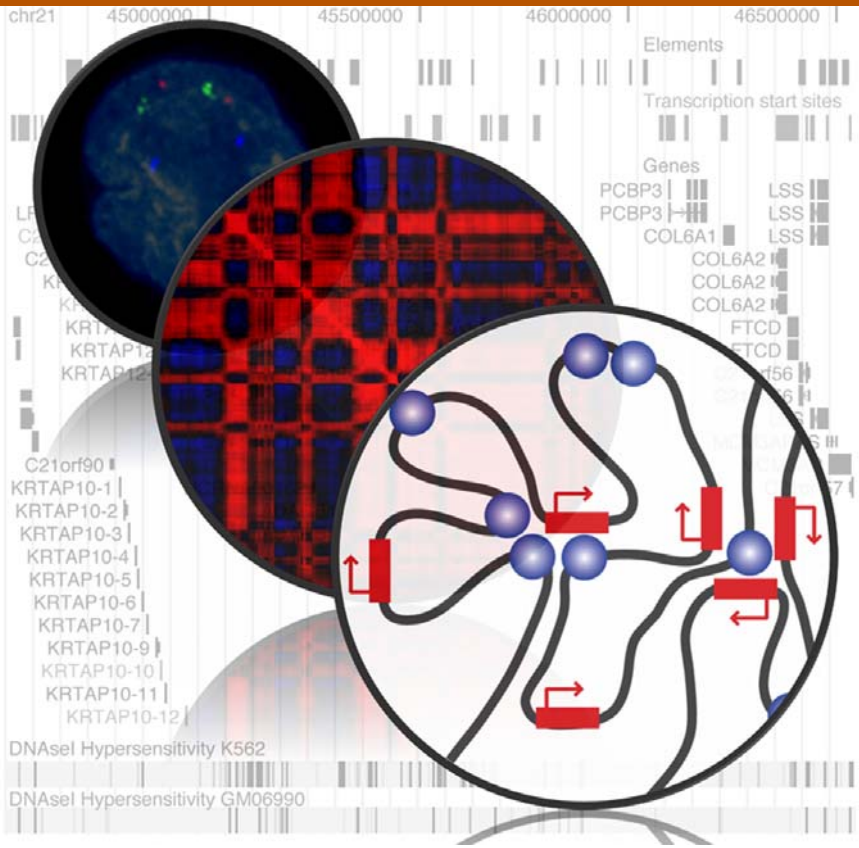
Abstracts of papers presented
at the 2010 meeting on

SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION

March 23–March 27, 2010



Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Abstracts of papers presented
at the 2010 meeting on

SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION

March 23–March 27, 2010

Arranged by

Bradley Bernstein, *Harvard Medical School*
Martha Bulyk, *Harvard Medical School*
Harmen Bussemaker, *Columbia University*
Marian Walhout, *UMass Medical School*

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

This meeting was funded in part by the **National Cancer Institute,** a branch of the **National Institutes of Health;** and the **National Science Foundation.**

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

Corporate Sponsors

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

Plant Corporate Associates

Monsanto Company
Pioneer Hi-Bred International, Inc.

Foundations

Hudson-Alpha Institute for Biotechnology

Cover: Figure by Nynke van Berkum, Dekker Lab, University of Massachusetts Medical School. See abstract #95, van Berkum et. al.

SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION
Tuesday, March 23 – Saturday, March 27, 2010

Tuesday	7:30 pm	Keynote Speaker
Tuesday	8:45 pm	**1** Transcription
Wednesday	9:00 am	**2** Transcriptional Regulatory Networks
Wednesday	2:00 pm	**3** Poster Session I
Wednesday	4:30pm	Wine & Cheese Party
Wednesday	7:30 pm	Keynote Speaker
Wednesday	8:45 pm	**4** Chromatin
Thursday	9:00 am	**5** Epigenetics
Thursday	2:00 pm	**6** Poster Session II
Thursday	7:30 pm	**7** Cis-regulatory Logic
Friday	9:00 am	**8** Post-transcriptional Regulation
Friday	2:00 pm	**9** Emerging Technology
Friday	6:00 pm	Banquet
Saturday	9:00 am	**10** Variation and Evolution

Mealtimes at Blackford Hall are as follows:
Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm

Bar is open from 5:00 pm until late

Printed on 100% recycled paper.

PROGRAM

TUESDAY, March 23—7:30 PM

KEYNOTE SPEAKER

Mapping genome-wide nucleosome dynamics
Steven Henikoff, Roger M. Deal, Christopher M. Weber.
Presenter affiliation: Fred Hutchinson Cancer Research Center, Seattle, Washington. 1

TUESDAY, March 23—8:45 PM

SESSION 1 TRANSCRIPTION

Chairperson: **T. Hughes,** University of Toronto, Canada

Variability in gene expression underlies incomplete penetrance
Alexander van Oudenaarden.
Presenter affiliation: Massachusetts Institute of Technology, Cambridge, Massachusetts. 2

Exploring relationships in data sets using spatial gene expression patterns for all *Drosophila* transcription factors
Erwin Frise, Ann S. Hammonds, Susan E. Celniker.
Presenter affiliation: Lawrence Berkeley National Labs, Berkeley, California. 3

A quantitative model of glucose signaling in yeast reveals an incoherent feed forward loop leading to a specific, transient pulse of transcription
Michael R. Brent, Sooraj Kuttykrishnan, Jeffrey Sabina, Mark Johnston.
Presenter affiliation: Washington University, Saint Louis, Missouri. 4

Genome-wide binding of MyoD in specified and differentiating muscle cells
Stephen J. Tapscott, Yi Cao, Zizhen Yao, Walter L. Ruzzo, Robert C. Gentleman.
Presenter affiliation: Fred Hutchinson Cancer Research Center, Seattle, Washington. 5

WEDNESDAY, March 24—9:00 AM

SESSION 2 TRANSCRIPTIONAL REGULATORY NETWORKS

Chairperson: **X. Shirley Liu,** Dana-Farber Cancer Institute, Boston, Massachusetts

WEDNESDAY, March 24—2:00 PM

SESSION 3 POSTER SESSION I

WEDNESDAY, March 24—4:30 PM

Wine and Cheese Party

WEDNESDAY, March 24—7:30 PM

KEYNOTE SPEAKER

WEDNESDAY, March 24—8:45 PM

SESSION 4 CHROMATIN

Chairperson: **R. Morse,** Wadsworth Center, Albany, New York

THURSDAY, March 25—9:00 AM

SESSION 5 EPIGENETICS

Chairperson: **J. Stamatoyannopoulos,** University of Washington, Seattle

Combined computational and experimental approaches pinpoint a GC-rich element sufficient for the recruitment of Polycomb Complexes in ES Cells.
Eric M. Mendenhall, Richard P. Koche, Thanh T. Truong, Vicky W. Zhou, Bradley E. Bernstein.
Presenter affiliation: Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts; Howard Hughes Medical Institute, Chevy Chase, Maryland; Broad Institute, Cambridge, Massachusetts. 100

Bimodality in gene expression levels correlates with an epigenetic module
Daniel Hebenstreit, Miaoqing Fang, Varodom Charoensawan, Daniel J. Turner, Alexander van Oudenaarden, Sarah A. Teichmann.
Presenter affiliation: MRC Laboratory of Molecular Biology, Cambridge, United Kingdom. 101

