often difficult as the SNPs may be located in regions far from the nearest known gene and can be found in genomic regions whose function is so far unknown. Thus it has been hypothesized that many of the causal variants act as long-range regulators of gene expression. For complex traits it has further been suggested that individual variation in chromatin modifications of coding and non-coding DNA sequence may underlie complex phenotypes in humans through regulation of gene expression. This study investigated the relationships between a large dataset of variants implicated in a wide variety of complex traits and various classes of functional annotation to explore the genomic landscape underlying trait-associated variants. We particularly examined the enrichment of specific classes of variants (e.g. cancer associated SNPs) in regions of known chromatin structure, such as histone methylations or acetylations. The results have implications for our understanding of chromatin structure in complex traits and may aid researchers in identifying causal variants of traits.

CHALLENGES IN THE COMPARATIVE ANALYSIS OF GENE EXPRESSION IN APES USING ILLUMINA DIGITAL GENE EXPRESSION

Martin Kircher[1], Esther Lizano[1,2], Thomas Giger[1], Svante Pääbo[1], Janet Kelso[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, 04103, Germany, [2]Centre for Genomic Regulation, Genetic Causes of Disease, Barcelona, 08003, Spain

Microarray studies of gene expression in brain, heart, liver, kidney, and testis have yielded insights into how differentially transcriptomes evolved between humans and chimpanzees. In particular, a gradation of selective constraints among the tissues (highest brain, lowest liver) has been described as well as the predominant consistency of patterns with a model of neutral evolution (1). However, biases from hybridization-based technologies might have impacted these results as microarrays are sensitive to sequence differences (from polymorphisms and divergence) in the regions to which probes are designed and limited to known and correctly annotated genomic features.

We have therefore taken advantage of the rapid developments in sequencing technologies to extend our studies of primate gene expression using a serial analysis of gene expression (SAGE) protocol for the Illumina Genome Analyzer platform.

During the analysis of the data we have addressed a range of technical and bioinformatic challenges. Technical artifacts include adaptor dimers and chimeric molecules from library preparation, incomplete restriction enzyme digestion resulting in scattered tag counts and tag counts from DNA carry-over (causing a potential transcript length effect and a wrong signal for anti-sense transcription). Analysis issues involve the ambiguity in alignments of ultra short reads (17nt tag sequences) to primate genomes, non-Poisson distributed variances of tag counts, and the dependence of results on the gene annotation used.

We conclude that although data generation and its technical artifacts may not introduce a species-specific bias, such biases can easily be introduced during data analysis if differences in annotation and genome quality among the genomes analyzed are not appropriately considered.

(1) Khaitovich, P. et al., Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309 (5742), 1850 (2005).

DRUG-INDUCED ALTERATIONS IN THE BRAIN TRANSCRIPTOME

Michal Korostynski, Marcin Piechota, Ryszard Przewlocki

Institute of Pharmacology PAS, Dept. of Molecular Neuropharmacology, Krakow, 31-343, Poland

Psychotropic drugs activate intracellular pathways in the brain, these pathways regulate the expression of genes that are important to the therapeutic activity as well as long-term adverse effects. To reveal the genetic networks activated by different classes of drugs we compared the effects of antidepressant (e.g. fluoxetine), analgesic (e.g. morphine), psychostimulant (e.g. cocaine) and antipsychotic drugs (e.g. haloperidol) on genomic profile in mouse (C57BL/6J) striatum. We applied a whole-genome microarray (Illumina WG-6) profiling to characterize time-course of transcriptome alterations following acute drug administration (1, 2, 4 and 8h after injection). Data analysis using two-factor (drug and time) ANOVA indicated list of drug-responsive genes ($P < 0.05$, adjusted for multiple comparisons). The data were stored as raw values, fold of change versus saline and P value of drug versus saline comparison in the database (available at genes2mind.org). To define molecular mechanisms involved in the transcriptional regulation of genes within the patterns we combined the bioinformatic, transgenic and pharmacological approaches. We identified major drug-regulated expression patterns that are formed by inducible transcriptional networks, as for example: (1) CREB/SRF-dependent genes that appears to be related to drug-induced neuronal activity, (2) the group of genes controlled at least in part via release of steroid hormones. Our results elucidates the networks of drug-induced genes that share common regulatory elements, functional relations and may provide novel diagnostic tools for prediction of drug effectiveness.
This work was supported by EU grant LSHM-CT-2004-005166, POIG DeMeTer 3.1 and NN405 274137 grants.

THE EUROPEAN GENOME-PHENOME ARCHIVE (EGA)

Ilkka Lappalainen, Jonathan Hinton, Vasudev Kumanduri, Michael Maquire, Pablo Marcin-Garcia, Paul Flicek

European Bioinformatics Institute, EBI, Hinxton, Cambridge, CB10 1SD, United Kingdom

The European Genome-phenome Archive (EGA), a service of the European Bioinformatics Institute, provides a permanent archive for all types of potentially identifiable genetic and phenotypic data. The EGA contains data collected from individuals for the purpose of medical or genetic research and whose consent agreements prevents open, public data distribution. The EGA follows strict protocols for information management, data storage, security and dissemination. Researchers with appropriate authorisation may access data from the over 402 studies and 5073000 individuals currently provided through the EGA web site at http://www.ebi.ac.uk/ega/.

The EGA has implemented a distributed data access policy whereby the data access decisions are made by a data access-granting organisation (DAO) and not by the EGA project. The DAO may be the same organisation that approved and monitored the initial study protocol or a designate of this approving organization such as a dedicated data access committee. In a typical case, the EGA will direct users to a project homepage where the user can apply for access that is then managed by the EGA.

Accepted data types include genotypes, structural variants and whole-genome sequence which are stored in optimised data structures. In addition, manufacturer-specific raw data formats from array-based genotyping and raw DNA sequence data arising from re-sequencing, transcriptomics or other assays may be deposited for archiving and distribution. The EGA also accepts any phenotype data associated with the samples. All deposited data must have a DAO approved data release policy that provides data access in accordance with the original consent agreements.

The EGA also provides an analysis infrastructure to add value to data submitted into our system. Our quality control applies to both samples and experimental data without altering the original data but allowing us to merge data collected using different technologies, phase submitted data or impute unobserved genotypes using public resources such as the 1000 Genomes project. The data are made available together with our partner DAOs in the most widely used formats to those users that have been granted access.

THE DGVARCHIVE FOR STRUCTURAL VARIATION DATA

Paul Flicek, Jonathan Hinton, Michael Maguire, Ilkka Lappalainen

European Bioinformatics Institute, EBI, Hinxton, Cambridge, CB10 1SD, United Kingdom

The DGVarchive (DGVa) is a new service from the European Bioinformatics Institute (EBI) offering archiving, accessioning and distribution of public structural variations from all species. The DGVa exchanges data with the companion archive, dbVar, at the NCBI. We will also work with the Toronto based Database of Genomic Variants (DGV) group to bring human population specific reference sets for public use. The data archived in the DGVa will be integrated to other resources at the EBI such as Ensembl.

Both DGVa and dbVar have implemented the same accessioning model. The submitter asserted structural variants (SV variants) are accessioned within the context of a study. Each structural variant requires at least one supporting structural variation (SSV variant) that can either be from a different sample used in the same study or from the same sample using a different detection method.

In June 2010, the DGVa included more than 88000 accessioned variants from 38 studies covering nine different species. The DGVa is available at the www.ebi.ac.uk/dgva/.

A PROTOTYPE JAVA API FOR THE ENSEMBL DATABASE SYSTEM

Trevor Paterson, Andy Law

The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Division of Genetics and Genomics, Roslin, EH25 9PS, United Kingdom

The Ensembl project provides an integrated software system for automated genome annotation, storage and retrieval. Annotated genome information is stored in archived releases of Ensembl schema databases and may be freely browsed over the Web using the Ensembl genome browser, accessed programmatically using a Perl API or retrieved via a BioMart web interface or web services. The Perl API allows efficient high-level programmatic access to data stored in the various database schemata for that release.

However, each new release of Ensembl requires an updated version of the Ensembl Perl API which is often incompatible with previous database releases. In addition, the user must know which part of the schema is to be accessed as there are currently four separate parts (Core, Compara, Variation and Genomes) to the API, which need to be updated at each release. Continued access to historical data sets contained within the archived database releases requires parallel installation of the corresponding Perl API code.

In addition, whilst perfectly suited as a scripting language for bioinformatic tasks which process large volumes of text-based data, Perl is not ideal for embedding in graphical interfaces nor for developing large-scale software applications. Provision of an equivalent Ensembl Java API would potentially facilitate the development of a wide range of novel Bioinformatics visualisation tools to access and analyze Ensembl data.

We report here the development of a prototype architecture for such an Ensembl Java API. The modular architecture of this prototype allows separation of data access functionality from the Java model objects that represent genomic features. Specifically separation of the data access configuration (the mapping of SQL statements to model objects) away from the data access and model objects allows the configuration module to control per schema changes in access code relatively simply. The software modules are published as separate Maven artifacts, which allows for controlled software updates as and when required.

This prototype API allows access to multiple versions of databases at Ensembl and EnsemblGenomes (current and recently archived), including the EnsemblBacteria collection databases within a single code module. It implements sufficient functionality to demonstrate data retrieval from current core databases, and has the ability to map transparently between archived database versions even where there has been a change in the database schema.

ACCURATE QUANTIFICATION OF GLOBAL MRNA EXPRESSION LEVELS BASED ON PAIRED-END RNA-SEQ DATA

Soohyun Lee[1], Chae Hwa Seo[1], Byungho Lim[2], Jin Ok Yang[1], Jeongsu Oh[1], Sanghyuk Lee[1]

[1]Korean Bioinformation Center, KRIBB, Daejeon, 305-806, South Korea, [2]Dept. of Biological Sciences, KAIST, Daejeon, 305-701, South Korea

Measuring gene expression level is crucial in many biology studies. High-throughput expression profiling offers a global insight, but usually suffers from an unreliable accuracy, due to systematic bias and/or limited performance of processing pipelines. We provide here a novel computational approach that allows accurate quantification of genes and transcript isoforms based on paired-end RNA-seq data. We performed comparison with several recently developed methods, from various angles including quantitative PCR, computer simulation, comparison with previously reported gene expression levels, stability of housekeeping gene expression and internal consistency between independently estimated gene expression levels versus isoform expression levels. Our method is based on a new concept, EUMA, expected uniquely mappable area for sequenced read pairs. Gene-level and transcript-level expression levels were estimated independently and later each was updated by taking into account the additional information from the other. We have generated the expression level tables based on the EUMA method for 13 human gastric cancer models, to provide a useful preliminary for researchers who use these cell lines.

A TALE OF GENOME, ANNOTATIONS, METABOLISM AND PHYLOGENOMICS: THE PEA APHID GENOME RESOURCES

F Legeai[1,2], S Colella*[3,4], J Huerta-Cepas[5], J-P Gauthier[1], A Vellozo[4,6], P Baa-Puyoulet[3], M-F Sagot[4,6], T Gabaldon[5], O Collin[2], H Charles[3,4], D Tagu[1]

[1]INRA, BiO3P, Rennes, 35653, France, [2]INRIA-IRISA, Centre Rennes-Bretagne-Atlantique, Rennes, 35000, France, [3]INRA-INSA, BF2I, Lyon, 69000, France, [4]INRIA, Bamboo, Lyon, 69000, France, [5]CRG, Centre for Genomic Regulation, Barcelona, 08003, Spain, [6]UCBL1, LBBE, Lyon, 69000, France

Pest aphids inflict damage on plants through the direct effects of their feeding and by vectoring debilitating plant viruses.They represent an original biological model for insects, in particular due to their unusual life-cycle (sexual and parthenogenetic reproduction) and to the key contribution of bacterial symbionts to their metabolism. The International Aphid Genomics Consortium (IAGC) has recently sequenced and annotated in a community effort the pea aphid genome. Setting up a centralized bioinformatics warehouse is crucial to organize and distribute all the available genomic resources, and to facilitate their handling by non-specialist bio-analyst. In that framework, we implemented three interconnected information systems.

AphidBase (http://www.aphidbase.org) is a comprehensive information system set up to safely centralize and promulgate data generated by the IAGC. It is constructed using software from GMOD including several Chado instances, genome browser, gene reports, an ontology navigator, Apollo and various other tools such as a blast and a full text search facilities. **Acypicyc** (http://acypicyc.cycadsys.org/) is a BioCyc database which integrates the latest gene annotations into the metabolic network reconstruction, using an automated data management system for Cyc databases (CycADS) and the 'Pathway tools' software (BioCyc). AcypiCyc is a key resource for computational systems biology research and complex genomic data analyses. Both databases are also connected to **PhylomeDB** (http://phylomedb.org/) a database for genome based phylogenetic analysis of the pea aphid. Orthology data from PhylomeDB have also been used for functional annotation of predicted genes.

At least 6 new pea aphid genomes will be available in the next couple of years. Furthermore, we are already collecting millions of RNA-Seq sequences, and data from various ChIP-Seq and peptides analysis projects. Thus, AphidBase requires an improved automated management in order to efficiently reduce both cost and time for the demanding manual curation. The development of powerful extraction and mining tools will also facilitate data analyses. Finally, we are now setting up an AphidAtlas to store and manage aphid morphological data. This resource will be linked to the full information arising from transcriptomic, proteomic, and metabolomic data.

A NEW PROCEDURE FOR DE NOVO CNV DETECTION IN COMPLEX PEDIGREE.

Louis-Philippe Lemieux Perreault[1,2], Gregor U Andelfinger[2,3], Philip Awadalla[2,3], Marie-Pierre Dubé[1,2]

[1]Montreal Heart Institute, Research Center, Montréal, H1T 1C8, Canada, [2]Université de Montréal, Faculté de Médecine, Montréal, H3T 1J4, Canada, [3]CHU Sainte-Justine, Research Center, Montréal, H3T 1C5, Canada

Copy number variations (CNVs) are heritable segments of DNA that are 1 kb or larger and are present at a variable copy number when compared to a reference genome. It is now acknowledged that the presence of CNV leaves a noticeable footprint in SNP genotyping data, such as mendelian inconsistencies, deviation from Hardy-Weinberg equilibrium and null SNP genotypes, resulting in loss of information for association and linkage studies. *De novo* CNVs are structural variants that appear in one generation, but are absent in the previous one and which are transmitted to following generations. The current methods used for the detection of *de novo* CNVs are imprecise as they generally rely on total genomic CNVs summarized as gains or losses without distinction for the number of copies or the underlying partition of the CNVs with respect to the two chromosomes. We propose a new approach to search for the presence of *de novo* CNVs in complex pedigrees using partitioned CNV genotypes obtained from high throughput SNP genotyping arrays. CNVs are first partitioned using allele-specific dosing (integrated genotypes) and by relying on pedigree information. More specifically, the partitioning uses information from first degree relatives in order to infer the type and dosage of each allele residing on both homologous chromosomes, making possible to follow the transmission of those alleles from one generation to the other. Then, inconsistencies in partitioned CNVs are found based on stretches of consecutive SNP and CNV data departing from Mendel's Laws. The *de novo* status of each deviant stretch is evaluated probabilistically according to evidence from pedigree transmissions throughout the following generations. In order to validate the procedure, gene-dropping simulations were performed by using normal SNP genotypes (presenting a genomic copy number of 2) from a real life dataset and randomly adding *de novo* CNVs in a complex pedigree. We have applied the method to a data set consisting of 393 French-Canadian individuals (42 complex pedigrees and 31 trios) segregating *Left Ventricular Outflow Tract Obstruction* (LVOTO).

THE USE OF ZEBRAFISH (*DANIO RERIO*) EMBRYOS IN A HIGH DEFINITION TRANSCRIPTOMIC EXPRESSION PROFILING APPROACH TO ECOTOXICOLOGICAL INVESTIGATIONS

Luca Lenzi[1], Ashley Sawle[2], Pia Koldkjaer[1], Suzanne Kay[1], Kevin Ashelford[1], Neil Hall[1], Andrew Cossins[1]

[1]Centre for Genomic Research, University of Liverpool, Liverpool, L69 7ZB, United Kingdom, [2]School of Biological Science, University of Liverpool, Liverpool, L69 7ZB, United Kingdom

The adoption of alternative models is often hindered by the limited scope and definition of the scientific outputs. The zebrafish embryo offers the convenience and ethical status of an *in vitro* model with the benefits of an integrated understanding of system-wide processes that comes from the assessment of whole organism responses. We have developed a strategy for an intensive assessment of ecotoxicants effect in the whole animal through the use of Serial Analysis of Gene Expression (SAGE) of zebrafish embryos.

Zebrafish embryos were semi-statically exposed to Diethyl Phthalate at EC_{10} over 96 hours, and the resulting effects on gene expression determined. Normally developed larvae at 96 hours old were used as controls. For each sample the obtained SAGE tag set was sequenced by SOLiD technology and the expression profile was determined.

Zebrafish genomic and transcriptomic datasets were used for the mapping process to resolve previously unknown transcripts and improve zebrafish annotation. In order to analyze the mapping onto the zebrafish genome an in house pipeline was developed.

Cluster and statistical analyses have been used to identify common and differentiated gene expression patterns and sophisticated system-wide pattern searching algorithms to identify biological pathways and processes affected

The resulting high resolution datasets allow expression patterns to be used as fingerprints to identify ecotoxicants, for high-throughput screening of chemicals, or for the investigation of modes of toxicity through association with altered regulation of response pathways. Initial testing with four chemicals, by high density microarray screening, has shown that this methodology is highly amenable to all three uses. The gene expression profiles for the different toxicants were distinctive, and analysis using Gene Ontology profiling, Reactome or network-based packages provided a high level of detail of affected processes. Ongoing research aims to substantially increase the predictive power of the technique, and to extend the project to use cell cultures.

COMPARISON BETWEEN RNASEQ MAPPING AND RNASEQ ASSEMBLY USING SIMULATED DATA

Mathias Lesche, Kay Prüfer, Janet Kelso

Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics/Bioinformatics, Leipzig, 04103, Germany

The Illumina Genome Analyzer is finding increasing application in gene expression studies. Many millions of reads (varying in length from 35-100 bp and with or without paired end information) can be generated relatively rapidly, enabling both transcript reconstruction and counting. Methods for transcript quantification rely heavily on the ability to accurately reconstitute transcripts from short reads. The parameters for this reconstruction include, but are not limited to, the read length, number of reads generated, repeat and duplication content of the genome, and complexity of the transcriptome. It remains an open question whether reference-genome mapping followed by transcript reconstruction, or read assembly followed by genome mapping, is a more efficient and accurate approach. It is also useful to address the question at which transcript sequencing depth assembly in the absence of a reference genome becomes possible.

Using simulated read data we have explored the performance of the next generation genome assemblers VELVET and ABySS, and compared these to the mapping-based approach of the Tuxedo suite. Our simulation includes sequencing error, different depth of sequencing (from 20 million to 120 million reads) and read length (50, 75 and 100 base pairs).

We will demonstrate how the outcome of the transcriptome reconstruction depends on the parameters chosen and compare the results between the assembly and mapping approach. We also investigate the effect of different assembly parameters on the quality of assembled transcripts. Our analysis deepens the understanding of the strength and weakness of both approaches.

CORRECTING 454 ASSEMBLIES USING SOLID MAPPING

Xuan Liu, Alistair C Darby, Gareth Weedall , Kevin Ashelford, Neil Hall

Centre for Genomic Research, School of Biological Sciences, The University of Liverpool, Liverpool, L69 7ZB, United Kingdom

The accuracy of *de novo* assemblies is important for downstream analysis, and is highly relied upon the accuracy of the reads from next generation sequencing platforms. The Roche 454 GSFLX genome sequencer provides the genome biologists with long reads and a range of paired end insert sizes, which can be *de novo* assembled to give good quality genome contigs and scaffolds. However, there are two main disadvantages with the 454 platform limiting the quality of 454 assemblies. The first is homopolymer repeat errors and the second is the relatively high cost per base, which makes deep coverage of a genome expensive compare to short read platform. The SOLiD platform provides a solution to both these problems if we can combine the 454 assemblies with the high accuracy and cheap coverage provided by the SOLiD sequencer.

In our research, we proposed a iterative methods for correcting Roche 454 *de novo* assemblies using automated SOLiD mapping. In the iterative process, SOLiD short reads are mapped against the 454 assemblies. All the possible homopolymer repeat errors in 454 assemblies are picked out and corrected according to the probability $P_x(C,Q,Po,F)$, where x is the possible error, C is the coverage, Q is the quality value, Po is the position and F is the frequency. The algorithm will accept the corrections only if the mapping coverage at each corrected base is improved. This iterative process is terminated when no more possible errors are identified.

As a test case, we have sequenced the genome of *Babesia divergens*, a protozoan parasite of cattle, with both Roche 454(30X fragment, 3KB and 8KB paired end reads libraries) and SOLiD(Fragment library 50 bp reads) technologies. The 454 sequencing was assembled with NEWBLER 2.6. We then used this 454 assembly as a reference for SOLiD mapping using BioScope 1.2.1. The results show that our approach can effectively correct homopolymer repeat errors in 454 sequences.

FEASIBILITY OF IDENTIFYING COORDINATE CONTROL OF GENE EXPRESSION FROM LARGE-SCALE TRANSCRIPTION FACTOR BINDING SITE DATA

Oscar Junhong Luo[1], Xiaohui Xie[2], Rohan B Williams[1]

[1]The Australian National University, The John Curtin School of Medical Research, Acton, 2601, Australia, [2]University of California, Irvine, Department of Computer Science, Irvine, CA, 92697

Higher eukaryotic genomes are dynamic and complex systems with gene expression finely tuned at different levels of control. Understanding the organization of transcriptional control remains an important goal in genome biology. Collections of defined transcription factor (TF) binding motifs, and their corresponding target genes, in a given species can be used as a "dictionary" to define groups of genes that have the potential of being under coordinated control. Here, we use a genome-wide of TF binding sites (TFBS) (MotifMap, Xie et al., 2009, containing 440 distinct motifs and 17238 corresponding target genes) to survey combinations of TFBS that may result in coordinated control of gene expression. To address redundancy, we collapse the 440 motifs into a reduced set of grouped-motifs, using similarities of Position Specific Scoring Matrices (PSSM). Target genes are then clustered, on the basis of similarity in grouped-motif binding sites, using hierarchical clustering, which identified 2362 clusters of target genes that were reproducible under a resampling based cluster validity test. Approximately, 95% of these stable clusters have more than 1 motif (range: 2 to 35, median: 6), annotated to all genes in cluster. The average cluster size was 4 (range: 2 to 172, median: 3), with most clusters (92%) being comprised of genes distributed across >1 chromosomes. Six clusters were enriched (*Bonferroni* corrected) for the GO terms: *regulation of RNA metabolic process, macromolecule metabolic process, regulation of transcription, G-protein coupled receptor protein signaling pathway, macromolecular complex assembly and immune system process*. We performed more specific investigation of clusters using ChIP-Seq data, and, using global gene expression surveys of patients with Huntington's disease (candidate gene, the DNA-binding protein, HTT), clusters that were highly scored on a GSEA analysis were characterized for known transcriptional co-activators of HTT. These findings demonstrate the feasibility of identifying putative coordinately trans-regulated clusters of genes using pre-defined motif dictionaries and simple multivariate approaches.

HCOP: A ONE STOP ORTHOLOGY SHOP

Michael J Lush, Susan M Gordon, Ruth L Seal, Matt W Wright, Elspeth A Bruford

HUGO Gene Nomenclature Committee, European Bioinformatics Institute, , Cambridge, CB10 1SD, United Kingdom

A standardised human gene nomenclature allows researchers to communicate unambiguously, and facilitates both text mining and biological data retrieval. Applying this nomenclature to orthologous genes in other species greatly extends this utility. In addition to naming human genes, the HUGO Gene Nomenclature Committee (HGNC www.genenames.org) promotes the use of the same name and symbol for orthologous genes in other species.

We have developed the HCOP (HGNC Comparison of Orthology Predictions) search tool which allows us to make a rapid survey of available orthology assertions for a given gene or group of genes. We believe HCOP to be of general use to the biomedical community and have made it freely available from .

HCOP is an orthology aggregator that gathers assertions between human and 14 species (chimp, macaque, horse, cow, mouse, rat, dog, chicken, opossum, duck billed platypus, zebrafish, fruitfly, C. elegans and S. cerevisiae) sourcing the assertions from a basket of 11 databases: Ensembl, Evola, HGNC, Homologene, Inparanoid, MGI, OMA, OPTIC, Treefam, UCSC and Zfin. The results are combined into a consensus dataset and links back to each of the databases that made the orthology assertions are provided.

A LINEAR LAYOUT FOR VISUALIZATION OF TRIPARTITE NETWORKS

Martin Krzywinski, Katayoon Kasaian, Olena Morozova, Inanc Birol, Jones Steven, Marco Marra

British Columbia Cancer Research Center, Canada's Michael Smith Genome Sciences Center, Vancouver, V5Z4S6, Canada

Visualizations of large networks are notoriously difficult to interpret. While a network's node and edge structure can be easily traversed algorithmically, a direct translation of this representation into a visual form is burdened with a very high cognitive load for the reader. In addition to the underlying network structure, the interpretability of these visualizations critically depends on the choice of layout algorithm, which should help to answer questions relevant to the data set. As the size and complexity of the network grows, a 2D layout occludes its structure and becomes increasingly opaque to interpretation.

To help visually interpret networks we present a layout approach in which nodes are placed on one of three linear segments. Node-to-segment assignment can be done on the basis of connectivity (e.g. out only, in only, in-out) or annotation (e.g. genes of interest, positively and negatively co-expressed genes). Within a segment nodes can be ordered arbitrarily, such as by connectivity or quantitative annotation (e.g. expression). Connected nodes are joined in the visualization by a Bezier curve.

This linear layout facilitates interpreting the connectivity distribution of nodes within each category. When the Bezier curves are drawn with transparency, overlapping curves clearly indicate the density of connections. Nodes of interest, together with associated connections can be easily highlighted. This highlighting can be extended to include next nearest neighbors.

Our approach is applicable to studying gene regulatory networks (out-only nodes are categorized as monitors, in-out as regulators, and in-only as workers). When the linear segments are interpreted as linear genomes, our layout is suitable for three-way alignment and conservation comparisons. For example, each segment can correspond to a modern species, with links connecting regions inherited from the same ancestral genome region.

We will demonstrate this layout with visualizations of three diverse examples: a gene regulatory network, evolutionary conservation in crucifers, and gene co-expression signatures in neuroblastoma.

THE MODENCODE DCC: CAPTURING DEEP METADATA FROM GENOME-SCALE EXPERIMENTS

N Washington[1], E Stinson[1], M Perry[2], P Ruzanov[2], S Contrino[3], R Smith[3], Z Zha[2], R Lyne[3], E Kephart[1], P Lloyd[1], G Micklem[3], S Lewis[1], L Stein[2]

[1]Lawrence Berkeley National Laboratory, Genome Division, Berkeley , CA, 94720, [2]Ontario Institute for Cancer Research, Informatics & Biocomputing, Toronto, M5G 0A3, Canada, [3]University of Cambridge, Dept of Genetics, Cambridge, CB2 3EH, United Kingdom

The model organism Encyclopedia of DNA Elements (modENCODE) project is an NHGRI project designed to fully describe all biologically informative features on the genomes of Drosophila melanogaster and Caenorhabditis elegans. The project is characterizing transcription factor binding sites, times and locations of transcription, ncRNAs, chromatin states, and DNA replication control. Notable features of the project are the diversity of the data, and the high level of integration required. The modENCODE Data Coordinating Center (DCC) must accommodate this multiplicity of biological types, experimental techniques, and data formats with the overarching goal being to produce a research resource, which will benefit model systems and comparative genomics researchers.