Programming chromatin states by long noncoding RNAs
Howard Y. Chang.
Presenter affiliation: HHMI, Stanford University, California. 102

THURSDAY, March 25—2:00 PM

SESSION 6 POSTER SESSION II

A protein-DNA interaction network regulating responses to reactive oxygen species
Lesley T. MacNeil, A. J. Marian Walhout.
Presenter affiliation: University of Massachusetts Medical School, Worcester, Massachusetts. 103

The impact of genetic polymorphism on the transcriptional pathway underlying glucocorticoid response
Joseph C. Maranville, Francesca Luca, Allison Richards, Matthew Stephens, Anna Di Rienzo.
Presenter affiliation: University of Chicago, Chicago, Illinois. 104

Genome-wide characterization of the transcriptome in ENCODE cell lines
Georgi K. Marinov, Brian A. Williams, Ali Mortazavi, Timothy E. Reddy, Jason Gertz, Richard M. Myers, Barbara Wold.
Presenter affiliation: California Institute of Technology, Pasadena, California. 105

THURSDAY, March 25—7:30 PM

SESSION 7 CIS REGULATORY LOGIC

Chairperson: **L. Mirny,** Massachusetts Institute of Technology, Cambridge

SESSION 8 POST-TRANSCRIPTIONAL REGULATION

Chairperson: **M. Walhout,** University of Massachusetts Medical School, Worcester

FRIDAY, March 26—2:00 PM

SESSION 9 EMERGING TECHNOLOGY

Chairperson: **B. Andrews,** University of Toronto, Canada

A combinatorial protein code defines new principal chromatin types in *Drosophila*
Ulrich Braunschweig, Guillaume J. Filion, Joke G. van Bemmel, Wendy Talhout, Jop Kind, Aranzazu Rosado, Ines de Castro, Wim Brugman, Ron Kerkhoven, Bas van Steensel.
Presenter affiliation: Netherlands Cancer Institute, Amsterdam, Netherlands. 194

Global correlation of transcription and DNA conformation reveals new modes of gene regulation
Fedor Kouzine, Ashutosh Gupta, Damian Wojtowicz, Matthew Vogt, Arito Yamane, Rafael Casellas, Teresa Przytycka, Craig J. Benham, David L. Levens.
Presenter affiliation: CCR, NCI, National Institutes of Health, Bethesda, Maryland. 195

Chromatin on the megabase scale—The fractal globule architecture and its physical properties
Erez Lieberman-Aiden, Maxim Imakaev, Nynke van Berkum, Louise Williams, Andreas Gnirke, Najeeb Tarazi, Job Dekker, Eric S. Lander, Leonid A. Mirny.
Presenter affiliation: Harvard-MIT, Cambridge, Massachusetts. 196

The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules
Davide Bau, Amartya Sanyal, Bryan R. Lajoie, Emidio Capriotti, Job Dekker, Marc A. Marti-Renom.
Presenter affiliation: University of Massachusetts Medical School, Worcester. 197

FRIDAY, March 26

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SESSION 10 VARIATION AND EVOLUTION

Chairperson: **M. Bulyk,** Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

MAPPING GENOME-WIDE NUCLEOSOME DYNAMICS

Steven Henikoff, Roger M Deal, Christopher M Weber

Fred Hutchinson Cancer Research Center, Basic Sciences Division and Howard Hughes Medical Institute, 1100 Fairview Avenue North, Seattle, WA, 98104

Eukaryotic gene expression occurs in the context of chromatin, and maintaining a region accessible to DNA-binding proteins for transcriptional regulation requires active processes that mobilize nucleosomes. Our approach to studying these processes has been to map nucleosome dynamics genome-wide, and we have introduced three different strategies to achieve this goal: (1) To measure relative levels of histone replacement across the genome, we have used biotin-tagging with chromatin affinity purification of the replication-independent histone variant, H3.3, which replaces replication-coupled H3 over the course of the cell cycle. (2) To map classical 'active' chromatin genome-wide we have applied salt fractionation to intact micrococcal nuclease-treated nuclei. (3) To map histone turnover directly we have developed a novel method based on metabolic labeling of proteins followed by affinity purification of newly synthesized histone core particles. These diverse methods applied to Drosophila chromatin provide a consistent view of dynamic processes at promoters and epigenetic regulatory elements that disrupt and turn over nucleosomes. In addition, our studies of the H2A.Z histone variant in diverse eukaryotes reveal how the subunit structure of a nucleosome might impact its stability and affect its functional properties. Taken together, our results suggest that promoters and regulatory elements are maintained in an accessible state by the transient destabilization of nucleosomes.

VARIABILITY IN GENE EXPRESSION UNDERLIES INCOMPLETE PENETRANCE

Alexander van Oudenaarden
Massachusetts Institute of Technology, Cambridge, MA, 02139

The phenotypic differences between individual organisms can often be ascribed to underlying genetic and environmental variation. However, even genetically identical organisms in homogeneous environments vary, indicating that randomness in developmental processes such as gene expression may also generate diversity. To examine the consequences of gene expression variability in multicellular organisms, we studied intestinal specification in the nematode Caenorhabditis elegans in which wild-type cell fate is invariant and controlled by a small transcriptional network. Mutations in elements of this network can have indeterminate effects: some mutant embryos fail to develop intestinal cells, whereas others produce intestinal precursors. By counting transcripts of the genes in this network in individual embryos, we show that the expression of an otherwise redundant gene becomes highly variable in the mutants and that this variation is subjected to a threshold, producing an ON/OFF expression pattern of the master regulatory gene of intestinal differentiation. Our results demonstrate that mutations in developmental networks can expose otherwise buffered stochastic variability in gene expression, leading to pronounced phenotypic variation.

EXPLORING RELATIONSHIPS IN DATA SETS USING SPATIAL GENE EXPRESSION PATTERNS FOR ALL *DROSOPHILA* TRANSCRIPTION FACTORS

Erwin Frise, Ann S Hammonds, Susan E Celniker
Lawrence Berkeley National Labs, BDGP/Genome Dynamics, 1 Cyclotron Road, Berkeley, CA, 94720

To analyze regulatory networks in multicellular organisms, it is essential to know both spatial and temporal patterns of genes expression. We are generating a 2D mRNA expression atlas to profile embryonic development of *Drosophila*. To utilize the image dataset for computational analysis, we converted images into standardized virtual representations of expression patterns that we called Triangulated Images (TIs). TIs provide an intuitive representation in which spatial expression can be visualized in context of other genes and analyzed using standard computational tools. We have demonstrated the utility of our approach in a systematic analysis of the gene expression dataset and have made the TIs available to the community (Nature Molecular Systems Biology, in press).

We have now extended the dataset to over 7,500 genes and completed collecting images of expression patterns for over 97% of the set of 711 sequence specific Transcription Factors (TF) identified by manual curation of genes with known or predicted DNA binding domains. The inclusion of nearly all TF expression patterns greatly enhances the use of our dataset for identifying components of regulatory networks. Here we present novel interactive visualization tools and strategies to organize and refine experimentally and computationally derived datasets.