The DCC uses a three-part strategy to collect and assemble extensive information on every experiment, including details of the experimental approach and rich sample descriptions (strains/cell line, stage, tissue, etc.): 1) To facilitate downstream queries and analysis of these data we developed strict metadata standards to capture details of each experiment, and created corresponding extensions to offer structured Wiki pages for collecting these metadata. 2) To manage the collection and validation of the raw and processed primary data we built a processing pipeline driven by user-guided web forms. 3) We developed an extended MAGE-TAB derivative that is used to identify the relationships between the originating samples and the resulting data (files) together with the associated metadata collected from the structured Wiki pages.

Ultimately, the data is made available to the community through multiple portals. modMine intermine.modencode.org provides users with flexible querying capabilities; it is backed by the GMOD Chado database. For graphical presentation users can browse modENCODE using GBrowse, to which we have added a number of new features. For bulk-access all verified data sets produced by the consortium are available at www.modencode.org. The data are also being incorporated into FlyBase and WormBase.

We present the rationale behind our design, the advantages and disadvantages that come with collecting thorough and deep metadata, and lessons learned. Our experience is applicable to both large data centers and small groups hosting data for the broader community.

SAVANT GENOME BROWSER

Marc Fiume[1], Vanessa Williams[1], Andrew Brook[1], Michael Brudno[1,2]

[1]University of Toronto, Computer Science, Toronto, M5S3G4, Canada,
[2]University of Toronto, Banting and Best Department of Medical Research, Toronto, M5S3E1, Canada

High Throughput Sequencing (HTS) technologies have made the generation of genomic data quicker and more affordable than ever before. Despite the increasing availability of sequence information, the analysis of these and other large genomic datasets remains a challenge. Visualization tools (e.g. UCSC Genome Browser and the Integrative Genomics Viewer) have been developed to help interpret genomic datasets that have been created by other external computational tools which manipulate, convert, and analyze raw data (e.g. Galaxy, SAMtools, algorithms for genetic variation detection, etc.). We present Savant, the Sequence Annotation Visualization and ANalysis Tool, a platform that enables the integration of genomic data processing, analysis, and visualization into a single user-friendly environment.

Savant is a fast and interactive genome browser that is capable of displaying sequence, read alignment, SNP, and other genomic datasets stored in standard file formats. The browsing environment is designed to be familiar to users of popular browsers such as UCSC, but it is augmented with additional features such as a bookmarking component for adding custom annotations, and a tabular view of the data that complements its visualization, and a modular docking framework that enables a custom organization of the interface.
In addition to being a stand-alone genome browser, Savant is also a platform for development and dissemination of computational tools (in the form of plugins) which can, for example, process and analyze genomic data in realtime within the visual environment. Developers can take advantage of a rich library of useful functions (e.g. for retrieving in-range or whole-genome data, making annotations, altering the UI, etc.). A catalog of developed plugins is provided on the Savant website, from which users can browse, download, and use in combination to perform very specific tasks.

We will present the key elements of the browser and demonstrate by example the power and utility of the plugin framework of Savant.

Savant is freely available at http://compbio.cs.toronto.edu/savant/.

EXTENDING WIKI SOFTWARE FOR COMMUNITY ANNOTATION

Daniel P Renfro[1], Deborah A Siegele[2], Nathan M Liles[1], Brenley K McIntosh[1], James C Hu[1]

[1]Texas A&M University, Dept. of Biochemistry & Biophysics, College Station, TX, 77843, [2]Texas A&M University, Dept. of Biology, College Station, TX, 77843

The growth of sequencing data is far outpacing the rate of biocuration. The model of having a large number of contributions from a small number of contributors is not meeting the demands of genome biology. One response has been to incorporate the community into the genome annotation process; our approach uses wiki software to accomplish this. We have built community-focused annotation resources using the Mediawiki software for two model bacterial species: *Escherichia coli* [EcoliWiki; http://ecoliwiki.net] and *Bacillus subtilis* [SubtilisWiki; http://subtiliswiki.net]. We have also used Mediawiki as the basis for the Gene Ontology (GO) Normal Usage Tracking System [GONUTS; http://gowiki.tamu.edu], a resource for user contribution of usage notes for GO terms and community annotation of any protein in UniProt. These 3 wikis utilize a series of template-driven pages and tables to impose the structure associated with biological databases onto the free-form wiki architecture. A custom extension, TableEdit, allows users to enter and edit tabular data without having to know the markup associated with wiki tables. This code also allows us to automate loading and retrieval of large and complex sets of data, as well as provide an interface for other extensions. In this way, we have implemented a number of features such as a) automated comparison of GO annotations for sets of genes, b) a graphic display of protein domains and motifs controlled by table entries, c) tables that extract data from other pages and update automatically, and d) monthly extraction and submission of GO annotations to the GO Consortium. We have also extended the functionality of the wiki to simplify common tasks such as adding references, tables, and new wiki pages. We are in the process of creating interfaces with other GMOD components, including the popular genome browser GBrowse and CHADO, a relational database schema.

GENOME MODEL: AN EXTENSIBLE SYSTEM FOR DETAILED TRACKING AND FLEXIBLE SCHEDULING OF GENOMIC ANALYSIS

David J Dooling[1,2], Craig S Pohl[1], Scott M Smith[1], George M Weinstock[1,2], Elaine R Mardis[1,2], Richard K Wilson[1,2]

[1]Washington University School of Medicine, The Genome Center, Saint Louis, MO, 63108, [2]Washington University School of Medicince, Department of Genetics, Saint Louis, MO, 63110

Many bioinformaticians and analysts are struggling to keep pace with the rate of raw sequence production now widely available through the several next-generation, massively-parallel sequencing platforms. The deluge of data places incredible burdens on the computational and storage resources of individual researchers, research groups, and even many departments. The Genome Center at Washington University has developed a high-throughput, fault-tolerant Analysis Information Management System (AIMS) called Genome Model capable of executing complex, interdependent analysis pipelines with detailed tracking and ensured reproducibility. This presentation will detail this system, its design principles and example pipelines and results. Initial investigations of distributed computing environments like grid and cloud resources will also be discussed.

REFLOW: A REVERSIBLE WORKFLOW FOR BUILDING AND MAINTAINING LARGE FUNCTIONAL GENOMICS DATABASES

Steve Fischer[1], Brian P Brunk[1], Jessica C Kissinger[2], Wei Li[1], Deborah F Pinney[1], David S Roos[1], Christian J Stoeckert[1]

[1]University of Pennsylvania, Center for Bioinformatics, Philadelphia, PA, 19104, [2]University of Georgia, Center for Tropical and Emerging Global Disease, Athens, GA, 30602

ReFlow is a reversible workflow system that manages the creation and ongoing maintenance of multi-species data warehouses. The Eukaryotic Pathogen Genome Database project (EuPathDB.org) developed it for scalability in the context of a rapidly-growing number of available species and genomic-scale datasets. ReFlow programmatically builds large genomics databases from scratch or, in reverse, targets any subset of data for automated deletion and refreshing. This has improved the freshness of our data, release turn-around time, and staff efficacy, issues that will increasingly confront many genomics databases.

We surveyed existing workflow and build systems (e.g., JBPM, Ergatis, Taverna, Ant and make) and found that, while powerful, none were tailored for the specialized task of building a genomics warehouse and keeping it up-to-date through release cycles. ReFlow's distinctive features include recursive undo, correcting or adding steps while running or after, global steps, parameterized subgraphs, subgraph references, pause/resume, and test mode.

The malaria parasite database PlasmoDB.org, for example, uses ReFlow to manage the integration of 240 datasets (genome sequences, SNPs, ESTs, SAGE tags, microarrays, RNA-seq, ChIP-chip, proteomics, etc) from eight species. The XML graph to build a new database has 5500 steps and runs (persistently) for a few weeks. It incorporates 100 reusable subgraphs across the eight organisms, such as those for the acquisition and loading of Interpro domains, or the analysis of RNA-seq samples.

Databases created by ReFlow can be maintained indefinitely, obviating the need for complete rebuilds. Outdated or incorrect data can be reversed and reloaded, with all dependent data updated as appropriate. For example when new genome sequence becomes available, a command can recursively undo the old sequence, deleting it and all dependent data (i.e., most data for that organism). Running forward, the workflow loads the new sequence and re-performs all downstream analysis and integration. It leaves other organisms untouched but recomputes comparative genomics results. Similarly, new genomes can be added dynamically, or the database can be synchronized to new Gene Ontology releases.

ReFlow's lightweight implementation is coded in Java and Perl, and is available at http://code.google.com/p/reflow-genomics

ANNOTATING GENES AND GENOMES WITH DNA SEQUENCES EXTRACTED FROM BIOMEDICAL ARTICLES.

Maximilian Haeussler, Casey Bergman

University of Manchester, Faculty of Life Sciences, Manchester, M139PT, United Kingdom

Increasing rates of publication and DNA sequencing make the problem of finding relevant papers for a particular gene or genomic region more challenging than ever. Efforts to link publications to genes by manual curation are limited in scope, and automatic linking to gene names using text-mining approaches remains an open area of research. Here we have developed a novel software system called text2genome that scans full-text biomedical articles, extracts DNA sequences and automatically maps them to genes and genome sequences. Overall, we find ~20% of open access articles in PubMed Central have extractable DNA sequences, the majority of which are short sequences that are not found in GenBank. Sequences extracted from articles can be mapped to genes and genomes very accurately, with 94.7% of species-article and 87.7% of gene-article associations predicted using text2genome matching those based on GenBank submissions. In addition to providing bidirectional links from articles to genes and organisms, our approach produces genome annotation tracks of the biomedical literature, thereby allowing researchers to use the power of modern genome browsers to access and analyze the biomedical literature in the context of genomic data. This work will facilitate research across many domains of biomedical research by integrating two of the most essential sources of biological data available to post-genomics researchers.

A COMPREHENSIVE APPROACH FOR SUPPORTING ACCESSIBLE, REPRODUCIBLE, AND TRANSPARENT COMPUTATIONAL RESEARCH IN THE LIFE SCIENCES

Jeremy Goecks[1], Anton Nekrutenko[2], James Taylor[1]

[1]Emory University, Departments of Biology and Math & Computer Science, Atlanta, GA, 30341, [2]Penn State University, Center for Comparative Genomics and Bioinformatics, University Park, PA, 16802

Increased reliance on computational approaches in the genomic sciences has revealed grave concerns about how accessible and reproducible computation-reliant results truly are. A recent investigation found that less than half of selected microarray experiments published in *Nature Genetics* could be reproduced, often due to missing raw data and analysis details (often computational).

Experiments that employ next-generation sequencing (NGS) will only exacerbate challenges in reproducibility due very large datasets and increasingly complex computational tools. To assess the difficulty of reproducing NGS experiments, we surveyed re-sequencing studies that have appeared in *Nature Genetics*, *Nature*, and *Science* in 2010. We broadly defined a re-sequencing study as any publication utilizing NGS where reads are compared against a reference genome in search of sequence variants. Almost 50% of surveyed studies did not provide access to the primary datasets. In these cases, we contacted corresponding authors asking for sequencing reads, associated quality scores, and analysis tools. In the majority of cases, we were able to obtain the primary data. The data, however, is only one component needed to reproduce a study. While most studies utilized commonly used packages such as Maq for aligning and SNP calling or MACS for ChIP-seq peak identification, no software versions or settings were given (with one exception), making exact reproduction of an analysis impossible for nearly all publications surveyed.

We describe a model for addressing the reproducibility of computational experiments and also the complementary goals of accessibility and transparency. This model supports accessibility, reproducibility, and transparency at every step of the computational science process, from early exploratory analysis to publication. We have implemented this model in Galaxy (http://usegalaxy.org), a popular, open web-based platform for genomic research. This model employs computational infrastructure that automatically tracks and manages provenance and provides support for capturing the context and intent of computational methods. Galaxy Pages provide users with a medium to communicate and share a complete computational analysis. Galaxy Pages are interactive, web-based documents, such as supplementary materials, that enable readers to easily move among the levels of details necessary to understand a published analysis. In addition, readers can copy methods and data from a Page into their workspace and reproduce an analysis or reuse the methods or data.

RNA WIKIPROJECT: COMMUNITY ANNOTATION OF RNA FAMILIES

Alex Bateman

Wellcome Trust Sanger Institute, Senior Investigator, Hinxton, CB10 1SA, United Kingdom

Community annotation has a checkered history littered with failed efforts at engaging scientific experts in annotation. This led many to write off this approach for biology. However, Wikipedia has shown that with the right engineering this approach can work.

This talk details our efforts to engage the community of RNA biologists to improve the annotation of the Rfam RNA families database as well as more generally contribute to the annotation of RNA in Wikipedia. I will discuss the successes and failures along the way.

Finally, I will discuss how the journal RNA Biology in collaboration with Rfam has pioneered a novel model of scientific publication where scientists are required to write a Wikipedia article to go alongside their scientific paper describing new families of non-coding RNAs. The wikipedia article undergoes full scientific peer review under the careful eye of the Assistant Editor-in-Chief Paul Gardner.

IDENTIFICATION OF CONSERVED MIRNAS IN PLANTS BASED ON EST ANALYSIS

Michal Szczesniak*[1], Lukasz Kaczynski*[1], Katarzyna Nuc[2], Przemyslaw Nuc[1], Izabela Makalowska[1]

[1]Adam Mickiewicz University, Faculty of Biology, Poznan, 61-614, Poland, [2]University of Natural Sciences, Department of Biochemistry and Biotechnology, Poznan, 60-637, Poland

Mature miRNAs are short (~21 nucleotides) single stranded RNAs that play multiple regulatory roles in a cell. In plants miRNAs are critical for regulation of developmental, stress related and a range of other physiological processes. Mature miRNAs are specifically cleaved from precursors that usually are 70-250 nucleotides long. miRNAs have two features that were essential to our study. 1) miRNA precursors, when folded, yield a characteristic hairpin structure. 2) Mature miRNAs are quite highly conserved between species. Keeping this in mind we searched for miRNAs in plant species that have none or relatively low number of known miRNAs and high number of EST sequences at the same time. We mapped all known mature miRNAs onto ESTs and identified sequences having a perfect match with the mature miRNA. The EST sequences were assembled into contigs and subsequently folded into a secondary structure. Next, the hairpin structures were evaluated to check if they meet basic criteria for plant miRNA precursors. In addition we mapped short RNA reads from deep sequencing to the hairpins, whenever this type of data was available. This analysis allowed us to track the pattern of miRNA and miRNA* excision from putative hairpin providing even stronger confirmation for a predicted hairpin to be a precursor of real miRNA. In silico analysis of EST data made possible to determine dozens of putative miRNA precursors in plants like cotton, apple, wheat, lotus, tomato, pine, and more.

THE TRANSPOSITION IN TRANSPOSITION (TINT) ALGORITHM AND THE CHRONOLOGY OF THE PRIMATE ALU RETROPOSON ACTIVITY

Gennady Churakov[1], Norbert Grundman[2], Andrej Kuritzin[3], Juergen Brosius[1], Wojciech Makalowski[2], Juergen Schmitz[1]

[1]University of Münster, Institute of Experimental Pathology, Münster, 48149, Germany, [2]University of Münster, Institute of Bioinformatics, Münster, 48149, Germany, [3]Saint Petersburg State Institute of Technology, Department of Physics and Mathematics, Saint Petersburg, 198013, Russia

Discernible transposed elements (TEs) occupy about half of our genome. They integrate into host DNA in waves of activity. In the face of increasing density, they frequently insert into each other. Nested insertions encrypt valuable historical information about the relative age of the elements, comparable to fossils in distinct layers of earth. As old fossils are absent in young layers, older inactive TEs are not inserted into younger elements. In contrast, young TEs are able to occupy all strata of older elements as well as those active at the same time. Hence TEs active at different historical periods display characteristic insertion profiles. Comprised as they are of a substantial fraction of TEs, mammalian genomes are ideally suited for such analyses.

Recently, we developed the Transposition in Transposition (TinT) algorithm, which uses RepeatMasker coordinates to compile interrupted and nested retroposons. The frequencies of fragmented versus nested elements are counted, assembled in a data matrix, and sorted by pre-selected retroposon types. This matrix applies a specific probabilistic likelihood model to calculate the relative integration period for each retroposon subtype in relation to all other subtypes. Additionally, we have developed web-based interface for the TinT application.

To demonstrate and test the web-based TinT method, we investigated the representative primate genomes and their well characterized, Alu dimeric elements. Because of the well-known evolutionary histories of both the species and their retroposons, primates represent an ideal test group for the TinT application. The lemur-specific elements seem to be most closely related to the AluJo elements, the only other dimeric Alu present in the grey mouse lemur. The New World marmoset-specific AluTa elements are derived from an AluS progenitor and the rhesus monkey specific AluR elements are most closely related to the AluY elements. The significant separations between the AluJ, AluS and AluY elements are well supported, indicating clearly resolved relationships. In contrast, the various AluS elements were active during the same period and they did not leave behind significant phylogenetic signals to clearly resolve their more detailed affiliations using presented method.

TURN DOWN THAT NOISE! A FINITE MIXTURE MODEL FOR CHIP-SEQ: QUALITY CONTROL, ANALYSIS, AND COMPARISON OF CHIP SEQUENCING DATA.

Rob Cohen, Rob Goor, Ian Fingerman, Lee McDaniel, Xuan Zhang, Greg Schuler

National Library of Medicine, National Center for Biotechnology Information, Bethesda, MD, 20894

Epigenomics is an emerging field of research aimed at understanding how—despite sharing a common genomic sequence—different cell types and lineages acquire distinct patterns of gene expression. ChIP-Seq, which combines chromatin immunoprecipitation with high throughput sequencing, is becoming an increasingly common method for assaying epigenomic states.

A popular approach for visualizing epigenomic data is to display the results as 'tracks' on a genome browser. Tracks represent continuous-value data aligned against the genome and enriched regions are depicted as peaks in the track.

Typical ChIP-Seq analysis performed to generate these 'tracks' entails dividing the genome into "bins" and counting the number of aligned reads in each bin. Any revealing patterns within a bin are lost. In order to capture these patterns, we constructed a geometric distribution mixture model to compute the probability that a read comes from a site of enrichment. The model parameters are used to quantify data quality. Because ChIP-Seq is an enrichment process, the majority of reads come from "noise" – non putative histone modification sites. By eliminating this noise, the model reduces the file size by 50-90% depending on data quality; the corresponding view reveals peaks that were before entrenched in noise.

Noise reduction also means data size reduction; the reduced data size serves as the basis for an extremely fast O(n) epigenome comparison algorithm, where n is the number of data points remaining after noise reduction. The comparison algorithm quickly finds regions of difference between arbitrary epigenomic tracks. For example, a comparison of H3K4ac between CD4+ T cells and human fetal lung cells took 4 seconds and the most significant differences included genes COL1A1 and KRT80 in fetal lung cells, and genes FGD3 and DEF6 in CD4+ T cells.

Tracks suffering from significant noise can be elucidated without relying on smoothing, which tends to eliminate spatial detail. The NCBI epigenomics page (http://www.ncbi.nlm.nih.gov/epigenomics/) uses these algorithms, along with the wealth of data stored in NCBI's GEO and SRA databases, to make tasks such as viewing, searching, and comparing epigenomes far easier for users.

NEW PERSPECTIVES IN ALTERNATIVE SPLICING FROM THE GENCODE GENEBUILD

Jonathan M Mudge, Adam Frankish, Gary Saunders, Tim Hubbard, Jennifer Harrow

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

Whilst alternative splicing undoubtedly increases the size of the human transcriptome, the actual extent to which this process increases metabolic complexity remains largely unclear; estimates into the size of the human proteome, for example, vary widely. The first step towards addressing such questions is to capture the full complement of alternatively spliced transcripts. Secondly, functional information must be discerned for each individual transcript; in particular, spliceforms encoding protein molecules must be distinguished from those that are non-coding, induce the nonsense mediated decay (NMD) pathway and / or represent spurious 'noise'. The HAVANA group use manual genome annotation to maximise the description of transcriptional diversity within the human GENCODE genebuild. Whilst labour intensive, manual analysis affords both a level of accuracy and a sophisticated degree of categorisation not possible when using automated genebuilding methodologies. We have compared the alternative splicing patterns of 300 human and mouse orthologous genes within the ENCODE pilot regions, allowing us to characterise both conserved splicing events and lineage-specific innovations. We observe that distinct patterns of splicing architecture and levels of exonic conservation are associated with both CDS-predicted and NMD-inducing transcripts. We will discuss the incorporation of RNAseq data into this annotation pipeline, as well as the prospects for utilising proteomics data in the confirmation of alternative translations. Finally, we will justify the use of GENCODE annotation in variation-based projects such as 1000 Genomes, where a knowledge of alternative splicing can profoundly affect the functional interpretation of SNPs identified within exonic regions.

GENOME-WIDE ANALYSIS OF DNA METHYLATION PROFILES IN A PRECLINICAL ANIMAL MODEL OF NONGENOTOXIC CARCINOGENESIS

Arne Mueller[1], Harri Lempiaeinen[1], Sarah Brasa[1], Remi Terranova[1], Jennifer Marlow[1], Roloff C Roloff[2], Michael Stadler[3], Olivier Grenet[1], Jonathan Moggs[1]

[1]Novartis Institute for Biomedical Research, Investigative Toxicology, Basel, 4057, Switzerland, [2]Friedrich Miescher Institute for Biomedical Research, Epigenetics Research Group, Basel, 4058, Switzerland, [3]Friedrich Miescher Institute for Biomedical Research, Computational Biology, Basel, 4058, Switzerland

Epigenetics describes heritable changes in gene function that occur in the absence of a change in DNA sequence. Specific patterns of epigenetic marks form the molecular basis for developmental-stage and cell-type specific gene expression patterns that are hallmarks of distinct cellular phenotypes. An emerging body of data suggest that epigenetic perturbations may also be involved in the adverse effects associated with some drugs and toxicants, including certain classes of non-genotoxic carcinogens. The working hypothesis for the mechanistic investigation outlined here is that epigenetic modifications, namely DNA methylation, will provide valuable insights into early molecular mechanisms associated with nongenotoxic carcinogenesis. The objective of the study is thus to evaluate the sensitivity and specificity of genome-wide and locus-specific DNA methylation assays in rodent tissues that consist of heterogeneous cell types (e.g. liver, kidney, blood). We develop bioinformatics approaches to integrate the epigenetic data with transcriptomics data. B6C3F1 mice were treated for 4 weeks with the nongenotoxic liver carcinogen phenobarbital. Methylated DNA from liver and kidney was isolated using the MeDIP assay for immunoprecipitation of 5-methylcytosine, and applied to promoter/CpG island microarrays from Nimblegen, which include all UCSC-annotated CpG islands and 1.8 kb promoter regions for all RefSeq genes. Transcriptomics was carried out on an Affymetrix platform. In liver, the target tissue of Phenobarbital, the promoter of Cyp2b10, a known member of Phenobarbital induced pathways, is de-methylated significantly after treatment and expression strongly increased. In kidney (non-target tissue) methylation and expression remains unaffected. The bioinformatics approach to integrate genome-wide epigenomics and transcriptomics profiling is promising for identifying early mechanism-based markers of nongenotoxic carcinogenesis and may ultimately increase the quality of cancer risk assessments for candidate drugs and ensure a lower attrition rate during late-phase development.

EVALUATION OF SEQUENCE ALIGNERS AND SNP CALLERS WITH SHORT READS GENERATED USING NEXT GENERATION SEQUENCERS

Tim Beck[1], Alex Kanapin[1], Richard de Borja[1], Bojan Losic[1], Quang Trinh[1], John McPherson[1], Lincoln Stein [1,2], Lakshmi Muthuswamy[1,2,3]

[1]Ontario Institute for Cancer Research, Bioinformatics and biocomputing, Toronto, M5G0A3, Canada, [2]Cold Spring Harbor Lab, Bioinformatics, Cold Sping Harbor, NY, 11724, [3]University of Toronto, Medical Biophysics, Toronto, M5G0A3, Canada

Robust alignment of short-reads generated by next-generation sequencers and detection of Single Nucleotide Polymorphism with high level accuracy are critical steps in utilizing the sequencing data effectively. In view of that, we bench- marked 8 aligners and 3 SNP callers based on a set of synthetic short-reads.

We developed a model to generate a set of 76bp synthetic short reads to test for the accuracy of genomic position alignment, tolerance to nucleotide base variations, and tolerance to experimental variations. The model included SNPs, insertions and deletions with size ranging from 1bp to 20bp and simulated experimental variables such as quality of bases and dependence of read depth on GC content of the genome.

We find that using a combination of aligners is likely to improve both accuracy and speed. All the SNP callers show a sensitivity greater than 75% for fewer than 3 SNPs.; however, have difficulty if there are more than 2 mismatches within a read. We report here the details of the accuracy of aligners and SNP callers.

DDBJ READ ANNOTATION PIPELINE: A CLOUD COMPUTING-BASED ANALYTICAL TOOL FOR NEXT-GENERATION SEQUENCING DATA.

Eli Kaminuma[1], Takako Mochizuki[1], Yuichi Kodama[1], Satoshi Saruhashi[1], Hideaki Sugawara[1], Kousaku Okubo[1,2], Toshihisa Takagi[1,2], Yasukazu Nakamura[1]

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, ROIS, Mishima, Shizuoka, 411-8510, Japan, [2]Database Center for Life Science, ROIS, Bunkyo, Tokyo, 113-0032, Japan

Ultra high-throughput sequencing becomes a nessessary tool in biological research area, due in part to rapidity, precision and cost-effectiveness of new-generation sequencer (NGS). In the year 2009, DNA Data Bank of Japan (DDBJ) started DDBJ Read Archive (DRA), an archive databank for raw data from NGS's. DDBJ Read Archive collaboratively exchanges data with the Sequence Read Archive (SRA) at NCBI and the European Read Archive (ERA) at EBI. To stimulate the analyses the raw sequence data stored in DRA, we are in construction the DDBJ Read Annotation Pipeline. The pipeline has three distinct features. First, analytical results may be easily submitted to DDBJ databases using a streamlined process, whereby map outputs are converted to DRA formats, and similarly the results of assembly/annotations are converted to DDBJ-based International Nucleotide Sequence Database Collaboration (INSDC) format (http://www.insdc.org/). Second, a web-based graphical user interface enables biologists without high-level bioinformatics expertise to analyze large amounts of raw sequencing data. Third, the use of cluster computing systems and computers with huge memory in DDBJ infrastructure allows for high throughput.