For example, to investigate the contribution of TF to gene regulation, we analyzed gene expression of downstream targets. We computationally predicted targets for the mesoderm specific TF *twist* (*twi*) and identified 9 distinct pattern categories at stage 4-6. All patterns overlap with *twi* suggesting that sole function of *twi* is to activate expression. Using these pattern categories, we screened our TF dataset for overlapping or excluded patterns to identify candidate TFs that may co-regulate these downstream targets. We applied a similar strategy using genes with known functions to identify new pathway components. We created a dataset of expression patterns for genes associated with the GO-term “eye development” with 12 distinct pattern categories, which provided a starting point to identify TFs with similar expression. We have shown in our systematic analysis that gene expression divides the stage 4-6 embryo into clearly defined spatial domains similar to an experimentally determined fate map. We have developed methods to identify patterns overlapping with those domains. Linking the expression domains with developmental fates will lead towards a functional description for previously uncharacterized TF.

A QUANTITATIVE MODEL OF GLUCOSE SIGNALING IN YEAST REVEALS AN INCOHERENT FEED FORWARD LOOP LEADING TO A SPECIFIC, TRANSIENT PULSE OF TRANSCRIPTION

Michael R Brent[1], Sooraj Kuttykrishnan[1], Jeffrey Sabina[2], Mark Johnston[3]

[1]Washington University, Center for Genome Sciences and Department of Computer Science, 4444 Forest Park Blvd., Saint Louis, MO, 63108, [2]Washington University, Center for Genome Sciences and Department of Computer Science, 4444 Forest Park Blvd., Saint Louis, MO, 63108, [3]Univ. of Colorado Denver, Department of Biochemistry and Molecular Genetics, 12801 E. 17th Ave., Aurora, CO, 80045

The ability to design and engineer organisms demands the ability to predict kinetic responses of novel regulatory networks built from well characterized biological components. Surprisingly, few validated, kinetic models of complex regulatory networks have been derived by combining models of the network components. A major bottleneck in producing such models is the difficulty of measuring *in vivo* rate constants for components of complex networks. We demonstrate that a simple, genetic approach to measuring rate constants *in vivo* produces an accurate kinetic model of the complex network that *Saccharomyces cerevisiae* employs to regulate the expression of genes encoding glucose transporters. The model predicts a transient pulse of transcription of *HXT4* (but not *HXT2* or *HXT3*) in response to addition of a low concentration of glucose to cells. We show that these predictions are correct. It also provides a mechanistic explanation: *HXT2-4* are regulated by a type 2, incoherent feed forward loop involving Rgt1 and Mig2. The efficiency with which Rgt1 and Mig2 repress each HXT gene determines which of them show a pulse of transcription in response to glucose. Finally, the model correctly predicts changes in the kinetics of the *HXT4* response when components of the feed forward loop are deleted.

GENOME-WIDE BINDING OF MYOD IN SPECIFIED AND DIFFERENTIATING MUSCLE CELLS

Stephen J Tapscott[1], Yi Cao[1], Zizhen Yao[2], Walter L Ruzzo[2], Robert C Gentleman[2]

[1]Fred Hutchinson Cancer Research Center, Division of Human Biology, 1100 Fairview Avenue North, Seattle, WA, 98109, [2]Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, 1100 Fairview Avenue North, Seattle, WA, 98109

In many cell lineages, the transcription factors that regulate terminal differentiation are expressed during the regulatory growth phase of the specified cell type, prior to expression of genes associated with terminal differentiation. For example, the myogenic transcription factor MyoD is expressed in replicating myoblasts but does not activate expression of muscle genes until the cell switches from a growth to a differentiation phase. We have used chromatin immunoprecititation and ultra-high throughput sequencing to identify genome-wide binding of MyoD in several skeletal muscle cell types. As anticipated, MyoD preferentially binds to an E-box sequence that resembles the in vitro selected site for a MyoD:E-protein heterodimer, and MyoD binding increases during differentiation at many of the regulatory regions of genes expressed in skeletal muscle. Unanticipated findings were that MyoD was constitutively bound to a majority of sites throughout the genome in both myoblasts and myotubes, and was associated with altered histone acetylation, but not transcription, at these sites. Therefore, in addition to regulating muscle gene expression, MyoD binds genome-wide and has the ability to broadly re-organize the genome in myoblasts, suggesting a previously unrecognized architectural role in cell specification.

BIOMARKERS BASED ON NETWORKS, NOT INDIVIDUAL LOCI

Trey Ideker

University of California, San Diego, Medicine, 9500 Gilman Drive, La Jolla, CA, 92093-0688

A biomarker is typically thought of as an individual locus, gene or protein. However, complex phenotypes observed during development and disease are rarely due to single proteins. Recently, we have shown that protein networks are a source of powerful biomarkers, and that in many cases these biomarker networks are more predictive than any individual gene. I will discuss use of protein-network-based biomarkers in two areas:

1. Markers of developmental state. With collaborators at RIKEN, we have assembled a large transcription factor protein interaction network by screening for interactions among all mammalian transcription factors. Through analysis of these data, we identify a subnetwork of 15 interacting homeobox transcription factors whose expression levels can predict the developmental fate of tissues during differentiation.

2. Improved power and interpretation of genome-wide association studies (GWAS). Protein networks may be the key to mining GWAS to understand complex diseases for which not one but many genetic loci play a role. We have recently used protein networks to translate GWAS into maps of functional interactions among protein complexes and pathways.

For professional distribution of our network-based technologies, we are developers of the Cytoscape platform, an Open-Source software environment for visualization and analysis of biological networks and models (http://www.cytoscape.org/).

FUNCTIONAL MODULARITY OF NUCLEAR HORMONE RECEPTORS IN A *C. ELEGANS* GENE REGULATORY NETWORK

H. Efsun Arda[1], Stefan Taubert[2,3], Colin Conine[1], Ben Tsuda[1], Marc Van Gilst[4], Reynaldo Sequerra[5], Lynn Doucette-Stamm[5], Keith R Yamamoto[2], A.J. Marian Walhout[1]

[1]University of Massachusetts Medical School, Program in Gene Function and Expression and Program in Molecular Medicine, Worcester, MA, 01605, [2]University of California, San Francisco, Department of Cellular and Molecular Pharmacology, San Francisco, CA, 94143, [3]Present Address: University of British Columbia, Department of Medical Genetics, Vancouver, V5Z 4H4, Canada, [4] Fred Hutchinson Cancer Research Center, Basic Sciences Division, Seattle, WA, 98109, [5]Agencourt Bioscience Corporation, Beverly, MA, 01915

Gene regulatory networks (GRNs) provide insights into mechanisms of differential gene expression at a systems level. GRNs that relate to metazoan development have been studied extensively. However, little is still known about the design principles, organization and functionality of GRNs that control post-developmental processes. We will describe the first metazoan GRN of *Caenorhabditis elegans* metabolic genes. This network is enriched for nuclear hormone receptors (NHRs). The NHR family has greatly expanded in nematodes: humans have 48 NHRs, but *C. elegans* has 284, most of which are uncharacterized. The GRN is highly modular and two modules predominantly consist of NHRs. Network modularity has been proposed to facilitate a rapid response to different cues. Since NHRs are metabolic sensors that are poised to respond to ligands, this suggests that *C. elegans* post-developmental GRNs evolved to enable rapid and adaptive responses to different cues by a concurrence of NHR family expansion and modular GRN wiring.