FUNCTIONAL INDEXING AND CURATION OF NEXT-GENERATION SEQUENCING DATA

Takeru Nakazato, Hidemasa Bono, Toshihisa Takagi

Database Center for Life Scinence (DBCLS), Research Organization of Information and Systems (ROIS), Tokyo, 113-0032, Japan

Various types of biological data are provided and archived in the public databases: nucleotide sequences in DDBJ/EMBL/GenBank International Nucleotide Sequence Database (INSD), gene expression in gene expression omnibus (GEO), and journal articles in PubMed. In Japan, Database Center for Life Science (DBCLS) has developed infrastructure for researchers to access and re-use these data easily by providing indexes as yellow pages of INSD and GEO, and constructing a portal site for life science databases and tools.
The next generation sequencing (NGS) technology is rapidly spread to approach whole genome sequencing, metagenomics, and transcriptomics. The tremendous size of these NGS data is also archived in short read archive (SRA) in NCBI, European read archive (ERA) in EBI, and DDBJ read archive (DRA) in DDBJ as well as nucleotide sequences and gene expression data. In order to make NGS data more searchable and re-usable, we have developed index site for NGS data. The deposited NGS data contains not only short read sequences but also conditions of experiment including a project title, species or cell line names of samples, and sequencing platforms as a metadata. The metadata consists of six files with XML format: submission, study, experiment, run, sample, and analysis. However, each submission has not all of those metadata because additional experiments or runs to be assigned to a previous project are often performed and reposited as a new submission. We therefore made connections among each type of corresponding metadata, and developed project list as a index site. We attempt to curate the metadata by correcting misspelling and disambiguate spelling variation. This index developed is used by another project in DBCLS, constructing/maintaining reference expression dataset, and will accelerate to the analysis of short read sequences.

PRACTICAL NGS ANALYSIS ON THE CLOUD WITH GALAXY AMIS: UNCOVERING MITOCHONDRIAL VARIATION

Enis Afgan[1,4], Hiroki Goto[2], Ian Paul[3], Kateryna Makova[2], James Taylor[1,4], Anton Nekrutenko[2,4]

[1]Emory, Biology, Atlanta, GA, 30322, [2]Penn State, Biology, Univesity Park, PA, 16802, [3]Penn State Medical Center, Pediatrics, Hershey, PA, 30322, [4]galaxyproject.org, galaxyproject, University Park, PA, 16802

We have developed a solution that allows experimentalists to perform large-scale analysis using cloud-computing resources with nothing more than a web browser (http://usegalaxy.org/cloud). Using our solution, a user without computational expertise can instantiate an analysis environment on a cloud, and can add storage and compute resources to this environment as needed. Because the solution is built on the Galaxy framework, analyses using this solution are accessible, transparent, and reproducible. Popular tools and workflows for analyzing sequence data from various types of experiment are built-in and ready to run.
To demonstrate the utility of this analysis solution we have sequenced mitochondrial genomes from multiple related individuals from three independent families (a total of 48 Illumina datasets). Our goal was quantify and track heteroplasmy in mitochondrial DNA. Our analysis involved standard mapping steps as well as custom developed tools to identification of nucleotide substitutions and indels in mixtures. 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. While performing the analyses we developed a series of Galaxy workflows that can used by anyone in the community to replicate our analyses exactly as they were performed initially. In addition we used Galaxy pages system to annotate and explain every step of each workflow as well as describe metadata associated with every of 48 illumina datasets used in this study.

DE NOVO ASSEMBLIES OF THE TASMANIAN DEVIL GENOMES

Zemin Ning[1], Elizabeth P Murchison[1], Ole B Schulz-Trieglaff[2], Matthew M Hims[2], Dirk Evers[2], Mike Stratton[1]

[1]The Wellcome Trust Sanger Institute, Genome Campus, Cambridge, CB10 1SA, United Kingdom, [2]Illumina Cambridge Ltd., Chesterford Research Park, Essex, CB10 1XL, United Kingdom

The Tasmanian devil (Sarcophilus harrisii) is a carnivorous marsupial now found in the wild only in the Australian island state of Tasmania. First observed in 1996, the Tasmanian devil facial tumour disease (DFTD) has rapidly spread through the Tasmanian devil population and is threatening to cause extinction of the species in the wild. The Tasmanian devil genome project aims to detect somatic mutations by sequencing normal and tumour samples, using the Illumina platform. For both normal and tumour genomes, de novo assemblies are needed in order to carry out data analysis, since there is no existing reference assembly.

We have developed the Phusion2 [1] pipeline to assemble large eukaryotic genomes using Illumina short reads. Read files from individual lanes are processed to generate kmer words at a given size. K-tuples are then merged and sorted into a table so that multiple occurring kmer words shared by different reads can be linked. A relation matrix is used to record the shared kmer words among all the reads. Setting a minimum threshold of shared k-tuples, billions of short reads can then be clustered into groups using kmer sharing information in the relational matrix. After obtaining small read clusters with a controllable size, we use an assembler either based on de Bruijn graph algorithms or traditional overlap assembler such as Phrap to generate contigs.

We present assembly procedures as well as the initial draft assemblies. For the normal genome, we first used 650 millions of 2x100bp pair-end reads with insert size 500bp, generated from a HiSeq 2000 machine. Given an estimated genome size at 3.3Gb, the data of short insert reads covers the genome at ~40X times. Using Phusion2, this set of data produced an assembly of 541,502 contigs with contig N50 at 15.3Kb and total assembled bases (>=150bp) at 3.08Gb. Adding one run of mate pair data of 194 million paired reads of 2x50bp with insert sizes from 3-10Kb, we obtained a scaffold assembly of 3.3Gb with N50 = 823Kb. Work is also underway to flow-sort the 7 devil chromosomes and each individual chromosome will be sequenced at 4-10X read coverage. Scaffold contigs can be assigned to individual chromosomes by aligning the flow-sorted chromosomal reads. Therefore a draft reference assembly can be obtained using this in silico chromosome map. The genomes of two DFTD cancers isolated from geographically distant devils have been sequenced at ~30X read coverage and reads are aligned against the reference genome assembly to identify genetic variants. Finally, two de novo assemblies of the tumor genomes are produced and compared with the draft reference assembly to detect chromosome translocations and structural rearrangements.

[1] Mullikin J.C. and Ning Z. (2003) The Phusion assembler. Genome Res., 13 , 81-90

THE IPLANT COLLABORATIVE : NEW TOOLS FOR INNOVATIVE THE GENOTYPE TO PHENOTYPE RESEARCH

Christos Noutsos[1], Matthew Vaughn[2], Doreen Ware[1], Christopher Jordan[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, [2]The University of Texas at Austin, Texas Advanced Computing Center, Austin, TX, 78758

Ultra High Throughput Sequencing (UHTS) offers the promise that identifying a gene or genes responsible for an observed Phenotype in a given Environment will become more efficient and less expensive, due to the massive sequence data generation it enables. Thus far, it has been a great challenge to extract the total variation of sequence data, as the developed methods are not fully capable of handling efficiently the large amount of data and their complexity. An additional challenge is maximizing the outcome from an experiment, while simultaneously keeping the data in a standard format, allowing repetition of the analysis by scientists from various scientific fields through comparative, evolutionary, or expression analysis. This can be facilitated by storing massive genome sequence in databases, which can be public, in a format readily applicable by traditional biologists not trained in management of such enormous amounts of data. In the iPlant Collaborative (iPlant), under the Genotype to Phenotype Grand Challenge project, we are building a platform where UHTS data can be uploaded or imported from the Sequenced Read Archive on NCBI, to perform variant analysis of expression or genomic sequence. The output format for the variant is VCF3.3 while for the transcript abundance Cufflinks is used to export data in the EXPR format. In the first phase, the platform will be able to perform detection of variants against maize and Arabidopsis genomes, but in later phases components in the infrastructure will be added that will allow de novo variant detection. The platform is being designed to allow data to be comparable across experiments and formats.

ITERATIVE APPROACHES TO GENERATE NEAR BASE PERFECT GENOME SEQUENCES

Thomas D Otto[1], Isheng J Tsai[1], Gary P Dillon[1], Chris Newbold[1,2], Matthew Berriman[1]

[1]Parasite Genomics, Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, United Kingdom, [2]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, OX3 9DS, United Kingdom

Second generation sequencing technology has had a major impact on biomedical research. Any experiment where sequence is the primary readout can now be carried out at a lower cost and at steadily increasing depth compared to capillary sequencing. Although read lengths are increasing, those of the Illumina and SOLID platforms are still too short to allow reliable, complete de novo assembly of complex eukaryotic genomes. We are developing several tools that use an iterative approach to significantly improve the quality of reference sequences and related re-sequenced genomes as well as to expand de novo assemblies. By performing local assemblies iteratively, contigs can be elongated and gaps diminished, using IMAGE (http://image2.sourceforge.net). iCORN (http://icorn.sourceforge.net) is then used to correct base errors and to provide genome wide assessment of base accuracy. Annotation can subsequently be transferred based on synteny using RATT (http://ratt.sourceforge.net). The suite of tools can be run independently or as a pipeline including an improved version of ABACAS (http://abacas.sourceforge.net) to order contigs against a reference. Currently we use these tools on more than 20 genomes including those of human and rodent malaria parasites, as well as other protozoan and metazoan parasites, and even mice. We will show how generating very accurate reference sequences, using iCORN, has also allowed us to evaluate SNP detection and construct error models for Illumina data. Close inspection of the Illumina reads aligned to the corrected reference genome reveals large numbers of clustered sequencing errors, primarily 3' to homopolymeric tracts. We will also show errors that are clustered around some simple sequence repeats, serving as a warning that particular care should be taken when analysing Illumina data that contains these features.

PREPARING AND ANALYSING EST'S FOR THE MAMMARY TISSUE OF SHEEP IN PRENATAL AND POSTNATAL PERIOD

Nehir Ozdemir Ozgenturk[1], Zehra Omeroglu[1], Kemal Oztabak[2], Cemal Un[3]

[1]Yildiz Technical University, Department of Biology, Istanbul, 90, Turkey,
[2]Istanbul University, Faculty of Veterinary Medicine, Istanbul, 90, Turkey,
[3]Ege University , Department of Biology, Izmir, 90, Turkey

Because the colostrum is essential nutrient for new born lamb, the colostrum secretion, which secrets during first 48 h after the parturition, play very critical role on sheep productivity and sheep husbandry. With a EST collection we want to figure out of the gene expression profiles of colostrum secretion for prenatal and postnatal period in Kivircik sheep which is an important local Turkish sheep according to their meet quality and milk productivity. In this study Kivircik sheep in farm of the Faculty of Veterinary science at University of Istanbul was used. The mammary tissues from the same sheep were taken by biopsy twice. End of the prenatal period just before 1 week of parturition first tissue was taken. 18-36 h during high period of the colostrum secretion after the parturition second mammary gland tissue was taken. From two cDNA libraries, two different EST collection was obtained and analayzed with Phred/Phrap, CAP3, BLAST.

After construction of two cDNA libraries, total 3072 colonies which are randomly selected from the two libraries were sequenced to establish two ESTs collection for mammary gland tissue in pre- and postnatal period. For analysis of the raw EST data, the low-quality sequences and the vector fragment was removed with Phred / Phrap computer programmes. Fragment assembly was done with the CAP3 software. Putative functions of all unique sequences were designated by gene homology based on BLAST. Total 429 ESTs which show over % 80 homolgy to known sequences of other organism in NCBI have been determined. Also according to BLAST result we have compared two EST collections and listed differences of between prenatal and postnatal gen expression profiles. This EST data are very valuable resource for functional genome studies of sheep.

GENOME SEQUENCING OF ALGAE AND GRASS: INITIAL RESULTS FROM THE DE-NOVO ASSEMBLY OF *PENIUM MARGARITACEUM* AND *LOLIUM PERENNE*

Frank Panitz[1], Jakob Hedegaard[1], Bernhard Borkhardt[2], Peter Ulvskov[3], Torben Asp[4], Christian Bendixen[1]

[1]Aarhus University, Genetics and Biotechnology - Molecular Genetics and Systems Biology, Tjele, 8830, Denmark, [2]Aarhus University, Genetics and Biotechnology - Cell Wall Biology and Molecular Virology, Frederiksberg C, 1871, Denmark, [3]Copenhagen University, Plant Biology and Biotechnology, Frederiksberg C, 1871, Denmark, [4]Aarhus University, Genetics and Biotechnology - Molecular Genetics and Biotechnology, Slagelse, 4200, Denmark

Penium margaritaceum is a green algae which can serve as a model organism for studying cell wall development and biofilm formation. The expected genome size of the *Penium* genome is about 275 Mb. We sequenced 239 M short reads (36 bp) using single read (SR) and paired-end (PE) libraries with 350 bp insert size, as well as 1.5 M long 454 reads (about 400 bp) using SR and 7 kb PE libraries. De-novo assembly was performed using two approaches: first, the Abyss PE program (Simpson et al. 2009) was used to analyse the short read data only, using different settings to find the 'optimal' k-mer value for assembly. Second, we performed hybrid assembly with the CLC GenomicsWorkbench (CLC bio, Denmark) using Illumina and 454 reads, after masking the latter for *Viridiplantae* repeats with RepeatMasker (www.repeatmasker.org). Taken together, the results show a total genome length smaller than expected indicating that many reads might fall into repetitive regions thus forming deep clusters which do not contribute to the total length of the assembly. Additionally PE sequences with larger insert sizes will have to be generated to assemble larger contigs and scaffolds.

Perennial ryegrass (*Lolium perenne*) has an estimated genome size of 2.3 Gb and is an ubiquitous grass species for which a large number of variations and genotypes are known. 451 M reads (100 bp) generated using Illumina PE libraries with 200 and 600 bp insert size were assembled with the CLC GenomicsWorkbench resulting in 691598 preliminary contigs. Read filtering and trimming parameters as well as repeat content are currently investigated with respect to contig depth/coverage and N50 contig length. In addition, larger insert size libraries for Illumina and 454 sequencing are under preparation.

Generally, assessing coverage and repeat content estimates will be important to determine the amount of sequence coverage needed to generate high-quality genome drafts. Moreover, BAC(end) sequences can be used to validate the assembly, in addition to transcriptome sequences, which also will help to annotate the genome. Establishing the genome resources will eventually contribute to study genetic variation and gene expression in these non-model species.

INSECTACENTRAL: FACILITATING COMPARATIVE GENOMICS WITH MORE THAN ONE MILLION INSECT PROTEINS AND THE *HELICOVERPA ARMIGERA* GENOME PROJECT

Alexie Papanicolaou[1,2,3], Karl H Gordon[3], Lars S Jermiin[3,4], David H Heckel[1]

[1]Max Planck Institute for Chemical Ecology, Entomology, Jena, 07745, Germany, [2]CSIRO, Entomology, Canberra, 2601, Australia, [3]University of Exeter, CEC-Biology, Penryn, TR10 9EZ, United Kingdom, [4]University of Sydney, School of Biological Sciences, Sydney, 2006, Australia

Next Generation Sequencing (NGS) has removed the sequencing bottleneck, which hampered sequence-based projects, and resulted in a bioinformatic bottleneck. For most laboratories, one of the immediate needs is the deployment of an analysis infrastructure, driven by one or more underlying databases. Laboratory- or species-specific solutions are no sustainable. In response, a general shift of the bioinformatic community towards collaborative bioinformatics utilizing stricter standards, species-neutral solutions and open-access frameworks. We present novel and robust bioinformatic solutions to the transcriptome analysis and dissemination bottlenecks. The genes4all module of the Drupal Content Management System interfaces with GMOD and Chado to provide a robust, easy-to-deploy solution. We used these tools and est2assembly to built the InsectaCentral database (http://insectacentral.org) containing all assemblies and deep annotation of all public insect transcriptomes, obtained from circa 200 species. Often, in transcriptome projects, researchers do not release their data, even though only a small fraction of it is of importance to them. Thus, a secure facility for pre-publication data allows us to immediately process and hold new NGS transcriptomes and plays a crucial role in convincing researchers to publicly release gene models for which the depositor holds no personal interest. We have used InsectaCentral to construct a single gene phylogeny of insect orders with 33 taxa identifying extensive compositional heterogeneity at both 1st and 3rd codon sites. Gene models, however, need not be derived from transcriptome data. We are also presenting, therefore, progress on sequencing and annotation of the *Helicoverpa* genome, the first non-model insect with a complex genome (400 Mb of sequence in 31 chromosomes, of which two-thirds is repetitive sequence) to be sequenced solely using NGS technologies.

GENOMIC CHARACTERIZATION OF *BORDETELLA PERTUSSIS* STRAIN 18323

Jihye Park[1], Ying Zhang[1], Stephen D Bentley[2], Julian Parkhill[2], Eric T Harvill[1]

[1]The Pennsylvania State University, Department of Veterinary and Biomedical Sciences, University Park, PA, 16802, [2]The Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

B. pertussis is the causative agents of whooping cough in humans. Despite the introduction of DTP vaccine in 1940s, pertussis is still one of the leading causes of vaccine-preventable deaths. Bacterial pathogens must have strategies to evade the immune response and compete with other bacteria for resources within the host. These include gene gain through horizontal transfer, gene loss, or gene changes. *B. pertussis* is believed to have evolved from *B. bronchiseptica*-like progenitors through genome reduction with little evidence of gene acquisition or horizontal gene transfer.

For more effective vaccination programs to prevent pertussis, it is important to understand genetic content of *B. pertussis* and how this pathogen is evolving to adapt and thrive in the human population. Importantly, there is some evidence that *B. pertussis* is rapidly evolving in response to vaccine-induced immune pressures. To date this evidence is limited to analysis of a small number of antigenic virulence factors. The newly available genomic sequences allow a more complete analysis of the relative variation in all *B. pertussis* genes between two very different strains.

The sequenced *B. pertussis* strain Tohama I is considered to be a reference strain and is used for the production of acellular pertussis vaccines. American Type Culture Collection (ATCC) strain 18323 has been differentiated from many *B. pertussis* stains by multiple typing methods, such as multi-locus sequence typing (MLST). This strain is highly virulent in the mouse intracerebral test and has been widely used as the challenge strain for potency testing of pertussis vaccines. Identified differences between 18323 and Tohama I include expression of the type III secretion system *in vitro*. We have sequenced and analyzed the genome of strain 18323 in comparison to that of Tohama I. These results will contribute to a better understanding of the recent evolution of *B. pertussis* and the specific mechanisms generating variation.

WHOLE-GENOME SEQUENCING AND COMPARATIVE ANALYSIS OF A MELANOMA CELL LINE AND THE METASTATIC TUMOR FROM WHICH IT WAS DERIVED

Stephen C Parker[1], Hatice Ozel Abaan[1], Isabel Cardenas-Navia [1], Praveen F Cherukuri [1], Pedro Cruz[1], Nancy F Hansen[1], Jamie K Teer[1], Subramanian S Ajay[1], Andrew L Young[1], Rachel L Goldfeder[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, Bethesda, MD, 20892, [2]National Institutes of Health, National Cancer Institute, Bethesda, MD, 20892

Quantifying and characterizing the genetic differences between laser-captured tumor cells, and a derived cell line from the same tumor has important implications for developing cancer therapies. To address this question, and as a pilot towards exhaustively cataloging all genetic variations associated with melanoma, we present the sequence and analysis of a melanoma cell line, the metastatic tumor it was derived from, and the patient-matched normal genomes.

Using the Illumina GAIIx and HiSeq 2000 platforms coupled with new bioinformatics methods, we generated genome builds with at least 95% of the bases covered 10X or greater in each sample. We identified copy number and single nucleotide variants (CNVs, and SNVs, respectively). Variant calls were verified using (i) an independent array platform, (ii) a targeted capture method, and (iii) with PCR followed by Sanger sequencing technology, allowing us to develop a robust algorithm for identifying variants from the whole genome builds.

We identify 148,593 novel somatic SNVs in the tissue-derived tumor genome. Melanoma cancers are known to have high mutation rates and this sample may be more accelerated due to the patients advanced condition. In the cell line-derived genome, we identify 173,629 novel somatic SNVs—45,602 are different compared to the tissue-derived tumor genome.

We observe the mutational signature that ultraviolet light bears on the tissue-derived tumor, and to a lesser extent, in the cell line-specific variants. Transcription-coupled nucleotide excision repair is likely responsible for the correlation between mutation accumulation and distance along transcripts. Mutations are depleted at non-genic conserved sites suggesting enhanced repair in such regions, and confirming their evolutionary and cellular importance.

We provide results that help identify the underlying genetic components of melanoma and define the differences between a tissue-derived tumor sample and the cell line created from it. Future plans include sequencing additional melanoma genomes to build tumor-specific signatures and better classify disease state. Such information can be used to guide the development of targeted therapeutics to help fight this devastating disease.

METHYLPIPE: AN R PACKAGE FOR THE ANALYSIS OF BASE-PAIR RESOLUTION DNA METHYLATION DATA.

Mattia Pelizzola, Ryan Lister, Joseph R Ecker

Salk Institute for Biological Studies, Genomic Analysis, La Jolla, CA, 92037

DNA methylation is a potentially heritable epigenetic modification of the genomic DNA typical of most eukaryotic organisms and critical for the regulation of gene transcription. It is important for the onset of cellular differentiation processes and varies according to age, diet and environment, and can be deregulated in diseases as cancer. Nowadays it is possible to generate genome-wide base-pair resolution DNA methylation maps but there is currently no comprehensive software available for handling the analysis of such datasets. MethylPipe is an R package that includes a series of objects and methods for the management, query, analysis and visualization of DNA methylation data and their integration with heterogeneous data types.
Ranged data (collection of genomic regions of interest), lists of ranged data and transcriptional units are the basic data types. Methods are available for comparing and combining ranged data. One or more ranged data objects can be quickly visualized on a USCS temporary track to compare them to online annotation sources. The latter can be easily incorporated into MethylPipe as additional ranged data objects. For any ranged data the absolute and relative methylation profiles can be determined and visualization routines allow displaying of mean profiles and heatmaps.
Several methods for the detection of differentially methylated regions are implemented, according to the number of samples and number of replicates. Finally, integrative heatmaps of heterogenous data types (for example DNA methylation, histone modifications, ChIP-seq transcription factor binding site profiles and genomic annotations) can be easily generated.
Standard Bioconductor structures are used as input and output, making the use of other R packages for preliminary or further analysis steps effortless. For ease of use, lower level functions are available to make the algorithms executable independently from the objects they were formerly designed for. MethylPipe methods that are more demanding in terms of computational resources are optimized in terms of memory efficiency and multi-processor support.
In conclusion, MethylPipe includes a series of objects and methods that can be used as building blocks for the creation of pipelines for the data analysis of epigenomics data and their integration with any kind of annotation or additional data type.

INITIAL *DE NOVO* TRANSCRIPTOME AND GENOME ASSEMBLIES OF *NOTHOBRANCHIUS FURZERI* – A NEW MODEL FOR AGEING RESEARCH

Andreas Petzold, Bryan Downie, Matthias Platzer, Kathrin Reichwald

Leibniz Institute for Age Research – Fritz Lipmann Institute, Genome Analysis, Jena, 07745, Germany

The turquoise killifish (*Nothobranchius furzeri*) is the vertebrate with the shortest known lifespan in captivity. Furthermore, laboratory strains derived from wild locations showing differences in yearly precipitation have differences in their lifespan in captivity (3 to 9 months). This makes *N. furzeri* a promising model organism for ageing research.

By RNA-seq we want to identify determinants for ageing in *N. furzeri*. One prerequisite for this analysis is an annotated reference sequence of its transcriptome. Towards this, we generated more than three million ESTs (131,808 Sanger reads, 3,518,547 454 FLX/Titanium reads) comprising a total of 1.2 Gb. Using the "Program for Assembling and Viewing ESTs" (PAVE), we assembled 81% of the data into 118,795 contigs with a total length of 87 Mb (N50: 854 bp, largest contig: 9,241 bp). Of these, 23,534 contigs are longer than 1 kb comprising a total of 38 Mb.

To assess complexity and completeness, we compared the *N. furzeri* transcript contigs to the most closely related fish species with an assembled genome available, the Japanese ricefish medaka (*Oryzias latipes*). Based on BLAST analyses against the medaka transcript sequences (ENSEMBL assembly HdrR, Oct 2005, database version 58.1j), we estimate that the current *N. furzeri* transcript catalogue contains 17,883 unique transcript contigs derived from 15,922 genes. These represent roughly 80% of presently annotated protein coding genes in medaka.

Furthermore, using whole-genome shotgun data generated by 454/Roche (4.7 Gb) and Solexa/Illumina (43 Gb) with an approximate total coverage of 24x and CLC Assembly Cell, we have *de novo* assembled ~40% of the *N. furzeri* genome (725 Mb, 394,000 contigs, N50: 2.9 kb). About 35% of the transcript contigs >1 kb align to unique genomic sequences allowing a first gene annotation.
Altogether, the presented *de novo* assemblies will provide the basis for gene expression and epi-/genetic variation analyses in *N. furzeri* of different strains and ages.

GENES2MIND.ORG: AN ONLINE RESOURCE FOR THE GENOMIC PROFILING OF PSYCHOACTIVE DRUGS

Marcin Piechota, Michal Korostynski, Wiktor Mlynarski, Ryszard Przewlocki

Institute of Pharmacology PAS, Molecular Neuropharmacology, Krakow, 31343, Poland

Molecular biology and genomics provide better understanding of the effects of psychoactive drugs on the central nervous system. The transcriptional complexity of the brain response to the drug require high-throughput approaches and data analysis tools. We describe here a database containing comparison of effects of various classes of psychotropic drugs (antidepressant, analgesic, psychostimulant and antipsychotic) on transcriptional alterations of ~20,000 genes in the mouse brain. The data were generated using whole genome Illumina microarrays at 1, 2, 4 and 8 hours time points after drug administration. The data are stored in a MySQL database, logic layer is Java based and AJAX (GWT) is utilized in the presentation layer. Implemented data analysis and visualization tools enables the identification of drug-specific genomic signatures and drug-related transcriptional modules. Our system is capable of performing multidimensional data analysis (PCA), co-expression analysis, determining genome signature similiarity and heatmap plotting. Fold change of candidate genes can be visualized with barplots. The genes2mind database provides an open data resource for a wide variety of neurobiological and pharmacological research studies. The detailed comparison between addictive and antidepressant drugs is presented as an example.
This work was supported by MSHE grants POIG DeMeTer 3.1 and NN405 274137.

SMALT – AN EFFICIENT AND ACCURATE MAPPER FOR DNA SEQUENCING READS

Hannes Ponstingl, Zemin Ning

The Wellcome Trust Sanger Institute, Sequencing Informatics, Hinxton, Cambridge, CB10 1SA, United Kingdom

A computer program is presented that facilitates the efficient and accurate mapping of DNA sequencing reads onto genomic reference sequences. Reads from a range of sequencing platforms, for example Illumina, Roche-454 or ABI-Sanger, can be processed.

The software employs a hash index of short words, less than 15 nucleotides long, sampled at equidistant steps along the genomic reference sequences. For each read, potentially matching segments in the reference are identified from seed matches in the index and subsequently aligned with the read using a banded Smith-Waterman algorithm.

The best gapped alignments of each read are reported including a score for the reliability of the best mapping. The user can adjust the trade-off between sensitivity and speed by tuning the length and spacing of the hashed words. A mode for the detection of split (chimeric) reads is provided. Multi-threaded program execution is supported.

Speed, sensitivity and error rates were assessed on paired sequencing reads generated computationally from the sequence of the entire human genome. Single base changes and short insertions and deletions of up to 10 nucleotides were introduced randomly at a rate of 1% with every 5th variation an insertion or deletion. A total of $4x10^6$ read-pairs were generated with an individual read length of 100 nucleotides. These simulated reads were mapped on a single Intel E5420 2.5 GHz CPU. The average speed was $1.0x10^6$ read-pairs per hour using 3.3 GB memory. 97.6% of the reads were confidently mapped with an error rate of 0.1%.