NEW INSIGHTS INTO GLOBAL TRANSCRIPTIONAL CONTROL IN EMBRYONIC STEM CELLS AND TUMOR CELLS

Peter B Rahl[1], Charles Y Lin[1,2], Amy C Seila[2], Ryan A Flynn[2], Christopher B Burge[2], Phillip A Sharp[2,3], Richard A Young[1,2]

[1]Whitehead Institute for Biomedical Research, Biology, 9 Cambridge Center, Cambridge, MA, 02142, [2]MIT, Biology, 9 Cambridge Center, Cambridge, MA, 02142, [3]Koch Institute, Biology, 9 Cambridge Center, Cambridge, MA, 02142

Recruitment of the RNA Polymerase II (Pol II) transcription initiation apparatus to promoters by specific DNA binding transcription factors is well recognized as a key regulatory step in gene expression. We have found that promoter-proximal pausing is a general feature of transcription by Pol II in embryonic stem (ES) cells, and thus an additional step where regulation of gene expression occurs. It has been proposed that some transcription factors recruit the transcription apparatus to promoters, while others recruit the Pol II kinase P-TEFb to effect promoter-proximal pause release. We found that c-Myc, which occupies a third of actively transcribed genes in ES cells and is a key regulator of cellular proliferation, binds P-TEFb and contributes to release of promoter-proximal paused Pol II at these genes. Many human tumor cells have mutations that lead to c-Myc overexpression. Our results suggest that such tumor cells have enhanced expression of proliferation and metabolic genes due to the role of c-Myc in regulation of Pol II pause release. This notion suggests new approaches to cancer therapy based on inhibition of c-Myc–mediated Pol II pause release.

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

Ido Amit[1,2], Manuel Garber[1], Nicolas Chevrier[1,2,3], Ana Leite[1], Yoni Donner[1], Tom Eisenhaure[2,3], Mitchell Guttman[1,4], Jen Grenier[1], Lisa A Schubert[5], David Root[1], Nir Hacohen[1,2,3], Aviv Regev[1,4]

[1]Broad Institute, of MIT and Harvard, 7 Cambridge Center, Cambridge, MA, 02142, [2]Massachusetts General Hospital, Center for Immunology and Inflammatory Diseases, 149 13th St., Charlestown, MA, 02129, [3]Harvard University, Harvard Medical School, 200 Longwood Ave, Boston, MA, 02115, [4] Howard Hughes Medical Institute, MIT, Department of Biology, 77 Massachusetts Ave, Cambridge, MA, 02140, [5]Nanostring, Technologies, 530 Fairview Ave N, Suite 2000, Seattle, WA, 98109

Deciphering the regulatory networks that control dynamic and specific gene expression responses in mammalian cells remains a major challenge. While models inferred from genomic data have identified candidate regulatory mechanisms, such models remain largely unvalidated. Here, we present an unbiased strategy based on systematic gene perturbation and innovative multiplex detection to derive regulatory networks in mammalian cells. We apply this approach to decipher the network that controls the transcriptional response to pathogens in primary dendritic cells (DCs), testing the regulatory function of 125 transcription factors, chromatin modifiers, and RNA binding proteins. Our approach accurately assigned 32 known regulators (e.g. NF_KB, IRFs, and STATs) to their target genes and discovered 68 additional functional regulators that were not previously implicated in this response. We quantify the contribution of each regulator to two major transcriptional programs (inflammatory and anti-viral), identifying a core network of two dozen key regulators and 76 fine-tuners, that uses a combination of coherent feed-forward circuits, dominant activation, and cross-inhibition to control response specificity. Among these we discover a tier of chromatin modifiers that specifically repress interferon beta 1 (IFNB1) expression upon bacterial but not viral stimulation, and a large circuit of cell cycle regulators that was co-opted to regulate the viral response. Our work establishes a broadly-applicable, comprehensive and unbiased approach to identifying the wiring and function of a regulatory network controlling a major transcriptional response in primary mammalian cells.

PROFILING THE HUMAN PROTEIN-DNA INTERACTOME IDENTIFIES UNCONVENTIONAL DNA-BINDING PROTEINS

Heng Zhu

Johns Hopkins University, Department of Pharmacology & HiT Center, 733 N Broadway, Baltimore, MD, 21212

Protein-DNA interactions (PDIs) mediate a broad range of functions essential for cellular differentiation, function, and survival. However, it is still a daunting task to comprehensively identify and profile sequence-specific PDIs in complex genomes. We have used a combined bioinformatics and protein microarray-based strategy to systematically characterize the human protein-DNA interactome. We identified 17,718 PDIs between 460 DNA motifs predicted to regulate transcription and 4,191 human proteins of various functional classes. Among them, we identified a large number of new PDIs for known transcription factors (TFs), as well as for previously uncharacterized TFs. Remarkably, we found that over three hundred proteins not previously annotated as TFs also showed sequence-specific PDIs, including RNA binding proteins, mitochondrial proteins, and protein kinases. One of such unconventional DNA-binding proteins (uDBPs), MAPK1, acts as a transcriptional repressor for interferon gamma-induced genes. Additional analysis of uDBPs in other families revealed that such moonlighting functions are probably widely spread in humans.

FROM BOLTING TO BLOOM: GENOME-WIDE CONTROL OF FLORAL MERISTEM INITIATION AND DIFFERENTIATION

Kerstin Kaufmann[1], Frank Wellmer[2], Jose M Muino[3], Thilia Ferrier[4], Samuel E Wuest[2], Vijaya Kumar[5], Pawel Krajewski[6], Elliot M Meyerowitz[5], Gerco C Angenent[3], Jose Luis Riechmann[4]

[1]Wageningen University and Research Centre, Molecular Biology, Droevendaalsesteeg 1, Wageningen, 6708PB, Netherlands, [2]Trinity College Dublin, Smurfit Institute of Genetics, Dublin 2, Dublin, Dublin 2, Republic Of Ireland, [3]Plant Research International, Bioscience, Droevendaalsesteeg 1, Wageningen, 6708PB, Netherlands, [4] Center for Research in Agricultural Genomics, CRAG, Jordi Girona 18-26, Barcelona, 08034, Spain, [5]California Institute of Technology, Division of Biology, Pasadena, CA, 91125, [6]Polish Academy of Sciences, Institute of Plant Genetics, Strzeszyńska 34, Poznan, 60-479, Poland, [7]Institució Catalana de Recerca i Estudis Avançats, (ICREA), Passeig Lluís Companys, 23, Barcelona, 08010, Spain

Floral homeotic genes act as master regulators in the initiation of floral meristems and the specification of identities of the different types of floral organs (sepals, petals, stamens and carpels). Most of these genes encode transcription factors of the MADS-box family, which assemble in a combinatorial fashion into floral organ-specific multimeric protein complexes. We used chromatin immunoprecipitation followed by ultrahigh-throughput sequencing (ChIP-SEQ) to obtain genome-wide DNA-binding patterns of two members of the MADS-box TF family that are key regulators in floral initiation and differentiation. Comparison of the ChIP-SEQ data with results of time-series gene expression microarrays suggests distinct phases in the direct regulation of target genes. The data suggest that these TFs act as global regulators and orchestrate different developmental pathways in the flowering process. Characterization of the target genes shows that both TFs integrate and modulate growth-related and hormonal pathways. Furthermore, our data suggest multiple molecular links with other MADS-box TFs, suggesting a complex intra-family regulatory network.

ELUCIDATING THE STRUCTURE OF THE TRANSCRIPTIONAL REGULATORY NETWORK WITH GENETIC INTERACTIONS

Brenda J Andrews[1,2,3], Michael Costanzo[1], Anastasia Barshinikova[1,2,3], Yungil Kim[4], Chad Myers[4]

[1]University of Toronto, Terrence Donnelly Centre for Cellular & Biomolecular Research, 160 College Street, Toronto, M5S 3E1, Canada, [2]University of Toronto, Department of Molecular Genetics, 1 Kings College Circle, Toronto, M5S 1A8, Canada, [3]University of Toronto, Banting & Best Dept of Medical Research, 160 College Street, Toronto, M5S 3E1, Canada, [4] University of Minnesota, Department of Computer Science and Engineering and Department of Electrical & Computer Engineering, 4-192 EE/CS Building 200 Union Street SE, Minneapolis, MN, 55455

Knock-out studies in yeast have revealed a surprising tolerance of single gene deletions, particularly among transcription factors, where the rate of gene essentiality is 8% compared to the genome background rate of 17%. We have used an automated yeast genetics approach called synthetic genetic array or SGA analysis to systematically address the issue of genetic redundancy. We constructed a genome-scale genetic interaction map by examining 5.4 million gene-gene pairs for synthetic genetic interactions, generating quantitative genetic interaction profiles for about 75% of all genes in *Saccharomyces cerevisiae*. These interactions include both negative and positive genetic interactions; negative interactions refer to a more severe fitness defect than expected, with the extreme case being synthetic lethality while positive interactions refer to double mutants with a less severe fitness defect than expected. A network based on genetic interaction profiles reveals a functional map of the cell in which genes of similar biological processes cluster together in coherent subsets and highly correlated profiles delineate specific pathways to define gene function. The global network identifies functional cross-connections between all bioprocesses, mapping a cellular wiring diagram of pleiotropy.

We explored the SGA regulatory network for motifs that might illuminate the structure of transcriptional regulatory pathways. We found that local motifs within the regulatory network (e.g. feed-forward, single-input, multi-input structures) are predictive of the frequency and type of genetic interactions between transcription factors and their targets. We also observed a surprisingly high frequency of positive interactions between pairs of transcription factors regulating the same target, particularly when the direction of regulation was coherent (activator/activator or repressor/repressor relationships), suggesting cooperativity among transcription factors may play a larger role in regulation than previously appreciated. Finally, we found that the position of target genes within the regulatory network and the number of different transcription factors regulating them was predictive of their genetic interaction degree across the genome, confirming a connection between the regulatory hierarchy and gene redundancy.

DEVELOPMENT OF A GENOME-WIDE DNA DAMAGE DETECTION ASSAY

Sheera Adar, Jason D Lieb

University of North Carolina at Chapel Hill, Department of Biology, Carolina Center for Genome Science, and Lineberger Comprehensive Cancer Center, 407 Fordham Hall, Chapel Hill, NC, 27599

The integrity of the genome is constantly challenged by both internal and external damaging agents. Unrepaired damage to DNA compromises gene expression and can lead to cell death. Inaccurate repair results in the formation of mutations, and enhances the risk for genetic disease and cancer development.

The packaging of DNA into chromatin can have both benefits and costs from the perspective of DNA repair. On one hand, tightly packed DNA may be more protected from damage. On the other, chromatin may hinder the detection and repair of damage once it has formed. Added complexity comes from the fact that specific chromatin modifications can play a role in the recruitment of DNA repair proteins and in the perpetuation and amplification of the DNA damage-response signaling network.

Recent advances in high throughput genomic tools have been applied to study chromatin structure and its association with DNA-related biochemical processes. These studies have focused primarily on transcription, and have shown that the pattern of histone modifications and chromatin structure in the nucleus influence and respond to the expression of genes and the developmental fate of a cell. The study of DNA damage formation and DNA repair, however, has largely been limited to standard biochemical and molecular biology tools. We describe the development of a new genome-wide DNA-damage detection assay based on the isolation of damaged genomic DNA followed by high throughput sequencing. Our assay will allow the identification of damage-prone sequences in the genome, and could be used to decipher the relationship between gene expression, chromatin structure, DNA damage formation and DNA repair.

COMPREHENSIVE MODELING OF MICRORNA TARGETS PREDICTS FUNCTIONAL NON-CONSERVED AND NON-CANONICAL SITES

Doron Betel[1], Anjali Koppal[2], Phaedra Agius[1], Chris Sander[1], Christina Leslie[1]
[1]MSKCC, Computational Biology, 68th street, New York, NY, 10065,
[2]Columbia University, Department of Computer Science, Amsterdam Ave, New York, NY, 10027

Accurate prediction of microRNA targets is a challenging computational problem, impeded by incomplete biological knowledge and the scarcity of experimentally validated targets. The primary determinant for regulation, near-perfect base pair-ing in the seed region of the microRNA (positions 2-7), gives poor specificity as a prediction rule. In an effort to reduce false predictions, most computational meth-ods restrict to perfect seed matches that are evolutionary conserved, despite experimental evidence that neither constraint holds in general. Here we present a new algorithm called mirSVR for predicting and ranking the efficiency of mi-croRNA target sites by using supervised learning on mRNA expression changes from microRNA transfection experiments. We use support vector regression (SVR) to train on features of the predicted miRNA::mRNA duplexes as well as contextual features without restricting to perfect seed complementarity or filtering by conservation. In a large-scale evaluation on independent transfection and inhibition experiments, mirSVR significantly outperformed existing target predic-tion methods for predicting genes that are deregulated at the mRNA or protein levels.

mirSVR effectively broadens target prediction beyond the standard restrictions of perfect seeds and strict conservation without introducing a large number of spuri-ous predictions. In particular, we found that mirSVR correctly identified functional but poorly conserved target sites, and that imposing a conversation filter resulted in a reduced rate of detection of true targets. mirSVR scores are calibrated to correlate linearly with the extent of downregulation and therefore enable accurate scoring of genes with multiple target sites by addition of individual site scores. Furthermore, the scores can be converted to an empirical probability of down-regulation, which provides a meaningful guide for selecting a score cutoff. The model successfully predicted genes that are regulated by multiple endogenous microRNAs – rather than transfected microRNAs whose concentrations are above physiological levels – when analyzing targets bound to human Argonaute (AGO) proteins as identified by AGO immunoprecipitation. Finally, we tested the usefulness of including non-canonical sites in the model by evaluating perform-ance on biochemically determined sites from recent PURE-CLIP experiments, <20% of which do not contain any perfect microRNA seed match. We found that mirSVR indeed correctly detected a significant number of these experimentally verified non-canonical sites.

HIGH RESOLUTION MODELS OF TRANSCRIPTION FACTOR-DNA AFFINITIES IMPROVE *in vitro* AND *in vivo* BINDING PREDICTIONS

Phaedra Agius[1], Aaron Arvey[1], William Chang[1], William Stafford Noble[2], Christina Leslie[1]

[1]MSKCC, Computational Biology, 68th street, New York, NY, 10065, [2]University of Washington, Department of Genome Sciences, NE Pacific Street, Seattle, WA, 98195

Accurately modeling the DNA sequence preferences of transcription factors (TFs), and using these models to predict *in vivo* genomic binding sites for TFs, are key pieces in deciphering the regulatory code. These efforts have been frustrated by the limited availability and accuracy of TF binding site motifs, usually represented as position-specific scoring matrices (PSSMs), which may match large numbers of sites and produce an unreliable list of potential target genes. Recently, protein binding microarray (PBM) experiments have emerged as a new source of high-resolution data on *in vitro* TF binding specificities. PBMs measure the binding of a fluorescently tagged TF to a carefully designed set of ~44K double-stranded DNA probes. How best to use this data to represent the TF's binding preferences is an open question. So far, PBM data has been analyzed via rank statistics on probe intensities from the TF binding experiment, so that individual sequence patterns (e.g. 8-mers or longer gapped patterns) are assigned enrichment scores (E-scores). This representation is informative but unwieldy because every TF is assigned a list of thousands of scored sequence patterns.