The speed combined with high sensitivity and low error rates across a range of sequencing platforms makes the program very useful for genomic re-sequencing projects.

The software is available via FTP from
ftp://ftp.sanger.ac.uk/pub/hp3/smalt.tgz

SEQUENCED PRIMATE GENOMES ALLOW *IN SILICO* DESIGN OF UNIVERSAL PRIMERS FOR PHYLOGENETIC STUDIES OF PRIMATES.

Joan U Pontius[1], Polina L Perelman[2], Pecon-Slattery Jill[2], O'Brien J Stephen[2]

[1]SAIC-Frederick, Laboratory of Genomic Diversity, Frederick, MD, 21702, [2]National Cancer Institute, Laboratory of Genomic Diversity, Frederick , MD, 21702

The availability of full and low-coverage genomic sequence from seven primate species: three Great Apes (*Homo sapiens, Pan troglodytes and Pongo pygmaeus*), an Old World Monkey (*Macaca mulatta*), the New World Monkey marmoset (*Callithrix jacchus*), as well as two Prosimians, gray mouse lemur (*Microcebus murinus*) and bushbaby (*Otolemur garnettii*) has allowed non-degenerate primers to be designed for 1,311 genomic regions that have diverged across these primates.

The strategy in primer design entailed finding putative orthologs between the species, and, for each set of orthologs, using multiple sequence alignment to generate a consensus sequence. Primer3 was used to design primers for non-degenerate regions within the consensus sequences. The candidate primer pairs were screened using electronic PCR, retaining those having a single product per genome as well as informative inter-species variation within the target sequence.

The final set of 1,311 primer pairs represents all human chromosomes and 957 human genes. The primer products have average length of 627 bp, totaling over 800 kb of sequence. The average number of mismatches and gaps between the chimp and human sequences averages 0.6 per 100 bp, while between bushbaby and human averages 9.7 per 100 bp. A sampling of 200 of these primers amplified using samples from 11 primate species, with a success rate of 99%, suggesting that amplification may also be successful in more distantly related species.

The design of primers has not only been valuable for studying inter-species variation among the primates, but also illustrates the value of genomic sequence from low-coverage genomes.

Funded by NCI Contract N01-CO-12400

LANDSCAPES OF INCONGRUENCE: A WINDOWING APPROACH TO DELINEATING PHYLOGENETICALLY INCONGRUENT REGIONS IN GENOMIC SEQUENCE DATASETS

Arjun B Prasad[1], NISC Comparative Sequencing Program[2], Eric D Green[3], James C Mullikin[1,2]

[1]Genome Technology Branch, National Human Genome Research Institute, NIH, Bethesda, MD, 20892-9400, [2]NIH Intramural Sequencing Center, National Human Genome Research Institute, NIH, Rockville, MD, 20892-9400, [3]National Human Genome Research Institute, NIH, Bethesda, MD, 20892-2152

As an evolutionary process, speciation leaves an impression on genome sequences allowing us to infer historical relationships between species. With the ever increasing depth and breadth of species' genomes represented in sequence datasets, the hope has been that the more difficult relationships between species would become better resolved. Interestingly, for many of the species relationships that have proved refractory to previous analyses, large genomic sequence datasets have not resolved the relationships. Large molecular sequence datasets have repeatedly revealed inconsistencies in phylogenetic trees across different data partitions, often with high confidence values. This incongruence may arise through methodological failure of the inference process, or by biological processes such as horizontal gene transfer, incomplete lineage sorting, and introgression.

To better understand the patterns and origin of incongruence, we have developed a method called PartFinder that uses likelihood ratios over sliding windows to visualize tree support changes across genomic sequence alignments. This fast and easily parallelized method allows for the examination of complex scenarios among many species. As a pilot project, we applied PartFinder to investigate incongruence in the *Homo-Pan-Gorilla* group using assembled high-quality BAC sequences from chimpanzee and gorilla generated by the NISC Comparative Sequencing Program and the homologous regions from the human reference sequence.

The majority of the sequence supports the accepted closest relationship of human and chimpanzee, but a significant portion supports the other two possible relationships. We also compared the results of PartFinder to a Bayesian hidden Markov model based on coalescent theory (Dutheil et al., 2009) using the same input data and find that the two methods largely agree. Applying PartFinder to other groups of mammals that have long proved difficult to resolve, such as the New World monkeys (Platyrrhini), also show interesting patterns of incongruence. These results help to illuminate why increasing amounts of DNA sequence do not always result in better resolved trees when viewed at a global level, and may point the way to understanding why high levels of incongruence are so often evident in the analysis of large molecular phylogenetic datasets.

COMPARATIVE GENOMICS BETWEEN *VOLVOX* AND *CHLAMYDOMONAS* PROVIDE INSIGHTS INTO THE EVOLUTION OF GREEN ALGAL MULTICELLULARITY

Simon Prochnik[1], James Umen[2], Aurora M Nedelcu[3], Armin Hallmann[4], Stephen M Miller[5], Ichiro Nishii[6], Jeremy Schmutz[7], Jane Grimwood[7], Daniel Rokhsar[1,8]

[1]DOE Joint Genome Institute, PGF, Walnut Creek, CA, 94598, [2]Salk Institute, Plant Biology, La Jolla, CA, 92307, [3]University of New Brunswick, Biology Department, Fredericton, EB3 5A3, Canada, [4]University of Bielefeld, Plant Biology, Bielefeld, 33615, Germany, [5]University of Maryland, Biological Sciences, Baltimore, MD, 21250, [6]Nara Women's University, Biological Sciences, Nara Prefecture, 630-8506, Japan, [7]HudsonAlpha, Institute for Biotech, Hunstville, AL, 35806, [8]UC Berkeley, Centre for Integrative Genomics, Berkeley, CA, 94720

The genomes of several green algae have been sequenced, most interestingly, those of closely-related *Volvox carteri* and *Chlamydomonas reinhardtii*, which represent the first pair of genomes to span the evolution of multicellularity outside the clade of animals and their much more distant unicellular ancestors. Comparisons within this pair of green algal genomes and their ~15,000 predicted protein coding genes apiece allow investigation of the genetic changes that accompanied the evolution of multicellularity within the green algal lineage. The availability of genomes of outgroup prasinophyte algae *Ostreococcus* and *Micromonas*, although reduced, aid comparisons to other more distant species, improving predictions of the ancestral gene set of the *Volvox*-*Chlamydomonas*. This shows that a core set of 1,835 genes are uniquely shared between the two chlorophyceae, and that, further, there is no evidence for any protein domains or large numbers of genes specific to *Volvox*. This suggests that the volvocine-algae-specific protein families provided the raw materials for changes that accompanied the evolution of multicellularity. This contasts with the large-scale invention of proteins that accompanied the appearance of metazoans.

DESIGN AND ANALYSIS OF STOCHASTIC PROFILING STUDIES

Franz Quehenberger

Medical University of Graz, Institute for Medical Statistics, Graz, A-8010, Austria

During the last years there has been considerable attention to non-genetic cell-to-cell variability. In population-averaged assays many components of metabolic, signalling, and transcriptional networks cannot be identified. It has been shown in several case studies that there are more than one stable state possible for a cell, resulting in multi-modal distributions of gene expression. It is presumed that such heterogeneities might play a role in tissue physiology.

Recently stochastic profiling of transcripts has been proposed (Janes et al. 2010). From an initial simulation study they conclude that 10 cells per sample are best trade-off between increase of measurement error and preservation of the variability of interest.

In this study their methods are revisited and improvements are suggested in order make them more efficient and more general. Finally recommendations for the design of new studies are given.

RGASP EVALUATION OF RNA-SEQ READ ALIGNMENT ALGORITHMS

Andre Kahles[1], Regina Bohnert[1], Paolo Ribeca[2], Jonas Behr[1], Gunnar Rätsch[1]

[1]Max Planck Society, Friedrich Miescher Laboratory, Tübingen, 72076, Germany, [2]Centro Nacional de Análisis Genómico, -, Barcelona, -, Spain

As the amount of high throughput sequencing (HTS) data is rapidly growing, the need for its fast and accurate analysis becomes increasingly important. Inside a wide spectrum of algorithms developed to align reads from RNA-seq experiments, algorithms capable of performing spliced alignments form a particularly interesting subgroup. The results of these techniques are very valuable for downstream transcriptome analyses. Unfortunately, most of the original publications were not accompanied by a comparison of alignment performance and result quality.

The RNASeq Genome Annotation Assessment Project (RGASP, carried out by the Wellcome Trust Sanger Institute) was launched to assess the current progress of automatic gene building using RNA-Seq as its primary dataset. Its goal was to assess the success of computational methods to correctly map RNA-Seq data onto the genome, assemble transcripts, and quantify their abundance in particular datasets. The input data originated from three model organisms (Human, Drosophila, and C. elegans) and comprised data from different sequencing platforms (Illumina, SOLiD and Helicos).

As part of RGASP, also alignments of a variety of different methods, including BLAT, GEM, PALMapper, SIBsim4, TopHat, and GSnap, were submitted; here we present the results of the analysis we performed on these submissions. Besides different descriptive statistical criteria, as sensitivity and precision of intron recognition, mismatch and indel distribution, we also compared the alignments among each other, e.g., with respect to the agreement of intron predictions and multiple mappings of reads. We further investigated the influence of different alignment filtering strategies to the alignment performance in general but also respective to downstream analyses as transcript prediction and quantification.

Our comparisons showed a great diversity in the behavior of the different alignment strategies with surprisingly small agreement between a subset of methods. We can show, that different filtering strategies influence the performance significantly and can drastically increase the precision of transcript prediction and transcript quantification. The evaluations of the transcript annotations derived from these alignments additionally allow us to correlate alignment accuracy with the preciseness of exon, transcript, and gene prediction. We will discuss specific features of the different alignment strategies that most influence the success of subsequent analysis steps.

The tools developed for this analysis are incorporated into the Galaxy instance at http://galaxy.fml.mpg.de. Further details will be available from http://www.fml.mpg.de/raetsch/suppl/srm-eval.

DEEP SEQUENCING OF *SCHMIDTEA MEDITERRANEA* REVEALS STRAIN-SPECIFIC TRANSCRIPT EXPRESSION

Alissa M Resch[1], Dasaradhi Palakodeti[1], Yi-Chien Lu[2], Michael Horowitz[1], Brenton R Graveley[1]

[1]University of Connecticut Health Center, Dept. of Genetics and Developmental Biology, Farmington, CT, 06032, [2]Weill Cornell Medical College, Dept. of Pathology and Laboratory Medicine, New York, NY, 10065

A central question in RNA biology is how complex patterns of transcript expression regulate key pathways in development. The freshwater planarian is a useful model for studying mechanisms of stem cell function and germline development; planarians exist as both sexual and asexual strains and possess pluripotent stem cells called neoblasts, which are responsible for their regenerative abilities and developmental plasticity.

We characterized the transcriptome of *S. mediterranea* using RNA-Seq data collected from sexual and asexual strains. Sixty-four million read sequences were aligned to genomic contigs and assembled into 19,503 transcripts from 17,682 genes. A GC content of 36% was estimated for coding regions. Functional annotation of transcript sequences revealed that 48% map to known homologs (evalue <1e-10) and 20% map to curated protein domains (evalue <1e-3). We surveyed the transcriptome for evidence of alternative splicing and discovered 798 novel alternative splice relationships in 408 genes, indicating that alternative splicing increases transcript diversity. A select group of transcripts were experimentally validated using RT-PCR, and results showed that 88% of predicted transcripts and 75% of predicted alternative splice forms were successfully validated.

Comparison of global transcript expression profiles between sexual and asexual strains revealed that ~15% of transcripts are differentially expressed, and that alternative splicing regulates strain-specific isoform expression. Significant differences in keyword enrichment were observed between strains; keywords associated with sexual reproduction were specifically found in transcripts expressed in the sexual strain. Experimental validation of specific examples of strain-specific expression showed that 84% of our predictions were accurate.

We compared transcript expression levels between irradiated and non-irradiated worms, and determined that ~5% of transcripts display neoblast-specific patterns of expression; transcript expression was up-regulated in sexual and asexual non-irradiated worms, but showed a marked decrease in irradiated strains. Analysis of RNA-Seq data taken from human embryonic stem (hES) cells revealed that a population of neoblast-specific transcripts are homologous to human sequences up-regulated in hES cells, demonstrating that transcript expression is evolutionarily conserved for a set of genes that regulate stem cell function.

HIERARCHICAL CLUSTERING OF METAGENOMIC DNA READS BASED ON OLIGONUCLEOTIDE COMPOSITIONAL BIASES

Oleg N Reva

University of Pretoria, Biochemistry, Pretoria, 0002, South Africa

New high throughput DNA sequencing technologies produce massive primary sequencing data. Complex biological samples such as metagenomes of environmental and pathogenic bacteria thus become amenable to analysis. Managing and analyzing of vast amounts of sequence data generated in sequencing projects calls for new reliable and high throughput programs. A computer program 'SeqWord MetaLingvo' was created to meet the challenge of a large-scale simultaneous analysis of multiple DNA reads generated by the next-generation sequencers and identify presence of pathogenic microorganisms or virulence-associated genomic islands in environmental and clinical samples. Binning and identification algorithms are based on multidimensional representation and unsupervised hierarchical clustering of DNA reads by analysis of their compositional biases. Several databases of standard oligonucleotide usage patterns have been created for all completely sequenced bacterial chromosomes, plasmids, viruses and mobile genomic islands. All these databases are portable and the database creation tools are provided with the program to allow updating and adapting the databases to achieve the best performance. The program is provided with a user-friendly GUI. It is scalable for analysis of large metagenomic datasets and may be run on a personal desktop computer. The program is available for download from the SeqWord project Web-site at http://www.bi.up.ac.za/SeqWord/metalingvo/. The analysis starts with reading the sequences of interest stored locally in FASTA or GenBank formats. First, the program calculates for the given DNA sequences the oligonucleotide usage (OU) patterns and creates an analytical workspace. The hierarchical clustering algorithm is used on this step to analyze the compositional variability of the OU patterns and to select the outermost patterns for a multidimensional clustering and representation of DNA sequences. Sets of OU patterns organized as hierarchical clusters next may be used for: i) saving the standard patterns to a new hierarchically structured database; ii) identification of unknown DNA sequences or clusters of DNA reads by calculating Mahalanobis distances to the stored standard OU patterns followed by calculation of corresponding statistical parameters to measure the likelihoods of the false-positive and false-negative identification; iii) clustering of DNA reads by compositional similarity and storing them to new FASTA files per cluster; iv) alignment-free phylogenetic analysis of individual genomes or the whole clusters.

EXTENSIVE INNOVATION IN THE EVOLUTION OF TRANSIENT RAB:EFFECTOR INTERACTIONS

Maria Luisa Rodrigues, Filipe Tavares-Cadete, José B Pereira-Leal

Instituto Gulbenkian de Ciência (IGC), Computational Genomics Laboratory, Oeiras, 2780-156, Portugal

In vesicular transport, a vital process for eukaryotic cells, molecules are transported from a donor compartment to their final destination via spherical or tubular membrane-enclosed compartments, generally known as vesicles. Rab proteins are central components of membrane trafficking machinery, mediating almost every step in vesicle transport. Rabs carry out their regulatory role by recruiting downstream molecules – the Rab effectors, which bind to their active, GTP-bound, form (reviewed in (1)). Trafficking pathways are believed to be an ancestral eukaryotic feature, as their major molecular components are found throughout the eukaryotic tree of life. Their evolution by gene duplication and divergence is clearly exemplified by Rabs, with more than 60 paralogues in humans (2). This suggests that duplication and divergence of existing Rab effectors could play a role in the evolution of new Rab:effector interactions, as previously shown for other protein-protein interactions (3, 4). Moreover, Rabs are ancient proteins, some of which predicted to be present in the last common eukaryotic ancestor, opening the question whether effector proteins and the corresponding Rab:effector interactions are also conserved.
In order to test if Rab:effector interactions emerge by duplication from existing effectors and if they are conserved in evolution, we compiled a dataset of Rab:effector interactions in *Homo sapiens* and *Saccharomyces cerevisiae*, from public databases and literature curation. We then performed a comparison of effector repertoires in a number of representative eukaryotic organisms, searching for orthologs. For the comparison of Rab:effector interactions in human and yeasts, we used two complementary protein family assignment methods to detect paralogous proteins - structural domains and sequence families.
This first systematic study on the evolution of Rab:effector interactions shows that Rab effectors rarely arise by duplication, that they are not conserved between orthologous human and yeast Rabs, and that they appear to emerge in a taxon-specific manner. This is supported by the observations that the majority of effectors binds a single Rab protein; that there is a low similarity between Rab interactomes; that orthologous Rabs interact with distinct sets of effectors and that phylogenetic profiles of Rabs and corresponding effectors are very different. Our results reveal that in a background of a system that evolves by duplication and divergence, critical innovation and specificity is achieved by the recruitment of novel components and functionalities.

(1) Stenmark, H. (2009) Nat Rev Mol Cell Biol 10, 513-25.
(2) Pereira-Leal, J. B., and Seabra, M. C. (2001) J Mol Biol 313, 889-901.
(3) Pereira-Leal, J. B., Levy, E. D., Kamp, C., and Teichmann, S. A. (2007) Genome Biol 8, R51.
(4) Pereira-Leal, J. B., and Teichmann, S. A. (2005) 15, 552-9.

HEARING THROUGH THE GRAPEVINE OF RNA-SEQ EXPRESSION PROFILES

Michael Sammeth, Mar Gonzàlez Porta, Roderic Guigó

Centre de Regulacio Genomica, Bioinformatics and Genomics Program, Barcelona, 08002, Spain

Over the recent years RNA-Seq has been demonstrated to be a powerful tool for the generic surveillance of the transcriptome a cell expresses. RNA-Seq readouts of the transcriptional status in a cell raise the field of personal genomics to a new level as they open the path for analyzing variability in gene expression and processing caused by the impact of individual differences in genetically inherited information. However, the sensitivity of thus obtained read counts depends crucially on the homogeneity of the sample, e.g, purified cell lines in contrast to complete tissues/organs or even entire organisms-and-the number of different individual sources the sample has been obtained from. In addition to these background fluctuations, bioinformatics analyzes are known to be complicated by impacts of technical nature on the distribution of reads along a gene, especially affecting attempts of deconvoluting a gene's expression into alternative products or allelic copies. Finally, suitable approaches are required for the comparisons of apples with pears--as imposed by the different nature of phenomena, e.g., the comparison between gene expression and alternative processing of genes, or, when assessing common tendencies in the variability observed when exposing different individuals to a common treatment.

Herein, we present such challenges as arisen by RNA-Seq datasets from different species, individuals, cell type mixtures and states/treatments, and we propose computational approaches to address the problems in order to deduce meaningful biological messages from the observed read tags. Our studies extend to the number and ranking observed for expressed genes in corresponding tissues/cell-lines between individuals in contrast to different tissues/cell-lines from heterogeneous sample sources. Against the background of technical fluctuations, we compare the multi-variate expression levels of RNA molecules between individuals and cell states, discovering many interesting messages from biology along the way.

QUANTIFICATION, ERROR MODELLING AND QUALITY CONTROL OF RNASEQ USING SPIKE-IN CONTROL SEQUENCES

Felix J Schlesinger[1,2], Carrie A Davis[1], Alexander Dobin[1], Chris Zaleski[1], Marc L Salit[3], Thomas R Gingeras[1]

[1]Cold Spring Harbor Laboratoty, Functional Genomics Laboratory, Cold Spring Harbor, NY, 11724, [2]Cold Spring Harbor Laboratoty, Watson School of Biological Sciences, Cold Spring Harbor, NY, 11724, [3]National Institute of Standards and Technology (NIST), Biochemical Science, Gaithersburg, MD, 20899

High throughput shotgun sequencing of cDNA molecules (RNAseq) is a powerful technique to reconstruct and quantify the transcriptome of cells at base-pair resolution with a large dynamic range. It provides digital counts of expression for exons, splice junctions and other elements. In order to compare these counts between different experiments and detect differential expression, however, some normalization is needed. Normalizing to the total number of sequenced reads is commonly done, but can lead to wrong calls for individual genes if systematic changes occur between the samples studied. Another approach is to use "spike-in" controls, artificial RNAs of known sequence and concentration that are added to the sample before library preparation.

In addition to the normalization of read counts, such controls also allow for quality control of the RNAseq process and detection of the various biases, which affect the different protocols and technologies in use. Such biases are caused by library preparation, especially in the reverse transcription, fragmentation and amplification steps, and by the sequencing process itself. RNA spike-ins allow measuring error rates, the strandedness of the results, sequence biases and the evenness of coverage, edge effects (representation of the ends of RNA molecules), overdispersion of reads compared to random sampling (caused in part by PCR overamplification), and the generation of artificial chimeric reads on diverse and realistic sets of sequences. Comparing the observed read counts to spike-in concentrations and between replicates we can measure the precision and reproducibility of quantification at different expression levels to estimate confidence intervals for transcript abundance measurements.

In this study we present results from a pool of 96 RNA spike-ins created at NIST with diverse sequences covering 6 orders of magnitude in abundance. This pool as used in 90 RNAseq libraries totalling 18 billion reads as part of the ENCODE project.

K-MER ANALYSIS TO REVEAL GENOMIC AMBIGUITIES IN HIGHLY REPETITIVE GENOMES

Thomas Schmutzer, Burkhard Steuernagel, Fabian Bull, Andreas Houben, Uwe Scholz, Nils Stein

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Cytogenetics and Genome Analysis, Gatersleben, 06466, Germany

Many agronomically important plants are encoded by large genomes which are composed of up to 90% repetitive DNA, like tandem repeats (TD) or transposable elements (TEs). While high-throughput sequencing technologies are promising methods to sequence larges genomes in an affordable way, the high number of repeats still hamper the subsequent data analysis. To identify *in silico* highly repetitive regions we used the 'k-mer frequency', a method which is based on the exact occurrence of short sequence seeds. A 'k-mer' (word of length k) that originates from a repeat structure has a significantly higher frequency in context of the complete genome.

In order to analyze the high copy DNA fraction of the rye (*Secale cereale* L.) genome (1C = 8 GB) whole genome shotgun (WGS) data including 1,1 Mio sequence reads were screened for 'k-mer frequencies' using an index structure based on enhanced suffix arrays. By deploying this set of 'k-mer frequencies' on any data, regions of possible interest (e.g. genes) can be detected by a low average 'k-mer frequency'. For high throughput screening of any related dataset with our frequency index we implemented a pipeline with a batch mode option. Corresponding index structures were applied to screen a subset of 54,408 sequence contigs resulting from 454-sequence reads of genomic rye DNA. The same approach was extended to 70,000 rye ESTs to validate low copy sequences. We discuss cut-offs of 'k-mer frequencies' which could be used to locate regions of interest due to their low or high repetitive nature. 3,531 regions have been evaluated regarding their complex constitution, and similarity to coding sequences of the *Brachypodium* genome (26,552 CDS). These identified regions originating from 550 contigs displayed a low complexity of ~79% in comparison to non-matching parts of the processed genomic rye contigs. Furthermore, we used 'k-mer frequencies' to discard "noisy" contigs from SNP detection. In consequence we were able to reach a more reliable set of sequence polymorphisms.

NextGenMap: Using high throughput hardware for high throughput sequencing

Fritz J Sedlazeck[1,3,4,5], Gregory B Ewing[2,3], Arndt von Haeseler[1,3,4,5]

[1]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna (CIBIV),, Vienna, A-1030, Austria, [2]Max F. Perutz Laboratories, Mathematics and BioSciences Group (MaBS), Vienna, A-1030, Austria, [3]University of Vienna, , Vienna, A-1010 , Austria, [4]Medical University of Vienna, , Vienna, A-1090, Austria, [5]University of Veterinary Medicine, , Vienna, A-1210, Austria

High throughput sequencing provides access to important information on genes, gene function and genetic variation of genomes. Nonetheless, the amount of information generated per run is enormous and will even increase in the next future. Thus, the demand for expensive high performance computers to analyze such data is also increasing. The unavailability of appropriate HPC infrastructure will therefore become a major bottleneck for future research and non-standard application, such as for instance sequencing projects species evolutionary distantly related to fully sequenced model organisms.
Here, we assess some of the current limitations of state-of-the art genome assembly programs. To this end, we present a novel evaluation approach, which does not only count wrongly mapped reads, but also takes into account the number of correctly mapped nucleotides with respect to a reference genome.
Quantifying accuracy in this way is particularly important when it comes to variation studies, such as, for instance, SNP detection.
Our new evaluation scheme is applied to two artificial (simulated) genomes that differ by 1% from the Arabidopsis genome and by 10% from the Drosophila genome, respectively. We assess the accuracy of the frequently employed programs SSaha2, Bowtie, BWA, Shrimp and Maq.
We also introduce a novel reference genome based assembly approach (NextGenMap), which deploys a graphics processing unit (GPU) as co-processor. We demonstrate that the GPU-based reference assembly algorithm outperforms a multicore general purpose CPU with respect to costs and run time. Thus reference genome-based assembly of short reads can be carried out without the need for an expensive HPC infrastructure.
NextGenMap achieves a speed up of more than 10 over SSaha2, and at the same time an increase of mapping accuracy of up to 8% (compared to SSaha2).

THE UNIPROT KNOWLEDGEBASE: A TWO TIER SYSTEM OF MANUAL AND AUTOMATIC ANNOTATION.

Harminder K Sehra[1], UniProt Consortium[1,2,3]

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge , CB10 1SD, United Kingdom, [2]Protein Information Resource, Georgetown University Medical Center, Washington, DC, 20007, [3]Swiss Institute of Bioinformatics, Centre Medicale Universitaire, Geneva, CH-1211, Switzerland

The UniProt Knowledgebase (UniProtKB) is a collaboration between the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). UniProtKB is composed of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot contains information manually extracted from literature and curator-evaluated computational analyses, providing accurate sequence and functional annotation along with cross-references to other databases. UniProtKB/TrEMBL contains high quality computationally analysed records sourced from the INSDC databases, Ensembl and shortly RefSeq, enriched with automatic annotation and classification.

Although manual annotation of proteins is invaluable for the scientific community, it is very time-intensive and, with UniProtKB/TrEMBL now containing in excess of 11 million entries, it is crucial to have an automatic annotation strategy for annotating these proteins in an efficient and scalable manner. This consists of two prediction systems which are referred to as UniRule and SAAS (Statistical Automatic Annotation System). UniRule provides a set of manually reviewed rules in order to predict function and family relationships while data-driven statistical modelling by SAAS is being developed in a complementary manner.

The details of our automatic annotation approach including how it has developed from the manual annotation approach and how it augments manual curation will be presented here.

GRAMENE COMPARA GENETREES: A PHYLOGENOMICS RESOURCE FOR PLANTS

William Spooner[1], Joshua C Stein[1], Sharon Wei[1], Liya Ren[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, 11720, [2]USDA-ARS, Plant, Soil, and Nutrition Research Unit, Ithica, NY, 14853

Comparative functional genomics allows researchers to trace evolutionary histories of genes and traits. We present the database and web-site of Gramene Compara GeneTrees. The previously published Ensembl method uses clustering to define gene families and phylogenetic reconstruction to define orthologs and paralogs. It was applied to the complete genomes of six grass-lineage and four eudicot species, generating 27,031 families. As expected many species-specific genes and families were identified and most were attributed to differentially expanded transposable elements. Approximately 77% of all genes was assigned an orthologous relationship and 76% paralogous. Secondary phylogenetic analyses of called orthologs were in agreement with the expected species tree, demonstrating internal consistency of the method. Concordance of InterPro annotation was evaluated when compared between rice and Arabidopsis orthologs. Synteny between rice and sorghum, which was preserved from both speciation and an ancient whole genome duplication event, was used to demonstrate good sensitivity and specificity of ortholog and paralog calls. The Gramene website environment is integrated with genome browsers, comparative maps and functional annotations, including gene ontology and InterPro. The Compara database adds a new level of tools to aid researchers in the making inferences of function and strategies for gene annotation.