Here we apply supervised learning methods to PBM data to learn compact and statistically more powerful models of in vitro TF binding preferences. These models can be readily used to scan intergenic regions for predicting *in vivo* binding sites. We used a novel 1st order Markov mismatch string kernel to represent probe sequence similarities, and we trained support vector regression (SVR) models to learn the mapping from probe sequences to PBM binding intensities. Using a large data set of yeast and mouse TFs, taken from three separate studies, for which PBM data for two independent probe designs is available, we found that our SVR models can better predict probe intensity than the current E-score method or PBM-derived PSSMs. Moreover, by using SVRs to score yeast intergenic regions, we were better able to predict *in vivo* occupancy as measured by ChIP-chip and ChIP-seq experiments than a previous occupancy scoring method based on E-scores or PSSM-based prediction. This flexible SVR framework can be broadly used in place of PSSMs to improve modeling of regulatory sequences.

INFERRING TRANSCRIPTIONAL AND MICRORNA-MEDIATED REGULATORY PROGRAMS IN GLIOBLASTOMA

Manu Setty, Aly A Khan, Aaron Arvey, Frank Neezen, Phaedra Agius, Christina Leslie

MSKCC, Computational Biology, 68th street, New York, NY, 10065

Large-scale cancer genome characterization projects, such as The Cancer Genome Atlas (TCGA) initiative, are currently generating multiple types of high-throughput molecular profiling data for large cohorts of tumors. Typical computational studies of these data sets seek to identify cancer subtypes, often by clustering samples by mRNA expression profiles. However, the problem of deciphering the gene regulatory programs that may underlie these different subtypes has received less computational attention. In this work, we present a statistical framework for inferring both common and subtype-specific transcriptional(TFs) and microRNA(miRs)-mediated dysregulation in cancer from multimodal tumor data sets.
We developed our approach to analyze a data set of 188 TCGA glioblastoma (GBM) tumor samples for which mRNA, miR, and copy number profiles were available. To identify potential TFs and miRs involved in dysregulation in GBM, we first used motif analysis to identify putative TF binding sites in promoters and miR sites in 3' UTRs. We then trained a sparse regression model to predict mRNA expression changes in each GBM sample relative to normal brain from the presence of these regulatory elements. Here, sparsity involves imposing a lasso constraint so that relatively few features contribute to the regression model. This constraint avoids overfitting and also leads to more interpretable results. By cross-validation on held-out genes, we found that these models do account for a significant part of the differential mRNA expression in GBM. Moreover, cross-validation performance improved when including copy numbers as a feature in the regression model. We also filtered miRs based on their differential expression relative to normal brain prior to training in order to restrict to more confident candidate regulators.
Analysis by the TCGA project has categorized samples into three well-defined classes: proneural, classical, mesenchymal and a poorly defined neuronal class, based on a combination of expression phenotype and clinical information. After removing neuronal samples, we found that our sample-specific models indeed clustered by subtype. Furthermore, statistical analysis of model parameters identified TFs and miRNAs that are deregulated across GBM subtypes (e.g. miR-124 and miR-155) or in a specific subtype. Interestingly, the REST TF, a repressor of neuronal genes in non-neuronal cells, was identified as upregulated in the mesenchymal class, while miR-9, which silences REST, was identified as downregulated. Our statistical framework provides a powerful tool for deriving mechanistic hypotheses about dysregulation of gene regulatory programs in cancer.

A DNA-TRANSPOSON BASED APPROACH TO FUNCTIONAL SCREENING IN NEURAL STEM CELLS

Ilaria Albieri[1], Marco Onorati[2], Giovanna Calabrese[3], Zoltan Ivics[4], Elena Cattaneo[2], Gian Giacomo Consalez[1]

[1]San Raffaele Scientific Institute, Neuroscience Division, via Olgettina,58, Milan, 20132, Italy, [2]University of Milan, Department of Pharmacological Science and Center for Stem Cell Research, via Viotti, 3/5, Milan, 20133, Italy, [3]University of Catania, Department of Physiological Science, Viale A.Doria, 6, Catania, 95125, Italy, [4] Max Delbruck Center for Molecular Medicine, MDC, Robert-Rössle-Straße 10, Berlin, 13125, Germany

Mouse Embryonic Stem (ES) cell-derived Neural Stem (NS) cells can be propagated in vitro, in presence of bFGF2 and EGF, as an adherent monolayer and expand indefinitely as an homogeneous population of radial glial-like progenitors capable of self renewal and symmetric cell division. Proliferating NS cells uniformly express Nestin, BLBP, Sox2 and other neural markers but lack the expression of Sox1, the earliest neuroephitelial marker. Long-term self-renewing NS cells display a prevalent ventral identity, respond to bone morphogenetic protein family molecules by activating a glial gene expression program and have been shown to terminally differentiate into GABAergic neurons. Their responses to developmental signals appear to be restricted to late neuronal and glial fate, revealing a late embryonic progenitor identity.
By manipulating the developmental plasticity and the differentiation potential of NS cells, considerable insight shoud be gained about the pathways and intrinsic determinants that play a main role in neural stem cell biology and differentiation. To this aim, we are developing a gain of function screen into a Sox1-reporter NS cell line (Ying et al., 2003; Conti et al., 2005) using a transposon-based approach. By transfection of NS cells with a Sleeping Beauty transposable element containg a Gal4/UAS system we are able to randomly transactivate genes around the transposon integration site, modyfing the expression profile of NS cells. In order to assess the suitability of our approach, a preliminary characterization has been performed to test the following features: (1) integration efficiency of SB transposon system in NS cell genome; (2) the average number of SB transposon integrations per clone; (3) the stability over time of transposon DNA integrations (4) the distribution of SB transposon integrations; (5) the ability of Gal4/UAS system to transactivate genes across the direct/inverted repeats (IR/DRs) that flank the SB transposable element. In the future, we will also attempt to affect the plasticity of this line by identifying and manipulating epigenetic signatures that restrict its developmental competence.

SMALL RNA REGULATION IS DEPENDENT ON TARGET CONCENTRATION

Aaron Arvey[1], Christina Leslie[1], Debora S Marks[2]

[1]Memorial Sloan Kettering Cancer Center, Computational Biology, 408 E69th St, Zuckerman 1133, New York, NY, 10021, [2]Harvard Medical School, Systems Biology, 200 Longwood Ave, Alpert 536, Boston, MA, 02115

According to basic chemical kinetics, regulation of mRNAs by small RNAs depends on the concentration of the partners. Therefore, the concentration of target sites for microRNAs or siRNAs should determine both the specificity and sensitivity of such regulation. However, while small RNA dosage has been extensively explored in many experiments, much less is known about the effects of the natural variation in the cell of concentration of target sites for each microRNA or siRNA. Our hypothesis is that the amount of competition between target sites (in mRNAs) for a limited number of active small RNAs (in RISC) should determine how much a small RNA can down-regulate each of its target mRNAs.