ENSEMBL GENOMES: EXTENDING ENSEMBL ACROSS THE TAXONOMIC SPACE

Daniel M Staines, Paul S Derwent, Gautier Koscielny, Paul J Kersey

European Bioinformatics Institute, Ensembl Genomes Group, Cambridge, CB10 1SD, United Kingdom

Ensembl Genomes (http://www.ensemblgenomes.org) is a portal offering integrated access to genome scale data from non-vertebrates species of scientific interest. Developed using the Ensembl genome annotation and visualization platform Ensembl Genomes consists of five sub-portals (for bacteria, protists, fungi, plants and invertebrate metazoa). Ensembl Genomes provides a common set of user interfaces (which include a genome browser, FTP, BLAST search, a query optimised data warehouse, programmatic access and a Perl API) for each species.

Data types incorporated include annotation of protein-coding and non-coding genes, cross-references to external resources, high-throughput experimental data (e.g. data from large scale gene expression studies and population genomics studies visualized in their genomic context). Comparative analysis, both within defined clades and across the wider taxonomy has been performed to generate gene trees and sequence alignments.

A wide variety of genomic variation data have been included for a number of different species, including data from large scale resequencing projects. For example, in Ensembl Plants, the *Arabidopsis thaliana* dataset contains data from screening using the Affymetrix 250k Arabidopsis SNP chip, from resequencing data (including the 1001 Genomes Project), and phenotype data from a GWAS study. In Ensembl Metazoa, data have been added for two populations of *Drosophila melanogaster* from the Drosophila Population Genome Project.

As of June 2010, there have been five releases of Ensembl Genomes. Ensembl Metazoa includes 15 genomes from 4 genera of *Diptera*, 6 nematode genomes, the body louse and the black legged tick; Ensembl Plants includes the genomes of 6 dicots and 2 monocots; Ensembl Fungi contains two yeast genomes, the mold *Neurospora crassa* and 9 *Aspergilli*; Ensembl Protists contains 3 *Plasmodium* genomes, the slime mold *Dictyostelium discoideum* and 2 diatom genomes; and Ensembl Bacteria includes over 180 genomes from 10 bacterial and archaeal clades. We aim to work with community authorities wherever possible, and plan to add additional model species in future releases.

We will describe the infrastructure of Ensembl Genomes, criteria for inclusion into the portal and plans for integrating further genomic data.

STATISTICAL TESTS FOR DETECTING DIFFERENTIAL RNA-TRANSCRIPT EXPRESSION FROM READ COUNTS

O. Stegle*[1], P. Drewe*[2], R. Bohnert[2], K. Borgwardt[1], G. Rätsch[2]

[1]Max Planck Institutes, AG Borgwardt, Tübingen, 72076, Germany, [2]Friedrich Mischer Lab, AG Rätsch, Tübingen, 72076, Germany

As a fruit of the current revolution in sequencing technology, transcriptomes can now be analyzed at an unprecedented level of detail. Established applications of this include the detection of differentially expressed genes across biological samples and the quantification of the abundances of various RNA transcripts. The next step is now to combine these concepts to identify differential expression on the level of transcripts within individual genes. Methods for this fine-grained testing of differential abundance are valuable tools to tackle key biological questions; for example the mechanisms of alternative splicing.

Here, we present a statistical testing-framework to address this important need. Most notably, our method can be applied in settings where the complete transcript annotation (TA) is available, but also when it is unknown or incorrect. Our approach is based on a kernel method, called Maximum Mean Discrepancy (MMD), directly testing for differences of the underlying read distributions, inferred from the observed reads. In our model, we incorporate the assumption that reads follow a Poisson distribution and account for biological or technical variability.
We show how existing TA, if available, can be exploited to define a maximal discriminative set of regions within a gene, further increasing the accuracy of the method.

We analyzed the proposed approach with and without TA, comparing to established methods based on transcript quantification. We looked at simulated read data as well as factual reads generated by the Illumina Genome Analyzer for four *C. elegans* samples (Barberan-Soler et al. 2009). In our analysis, the MMD test identified differential transcript expression considerably better than methods based on transcript quantification (45% vs. 30% at 1% FPR). Even more striking, in the absence of knowledge about the TA, the MMD test was still able to identify 75% of the true differential cases. Our method is therefore well suited to analyze RNA-Seq experiments where other approaches fail, namely when the TA is incomplete or entirely missing.
We further investigated the MMD test on the data from (Hillier et al. 2009), comparing to a second study (Barberan-Soler et al. 2009) of 352 genes with confirmed alternative splicing events in the early development stages of *C. elegans*. Even when not making use of TA, our method was able to detect between 40% and 85% of the transcripts with at least one log fold change between developmental stages in (Barberan-Soler et al. 2009). This result becomes even more striking when taking the TA into account.

GENOME-WIDE IDENTIFICATION OF FUNCTIONAL ELEMENTS IN HUMAN USING A NOVEL APPROACH INVOLVING ANALYSIS OF OVER-REPRESENTED SEQUENCE MOTIFS

Todd D Taylor[1], Ramkumar Hariharan[1,2], Reji Simon[2]

[1]RIKEN Advanced Science Institute, MetaSystems Research Team, Yokohama, 230-0045, Japan, [2]Rajiv Gandhi Center for Biotechnology, Translational Cancer Research Laboratory, Kerala, 695014, India

The goal of this study was to identify short functional elements in the human genome that might have been *"overlooked"* by earlier evolutionary conservation based and other rigorous methods that have analyzed limited sub-sections of the genome. We developed a novel approach that consists of systematically identifying and analyzing all over-represented k-mers ($1 \leq k \leq 20$) in the entire human genome. Over-represented motifs that show perfect intra-species conservation may represent regulatory elements or important structural elements. Specifically, we computed sequence elements that were statistically overrepresented in defined genomic locations including non-overlapping upstream/downstream regions of known genes, non-overlapping upstream/downstream regions of non-coding RNA regions, coding exonic regions, introns, intergenic regions that are less than 1kb apart, intergenic regions that are more than 1kb apart or non-annotated regions, and other repetitive regions of the human genome. We identified several thousand statistically over-represented motifs in the human genome. Characterization of these motifs *"re"*-discovered a significant proportion of previously known transcription factor binding sites from TRANSFAC and almost all known repetitive elements. Additionally, we were able to annotate hundreds of the motifs using data from ChIP-ChIP and ChIP-Seq experiments available in the literature. For some of the remaining *"novel"* enriched sequence elements in the genic upstream/downstream, non-coding RNA upstream/downstream and intronic regions, we assigned candidate functions based on the genes adjacent to these elements. We identified dozens of sequence elements that flank clusters of genes sharing an enriched Gene Ontology (GO) term. A similar analysis of palindromes revealed that many of them are highly enriched in almost all the investigated locations. Following functional annotation with information from known datasets, a subset of the novel palindromes could be assigned putative function based on GO annotations of adjacent genes. While more detailed analysis is pending, we believe that our approach to functional element discovery, which is independent of comparative genomics methods and makes full use of the entire genome sequence, can complement existing methods in a powerful way.

MAPPING QUALITY VALUES FOR NEXT GEN SEQUENCING READS AND THEIR PREDICTIVE VALUE IN SMALL VARIANTS, ALTERNATIVE SPLICE EXON JUNCTIONS AND NOVEL GENE FUSION DETECTION.

Sowmithri Utiramerur, Zheng Zhang, Xing Xu, Eric Tsung, Caleb J Kennedy, Onur Sakarya, Dumitru Brinza, Fiona C Hyland, Asim Siddiqui

Life Technologies, Genomic Systems Division, Foster City, CA, 94404

Next generation sequencing (NGS) technologies have brought great promise to the use of DNA/RNA sequencing in a variety of biological applications, including resequencing, DNA Methylation, whole-transcriptome and cancer genomics. The reads generated by NGS instruments are relatively short and can be difficult to align uniquely and unambiguously to a reference genome, especially in the case of large, complex genomes (such as human) that contain repetitive and homologous regions. The confidence scores associated with alignments are important for reliable structural variant detection. Several mapping tools such as MAQ, BFAST, SHRiMP and BWA compute mapping quality values (mqv) to represent the accuracy of alignments, however, the algorithms do not include all types of read pairs and alignment types.

We introduce a inference based statistical approach to calculate mqvs for different library types such as single fragment, mate-pair and paired-end, which makes use of all paired read information including insert size distribution, read orientation, and strand and gene-id annotations. The algorithm is implemented in Bioscope software.

The accuracy and predictive value of the mqvs calculated using this algorithm was tested using both a simulated dataset of human reference chromosome 1 and an actual entire genome dataset from HuRef sample. The simulated reads with overall polymorphism rate of 0.001, were mapped back to the Hg18 reference using different alignment tools and the resulting mqvs were compared. Bioscope mqvs better represents the phred-scale alignment probability for all the different library types.

Variant calls on the HuRef sample dataset were made using Bioscope. SNP and small InDel calls were compared to those in the dbSNP (b129) database. Using a higher mqv cutoff for underlying reads used as evidence, we were able to reduce the false positive rate of SNP and small InDel detection by as much as 5% and 40% respectively, without loosing sensitivity. For whole-transcriptome analysis, exon junctions and gene fusions predicted in UHR and MC7 datasets were validated using real-time PCR assays and the false positive rate of gene fusion calls was reduced 12% by applying an mqv threshold.

In summary, our algorithm is shown to accurately represent phred quality scores and is comprehensive to include different alignment types and library types. The predictive value of mqvs is demonstrated directly and by the improved efficiency of variant calls and gene fusion calls made using Bioscope. Together with the base quality values of individual bases in a read, mqvs can be used to improve the efficiency of rare-allele detection in cancer genomics research.

SINGLE CELL ANALYSIS REVEALS EVOLUTIONARY ROBUSTNESS AND CHANGE IN MSX1 PROMOTER COMMUNICATION

Keith W Vance[1], Dan J Woodcock[2], Sascha Ott[2], Chris P Ponting[1], Georgy Koentges[3]

[1]MRC Functional Genomics Unit, University of Oxford, Oxford, OX1 3QX, United Kingdom, [2]Warwick Systems Biology, University of Warwick, Coventry, CV4 7AL, United Kingdom, [3]Dept of Biological Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom

The conservation of genetic hierarchies and programmes contrasts with significant differences in developmental speed and size of embryos. Communication between promoters and cis-regulatory modules (CRMs) is expected to be at the heart of regulatory conservation and change and it has remained elusive how much of this is encoded in the DNA sequence itself. Here we assay the robustness and change of cis-regulatory communication in hybrids of homologous Msx1 fugu and mouse CRMs and promoter DNA. We develop a new stochastic model that estimates DNA copy-number independent transcription rates from fluorescent reporter time course measurements in single cells. We find that promoter sequences define basic transcriptional rate distribution modes, while CRMs shift their means in a stereotypic manner. Our hybrid experiments also reveal global evolutionary changes in promoter communication that we can assign to specific components. We detect an unusually constant coefficient of variation across all transcription rates. This suggests a generic noise buffering that permits evolutionary change in copy number or allelic expression without immediate deleterious consequences.

ASSESSMENT OF EFFICIENCY IN DIRECTED SEQUENCING STRATEGIES

Jason Walker, Todd Wylie, Jasreet Hundal, Ryan Demeter, Vincent Magrini, Daniel C Koboldt, Elaine R Mardis, Richard K Wilson

Washington University School of Medicine, The Genome Center, St. Louis, MO, 63108

Directed sequencing strategies (e.g., whole exome capture, RNA-seq, Methyl-Seq) employing Next Generation Sequencing (NGS) have proven to be cost-effective solutions for identifying disease-related germline and somatic mutations in both coding and non-coding portions of the human genome,. Efficiency in generating sequence to adequately cover Regions Of Interest (ROI) is one of the greatest factors contributing to detection of genomic variation in directed sequencing efforts.

Efficiency is measured by the amount of sequence aligned on and off target to the defined ROIs. Coverage is based on both the breadth and depth of sequence reads overlapping the ROIs . Those overlapping reads and the resulting coverage can then be used to support genomic variation including SNVs, indels, structural variation, gene expression, methylation status, epigenetic factors, etc. We have developed tools to generate efficiency and coverage metrics for the purpose of comparison of both analysis and sequencing approaches.

SNV concordance with known genotypes indicates the overall power of an exome sequencing approach to confidently detect novel variant candidates. Here we present an automated analysis approach for early characterization of efficiency in sequencing defined genomic targets within a high-throughput environment, as well as providing comparative power between disparate protocols and platforms. As an example, we evaluated the efficiency of exome capture across 90 matched pairs of ovarian cancer samples and the relationship of the sequencing efficiency to genotype concordance.

COMPUTATIONAL SIMULATION OF EVOLUTION IN THE RIBOSOMAL DNA TANDEM ARRAY IN YEAST

Claire West[1,4], Steve James[1], Donald MacKenzie[2], Robert Davey[3], Jo Dicks[4], Ian N Roberts[1]

[1]Institute of Food Research, National Collection of Yeast Cultures, Norwich, NR4 7UA, United Kingdom, [2]Institute of Food Research, Integrated Biology of GI Tract, Norwich, NR4 7UA, United Kingdom, [3]The Genome Analysis Centre, BioInformatics, Norwich, NR4 7UH, United Kingdom, [4]John Innes Centre, Computational and Systems Biology, Norwich, NR4 7UH, United Kingdom

Ribosomal RNA genes (referred to as ribosomal DNA or rDNA) are encoded in a tandem array of repeating units, the number of which varies between organisms. In the yeast *Saccharomyces cerevisiae* there are ~140 repetitive elements on chromosome XII. Until recently, it was assumed that the sequences of all array units were identical. However, the Saccharomyces Genome Resequencing Project (SGRP) enabled the discovery and quantification of variation amongst array units in several yeast strains. In particular, a new type of microheterogeneity was recorded, where a subset of repeats have a Single Nucleotide Polymorphism (SNP) while others do not. These variations are referred to as partial SNPs, or pSNPs[1].

The rDNA tandem array is believed to evolve through a process of concerted evolution, which over time homogenises the sequences between array units, though as we now know not perfectly. Key mechanisms thought to play a role in concerted evolution are Unequal Sister Chromatid Exchange (USCE) and Gene Conversion (GC). We wish to understand the process of rDNA tandem array evolution more deeply, examining the balance and rate of these different evolutionary mechanisms with regard to experimental datasets, and to discover the way in which they lead to variation, such as pSNPs, that we have observed.

We have begun to analyse the evolution of the rDNA tandem array through the development of a computational simulation tool. We focus initially on mitotic events, where the sister chromatid is preferentially used as the repair template for double strand breaks. Here, we present our first results from a carefully designed simulation experiment, following the fixation and loss of polymorphic units within an array during simulated concerted evolutionary events. In the longer term, we plan to develop our models further within a mathematical framework. In particular, we aim to use them in the estimation of finely scaled phylogenies, enabling discrimination between closely related yeast strains, and to improve understanding of repetitive sequence dynamics and genome stability.

1) James SA, O'Kelly MJ, Carter DM, Davey RP, van Oudenaarden A, Roberts IN (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. *Genome Research* 19:626-635 doi:10.1101/gr.084517.108

PROTPAL: PHYLOGENETIC RECONSTRUCTION OF ANCESTRAL DNA AND PROTEINS

Oscar Westesson, Ian Holmes

UC Berkeley, Bioengineering, Berkeley, CA, 94720

We present a method for likelihood-based phylogenetic reconstruction of ancestral protein and DNA sequences. We use string transducers which allow unified probabilistic modeling and reconstruction of insertion, deletion, and substitution events. By utilizing a series of flexible and tunable constraints, we can efficiently approximate the maximum likelihood (ML) reconstruction of evolutionary events and sequences. Transducers can be thought of as finite state machines that mimic the action of evolution; a sequence is "absorbed" as input, undergoes substitution, insertion, and deletion events, and a new sequence is "emitted" to the output tape. As transducers model the distribution of emitted sequences conditional on the absorbed sequence, it is natural to chain together several machines, where the output of one transducer is fed as input to another. Here we place one transducer on each branch of a binary tree, duplicating the output tape of each transducer and feeding it to each of its child branches. This ensemble of transducers naturally yields a probability distribution for sequences at internal nodes. To improve tractability, we employ the technique of progressive profile reconstruction put forth in the DNA reconstruction program Ortheus [1]: proceeding from leaves to root, a profile of plausible reconstructed sequences is stored at each internal node. Profiles are joined as the reconstruction progresses up the tree, implicitly discarding some candidate reconstructions as more evidence becomes available. Upon reaching the root, the ML path through the profiles can be traced back down the tree. The result is complete reconstruction of ancestral sequences and the evolutionary events leading to extant sequences. Inquiries concerning ancestral sequences or estimates of indel and substitution rates can be addressed with a reconstruction as a starting point, and alignment annotation tasks such as gene-finding can be incorporated into the reconstruction algorithm.

Our implementation of this method is capable of reconstructing ancestral DNA and proteins and is freely available under the GPL at www.biowiki.org/ProtPal. In simulation studies we find that our method accurately reconstructs significantly more characters than using a fixed alignment and Felsenstein's pruning algorithm. We provide a precise mathematical description of the algorithms pertaining to our method with the hope that this will enable the concepts of progressive reconstruction and sequence profiles to be used for other methods which could benefit from an alignment-independent approach, or whose properties make exact inference intractable.

1. Paten B, et al. Genome Res. 2008 Nov;18(11):1829-43.

SOLANACEAE GENOMICS RESOURCE

Brett R Whitty, C. Robin Buell

Michigan State University, Plant Biology, East Lansing, MI, 48824

The plant family Solanaceae is a large, morphologically diverse clade that includes a number of agriculturally significant crop species (notably Potato, Tomato, Tobacco, Chili Pepper, Eggplant and Petunia). Ongoing projects in the Solanaceae are increasingly depositing a wealth of genomic and transcriptomic data to the public sequence databases; including forthcoming data from whole genome sequencing projects in Potato and Tomato.

With funding from the National Research Initiative (NRI) Plant Genome Program of the USDA National Institute of Food and Agriculture (NIFA), we have developed the Solanaceae Genomics Resource (http://solanaceae.plantbiology.msu.edu), a web-accessible suite of databases and tools that provides a centralized repository of sequence data, and the results of value-added bioinformatic analyses on this data, for enabling the Solanaceae community in breeding and research.

For draft genomic sequences in the clade, we provide a consistent set of novel gene predictions and sequence analysis to supplement any existing public annotation data. Using transcript sequences and predicted gene models we have identified putative orthologs, paralogs, SNPs, and lineage-specific genes within the Solanaceae allowing intra- and inter-species comparisons. As well, we have identified homologs of Solanaceae species within a number of model dicot species allowing users to leverage the resources from these model species and apply them to studies in Solanaceae. Our analysis pipelines are run on all publicly available sequence data from Solanaceae species, including all varieties of transcript-derived and genomic sequence. Our results are accessible through the open source Generic Model Organism Database (GMOD) Gbrowse genome viewer, and through custom views and displays. Overall, we provide a robust and integrated comparative genomics resource permitting data-mining of Solanaceae sequences by the community, publicly accessible through a unified, user-friendly web portal.

SVMERGER: AN EXTENDABLE PIPELINE TO BUILD A COMPREHENSIVE CATALOGUE OF STRUCTURAL VARIATION (SV) BY INTEGRATION OF MULTIPLE SV DISCOVERY TOOLS AND METHODS, AND ITS APPLICATION TO 17 INBRED MOUSE STRAINS

Kim Wong[1], Binnaz Yalcin[2], Thomas Keane[1], Jim Stalker[1], Richard Mott[2], Richard Durbin[1], Jonathan Flint[2], David Adams[1]

[1]Wellcome Trust, Sanger Institute, Cambridge, CB10 1SA, United Kingdom, [2]Wellcome Trust, Centre for Human Genetics, Oxford, OX3 7BN, United Kingdom

Recently, a number of computational pipelines have been developed to call different types of structural rearrangements from paired-end Illumina data. However, no single software package can detect all of the different types of SVs (insertions, deletions, inversions, duplications, and translocations). Therefore, we have developed a computational pipeline which enables the merging of results from several existing SV calling software, in order to generate a more complete catalogue of structural variation. The current set of SV callers in our pipeline are: BreakDancer (Chen, K. et al. 2009), Pindel (Ye, K. et al., 2009), CND (Simpson, JT. et al., 2009), and two in-house large insertion finders (SEcluster and RetroSeq), which use anomalous read pairs to find deletions, insertions, inversions and translocations, split-read mapping to find deletions and small insertions, a hidden Markov model (HMM) to call duplications and deletions, and a clustering algorithm for single end mapped reads to identify large insertions, respectively. SVMerger can also perform *de novo* assemblies using reads that have mapped proximal to the predicted SV breakpoints. Realignment of the contigs to the reference genomee is then used to computational confirmation of the existence of the SV, and to identify the SV breakpoints at the nucleotide level. Our pipeline is easily extendable to include more SV callers as they become available.

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 1*29S1/SvImJ*) to between 20-35x coverage, using paired-end Illumina sequencing. Using our SVMerger pipeline, we have generated a comprehensive catalogue of SVs in all 17 of these strains. As expected, the wild-derived strains show much greater genomic variation, relative to the reference genome. Structural variations in the wild-derived strains affect approximately 4-8 times more genes than SVs in laboratory strains. A number of the SVs appear to be complex, such as inversions containing a deletion, and deletions with small insertions. We are currently validating these complex SVs, in addition to deletions, insertions, duplications and inversions, with PCR and sequencing.

GSTRUCT: A PIPELINE FOR DE NOVO GENE STRUCTURE PREDICTION FROM RNA-SEQ DATA

Thomas D Wu

Genentech, Inc., Bioinformatics and Computational Biology, South San Francisco, CA, 94080

We have developed a pipeline for de novo gene structure prediction from RNA-Seq data. This pipeline is designed to exploit features of our short read alignment program GSNAP (Wu and Nacu, Bioinformatics 26, 2010, 873-881), which is designed to find splicing events within short reads, although it may also potentially be applicable to alignment results from other programs. According to preliminary results reported from round 2 of the RGASP competition (Kokocinski, ISMB, 2010, an early version of our pipeline produced the highest accuracy in the human and Drosophila datasets, with specificity at the transcript level that was 20 percentage points better than the next best pipeline. Our pipeline, named GSTRUCT, makes inferences primarily from RNA-Seq data, and therefore differs from other strategies that use ab initio gene finding as a major component. However, GSTRUCT can also optionally make use of known splice sites, if available, to facilitate the search for novel genes and novel splicing events.

We believe that features that accounted for our performance included novel algorithms that we have developed for filtering and analyzing splicing evidence, and methods for identifying gene boundaries. Since the RGASP competition, we have been working to streamline and improve our pipeline. Some of our improvements have required new features in GSNAP, including the ability to detect short-end splicing and to identify terminal alignments. Other improvements include steps to infer increase the yield of splicing evidence, to infer missing splices, and to identify alternate splicing and structural variations. Early applications of our pipeline have revealed alternate splicing events that correspond to known annotation, despite low levels of splicing evidence. Future applications should provide a more comprehensive view of the full complement of gene structures in transcriptomes.

GENEVAR: A PLATFORM OF DATABASE AND WEB SERVICES FOR THE INTEGRATION AND VISUALIZATION OF SNP-GENE ASSOCIATIONS IN EQTL STUDIES

Tsun-Po Yang[1], Antigone S Dimas[1,2,3], Emmanouil T Dermitzakis[1,2], Panos Deloukas[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom, [2]University of Geneva Medical School, Department of Genetic Medicine and Development, Geneva, CH-1211, Switzerland, [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, United Kingdom

Genevar (GENe Expression VARiation) is a platform of database and web services designed to integrate multiple datasets, and provides analysis and visualization of associations between sequence variation and gene expression on a universal Java interface. Genevar allows researchers to investigate eQTL associations within a gene locus of interest in real time. The database and application can be installed on a standard laptop in database mode and, in addition, on a server to share discoveries among affiliations or the broader community over the internet via web services protocols.

The main advantage of this innovative system design is that users can switch between public services and their local data on the same interface. Default services at the Sanger Institute currently contain gene expression profiling and genotypic data from the following dataset: three cell types derived from umbilical cords of 75 Geneva GenCord individuals (Dimas et al., 2009).

Genevar has two main functionalities in cis-eQTL analysis: (i) identifying eQTLs in genes of interest, and (ii) observing SNP-gene associations surrounding SNPs of interest. Additional features include SNP-probe association plots and external links to three major genome browsers. Either cis- or trans-eQTLs can be plotted in the SNP-probe association plot module. Mapping results are listed in tree nodes in a structural manner, and information can be saved as PNG diagrams or exported as tab-delimited lists for further use in presentations or publications.

Future work will include modified visualization for displaying next-generation sequence data, e.g. RNA-Seq; and implementation of methylation modules to interrogate epigenomic data.

Availability: http://www.sanger.ac.uk/software/analysis/genevar/

GRAMENE: A RESOURCE FOR COMPARATIVE PLANT GENOMICS

Ken Youens-Clark[1], Ed Buckler[2,3], Terry Casstevens[4], Charles Chen[4], Genevieve DeClerck[4], Palitha Dharmawardhana[5], Pankaj Jaiswal[5], A S Karthikeyan[4], Susan McCouch[4], Liya Ren[1], William Spooner[1], Joshua Stein[1], Jim Thomason[1], Sharon Wei[1], Doreen Ware[1,3]

[1]Cold Spring Harbor Lab, Ware Lab, Cold Spring Harbor, NY, 11724, [2]Institute for Genomic Diversity, Cornell University, Ithaca, NY, 14853, [3]USDA-ARS NAA Plant, Soil & Nutrition Laboratory Research Unit, Cornell University, Ithaca, NY, 14853, [4]Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, 14853, [5]Oregon State University, Dept of Botany and Plant Pathology, Corvallis, OR, 97331

Gramene (http://www.gramene.org) is a curated data resource for comparative genome analysis a wide variety of plants. The database integrates information about genomic sequence, genes, proteins, biochemical pathways, maps and markers, QTL, germplasm, and genetic and phenotypic diversity. To index and associate these different data types, Gramene makes extensive use of ontologies (controlled vocabularies) including those for plant structures and growth stages, traits and phenotypes, gene function, biological processes, cellular components, and environments. Online tutorials and help documents provide users with an overview of how to conduct a wide variety of operations on the database. All data in Gramene is publicly-available, and all code is open source.

In the 31st release of Gramene (May 2010), our genome browser was updated to Ensembl version 58 and is host to fourteen genomes with new annotations, Fgenesh gene predictions, gene trees, whole genome alignments, and synteny views. The genetic diversity databases for rice, Arabidopsis, and maize added new data sets, and there are helpful links to start external analysis tools such as TASSEL and Flapjack from our website. Additionally, our SNP query tool was improved to show gene loci overlapping SNP positions. We added three new genetic maps, updated our RicyCyc database to use the MSU6 assembly, added a new PoplarCyc mirror, redesigned our home page, and greatly expanded DAS options.

Gramene is supported by a grant from the NSF and represents a collaborative effort between Cold Spring Harbor Laboratory, the Department of Plant Breeding and Genetics at Cornell University, Ensembl Genomes, and various national and international projects dedicated to cereal genomics and genetics research.

TRANSCRIPTIONAL ANALYSIS OF MELATONIN REGULATED GENES IN THE SHEEP PARS TUBERALIS USING NEXT GENERATION TRANSCRIPTOME SEQUENCING

Le Yu[1], Sandrine M Dupré[2], Bob Paton[1], Alan S McNeilly[3], Andrew S Loudon[2], David W Burt[1]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies , University of Edinburgh, Roslin Institute, Division of Genomics and Genetics, Edinburgh, Eh25 9PS, United Kingdom, [2]University of Manchester, Faculty of Life Sciences, Manchester, M13 9PL, United Kingdom, [3]Queen's Medical Research Institute, Medical Research Council Human Reproductive Sciences Unit, Edinburgh, EH16 4TJ, United Kingdom

Seasonally breeding mammals use photoperiod, encoded by rhythmical production of the nocturnal pineal hormone melatonin, as a critical cue to drive hormone rhythms and synchronise reproduction to the most optimal time of year. Melatonin acts directly on the pars tuberalis (PT) of the pituitary, regulating expression of thyrotropin, which then relays messages back to the hypothalamus to control reproductive circuits. To understand the molecular control by melatonin in the PT, we undertook detailed gene expression analysis in this target organ using Next Generation Trancriptome Sequencing (RNA-Seq). RNAs were analysed from sheep treated or not with melatonin pellets.

There were three issues surrounding our data analysis. (1) The choice of reference sequence. Since the sheep genome sequence is not completed, we used the phylogenetically close Bos Taurus genome as the reference for gene annotation of sheep RNA seq tags. (2) The annotation of the bovine genome is still conservative, likely missing exons/genes. To tackle this issue, we used data from both bovine and human databases to annotate our sheep data. (3) The RNA-Seq data across exon junctions is difficult to map to the reference genome. We therefore used the Abyss de-novo assembler to create transcriptome from RNA-Seq reads, which was used for exon/gene discovery and design of cDNAs.

RNA-Seq tags were processed in several steps: (1) filter tags using quality scores, (2) assign each to a region of the reference genome by parallel computing, (3) cross reference tags to Ensembl/Entrez gene annotations and (4) calculate counts of tags for each gene. Next we used the Bioconductor package edgeR to identify differentially expressed genes regulated by melatonin. Moreover, we used Tophat to map splice junction between exons for RNA-Seq reads, then used Cufflinks to calculate the relative abundance of isoforms.

The results identified many putative targets of melatonin regulation. To examine the regulation of these genes we will use chip-seq and specific promoter-reporter assays to define the role of transcription factors. This requires we isolate and analyse the sheep promoter sequences. We used the raw sheep genome sequences from ARK-Genomics to create local assemblies using bovine/human genes as model templates and assembled using the Mosaik assembler.

REVISITING CO-EVOLUTION THEORY OF THE GENETIC CODE FROM A WHOLE-GENOME PERSPECTIVE

Chi-Shing Yu[1], Kay-Yuen Yim[1], Wai-Kin Mat[2], Tze-Fei Wong[2], Ting-Fung Chan[1]

[1]The Chinese University of Hong Kong, Department of Biochemistry, Shatin, , Hong Kong, [2]Hong Kong University of Science and Technology, Department of Biochemistry, Clear Water Bay, , Hong Kong

Various examples of altered genetic codes were discovered in nature, such as those used in non-plant mitochondria and *Candida albicans*. This leaves us wonder how does the genetic code evolve and why the canonical genetic code is almost universal. By employing a "top-down" approach, a tryptophan auxotroph of *Bacillus subtilis* was previously evolved for the ability to incorporate unnatural amino acids including 4-fluorotryptophan (4fW), 5-fluorotryptophan (5fW) and 6-fluorotryptophan (6fW). Mutants with the ability to utilize the toxic amino-acid analog 4fW, 5fW and 6fW were isolated. Displacement mutants were selected such that 4fW can support indefinite growth, yet canonical tryptophan turns into an inhibitory analogue. Proteomic analysis revealed that the amino acid analog is indeed being assimilated rather than metabolized. This provides experimental support that genetic code is mutable as predicted by the Co-evolution theory, which postulates that the code was assigned in parallel to the evolution of amino-acid biosynthetic pathways.

In the ongoing project, we try to probe into the mechanism of codon reassignment through comparative analysis of the mutated Bacillus subtilis strains at different stages of screening in two dimensions: Genomics and Transcriptomics. Mutated strains at different stages of screening provide snapshots along the path to the altered genetic code. By combining high-throughput genome sequencing and transcriptome analysis using RNA-sequencing, we set out to construct a minimal gene set that is essential for changing a genetic code.

Preliminary genomics analysis shows that hundreds of mutations were found in different protein coding sequences. A pathway mutations explorer was developed to visualize the mutation level in all pathways as defined in the KEGG database. Various pathways with high mutation level under toxic amino-acids stress were identified. As no significant changes were found in codon usage frequency, and no mutations were found in both Trp-tRNA and tryptophanyl-tRNA synthetase, the key factors leading to altered genetic code are yet to be discovered.

Insights gained from this study should allow a better understanding of key factors in genetic code determination. By removing this apparent "barrier" to genetic codon alterations, non-standard amino acids could be incorporated into protein sequences with a much higher level of control, and opens up new research paths into the exciting field of synthetic biology.

HOW TO GENERATE AND PROCESS IN EXCESS OF 18 BILLION RNA-SEQ READS IN 2 MONTHS AND LIVE TO TELL ABOUT IT.

Carrie A Davis, Chris Zaleski, Sonali Jha, Alex Dobin, Felix Schlesinger, Wei Lin, Jorg Drenkow, Kimberly Bell, Huaien Wang, Lei-Hoon See, Megan Fastuca, Thomas Gingeras

Cold Spring Harbor Labs, Functional Genomics, Cold Spring Harbor, NY, 11979

ENCODE Consortium

The structural and functional complexity of eukaryotic genomes is still poorly understood in spite of several existing genome sequences. The ENCODE Consortia, a collection of labs specializing in diverse functional genomics assays was formed to help remedy this. Over time, this collection of data types, all generated in common genetic backgrounds permits a higher order view of genome structure and function.

The eukaryotic Transcriptome represents a highly dynamic and structurally dissimilar product of the inherited genome. In order to dissect the nonlinear relationship between RNA -> DNA we have focused on generating RNA-Seq libraries compatible with the Illumina GAIIx platform. Moreover, recent improvements in Illumina sequencing chemistry and software have dramatically increased the number, length and quality of reads and demanded concomitant development of computational resources to store, map and analyze the data.

To date, we have generated over 90 Poly-A(+) and Poly-A(-) RNA-Seq Pair-end 76 base datasets totaling 18 billion reads from ENCODE cell lines. 22 terabytes of data in 2 months. Here we present the process we have established to do this, an assessment of the computational resources required and quality control methods we are using to ensure production of high quality data at a rapid rate. Whenever possible we are aiming to incorporate these tools into Galaxy to bring these resources to a larger audience.

See also companion abstracts by Alex Dobin (STAR, split read mapper) and Felix Schlesinger (The use of RNA spike-ins).

HISTONE MODIFICATION PROFILE CLASSIFIES TISSUE/CELL-TYPE SPECIFIC GENES AND HOUSE KEEPING GENES

Zhihua Zhang[1,2], Michael Q Zhang[1,2]

[1]University of Texas at Dallas, Department of Molecular Cell Biology, Richardson, TX, 75080, [2]Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724

Gene expression is regulated both at the sequence level and through the modification of chromatin. Previous studies have addressed tissue/cell-type specific regulation by cis-regulatory modules, but the effect of chromatin on tissue/cell-type specific regulation is still unclear. Here, we elucidate the relationship between histone modification and tissue/cell-type specific regulation in human CD4+ T cell. A classifier was built up by histone modification information alone, and it successfully differentiated CD4+ T cell specific genes from housekeeping genes. H3K4me3, H4K20me1 and H2AK9ac were the most predictive marks for CpG-related genes, while H3K4me3, H3K27me3, H3K79me3, H3R2me1, and H3K23ac were the most predictive marks for non-CpG-related genes. The histone modification marks located in gene bodies show similar predictive power as which in promoter regions for CD4+ T cell specific regulation. We successfully applied our classifier to microRNA genes, which suggests that microRNA's tissue/cell-type regulation may share similar mechanisms to that of protein-coding genes.

UNSUPERVISED INFERENCE OF CHROMATIN DOMAIN STRUCTURE FROM MULTIPLE FUNCTIONAL GENOMICS DATA SETS

William S Noble

University of Washington, Genome Sciences, Seattle, WA, 98109

Sequence census methods such as ChIP-seq have begun to produce an unprecedented amount of genome-anchored data. Numerous techniques exist to analyze the results of single assays, but genomics research has lacked an integrative method to identify patterns from multiple experiments simultaneously, while taking full advantage of the high-resolution data now available. We have developed such a method, which employs a dynamic Bayesian network to discover joint patterns across different experiment types.

We have applied this method to ENCODE chromatin data for the human chronic myeloid leukemia cell line K562, including ChIP-seq data on covalent histone modifications and transcription factor binding, and DNaseI and FAIRE readouts of open chromatin. In an unsupervised fashion, we have identified patterns associated with transcription initiation and elongation, as well as transcriptional repression, finding the locations of genes without using DNA sequence, mRNA or multispecies conservation data. We have also successfully performed a semi-supervised identification of enhancer regions across the genome using a curated set of seed enhancers. In both the unsupervised and semi-supervised cases, the method identifies patterns that elucidate the relationships among open chromatin, various transcription factors and histone modifications.

THE NOTCHOME SEGREGATES BREAST CANCER CELL-LINES INTO DISCRETE SUBSETS

Srinka Ghosh, Lisa Choy Tomlinson, Thijs Hagenbeek, Zora Modrusan, Somasekar Seshagiri, Christian Siebel

Genentech Inc., Research, South San Francisco, CA, 94080

The Notch pathway (Notchome) is an evolutionarily conserved intercellular signaling mechanism that plays a critical role in cell-communication and cell fate. The key players of the pathway, in human, are a set of trans-membrane proteins: the notch receptors (N1-N4) associated with the signal-receiving cells and the delta-like and jagged ligands associated with the signal-sending cells. In the canonical model, activation of notch receptors by its ligand triggers a cascade of proteolytic cleavages releasing the notch intra-cellular domain (ICD). The ICD is trans-located to the nucleus where it acts in a transcriptional complex to induce a set of downstream targets. Since the cleavage is Gamma-secretase(GS) driven, modulation of the signaling pathway by GS inhibitors holds potential for a therapeutic approach. The primary challenge in this endeavor, however, is identifying the elements of the Notchome. Based on the literature and internal studies – the Notchome signature varies based on tissue-types; it has only been experimentally defined for the case of T-cell acute lymphoblastic leukemia. Using microarray and RNA-seq technologies we are currently experimentally defining the signature specific to breast cancer. We are also taking advantage of the gamma-secretase and antagonistic inhibitors to modulate the pathway.

The basic interaction-model, as described, comprise of a trio: the receptor, ligand and downstream transcriptional targets. The core of the notchome research constitutes identification of the trio in indication-specific tumorigeneis. We have performed gene expression studies in cell-lines (Affymetrix, HGU133P) spanning the breast cancer subtype spectrum. The analytical approach involved filtering of probe sets based on annotation. This has enabled estimation of transcript-specific differential expression for the core notchome genes (compiled from literature). Unsupervised clustering was performed using the probe-set (in contrast to summarized gene) expression metric; this facilitated a discriminant analysis across multiple isoforms. Given the changing face of the Notchome, it is imperative to perform a whole genome differential regulation analysis, in parallel to a core-Notchome one. The analyses have identified genes, which are key differentiators for breast basal and luminal subtypes. The cell-line studies have both recapitulated known targets as well as identified novel ones. Jag1 appears to be the dominant ligand differentiating the subtypes. The Notch3 receptor has also been implicated. The whole-genome data revealed statistically significant response from Delta and Notch-like epidermal growth factor-related receptor (DNER) gene.

COMPUTATIONAL ANALYSIS OF THE BINDING AFFINITIES OF 142 TRANSCRIPTION FACTORS OF CIONA INTESTINALIS AS DETERMINED BY HIGH-THROUGHPUT SELEX.

Edwin Jacox[1], Kazuhiro R Nitta[1], Renaud Vincentelli[2], Daniel Sobral[1], Agnès Mistral[2], Jussi Taipale[3], Yutaka Satou[4], Christian Cambillau[2], Patrick Lemaire[1]

[1]CNRS, IBMDL, Marseille, 13288, France, [2]CNRS, AFMB, Marseille, 13288, France, [3]Karolinska Inst, Biosciences and Nutrition, Huddinge, 57, Sweden, [4]Kyoto Univ, Zoology, Kyoto, 606-8502, Japan

Using high-throughput Systematic Evolution of Ligands by EXponential enrichment (SELEX), we have determined the *in vitro* binding affinities for 142 transcription factors of the ascidian *ciona intestinalis*, a close relative to the vertebrates. In the SELEX method, a tagged recombinant protein (the transcription factor) is incubated in solution with double-stranded oligonucleotides, comprised of two constant ends and a central portion of random bases (we used 18 or 20 bases), which represent putative binding sites. Bound oligonucleotides are pulled down, amplified by PCR, and then sequenced by high-throughput methods. This procedure is repeated for 5-7 rounds. The results of each experiment are hundreds to hundreds of thousands of 18- or 20-mers that likely bind the transcription factor. This extensive collection of DNA-binding specificities covers a third of the predicted *Ciona* transcription factors, the largest fraction for any species to date.

Position weight matrices (PWMs) are typically used to represent the binding affinities of transcription factors to DNA. The construction of PWMs using nucleotides counts can be inaccurate because it assumes an incorrect model of binding probabilities. In addition, PWMs cannot represent complex binding sites with interdependencies between positions. Here, we advocate a simpler method using the enrichment of hexamers or octamers in bound sequences, whose accuracy and improvement over traditional PWMs we confirmed using gel-shift and chromatin immunoprecipitation assays. This representation is also significantly simpler to implement since it does not rely on aligning the bound portion of the selected oligos.

Preliminary analyses suggest that most transcription factors have a simple binding mode, resulting in the rapid enrichment of a set of hexamers during SELEX. In contrast, some factors showed a richer binding specificity, with different hexamer sets representing either monomers and dimers or multiple binding modes.

Recent *in vitro* studies using high-throughput SELEX and protein-binding microarrays have described the DNA-binding specificities of hundreds of human and mouse proteins. A comparison of these binding specificities with *Ciona* orthologs using the enrichment of octamers serves as a validation of our results and the differences show the evolution of DNA-binding specificities within chordates.

BEYOND HEURISTICS: A GENERIC STATISTICAL TOOL FOR THE RIGOROUS ANALYSIS OF *-SEQ ASSAYS

Nathan P Boley, James B Brown, Peter J Bickel

University of California, Berkeley, Statistics, Berkeley, CA, 94720

Assays based upon next generation sequencing technologies (*-seq assays) are widely used in the genomics community. As these assays mature and attempt to probe more subtle biological phenomenon, new tools based upon powerful statistical techniques will be needed to provide confidence in the resulting biological conclusions. To date, *-seq assay analysis tools can be split into two distinct classes, mapping and quantification. Mapping tools attempt to match each read with a genomic location, whereas quantification tools infer biological features from the 'mapped' reads. The results of the mapping are often highly dependent on tuning parameters without ever providing a notion of confidence; the analysis tools typically take the provided mappings as gospel. Our approach is different. We make the known physical and biochemical properties of the assay and biological properties of the feature assayed an integral part of part of the mapping process and then on the basis of our assay model, set confidence limits on our mappings that can then be made an integral part of downstream analysis, analytical or biological. We aim to unify those aspects of *-seq assays essential for accurate mapping and downstream biological analysis (e.g. peak calling, transcript quantification) in a single software suite. We call our working prototype Statmap (available at encodestatistics.org). Here, I provide a brief overview of the Statmap software and its functionality in various use cases. In particular, I will discuss the use of Statmap to perform integrative analysis on CAGE and RNA-seq data to identify promoter elements, and to quantify signal from ChIP-seq data.

H3K4ME3 EPIGENOMES OF NORMAL AND DISEASED HUMAN PREFRONTAL NEURONS

Hennady P Shulha[1], Iris Cheung[2], Jie Wang[1], Schahram Akbarian[2], Zhiping Weng[1]

[1]University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA, 01605, [2]University of Massachusetts Medical School, Department of Psychiatry, Brudnick Neuropsychiatric Research Institute, Worcester, MA, 01604

Little is known about the regulation of neuronal and other cell-type specific epigenomes from the brain. We map the genome-wide distribution of trimethylated histone H3K4 (H3K4me3), a mark associated with transcriptional regulation, in neuronal and nonneuronal nuclei collected from prefrontal cortex (PFC) of dozens of normal individuals as well as patients with neurological and psychological disorders. Massively parallel sequencing identified H3K4me3 enriched regions (peaks) that are unique to different age groups as well as diseases, revealing developmentally dependent and disease impacted epigenomes in human prefrontal neurons.

HIGH RESOLUTION PEAK CALLING FOR CHROMATIN IP SEQUENCING

Xin Feng[1,2,3], Lincoln Stein[2,3]

[1]Stony Brook University, Biomedical Engineering, Stony Brook, NY, 11794, [2]Ontario Institute for Cancer Research, Informatics and Biocomputing Platform, Toronto,ON, M5G 0A3, Canada, [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724

Chromatin immunoprecipitation sequencing (ChIP-seq) provides an extraordinary window into the dynamics and regulation of chromatin. A key step in ChIP-seq is converting the signal graph of sequencing coverage into a series of discrete peaks that represent the regions bound by the antibody used in the immunoprecipitation. Although many high-quality peak-calling algorithms are available, we find that even the best of them have difficulty resolving the complex peak structures that arise from multiple close antibody binding sites. In this paper, we introduce a new algorithm that improves upon the current generation of peak callers, allowing researchers to explore complex peak structures within regions of interest.

The algorithm, which we call PeakRanger, works by identifying broad regions of enriched binding, and then using topological clues to identify punctate peaks within the bound regions. This allows it to distinguish peaks that are within a few hundred bases of each other.

We first tested the resolution of our algorithm; Benchmarking with a semi-synthetic dataset shows that our algorithm has almost double the resolution of two other major peak callers: PeakSeq and MACS. We then reanalyzed a previously-published H3K4Me3 histone mark profile, and showed that PeakRanger's results are a close match to the published results. We next analyzed a CTCF ChIP-seq dataset. While the vast majority of sites are well separated from each other, as described in the literature, we find a small number of "twin-peak" regions that do not seem to be artefactual.

The algorithm is packaged in a fast and user configurable software package called "PeakRanger", which is available from http://www.modencode.org/software/Ranger/. Several pre-compiled processing pipelines also ship with PeakRanger. They are capable of producing WIG and GFF files in one click from raw FASTQ files. The package will also be available as a cloud-computation-ready virtual image from Amazon EC.

INFERENCE AND VISUALIZATION OF NETWORKS OF CO-EXPRESSED GENES ACTIVE DURING HAEMATOPOIESIS.

Tobias J Sargeant, Carolyn A de Graaf, Tracey Baldwin, Douglas J Hilton

The Walter and Eliza Hall Institute of Medical Research, Molecular Medicine, Melbourne, 3052, Australia

Haematopoiesis is a complex process of differentiation and lineage commitment involving a plethora of cell types from haemopoietic stem cells to a wide variety of terminal cell types. Both steady state maintenance of this process and efficient reaction to injury and illness are critical for survival. We are interested in the molecular mechanisms underpinning this process.

In order to study the regulation of key genes and networks involved in this process, we have constructed an atlas of gene expression consisting of over 150 microarrays for 47 FACS sorted cell types, covering stem cells, progenitor cells and mature cells for the various cell lineages. For all but a few extremely rare cell populations, samples have been collected from a single inbred mouse strain, and hybridised in a small number of batches to a single array platform in order to minimise unwanted technical and biological variability. Having collected the raw data for this expression atlas, we wish to make it available to biologists in a comprehensive, intuitive and informative manner in order to allow them both to understand the biology of the unperturbed system, and to gain insight into other experiments exhibiting a haematopoietic phenotype.

We have applied minimum spanning trees to the problem of clustering and visualising gene and cell co-expression data. Cell types clustered in this manner more directly mirror biological knowledge regarding ordering of differentiation than do other clustering methods, correctly reconstructing the maturation steps of a number of lineages. Networks of genes constructed using minimum spanning trees have proven useful as a representational tool on top of which we have overlaid ontology, drug target and phenotypic information, as well as experimental data from mutant-vs-wildtype expression studies. We have provided access to these visualizations to biologists through a HTML5 canvas-based interface. Taking the union of almost-minimum spanning trees produces a network that has proven useful as s substrate for automatically identifying signalling pathways coordinately regulated during haematopoiesis.

We are using the data contained in this haematopoietic expression atlas, together with publicly available ChIP-seq and protein protein interaction data to attempt to understand the process of lineage commitment, to guide shRNA screens for key components of differentiation and activation, and to identify potential drug targets in cell types of interest.

INVESTIGATING GENOMIC AND EPIGENETIC SIGNATURES OF ACTIVE CENTROMERE SEQUENCES WITHIN A HUMAN DIPLOID GENOME

Karen E Hayden, Sayan Mukherjee, Nicolas Altemose, Huntington F Willard

Duke University, Institute for Genome Sciences and Policy, Durham, NC, 27708

Centromeres are essential for faithful segregation of chromosomes during mitosis and meiosis. Despite the critical functional role of centromeres, efforts to study their underlying genomic and epigenetic marks are challenged by the repetitive nature of satellite sequences, leaving these genomic regions incomplete and largely unexplored in complex genomes. To address these challenges we have developed a novel strategy to specifically target these underrepresented sites in the genome, providing a preliminary reference sequence database for each chromosomal gap region to investigate signatures of enrichment of known centromere proteins.

To construct a comprehensive database of centromeric sequences, we identified sequence reads containing 523Mb of alpha satellite sequences, 138Mb of interspersed repeats, and 12Mb of non-repetitive sequence(~8x coverage, estimated 3.3% of the genome) from a single male donor, HuRef. Clustering of 1.5M different satellite monomers from this dataset predicted 75 higher order repeats defined by highly homogeneous multimeric repeat units, each of which could be assigned to specific chromosomes through the use of an additional 344Mb alpha satellite from 15 flow-sorted chromosomes, experimental FISH/STS mapping, and paired read support. Currently, our strategy is able to describe sequence composition spanning from p-arm to q-arm for 11 of the 24 human chromosomes, with remaining chromosomes only linking to one or the other chromosome arm due to adjacent satellite families and/or pericentromeric segmental duplications. Ten previously unmapped scaffolds (3.04Mb) have been mapped in somatic cell hybrid panels confirming paired read predictions to individual centromeric gaps within the current assembly.

In order to identify the specific centromeric sequences underlying kinetochore assembly, ChIPSeq was performed, leading to identification of 10.9Mb of alpha satellite DNA associated with the centromere-specific histone H3-variant CENP-A, an epigenetic mark capable of promoting kinetochore assembly. Enrichment of CENP-A was demonstrated for 21/75 alpha satellite arrays, with only one active array per chromosome, as expected from previous findings.

Our data provide the first global assessment of sequences associated with active human centromeres within a single genome and establish a genomic and epigenetic foundation for future work to complete sequence assembly across centromere gaps in the human genome assembly and to evaluate functional or sequence variability between individuals or populations.

A MACHINE LEARNING FRAMEWORK FOR INTEGRATIVE ANALYSIS OF ENCODE DATA

Anshul Kundaje[1], Arend Sidow[2,3], Serafim Batzoglou[1]

[1]Stanford University, Computer Science, Stanford, CA, 94305, [2]Sanford University, Genetics, Stanford, CA, 94305, [3]Stanford University, Pathology, Stanford, CA, 94305

The ENCODE Project is generating genome-wide maps of transcription factor (TF) binding sites, DNA methylation and open chromatin sites, chromatin modifications and nucleosome positioning data in diverse cell lines. We present a supervised machine learning approach to integrate these diverse datasets and learn predictive models of in-vivo TF binding.

We formulate the learning problem as a binary classification task: to accurately predict the binding or non-binding of a TF at all genomic locations, using a large set of features. The sequence-based features are based on density and scores of known motifs, k-mer frequencies and sequence conservation scores. We use ChIP-seq binding profiles and binding motifs of other TFs to model the influence of potential cofactors. The learning algorithm also includes a de-novo motif discovery engine. Features extracted from the ENCODE functional genomics signal tracks account for variation in signal strength and the diversity of local signal shapes and patterns around (un)bound loci. Other context-specific features include DNA flexibility scores and proximity to genes and transcriptional activity. The training data consists of a positive set of bound loci obtained from ChIP-seq data, and a negative set of informative unbound locations sampled from the genome. Our core learning algorithm is based on an ensemble method known as Boosting. The algorithm learns a non-linear combination of a sparse set of predictive features. The final model is a margin-based version of decision trees that can be queried to obtain specific sets of features that are strongly associated with a single TF binding location or a collection of sites.

We learn binding models for a large collection of transcription factors including Myc, c-Fos, E2F, CTCF, Pol2, Nrsf and p300. We show that our models are highly predictive (auROC = 0.7 - 0.9) and that the inclusion of ENCODE track features contributes significantly to improving accuracy by > >20% over using sequence features alone. We find that promoter localized TFs tend to have a shared set of predictive functional signal profiles and the sequence-specific motifs account for TF specificity. We find that the signal-shape based features (a variety of local patterns of chromatin marks) are strongly associated with TF binding. For TFs such as CTCF that also bind distal to genes we decipher subpopulations of sites, each with its own unique set of functional patterns. We analyze the interaction of sequence conservation, presence or lack of consensus motifs and histone modifications to provide a richer, integrated view of TF binding. Our framework can also be for integrative analysis of other classes of genomic elements such as transcription start sites and splice sites.

DISTINCTIVE EVOLUTION OF CTCF-BINDING EVENTS AMONG MAMMALS

Petra C Schwalie*[1], Dominic Schmidt*[2,3], Michael D Wilson[2,3], Gordon D Brown[2], Benoit Ballester[1], Duncan T Odom[2,3], Paul Flicek[1,4]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom, [2]Cancer Research UK, Cambridge Research Institute, Cambridge, CB2 0RE, United Kingdom, [3]University of Cambridge, Department of Oncology, Cambridge , CB2 0XZ, United Kingdom, [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, United Kingdom

*authors contributed equally

The evolution of the global chromatin organizer CCCTC-binding factor (CTCF) on the level of DNA-binding has not previously been characterized. We experimentally determined the genome-wide occupancy of CTCF in livers of human, mouse, rat and dog. We find that CTCF-binding in sharp contrast to canonical transcription factors shows strong conservation, with over 20% of the bound regions being common among the four analyzed species. Closely related species, such as mouse and rat, share as much as half of the bound regions. For those sites that are lost in a given lineage, we identify the molecular mechanisms of that loss at the level of sequence substitutions and small indels.