To test this hypothesis we considered all genes targeted by two different transfected microRNAs in the same cell type in ~50 experiments and whether the total concentration of all targets for a microRNA in a cell affects the amount of down-regulation on the average individual target. Specifically, for each set of genes targeted by a pair of microRNAs, miR-A and miR-B, we compared the difference in mean log expression change in the transfections of miR-A and miR-B and the difference in total target abundance for these two microRNAs (estimated from RNA-seq data). Consistent with our hypothesis, we found a significant correlation between these two differences ($p < 1e\text{-}5$, conservative empirical p-value). That is, greater target abundance correlates with smaller down regulation. Similarly, we also counted the number of target sites per miRNA and the number of mRNA transcripts and discovered a correlation between the number of mRNA targets in the cell and the mean amount of down-regulation for all the predicted targets of the microRNA (Spearman correlation 0.58, $p<0.01$).

Better understanding target concentration will shed insight into several critical problems to the community, including better understanding of (i) the varying levels of regulation by miRNAs, (ii) the effects of miRNA transfection studies, and (iii) designing siRNA screens and siRNA therapeutics.

FINE TUNING OF SIGNALING NETWORKS BY MICRORNAS

Avraham Roi, Yarden Yosef

Weizmann Institute of Science, Biological Regulation, Hertzel 1, Rehovot, 76100, Israel

microRNAs are involved in cancer and in signal transduction, but understanding their roles in signaling systems is still lacking. Here we study signal transduction by the epidermal growth factor (EGF), a process overly active in tumors, with the aim of deciphering the cellular functions of microRNAs and their interfaces with transcription factors (TFs). Employing genome-wide analyses of mRNAs and microRNAs, we uncovered a remarkably dynamic circuitry, densely populated by recurring elements incorporating microRNAs and TFs that converge on a common predicted gene targets. Furthermore, experimental data will be presented to exemplify these behaviors in the context of a small network of microRNAs and TFs. This network is further shown to be implicated in oncogenic processes, in association with clinical and molecular parameters of aggressive disease.

QUANTITATIVE GENOME-WIDE ANALYSIS OF CHROMATIN-REMODELING DURING 3T3-L1 ADIPOCYTE DIFFERENTIATION

Songjoon Baek[1], Rasmus Siersbaek[2], Ronni Nielsen[2], Sam John[1], Myong-Hee Sung[1], Susanne Mandrup[2], Gordon L Hager[1]

[1]National Cancer Institute, NIH, Laboratory of Receptor Biology and Gene Expression, Center for Cancer Research, 41 Library Dr., Bethesda, MD, 20892, [2]University of Southern Denmark, Department of Biochemistry and Molecular Biology, Campusvej 55, Odense M, DK-5230, Denmark

It is now well established that the chromatin template in which the DNA is imbedded plays a major role in the regulation of gene expression. In addition to serving as a scaffold for DNA, histone proteins, which constitute the core of the chromatin template, are subject to extensive modifications (e.g. acetylation, methylation, and phosphorylation), which may serve as docking sites for transcriptional regulators. In addition, histone modifications and remodeling factors determine the way the DNA wraps around the histones and the position of histones along the DNA, thereby regulating the accessibility of the DNA template for sequence-specific transcription factors. The susceptibility of DNA to DNaseI cutting can be used as a measure of the accessibility of the chromatin template. Chromosomal regions that have an open chromatin structure are hypersensitive to DNaseI cutting and are consequently called DNaseI hypersensitive sites (DHS).
The peroxisome proliferator-activated receptor γ (PPARγ) is a key transcriptional regulator of adipocyte differentiation and recent genome-wide binding profiles of PPARγ have implicated this receptor in many aspects of adipocyte biology. In this study we investigate how the chromatin landscape changes during adipocyte differentiation in terms of DNaseI sensitivity and correlate these changes to the binding profiles of PPARγ and other transcription factors.
The analysis of DHS transitions on the genome scale is central to understanding chromatin organization during these developmental processes. We present a quantitative algorithm for precise identification of DHS from short reads generated from DNase-Seq experiments. This algorithm will identify putative binding sites with a statistical score and compensates for the genomic variation in the mappability of sequences and local biases in the genome. The significance of enrichment is determined by thresholds based on false discovery rate of randomly generated tags. MEME motif-discovery tool-based analyses and genomic annotations of the enriched regions provide further insight into the chromatin-remodeling events during 3T3-L1 adipocyte differentiation.

INTRONIC ANTISENSE NONCODING RNA AS A CANDIDATE REGULATOR OF RASSF1 GENES.

Felipe C Beckedorff[1], Renan Crocci-Souza[1], Helder I Nakaya[1], Daniela T Soltys[2], Carlos F Menck[2], Eduardo M Reis[1], Sergio Verjovski-Almeida[1]
[1]University of São Paulo, 1Departamento de Bioquímica, Av Prof. Lineu Prestes, 748, São Paulo, 05508-900, Brazil, [2]University of São Paulo, 2Departamento de Microbiologia, Av Prof. Lineu Prestes, 1374, São Paulo, 05508-900, Brazil

Recent improvements in high-throughput gene expression analysis have led to the discovery that noncoding RNAs (ncRNAs) represent the majority of human transcriptional units. Uncovering this pervasive transcription has been followed by characterization of many novel functional regulatory ncRNAs; many classes of ncRNAs are now extensively studied and their regulatory roles are being broadly recognized. In this context, our group has used in silico approaches based on genomic mapping and clustering of ESTs, together with microarray experiments using combined intron/exon oligoarrays, to catalogue about 78,000 long intronic ncRNAs; most of them possibly represent novel ncRNAs. We were also able to point to intronic regions as key sources of potentially regulatory ncRNAs, especially because transcription arises from introns of protein-coding genes significantly enriched in “regulation of transcription” function. In this work, we describe a novel ncRNA transcribed from the opposite strand of the mitosis-suppressor gene, RASSF1. This antisense RNA was confirmed by strand-specific RT-PCR experiments in two human cell lines, namely HeLa and HEK293, and it was mapped to an intron of RASSF1A gene just upstream to the RASSF1C gene transcription start site. This novel antisense gene was designated as RASSF1 AS RNA. The RNA levels of RASSF1A genes and RASSF1 AS RNA in the two cell lines were determined by real-time PCR experiments. Interestingly, a negative correlation between the expression levels of antisense RNA and RASSF1A gene was found, suggesting that the AS RNA could regulate the expression of RASSF1A gene. To address this question, we over-expressed RASSF1 AS RNA in HeLa cells. Two independent transfection assays were performed and, in both cases, the overexpression of RASSF1 AS RNA promoted a decrease in the RNA levels of RASSF1A transcripts. In order to find if the down-regulation of RASSF1A gene promoted an effect at the cellular level, we performed cell-proliferation assays in HeLa with or without overexpression of RASSF1 AS RNA and detect that cells overexpressing the ncRNA proliferated more rapidly with a median 1.17 fold increase in proliferation rate. We have also identified a putative promoter sequence for RASSF1 AS RNA by in-silico analysis and cloned this region upstream to the luciferase gene in a promoter-reporter vector. The activity of the intronic antisense RNA promoter was then confirmed in HeLa cells. This study sheds further light on new mechanisms that involve ncRNAs as important players in regulating transcription.