A detailed analysis of the CTCF motif revealed characteristics of different classes of binding sites at sequence, conservation and functional level. The majority of the conserved CTCF binding events exhibit a single aligned motif that is on average closer to the consensus than the motifs found in species-unique binding events. Additionally, there seems to be a relationship between motif and binding strength, which is also related to the conservation of the bound region. Analysis of CTCF binding sets after CTCF knockdown shows that evolutionarily conserved sites are, in fact, more likely to be bound when the concentration of the protein is reduced.

We compared the tissue-matched CTCF sites from multiple species with identified CTCF binding sites in other tissues from the same species to assess which CTCF sites appear to be associated with tissue-specific as well as species-specific functions. Interestingly, over 30% of the species-specific binding sites in both rat and mouse are located in close proximity to rodent-specific B2 repeats, indicating an expansion of the CTCF sites in the rodent lineage associated with a retroposition mechanism.

Our study reveals the nature of CTCF binding evolution and that global chromatin structure as measured by CTCF binding is under stronger evolutionary selection compared to transcription factor binding events.

RETROTRANSPOSONS IN THE ORANGUTAN (*PONGO*) 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 Consortium[5]

[1]Louisiana State University, Biological Sciences, Baton Rouge, LA, 70803, [2]Louisiana State University, Computer Sciences, Center for Computation and Technology (CCT), Baton Rouge, LA, 70803, [3]Zoological Society of San Diego, Conservation and Research for Endangered Species (CRES), San Diego, CA, 92112, [4]Institute for Systems Biology, Systematics, Seattle, WA, 98103, [5]Washington University School of Medicine, The Genome Center , St. Louis, MO, 63108

Orangutans (*Pongo*) 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 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 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 (*Pongo pygmaeus*) and 19 Sumatran – using polymorphic retrotransposon markers. These elements were selected from the orangutan draft genome and also from Illumina paired-end reads from a Bornean orangutan. The orangutans of Bornean origin 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 distinction of Sumatran from Bornean orangutans supports the relatively recent notion that Bornean and Sumatran orangutans represent separate species.

MINING THE ALLOHEXAPLOID WHEAT GENOME FOR USEFUL SEQUENCE POLYMORPHISMS.

Rachel Brenchley[1], Rosalinda D'Amore[1], Gary Barker[2], Keith Edwards[2], Michael Bevan[3], Anthony Hall[1], Neil Hall[1]

[1]University of Liverpool, School of Biological Sciences, Liverpool, L69 7ZB, United Kingdom, [2]University of Bristol, School of Biological Sciences, Bristol, BS8 1UG, United Kingdom, [3]John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, United Kingdom

The 17GB allohexaploid genome of bread wheat (*Triticum aestivum*) is one of the largest and most complex plant genomes and contains predominantly repetitive sequence (an estimated 80-90%).

Bread wheat is a major source of nutrition for humans and domesticated animals. It is also of fundamental importance to world agriculture, with ~550million tonnes harvested in 2007. Current productivity levels cannot meet an ever-increasing demand and it has been estimated that in Europe, wheat production must double to keep pace with demand and maintain prices. This project aims to accelerate the process of identifying genetic markers for key phenotypic traits in wheat and will provide valuable data for wheat breeders in the U.K.

We have sequenced the wheat genome (var. Chinese spring) to a depth of 5X using whole genome shotgun sequencing on the 454 GS-FLX Titanium platform. To produce a gene-rich reference sequence and identify homeologous (intra-varietal) SNPs, we developed a bioinformatics pipeline to extract low copy number and genic regions from the 5X data. The pipeline compares cDNA and genome data from related species with the wheat genome shotgun sequences in order to extract putative genes. We have also characterised the repetitive elements using comparative and *de novo* strategies.

In addition, we are also sequencing 4 U.K wheat varieties using the SOLiD 4 platform. These will be mapped to the Chinese spring reference to locate inter-varietal SNPs. Here we present our initial assembly, analysis of the genic content of wheat and preliminary SNP data.

MAPPING THE EVOLUTION OF TRANSCRIPTION FACTOR BINDING

Dominic Schmidt[1], Michael Wilson[1], Benoit Ballester[2], Petra C Schwalie[2], David Thybert[2], Klara Stefflova[1], Michelle Ward[1], Duncan Odom[1], Paul Flicek[2]

[1]Cambridge Research Institute, Cambridge, CB2 0RE, United Kingdom, [2]European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom

Although the genomic locations of most transcription factor binding events change rapidly over evolutionary time due to sequence turnover in noncoding regions of the genome, gene expression and overall tissue function is apparently largely unchanged. We have explored this apparent paradox and the evolutionary dynamics of transcription factor binding through a series of experiments mapping the binding of conserved transcription factors in homologous tissues in multiple vertebrate species. Our experiments use ChIP-sequencing with highly conserved factors and liver tissue from each of the species considered.

We find that the majority of transcription factor binding is species specific and that fraction of sites that are consistently occupied over evolutionary time is extremely small for most factors. We observe that significant binding in repetitive regions of the genome is common for all factors and confirm that DNA binding preferences for each tested factor are highly conserved. In many cases, are able to refine the known binding motif. We compare binding locations using both whole genome and pairwise alignments. The inclusion of species at various evolutionary distances suggest a predictable decay of binding over evolutionary time driven both by apparently random changes and retention of important functional binding regions. However, apparently functional binding regions do not exhibit higher levels of sequence constraint than the set of all binding events. Direct comparisons of aligned sequence motifs allow us to dissect the contribution of sequence substitutions and small indels to the loss and gain of binding events and provide a view into the sequence evolution driving the locations of transcription factor binding.

ANALYSIS OF THE PRIMATE GENOMES USING THE 5-WAY EPO MULTIPLE ALIGNMENTS

Kathryn Beal*, Stephen Fitzgerald*, Paul Flicek, Javier Herrero

EBI, Vertebrate Genomics, Hinxton, CB10 1SD, United Kingdom

We have aligned 5 primate genomes: human, chimpanzee, gorilla, orang and macaque using our EPO (Enredo-Pecan-Ortheus) pipeline. Enredo [1] builds a graph to detect segments of collinear sequences. Importantly, the graph allows for any number of duplications in any species. We align these segments using Pecan [1], a consistency-based multiple aligner for genomic sequences. Pecan has recently been shown to be one of the best aligner of its class [2,3]. Ortheus [4] then uses Pecan alignments to infer the ancestral states of every base in the extant species.

Primate EPO alignments cover 89%, 83%, 90%, 79% and 85% of the human, chimpanzee, gorilla, orang and macaque genomes, respectively. Relative assembly quality can be estimated by looking at specific patterns of duplications or deletions. For instance, the human genome stands out in the span of species-specific duplications that can be resolved: these cover 40.1Mb of the human genome while they only account for up to 14.8 Mb (gorilla) in the other primates. Looking at species-specific deletions, we find that the orang is missing ca. 70Mb of sequence, approximately 100 times more than the human sequence. Incomplete areas of the orang assembly will account for such an excess of deletions.

Many duplications in the human-chimp ancestor have been retained to date. 13.6 MB of sequence is in duplicated regions between human and chimp only, while as little as 2.0Mb is duplicated in all hominine or 1.8Mb of sequence appear specifically duplicated in human and gorilla.

Ortheus predicts ancestral states, which can be used to call ancestral alleles for SNPs. We classify the ancestral predictions by comparing the ancestral (a), the sister (s) and the closest ancestor of the ancestral (p) sequences. High-confidence calls will correspond to cases where all three states agree, low-confidence calls when the ancestral state (a) agrees with one of the other two (s or p) only. We filter cases where neither (s) nor (p) agree with (a). We have analysed the ancestral predictions for both the human and the gorilla genome. As expected, we find the same amount of high-confidence (ca. 80%), low-confidence (ca. 4%) and unreliable ones (< 0.5%) in both cases.

In summary, we use Enredo to gain insight in the patterns of duplications in the primate genomes, Pecan to obtain high-quality global alignments and Ortheus to infer genome-wide ancestral states.

1. Paten B et al. Genome Res. 2008 Nov;18(11):1814-28.
2. Kim J, Sinha S. BMC Bioinformatics 2010, 11:54
3. Chen, Tompa. Nat Biotech 2010, 28:567-572
4. Paten B, Herrero J,et al. Genome Res. 2008 Nov;18(11):1829-43.

SNOWMAN: HIGH THROUGHPUT PHYLOTYPING, ANALYSIS AND COMPARISON OF MICROBIAL COMMUNITIES

Gernot Stocker*[1,2], Renè Snajder*[2], Johannes Rainer[3,4], Slave Trajanoski[5], Gregor Gorkiewizc[6], Zlatko Trajanoski[1], Gerhard Thallinger°[2]

[1]Innsbruck Medical University, Division for Bioinformatics, Innsbruck, 6020, Austria, [2]Graz University of Technology, Institute for Genomics and Bioinformatics, Graz, 8010, Austria, [3]Innsbruck Medical University, Division for Molecular Pathophysiology, Innsbruck, 6020, Austria, [4]Tyrolian Cancer Research Institute, ., Innsbruck, 6020, Austria, [5]Medical University of Graz., Institute of Pathology, Graz, 8036, Austria, [6]Medical University of Graz, Center for Medical Research, Graz, 8010, Austria

The number and scope of microbial community studies from diverse environments have increased exponentially with the recent introduction of next generation sequencing technologies. To cope with this increased data volume and to provide an easy to use analysis platform, we have developed SnoWMAn (**http://snowman.genome.tugraz.at**), a powerful web server for analysis of amplicon sequence data generated by microbiome studies. It integrates the complete analysis workflow covering sample splitting, sequence filtering and alignment, clustering, taxonomic classification, diversity estimation, sample comparison and visualization of the results. SnoWMAn is easily extendable to new marker genes and allows reproduction of results of previous analysis runs. To accommodate diverse analysis approaches, currently five different analysis pipelines are available, which allow either sequence independent or taxonomy independent phylotyping. Two of them are novel, in-house developed pipelines (based on BLAT and UCLUST[1]) and three of them are well established pipelines (JGast[2], Mothur[3] and RDP[4]) which are seamless integrated in our web interface.
Diversity of microbial populations is estimated by rarefaction analysis and microbiomes of different samples can be compared by principal component analysis, Venn diagrams and charts with the sequence distribution on taxonomic classification. The analysis process consists of three simple steps: (i) data upload, (ii) pipeline selection with parameter definition, and (iii) visualization of the results. Result files and charts can be downloaded for further analysis. Confidentiality of the input data and results can be secured by creating an account using the integrated authentication system.

[1] Edgar RC: UCLUST [http://www.drive5.com/usearch/].
[2] Schloss PD, et al; Appl. Environ. Microbiol 2009, 75:7537-7541.
[3] Cole JR, et al; Nucleic Acids Research 2009, 37:D141-D145.
[4] Hamp TJ, et al; Appl. Environ. Microbiol 2009, 75:3263-3270.

COMPARATIVE GENOMICS OF ALL 24 DESCRIBED SPECIES WITHIN THE GENUS *CAMPYLOBACTER*

William G Miller, Craig T Parker, Emma Yee, Robert E Mandrell

USDA, Agricultural Research Service, Produce Safety & Microbiology, Albany, CA, 94710

The 31 taxa that comprise the genus *Campylobacter* represent organisms that colonize a wide variety of hosts and occupy multiple environmental niches and habitats. *Campylobacter jejuni* e.g. has been isolated from a substantial number of different warm-blooded animals and birds; some campylobacters have been isolated also from reptiles. Campylobacters have been implicated also in disease in both livestock and humans, primarily causing enteritis or septic abortion in the former and gastroenteritis in the latter. Several *Campylobacter* species, predominantly *C. jejuni*, have been isolated from food, milk and water; thus, several campylobacters are considered food-borne pathogens. The genomes of 12 *Campylobacter* taxa have been previously sequenced to draft level or completion by JCVI, in collaboration with the USDA. To address the genotypic diversity that underlies the variation within *Campylobacter*, we have sequenced to draft level the genomes of 18 additional *Campylobacter* strains, thereby acquiring genomic data on all published taxa within the genus, including both subspecies of *C. lari* and *C. hyointestinalis* and all three biovars of *C. sputorum*. We also propose the existence of a new, second, clinically-relevant subspecies of *C. upsaliensis*. The average size and G+C% of each genome within the genus is 1.78 mb (range 1.46 mb – 2.51 mb) and 33.1% (range: 27.3% - 46%), respectively. BLASTP analysis of the predicted *Campylobacter* proteomes indicates a core *Campylobacter* gene set of approx. 400 genes. As expected, the majority of these core genes are involved in basic biological functions, such as replication, transcription and translation; however, several core proteins have only a general function or have no defined function. BLASTP analysis also indicates that, although the average genome contains ~1700 genes, the minimum Campylobacter gene pool is approx. 12,000 genes. This large gene pool underscores further the diversity within the genus; many of these variable genes are involved in signal and energy transduction, metabolism, transport and biosynthesis of surface structures. Analysis of the *Campylobacter* genomes will provide further insights into evolution, host and environmental adaptation and pathogenicity, and will be used in the development of improved typing and detection methods.

STATISTICAL CHALLENGES IN RNA-SEQ

A Roberts, C Trapnell, L Pachter

University of California-Berkeley, ,, Berkeley, CA, 94708

In Trapnell et al. (2010) we described the Cufflinks suite of tools for the assembly of RNA-Seq reads and quantification of transcript abundances. Here we report on improvements of our initial methods, specifically in addressing the statistical challenges in RNA-Seq analysis. We report on a novel method for leveraging replicates in differential expression analysis and on the estimation of positional and sequence-specific bias parameters together with expression estimates. We will also discuss other improvements to the Cufflinks software that allow us to address the entire range of problems associated with RNA-Seq analysis.

DE NOVO RNA-SEQ-BASED GENOME ANNOTATION

Jonas Behr[1], Georg Zeller[2], Gabriele Schweikert[1,3,4], Lisa Hartmann[1], Lisa Smith[3], Gunnar Rätsch[1]

[1]Max Planck Society, Friedrich Miescher Laboratory, Tübingen, 72076, Germany, [2]EMBL, Bork Lab, Heidelberg, 69117, Germany, [3]Max Planck Institute, Developmental Biology, Tübingen, 72076, Germany, [4]Max Planck Institute, Biological Cybernetics, Tübingen, 72076, Germany

We have developed a novel system for accurate *de novo* genome annotation based on RNA-seq experiments, which does not require, but benefits from, an existing genome annotation. First we construct a preliminary gene set for highly expressed genes that are well-covered with RNA-seq reads. In a second step, we train predictors for genomic signals on the preliminary gene set. In the third step we train transcript predictors, employing the preliminary gene models while taking advantage of the RNA-seq read coverage and genomic signal predictions. Here, we considered an extension of the gene finding system *mGene* [1], which assumes coding transcripts, and a novel unbiased prediction method that makes very few assumptions. Both methods predict alternative transcripts for which the abundance is estimated using *rQuant* [2].
We tested the proposed system for the *C. elegans* genome using strand-specific paired-end reads (Illumina; 76nt). The *ab initio* mGene-based system trained on the WS199 annotation achieves an average transcript-level F-score of 47% (50% on top half expressed genes). We achieve a similar performance (44%; 57% on top half) when we train the system only the preliminary gene set (i.e., not at all using the existing genome annotation). If we use the RNA-seq reads and train on the existing annotation, we achieve 54% (58% on top half), and can therefore take advantage of the previous annotation. The unbiased transcript prediction method achieved a lower accuracy (21%; 36% on top half). It compares, however, favorably to *Cufflinks* [3] that achieved 11% (19% on top half). The proposed system annotates new genomes completely *de novo*. It also has benefits for the improvement of annotations of well-analyzed genomes, for instance, for *C. elegans* the CDS length of WS199 transcripts is bimodal with a second peak of 1knt which very likely is an annotation artifact replicated by many gene finders [1]. RNA-seq-based de novo genome annotation avoid such long-term annotation biases, benefit from genomic signal predictions and very accurately predict genome annotations.
We currently work on comparisons with *Scripture* and on the mouse and other nematode genomes. Large parts of the systems are already available in the Galaxy instance available at http://galaxy.fml.mpg.de.

[1] G. Schweikert et al. Genome Research, 19:2133-2143, 2010.
[2] R. Bohnert and G. Rätsch. NAR, June 2010.
[3] C. Trapnell et al. Nature Biotech, 28:511-515, 2010.

RGASP: RNASEQ GENOME ANNOTATION ASSESSMENT PROJECT

J Harrow[1], F Kokocinski[1], J Abril[2], T Steijger[1], G Williams[1], A Mortazavi[3], M Gerstein[5], A Reymond[6], T Gingeras[7], B Wold[3], R Guigo[4], T Hubbard[1]

[1]Wellcome Trust Sanger Institute, Informatics, Hinxton, CB10 1HH, United Kingdom, [2]Universitat de Barcelona, Genetics , Barcelona, 08028, Spain, [3]California Institute of Technology , Biology, Pasadena, CA, 91125, [4]CRG, informatics, Barcelona, 08003 , Spain, [5]Yale University, Informatics , New Haven, CT, 06520, [6]University of Lausanne, Genetics, Lausanne , 1015, Switzerland, [7]Cold Spring Harbor Laboratory, Functional genomics , New York, NY, 11797

RNASeq data is revolutionizing eukaryotic transcriptomics, highlighting the extent different loci are expressed and alternatively spliced. Following the successful format of the EGASP workshop in 2005 (Guigo et al., 2006), the RNASeq Genome Annotation Assessment Project (RGASP) was launched to benchmark the current progress of automatic gene building programs using RNASeq as its primary dataset. The goals of this community effort are to assess the success of computational methods to correctly map RNASeq data onto the genome, assemble transcripts and quantify their abundance. The analysis of RNAseq data from three different organisms (Human, Drosophila and C.elegans) were compared, since high quality genome annotation was available for each but genome size and transcript diversity differed between organism.

Eighteen groups submitted transcript predictions which were evaluated against a filtered “expressed “ annotation datasets of the three organisms. Quantification predictions on transcript level have been analyzed on a subset of 100 loci selected for each organism using at least two alternative transcripts per locus. They are compared against Nanostring experimental results to avoid any amplification bias introduced by RT-PCR. In addition some submitters were able to supply BAM alignment files so the read alignments against the genome and annotation could be analyzed in parallel. Approximately 200 novel transcript predictions not in the worm and human reference annotations were targeted for experimental verification using RT-PCR. In general, fly and worm predictions were better matching the annotation than those for human datasets, mainly reflecting the transcriptional complexity. The best methods could predict at least one transcript correctly within 70% of highly expressed worm genes however this decreased to 55% of human genes. The multiple transcript accuracy within highly expressed genes was reduced to 60% worm and 34% in human. Low expressed genes are not as well predicted as highly expressed ones and the accuracy is greatly reduced. The spike-ins showed that the relative quantification is good between the methods although the absolute values vary significantly.

RNA-SEQ UNCOVERS THE INFLUENCE OF COPY NUMBER VARIANTS ON TRANSCRIPTOME DIVERSITY

Emilie Ait Yahya Graison, Alexandre Reymond

University of Lausanne, Center for Integrative Genomics, Lausanne, 1015, Switzerland

Copy number variation (CNV) of DNA segments has been identified as a major source of genetic diversity, but a comprehensive understanding of the phenotypic effect of these structural variations is only beginning to emerge. Our group and others established extensive maps of CNVs in wild mice and inbred strains. These variable regions cover ~11% of their autosomal genome. CNVs are suggested to shape tissue transcriptomes on a global scale and thus represent a substantial source for within-species phenotypic variation. The recently emerged RNA-seq method has brought transcriptome analysis to a new level, because it addresses both gene expression at nucleotide resolution level and alternative splicing events simultaneously. We used these advantages to unravel the effects of CNVs on expression at the nucleotide rather than locus level. We generated by ultra high-throughput sequencing on Illumina Genome Analyzer >450 millions RNA-seq reads from brain and liver of three mouse inbred strains (129S2, DBA/2J (D2) and C57BL/6J (BL6)) to monitor expression changes of transcripts that map within and outside genomic regions that vary in copy numbers. We used TopHat v1.0.13 and Cufflinks v0.8.2 to map, assemble and estimate the abundance of the assembled isoforms, respectively. When compared to the BL6 reference sample, 13,656 isoforms (representing 9,445 genes) and 11,754 splicing variants (9,034 genes) are differentially expressed in 129S2 and D2 brain, respectively. Among these, we observed a significant enrichment for CNV genes (p-value=2.55e-06 for 129S2, p-value=9.405e-09 for D2) meaning that alternative transcripts that derive from genes varying in copy numbers show significantly more differential expression between strains. This confirms previous observations made on expression arrays but at exon-level, a resolution never achieved before. We also tested difference in splicing events between strains and showed that 371 and 320 genes exhibit significant differential splicing between isoforms in 129S2 and D2 brain samples, respectively, compared to the BL6 reference. However, these significant differences in the distribution between isoforms show no enrichment in CNV genes.
This study provides a unique opportunity to extensively gauge the influence of CNVs on the transcriptome complexity and regulation.

INTEGRATIVE ANALYSIS OF CHIP-SEQ AND RNA-SEQ ENCODE TIER 1 AND TIER 2 DATA USING SELF-ORGANIZING MAPS

Ali Mortazavi[1], Shirley Pepke[1], Georgi Marinov[1], Richard M Myers[2], Barbara Wold[1]

[1]California Institute of Technology, Division of Biology, Pasadena, CA, 91106, [2]HudsonAlpha Institute for Biotechnology, Myers Group, Huntsville, AL, 35806

Transcription is the primary output of gene regulatory networks. In these networks, RNA polymerase and its cofactors integrate a variety of disparate inputs from site-specific and general transcription factors that are bound at enhancers and promoters. ENCODE uses ChIP-seq for multiple factors, cofactors and chromatin marks, plus RNA-seq to measure and define the inputs and outputs of these physical networks for diverse cell types and cell states. When we have assembled these diverse data, integrative analysis of the resulting high-dimensional data matrix becomes limiting for extracting relationships among data types and building network models. Self-organizing maps (SOMs) are an unsupervised machine learning-method to cluster and to visualize high-dimensional data in a two dimensional map. A useful property of SOMs for modeling network relationships is that additional datasets can be mapped onto a trained SOM with ease to identify further relationships. We are using large, fine-grained self-organizing maps constructed from ENCODE Tier 1 and Tier 2 ChIP-seq and RNA-seq datasets to cluster (1) promoters and (2) the genome into thousands of coherent units and then identifying units representing co-regulated regions within one or more cell types. Mining of SOM units and clusters of units, when combined with perturbation experiments, suggests a path forward for probing genome-scale network structure and function.

INTEGRATED ANALYSIS OF MULTIPLE NEXT GENERATION SEQUENCING DATASETS WITH APPLICATION TO GENE FUSION DISCOVERY

Andrew W McPherson[1,3], Fereydoun Hormozdiari[3], Chunxiao Wu[2], Iman Hajirasouliha[3], Faraz Hach[3], Deniz Yorukoglu[3], Anna Lapuk[2], Stas Volik[2], Sohrab Shah[1], David Huntsman[1], Colin Collins [2], Cenk Sahinalp[3]

[1]BC Cancer Agency, Centre for Translational and Applied Genomics, Vancouver, V5Z 4E6, Canada, [2]Vancouver Prostate Centre, Prostate Research Facility, Vancouver, V5Z 1M9, Canada, [3]Simon Fraser University, Computer Science, Burnaby, V5A 1S6, Canada

Background: Second generation RNA and genome sequence data each provide an alternate viewpoint into the rearrangement component of a cancer genome. However, there is currently no computational method that simultaneously leverages both types of data for the accurate prediction of expressed genomic rearrangements.

Methods: We propose an algorithmic framework for the integrated analysis of multiple sequencing datasets and use that framework to do a combined analysis of genome and transcriptome sequence data from the same prostate cell line. We consider all mappings of paired end RNA-Seq reads to cDNA and unspliced gene sequences, and all mappings of paired end genome sequence reads to unspliced gene sequences. We would like to find an assignment of multimapped reads to specific alignment locations that minimizes a linear combination of weighted structural variation differences between each sequencing dataset. The weights considered may be motivated by global information about the evolutionary distance between each set of sequences or local information about the likelihood of each specific structural variation. We show that this problem is NP-complete, and provide a heuristic solution.

Results: We have performed an integrated analysis of the RNA-Seq and genome sequencing datasets from the LNCap derivative cell line C42 in an attempt to discover novel gene fusions caused by genomic structural variation. We show that using our method we are able to rediscover 4 gene fusions previously reported in LNCap. We also predict 2 novel fusions and validate these fusions using RT-PCR of cDNA and genomic DNA. Further, we identify as potential read-through events 6 gene fusions between adjacent genes with no genomic support.

Conclusions: We show that an integrated analysis of RNA-Seq and genome sequencing datasets results in more accurate prediction of gene fusions produced by genomic rearrangements. The framework we propose represents a general approach that may be extended to the analysis of structural variation in multiple related sequencing datasets. The proposed framework has many applications, including combined analysis of transcriptome and genome tumour data, primary and metastatic tumour genome data, and genome data from related individuals.

PROFILING THE TRANSCRIPTOME OF HUMAN BRAIN REGIONS USING HIGH-THROUGHPUT CAPPED ANALYSIS OF GENE EXPRESSION (CAGE) SEQUENCE ANALYSIS

Luba M Pardo[1], Patrizia Rizzu[1], Margherita Francescatto[1], Takahashi Hazuki[2], Morana Vitezic[2], Nicolas Bertin[2], Carsten Daub[2], Piero Carninci[2], Peter Heutink[1]

[1]Medical Genomics, Clinical Genetics, Amsterdam, 1081 BT, Netherlands, [2]Riken Omics Science Center, Omics Science Center, Yokohama, 230-0045 Kanagawa, Japan

We are analyzing the transcriptome of post-mortem tissue from different aged human brain regions using CAGE to identify Transcription Start Sites (TSS) and their promoter regions. We prepared 25 CAGE libraries from total RNA isolated from 5 brain regions (caudate nucleus, frontal lobe, hippocampus, putamen and temporal lobe) from 5 subjects who died from non-neurological conditions. Mapping, expression normalization and clustering of the tags were carried out using automated pipelines (OSC, RIKEN). More than 14 million CAGE tags were mapped to unique positions in the human genome. We estimated differences in expression between regions using both global and pairwise tests based on uniquely mapped CAGE tags (TSS) present in at least 3 libraries. Over 76% and 47% of TSS originated from known RefSeq transcripts and from the promoter regions (-300 to +100 bp) around RefSeq TSS, respectively. Many of the 'intergenic' TSS were represented by high tag counts, which suggests that these are not the result of background transcription. We divided the TSS into 3 groups according to their level of expression: highly expressed (HE; top 25%), moderately expressed (ME, middle 50%) and lowly expressed (LE; bottom 25%). Most (93%) HE TSS were derived from known genes. Functional annotation analysis showed that HE genes were significantly overrepresented in categories such as metabolic processes, were located more often in mitochondria and were more often transcription factors. Unlike HE TSS, only 70% and 24% of ME and LE TSS, respectively were derived from known genes. We found that 21% of all TSS were differentially expressed (DE) across brain regions (1% FDR). The hippocampus accounted for most differentially expressed TSS and had the largest group of region specific TSS. Examples of these were TSS mapping to the RGL1, SOX5, and ITM2B genes. The most dissimilar regions were the hippocampus and caudate nucleus. DE genes between caudate and hippocampus included calcium channel genes (e.g. CACNA2D1, calmodulins) and transcription factors (e.g. TCF4 and CAMTA1). 19% of DE hippocampal TSS, mapped to intergenic regions. Our results shows that genes involved in metabolic processes are highly expressed throughout all brain regions. 19% of hippocampal TSS may represent novel promoters. ITM2B a gene involved in dementia was one of the most DE genes in hippocampus. This suggests that ITM2B is an interesting target for expression studies of aging and cognitive decline.

EVALUATION OF METHODS FOR FULL-LENGTH TRANSCRIPT RECONSTRUCTION FROM RNA-SEQ

Qiandong Zeng[1], Brian Haas[1], Moran Yassour[1,2], Manfred Grabherr[1], Nick Rhind[3], Chad Nusbaum[1], Aviv Regev[1,4,5]

[1]Broad Institute, Genome Sequencing and Analysis Program, Cambridge, MA, 02142, [2]Hebrew University, School of Engineering and Computer Science, Jerusalem, 91904, Israel, [3]University of Massachusetts, Medical School, Worcester, MA, 01605, [4]M.I.T., Department of Biology, Cambridge, MA, 02139, [5]HHMI, HHMI, Cambridge, MA, 02139

High throughput transcriptome sequencing using next generation sequencing technologies (termed RNA-Seq) has the potential to revolutionize the annotation of eukaryotic gene structure annotation. To fully realize this potential requires a method that takes tens of millions of short (30-75 base) reads, resolves all introns and exons, and reconstructs them into all full-length transcripts at single nucleotide resolution. Several tools that tackle this challenge are under development including ABySS, Cufflinks, Scripture, and Oases. Broadly, these tools follow two distinct strategies: (1) *de novo* assembly followed by spliced transcript alignment ("assemble, then align", e.g. ABySS, Oases), and (2) short read spliced alignment to a reference genome, followed by transcript reconstruction via alignment assembly ("align, then assemble", e.g. Cufflinks, Scripture). Each of the methods has important potential, but their relative performance in genome annotation has not been compared to date.

We have developed a systematic approach to evaluate the theoretical potential of these distinct methods and the actual performance of algorithms implementing them. We used the reference genome of *Schizosaccharomyces pombe* and strand-specific RNA-Seq data. We tested three methods: Cufflinks (mapping first), ABySS (assembly first), and a novel method that couples both aspects, by multiply iterating between alignment and *de novo* assembly. Specifically, in the integrative method, we (1) align the reads to the genome using TopHat, (2) identify disjoint regions of strand-specific sequence coverage, (3) collect reads for individual regions, (4) use a novel *de novo* assembly method that we call Inchworm Recursive Kmer Extender (IRKE) to assemble these reads independently for each region, (5) align the assembled sequences to the genome using BLAT, and (6) assemble the genome alignments into transcript structures with PASA.

We have devised a new *de novo* annotation method, IRKE, that combines the advantages of the previously reported methods, advancing the utility of RNA-Seq for eukaryotic genome annotation. In addition, our method leverages parallel data processing for the typically memory intensive *de novo* assembly steps, resulting in a minimal memory footprint precluding the need for specialized hardware.

Participant List

Dr. Emilie Ait Yahya Graison
University of Lausanne
emilie.aityahyagraison@unil.ch

Dr. Subramanian Ajay
National Institutes of Health
ajayss@mail.nih.gov

Ms. Sonja Althammer
Computational Genomics, Universitat Pompeu Fabra
sonja.althammer@upf.edu

Mr. Raymond Auerbach
Yale University
raymond.auerbach@yale.edu

Dr. Senduran Balasubramaniam
Wellcome Trust Sanger Institute
sb10@sanger.ac.uk

Ms. Anett Balla
Astrid Research Inc.
SILVERCENTRUM@OTPTRAVEL.HU

Ms. Karina Banasik
Hagedorn Research Institute
kabs@hagedorn.dk

Dr. Joachim Baran
University of Manchester
joachim.baran@manchester.ac.uk

Mr. Serge Batalov
GNF
sbatalov@gnf.org

Dr. Alex Bateman
The Sanger Centre
agb@sanger.ac.uk

Dr. Mark Batzer
Louisiana State University
mbatzer@lsu.edu

Prof. Serafim Batzoglou
Stanford University
serafim@cs.stanford.edu

Ms. Jennifer Becq
Illumina Cambridge Ltd
yhandy@illumina.com

Dr. Amir Ben-Dor
Agilent LAboratories
amir_ben-dor@agilent.com

Dr. Eva Berglund
Uppsala University
eva.berglund@medsci.uu.se

Dr. Inanc Birol
BC Cancer Agency
ibirol@bcgsc.ca

Prof. Hidemasa Bono
Research Organization of Information and Systems
bono@dbcls.rois.ac.jp

Dr. Gerard Bouffard
NIH/NHGRI
bouffard@mail.nih.gov

Mr. Tyler Bray
The University of Chicago
tyler.s.bray@gmail.com

Dr. Rachel Brenchley
The University of Liverpool
parsonsl@liv.ac.uk

Dr. Stephen Bridgett
Edinburgh University
stephen.bridgett@ed.ac.uk

Dr. James Brown
UC Berkeley
benbrownofberkeley@gmail.com

Dr. Michael Brudno
University of Toronto
brudno@cs.toronto.edu

Dr. Remy Bruggmann
ETH Zurich
remy.bruggmann@fgcz.ethz.ch

Dr. David Burt
Roslin Institute & R(D)SVS University of Edinburgh
dave.burt@roslin.ed.ac.uk

Dr. Mario Caccamo
The Genome Analysis Centre
mario.caccamo@bbsrc.ac.uk

Dr. Scott Cain
Ontario Institute for Cancer Research
scott@scottcain.net

Dr. Guillermo Carbajosa
BICMS, Queen Mary University
g.carbajosa@qmul.ac.uk

Dr. Tim Carver
Wellcome Trust Sanger Institute
tjc@sanger.ac.uk

Dr. Tanita Casci
Nature Reviews Genetics
t.casci@nature.com

Mr. Timothee Cezard
The University of Edinburgh
tcezard@staffmail.ed.ac.uk

Ms. Emily Chambers
Human Genetics Unit
emily.chambers@hgu.mrc.ac.uk

Ms. Suhua Chang
Institute of Psychology, CAS
changsh@psych.ac.cn

Dr. Jitender Cheema
John Innes Centre, Norwich, NR4 7UH, UK
Jitender.cheema@bbsrc.ac.uk

Mr. Kevin Cheeseman
INRA - Genomic Vision
k.cheeseman@genomicvision.com

Prof. Kevin Chen
Rutgers University
kcchen@biology.rutgers.edu

Ms. Yuan Chen
Wellcome Trust Sanger Institute
yuan@sanger.ac.uk

Dr. Feng Chen
Washington University School of Medicine
fchen@dom.wustl.edu

Dr. Lei Chen
Washington University School of Medicine
lchen@watson.wustl.edu

Dr. Jer-ming Chia
Cold Spring Harbor Laboratory
chia@cshl.edu

Dr. Frederic Choulet
INRA
frederic.choulet@clermont.inra.fr

Mr. Aaron Chuah
Cold Spring Harbor Laboratory
achuah@cshl.edu

Dr. Michele Clamp
Bioteam, Inc.
michele@bioteam.net

Dr. Stefano Colella
INRA-Institut National de la Recherche Agronomique
stefano.colella@lyon.inra.fr

Dr. Rafael Contreras-Galindo
University of Michigan
rafaelc@umich.edu

Dr. Miklos Cserzo
Semmelweis University
miklos.cserzo@eok.sote.hu

Dr. Craig Cummings
Life Technologies
craig.cummings@lifetech.com

Dr. Petr Danecek
Wellcome Trust Sanger Institute
pd3@sanger.ac.uk

Dr. Nishadi De Silva
Wellcome Trust Sanger Institute
nds@sanger.ac.uk

Dr. Zuoming Deng
Celera Genomics
zdeng@completegenomics.com

Dr. Benjamin Dickins
The Pennsylvania State University
ben@bx.psu.edu

Dr. Jo Dicks
John Innes Centre
jo.dicks@bbsrc.ac.uk

Mr. Mark Diekhans
University of California, Santa Cruz
markd@soe.ucsc.edu

Dr. Li Ding
Washinton University School of Medicine
lding@watson.wustl.edu

Dr. Alex Dobin
Cold Spring Harbor Laboratory
dobin@cshl.edu

Dr. Ian Donaldson
University of Manchester
ian.donaldson@manchester.ac.uk

Dr. David Dooling
Washington University School of Medicine
ddooling@wustl.edu

Dr. Bryan Downie
Fritz Lipmann Institute for Age Research
bdownie@fli-leibniz.de

Dr. Richard Durbin
Wellcome Trust Sanger institute
rd@sanger.ac.uk

Mr. Sebastian Eck
Helmholtz Zentrum München
sebastian.eck@helmholtz-muenchen.de

Dr. Matthew Eldridge
Cancer Research UK
matthew.eldridge@cancer.org.uk

Mr. Andre Faure
Europeon Bioinformatics Institute
andrefau@ebi.ac.uk

Mr. Xin Feng
SUNY SB, OICR, CSHL
drestion@gmail.com

Mr. Steve Fischer
University of Pennsylvania
stevef@pcbi.upenn.edu

Mr. Marc Fiume
University of Toronto
mfiume@cs.toronto.edu

Dr. Paul Flicek
European Bioinformatics Institute
flicek@ebi.ac.uk

Dr. Christof Francke
TI Food and Nutrition (@RUNMC)
c.francke@cmbi.ru.nl

Dr. Michael Gardner
EMBL - EBI
mgardner@ebi.ac.uk

Dr. Philippe Gautier
MRC - Human Genetics Unit
philippe.gautier@hgu.mrc.ac.uk

Dr. Srinka Ghosh
Genentech Inc
ghosh.srinka@gene.com

Dr. Thomas Gingeras
Cold Spring Harbor Laboratory
gingeras@cshl.edu

Mr. Lorenzo Giordani
Sanford-Burnham Medical Research Institute
lgiordani@sanfordburnham.org

Mr. Simon Girard
Center of Excellence in Neuromics
simon.girard.3@umontreal.ca

Mr. Dominik Glodzik
MRC HGU
Dominik.Glodzik@hgu.mrc.ac.uk

Dr. Jeremy Goecks
Emory University
jeremy.goecks@emory.edu

Ms. Bornali Gohain
Tea Research Association, Jorhat, Assam, India
bornalig6@gmail.com

Dr. Tanya Golubchik
University of Oxford
golubchi@stats.ox.ac.uk

Mr. Balacz Gor
Astrid Research Inc.
silvercentrum@otptravel.hu

Ms. Allison Griggs
Broad Institute
agriggs@broadinstitute.org

Dr. Mihail Halachev
University of Birmingham
m.halachev@bham.ac.uk

Dr. Kazuo Hara
University of Tokyo
haratky@gmail.com

Dr. Christopher Harris
Washington University
charris@genome.wustl.edu

Dr. Jennifer Harrow
Wellcome Trust Sanger Institute
jla1@sanger.ac.uk

Dr. Rachel Harte
University of California Santa Cruz
hartera@soe.ucsc.edu

Dr. Fritz Hauser
Project Segmenta
segmenta@gmx.net

Dr. Maximilian Haussler
University of Manchester
maximilian.haussler@manchester.ac.uk

Mr. Edwin Hauw
Pacific Biosciences
ehauw@pacificbiosciences.com

Ms. Karen Hayden
Duke University
kehayden@gmail.com

Ms. Shirley He
University of Michigan
xghe@umich.edu

Dr. Dale Hedges
University of Miami
dhedges@med.miami.edu

Dr. Andreas Heger
MRC Functional Genomics Unit
andreas.heger@dpag.ox.ac.uk

Dr. Javier Herrero
Herrero Sanchez
jherrero@ebi.ac.uk

Dr. Ian Holmes
University of California at Berkeley
ihh@berkeley.edu

Mr. Carson Holt
University of Utah
carson.holt@genetics.utah.edu

Mr. Chris Hunter
EMBL-EBI
ksmith@ebi.ac.uk

Dr. Chris Illingworth
Wellcome Trust Sanger Institute
ci3@sanger.ac.uk

Ms. Camilla Ip
University of Oxford
camilla.ip@stats.ox.ac.uk

Dr. David Jackson
Wellcome Trust Sanger Institute
david.jackson@sanger.ac.uk

Dr. Edwin Jacox
IBDML
jacox@ibdml.univ-mrs.fr

Dr. David Jaffe
Broad Institute of MIT and Harvard
jaffe@broadinstitute.org

Mr. Shibu John
MRC Clinical Sciences Centre, Imperial College
shibu.john@csc.mrc.ac.uk

Mr. Jeff Johnston
Stowers Institute for Medical Research
jjj@stowers.org

Prof. Victor Jongeneel
University of Illinois
vjongene@illinois.edu

Dr. Fourie Joubert
University of Pretoria
fourie.joubert@up.ac.za

Ms. Claire Jubin
Institut Curie
claire.jubin@curie.fr

Ms. Johanne Justesen
Hagedorn Research Institute
jmju@hagedorn.dk

Dr. Olof Karlberg
Uppsala University
olof.karlberg@medsci.uu.se

Dr. Lennart Karssen
Erasmus Medical Centre
l.karssen@erasmusmc.nl

Dr. Masahiro Kasahara
University of Tokyo
mkasa@cb.k.u-tokyo.ac.jp

Dr. Arek Kasprzyk
Ontario Institute for Cancer Research
arek.kasprzyk@oicr.on.ca

Dr. Yoshihiro Kawahara
National Institute of Agrobiological Sciences
y.kawahara@affrc.go.jp

Dr. James Kent
University of California Santa Cruz
kent@soe.ucsc.edu

Dr. Mugdha Khaladkar
University of Pennsylvania
mugdhak@pcbi.upenn.edu

Ms. Alida Kindt
MRC HGU
alida.kindt@hgu.mrc.ac.uk

Mr. Martin Kircher
Max Planck Institute for Evolutionary Anthropology
martin.kircher@eva.mpg.de

Dr. Michal Korostynski
Insitute of Pharmacology, PAS
michkor@if-pan.krakow.pl

Dr. David Kovalic
Monsanto Co.
david.k.kovalic@monsanto.com

Mr. Martin Krzywinski
BC Cancer Agency
martink@bcgsc.ca

Dr. Robert Kuhn
UC, Santa Cruz
kuhn@soe.ucsc.edu

Ms. Swathi Kumar
Pennsylvnia State University
sak980@psu.edu

Dr. Ilkka Lappalainen
EMBL - EBI
shelley@ebi.ac.uk

Dr. Adam Lauring
University of California, San Francisco
Adam.Lauring@ucsf.edu

Dr. Andy Law
The Roslin Institute
andy.law@roslin.ed.ac.uk

Dr. Soohyun Lee
KRIBB, Korea
duplexa@gmail.com

Mr. Fabrice Legeai
INRA
fabrice.legeai@rennes.inra.fr

Mr. Louis-Philippe Lemieux Perreault
Montreal Heart Institute - Université de Montréal
louis-philippe.lemieux.perreault@statgen.org

Dr. Luca Lenzi
The University of Liverpool
parsonsl@liv.ac.uk

Mr. Mathias Lesche
Max Planck for Evolutionary Anthropology
mathias_lesche@eva.mpg.de

Dr. Wulf Dirk Leuschner
Sanofi-Aventis
wulfdirk.leuschner@sanofi-aventis.com

Dr. Xuan Liu
The University of Liverpool
parsonsl@liv.ac.uk

Dr. Bojan Losic
Ontario Institute for Cancer Research
bojan.losic@oicr.on.ca

Mr. Oscar Junhong Luo
The Australian National University
Oscar.Luo@anu.edu.au

Dr. Michael Lush
HGNC
mjlush@ebi.ac.uk

Dr. Dan MacLean
The Sainsbury Laboratory
dan.maclean@tsl.ac.uk

Mr. Anup Mahurkar
University of Maryland, Baltimore
amahurkar@som.umaryland.edu

Prof. Izabela Makalowska
Adam Mickiewicz University
izabel@amu.edu.pl

Dr. Wojciech Makalowski
University of Muenster
wojmak@uni-muenster.de

Dr. Kateryna Makova
Penn State University
kdm16@psu.edu

Dr. Jonathan Mangion
Life Technologies
jonathan.mangion@lifetech.com

Dr. Elliott Margulies
National Institutes of Health
elliott@nhgri.nih.gov

Mr. Lee McDaniel
NCBI
Lee.Mcdaniel@nih.gov

Mr. Andrew McPherson
Simon Fraser University
andrew.mcpherson@gmail.com

Dr. Gos Micklem
University of Cambridge
g.micklem@gen.cam.ac.uk

Mr. Richard Mitter
Cancer Research UK
richard.mitter@cancer.org.uk

Dr. Stephen Montgomery
University of Geneva
stephen.montgomery@unige.ch

Dr. Ali Mortazavi
California Institute of Technology
alim@caltech.edu

Dr. Jonathan Mudge
The Wellcome Trust Sanger Institute
jm12@sanger.ac.uk

Dr. Arne Mueller
Novartis
arne.mueller@novartis.com

Dr. Lisa Murray
Illumina Cambridge Ltd
yhandy@illumina.com

Dr. Lakshmi Muthuswamy
Ontario Institute for Cancer Research
Lakshmi.Muthuswamy@oicr.on.ca

Dr. Yasukazu Nakamura
DDBJ, National Institute of Genetics
yanakamu@genes.nig.ac.jp

Dr. Takeru Nakazato
Research Organization of Information and Systems
nakazato@dbcls.rois.ac.jp

Dr. Anton Nekrutenko
Penn State University
anton@bx.psu.edu

Dr. Debbie Nickerson
University of Washington
debnick@u.washington.edu

Dr. Zemin Ning
The Wellcome Trust Sanger Institute
zn1@sanger.ac.uk

Dr. William Noble
The University of Washington
noble@gs.washington.edu

Dr. Christos Noutsos
Cold Spring Harbor Laboratory
cnoutsos@cshl.edu

Dr. Harry Noyes
University of Liverpool
harry@liv.ac.uk

Dr. Joseph Omololu-Aso
Obafemi Awolowo University
pastjoe2003@yahoo.com

Dr. Thomas Otto
Wellcome Trust Sanger Institute
tdo@sanger.ac.uk

Mr. Francis Ouellette
Ontario Centre for Cancer Research
Francis@oicr.on.ca

Dr. Nehir Ozdemir Ozgenturk
Yildiz Technical University
nehirozdemir@yahoo.com

Prof. Lior Pachter
UC Berkeley
lpachter@math.berkeley.edu

Dr. Frank Panitz
Aarhus University
frank.panitz@agrsci.dk

Mr. Alexie Papanicolaou
CSIRO Entomology
alpapan@gmail.com

Dr. Luba Pardo
Vrije Universiteit Medisch Centrum
lm.pardo@vumc.nl

Ms. Jihye Park
Penn State University
jup139@psu.edu

Dr. Stephen Parker
National Institutes of Health
stephen.parker@nih.gov

Dr. Craig Parker
USDA, Agricultural Research Service
craig.parker@ars.usda.gov

Mr. Shiran Pasternak
Cold Spring Harbor Laboratory
shiran@cshl.edu

Dr. Mattia Pelizzola
Salk Institute For Biological Studies
mpelizzola@salk.edu

Mr. Andreas Petzold
Fritz Lipmann Institute for Age Research
andpet@fli-leibniz.de

Mr. Marcin Piechota
Institute of Pharmacology PAS
piechota.marcin@gmail.com

Dr. Hannes Ponstingl
Wellcome Trust Sanger Institute
hp3@sanger.ac.uk

Dr. Joan Pontius
SAIC-NCI-Frederick
pontiusj@mail.nih.gov

Mr. Arjun Prasad
National Human Genome Research Institute
aprasad@nhgri.nih.gov

Dr. Simon Prochnik
DOE - Joint Genome Institute
seprochnik@lbl.gov

Dr. Franz Quehenberger
Medical University of Graz
franz.quehenberger@medunigraz.at

Dr. Gunnar Raetsch
Max Planck Society
Gunnar.Raetsch@tuebingen.mpg.de

Dr. Hubert Rehrauer
University of Zurich
hubert.rehrauer@fgcz.ethz.ch

Mr. Daniel Renfro
EcoliWiki, Texas A&M University
bluecurio@gmail.com

Dr. Alissa Resch
University of Connecticut Health Center
resch@uchc.edu

Dr. Oleg Reva
University of Pretoria
oleg.reva@up.ac.za

Dr. Maria Luisa Rodrigues
Instituto Gulbenkian de Ciência (IGC)
mlrodrigues@igc.gulbenkian.pt

Prof. Daniel Rokhsar
DoE Joint Genome Institute
dsrokhsar@lbl.gov

Ms. Pamela Russell
Broad Institute of MIT and Harvard
prussell@broadinstitute.org

Dr. Roslin Russell
Cancer Research UK
roslin.russell@cancer.org.uk

Dr. Michael Sammeth
Centre de Regulació Genòmica (CRG)
micha@sammeth.net

Mr. Scott Sammons
Centers for Disease Control
ssammons@cdc.gov

Dr. Tobias Sargeant
The Walter and Eliza Hall Institute
sargeant@wehi.edu.au

Mr. Johann Schlesinger
Cold Spring Harbor Laboratory
schlesin@cshl.edu

Mr. Thomas Schmutzer
IPK Gatersleben
schmutzr@ipk-gatersleben.de

Ms. Petra Schwalie
EMBL-EBI
schwalie@ebi.ac.uk

Mr. Fritz Sedlazeck
Max F. Perutz Laboratories GMBH
fritz.sedlazeck@univie.ac.at

Ms. Harminder Sehra
EMBL-EBI
mindi@ebi.ac.uk

Dr. Colin Semple
MRC Human Genetics Unit
colins@hgu.mrc.ac.uk

Dr. Taner Sen
USDA-ARS/Iowa State University
taner.sen@ars.usda.gov

Mr. Uemit Seren
GMI (Gregor Mendel Institute)
uemit.seren@gmi.oeaw.ac.at

Mr. Richard Shaw
Illumina Cambridge Ltd
yhandy@illumina.com

Dr. Jill Shen
Novartis Institute of Biomedical Research
jill.shen@novartis.com

Dr. Mei-Mei Shen
Affymetrix
mei-mei_shen@affymetrix.com

Dr. Elena Shumay
Brookhaven National Laboratory
eshumay@bnl.gov

Dr. Jared Simpson
Wellcome Trust Sanger Institute
js18@sanger.ac.uk

Dr. Surendra Singh
Wolverhampton City Primary Care Trust
spd@mediware.it

Dr. Jon Sorenson
Pacific Biosciences
jsorenson@pacificbiosciences.com

Dr. Anastassia Spiridou
Illumina
aspiridou@illumina.com

Dr. William Spooner
Cold Spring Harbor Laboratory
wspooner@cshl.edu

Dr. Prashant Srivastava
MRC cllinical sciences, Imperial College
prashant.srivastava@imperial.ac.uk

Mr. Dan Staines
EMBL - EBI
shelley@ebi.ac.uk

Dr. Oliver Stegle
Max Planck Institutes Tuebingen
oliver.stegle@tuebingen.mpg.de

Dr. Kristian Stevens
University of California Davis
kastevens@ucdavis.edu

Dr. David Stewart
Cold Spring Harbor Laboratory
stewart@cshl.edu

Mr. Aengus Stewart
London Research Institute CRUK
aengus.stewart@cancer.org.uk

Dr. Gernot Stocker
Innsbruck Medical University
gernot.stocker@i-med.ac.at

Dr. Hannah Stower
Genome Biology
hannah.stower@genomebiology.com

Dr. Tim Strom
Helmholtz Zentrum München
timstrom@helmholtz-muenchen.de

Dr. Juliesta Sylvester
The University of Chicago
jesylvester@mac.com

Dr. Peter Taschner
Leiden University Medical Center
P.Taschner@lumc.nl

Dr. Todd Taylor
RIKEN Advanced Science Institute
taylor@riken.jp

Dr. Thorgeir Thorgeirsson
University of California, Santa Cruz
thor@soe.ucsc.edu

Ms. Pin Tong
UCD
tongpin2008@gmail.com

Dr. Zsolt Torok
Astrid Research Inc.
silvercentrum@otptravel.hu

Dr. Carolyn Tregidgo
Illumina
ctregidgo@illumina.com

Dr. Quang Trinh
Ontario Institute for Cancer Research
renata.musa@oicr.on.ca

Ms. Urmi Trivedi
Univerity of Edinburgh
urmi.trivedi@ed.ac.uk

Ms. Edit Tukacs
Astrid Research Inc.
SILVERCENTRUM@OTPTRAVEL.HU

Dr. Samra Turajlic
The Iinstitute of Cancer Research
samra.turajlic@icr.ac.uk

Mr. Remco Ursem
Rijk Zwaan Breeding BV
r.ursem@rijkzwaan.nl

Mr. Sowmi Utiramerur
Life Technologies
sowmi.utiramerur@lifetech.com

Dr. Keith Vance
University of Oxford
keith.vance1@ntlworld.com

Mr. Ismael Vergara
Simon Fraser University
iav@sfu.ca

Mr. Jason Walker
Washington University School of Medicine
jwalker@watson.wustl.edu

Dr. Jing Wang
Institute of Psychology, CAS
lizhao@psych.ac.cn

Dr. Zhiping Weng
University of Massachusetts Medical School
Zhiping.Weng@umassmed.edu

Ms. Claire West
Institute of Food Research
claire.west@bbsrc.ac.uk

Mr. Oscar Westesson
UC Berkeley
oscar.westesson@gmail.com

Dr. Simon White
Wellcome Trust Sanger Institute
sw4@sanger.ac.uk

Mr. Nava Whiteford
Oxford Nanopore Technologies Ltd
esther.bartlett@nanoporetech.com

Mr. Brett Whitty
Michigan State University
whitty@msu.edu

Dr. Kim Wong
Wellcome Trust Sanger Institute
kw10@sanger.ac.uk

Dr. Thomas Wu
Genentech, Inc.
twu@gene.com

Mr. Tsun-Po Yang
Wellcome Trust Sanger Institute
tpy@sanger.ac.uk

Dr. Ken Youens-Clark
Cold Spring Harbor Lab
kclark@cshl.edu

Mr. Allen Chi-Shing Yu
The Chinese University of Hong Kong
allenyu@cuhk.edu.hk

Dr. Le Yu
Roslin Institute & R(D)SVS, University of Edinburg
le.yu@roslin.ed.ac.uk

Mr. Christopher Zaleski
Cold Spring Harbor Laboratory
zaleski@cshl.edu

Dr. Qiandong Zeng
Broad Institute
qzeng@broadinstitute.org

Dr. Daniel Zerbino
UC Santa Cruz
dzerbino@soe.ucsc.edu

Dr. Zhihua Zhang
University of Texas at Dallas
IZHANG@utdallas.EDU

Ms. Yanju Zhang
Leiden University Medical Center (LUMC)
Y.Zhang@lumc.nl

Dr. Junjun Zhang
Ontario Institute for Cancer Research
junjun.zhang@oicr.on.ca