A CONSENSUS OF CORE PROTEIN COMPLEX COMPOSITIONS FOR SACCHAROMYCES CEREVISIAE

Joris Benschop, Nathalie Brabers, Dik van Leenen, Philip Lijnzaad, Frank Holstege, Patrick Kemmeren

University Medical Center Utrecht, dpt of Physiological Chemistry, Universiteitsweg 100, Utrecht, 3584CG, Netherlands

Analyses of biological processes would benefit from accurate definitions of protein complexes. High-throughput mass-spectrometry data offers the possibility of systematically defining protein complexes, however the predicted compositions vary substantially depending on the algorithm applied. We determine consensus compositions for 409 core protein complexes from Saccharomyces cerevisiae, by merging previous predictions with a new approach. Various analyses indicate that the consensus is comprehensive and of high quality. For 85 out of 259 complexes not recorded in GO, literature search revealed strong support in the form of co-precipitation. New complexes were verified by an independent interaction assay and by gene expression profiling of strains with deleted subunits, often revealing which cellular processes are affected. The consensus complexes are available in various formats including a merge with GO, resulting in 518 protein complex compositions. The utility is further demonstrated by comparison with binary interaction data to reveal interactions between core complexes.

GENE EXPRESSION PROFILING OF BOTH PROTEIN-CODING AND NON-CODING RNA TRANSCRIPTS FROM SMALL AMOUNTS OF TOTAL RNA USING A SINGLE MICROARRAY DESIGN

Anne Bergstrom Lucas[1], Mitchell Guttman[2], Peter Tsang[3], Bo Curry[3], Vinayak Kulkarni[1], Stephanie Fulmer-Smentek[1], Anya Tsalenko[3], Sharoni Jacobs[1], John Rinn[2], Laurakay Bruhn[3]

[1]Agilent Technologies, Genomics R&D, 5301 Stevens Creek Blvd., Santa Clara, CA, 95051, [2]Broad Institute of MIT and Harvard, Department of Biology, 7 Cambridge Center, Cambridge, MA, 02142, [3]Agilent Technologies, Agilent Labs, 5301 Stevens Creek Blvd., Santa Clara, CA, 95051

Recently, thousands of large intergenic non-coding RNA (lincRNA) transcripts were identified in mouse and human cells from genome-wide chromatin-state maps. Through the comparison of lincRNA expression profiles to known protein-coding gene pathways, lincRNAs have been implicated in diverse biological processes such as stem cell pluripotency, innate immunity, and cell cycle regulation. In addition, many lincRNAs have been shown to play a role in these pathways through regulation of gene expression. To enable systematic profiling of all lincRNAs and protein-coding genes, we have developed human and mouse microarrays comprised of all known protein-coding mRNAs and lincRNAs allowing for simultaneous detection from a single sample. In validating these new array designs, we labeled mouse RNA samples using a new labeling kit, allowing us to start with total RNA input amounts as low as 10 nanograms. The new protocol uses a single round of amplification by *in vitro* transcription without purification of the cDNA product, yielding Cy-labeled cRNA in less than one day. Thus, these new array designs and labeling procedure allow rapid monitoring of protein-coding and non-coding genes from small amounts of starting material. RNA samples from a variety of mouse tissues were labeled and applied to the new mouse arrays to detect differences in coding and non-coding gene expression profiles. Comparisons of probe signals from technical replicate samples demonstrated high reproducibility with wide dynamic ranges and high sensitivity. Using this approach, we are able to identify novel differentially expressed lincRNAs in diverse cellular processes including stem cell differentiation and innate immunity. In addition, many of these lincRNAs strongly associate with key genes known to regulate these biological processes. Here we show examples of how profiling mRNA and lincRNA from the same sample can allow researchers to further define the role of lincRNAs in gene regulation. Together, these results highlight the power of simultaneously profiling lincRNAs and protein-coding genes on the same array.

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

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

[1]Duke University, Institute for Genome Sciences & Policy, 101 Science Dr., Durham, NC, 27708, [2]University of Texas at Austin, Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, 1 University Station, Austin, TX, 78712, [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

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

ALTERED GENOMIC TARGETING OF THE ONCOGENIC TRANSCRIPTION FACTOR TAL1/SCL IN T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

Carmen Palii[1], Carolina Perez-Iratxeta[1], Zizhen Yao[2], Yi Cao[2], Harold Atkins[1], David Allan[1], Jeffrey F Dilworth[1], Robert Gentleman[2], Steven Tapscott[2], Marjorie Brand[1]

[1]OHRI-Sprott Center, Regenerative Medicine, Smyth Road, Ottawa, K1H 8L6, Canada, [2]FHCRC, Basic Sciences, Fairview Avenue, Seattle, WA, 98109

The transcription factor TAL1/SCL is a dual role protein. While its expression is required for cell differentiation in the erythroid lineage, TAL1 is an important contributor to oncogenic transformation in the T-cell lineage, leading to T-cell acute lymphoblastic leukemia (T-ALL). Currently, the extent to which the cellular environment affects the function of this transcription factor is unclear.

Here, we have used a comparative strategy combining gene expression profiling after knockdown and ChIP-sequencing in leukemic T-cells versus non-leukemic erythroid cells to identify the molecular determinants that distinguish the role(s) of TAL1 in these two human cell-types. For these studies, the leukemic context is represented by the TAL1-expressing jurkat cell line, which was originally derived from a T-ALL patient and by TAL1-expressing blasts from T-ALL patients. To study TAL1 in normal conditions, we used primary erythroid cells differentiated ex vivo from human hematopoietic stem cells.

We found that wild-type TAL1 displays different DNA binding profiles in these hematopoietic cell types. In addition, TAL1 is preferentially targeted to distinct composite DNA binding motifs depending on the cellular environment. Unexpectedly, the alteration of TAL1 genomic binding selectivity in a T-ALL environment provides this factor with access to the transcriptional regulatory network of important T-cell regulators leading to a block in T-cell differentiation and oncogenesis.

Taken together, these findings reveal an unexpected contribution of the cellular environment in modulating the activity of a transcription factor via alternate genomic targeting, and underscore how changing the cellular context can render a transcription factor oncogenic.

INTEGRATING HETEROGENEOUS DATASETS TO PREDICT ACTIVE PROMOTERS, REGIONS OF REGULATORY IMPORTANCE, AND CHARACTERIZE GENE REGULATORY MECHANISMS IN DROSOPHILA

Christopher A Bristow[1,2], Pouya Kheradpour[1], Charlie Frogner[3], Rachel S Sealfon[1], Tomaso Poggio[3], Manolis Kellis[1,2]
[1]MIT, CSAIL, 77 Massachusetts Ave, Cambridge, MA, 02139, [2]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02139, [3]MIT, Center for Biological and Computational Learning, McGovern Institute, 77 Massachusetts Ave, Cambridge, MA, 02139

Drosophila provides a key model system for studying the control of gene expression during development. The modENCODE project has generated chromatin mark profiles across a developmental time-course and is accumulating a large compendium of transcription factor (TF) binding experiments with the goal of characterizing gene regulation. In addition, extensive spatial and temporal mRNA expression data are available as 6000 genes have been profiled by *in situ* hybridization during embryogenesis. These recent experimental studies provide new data that offers key insights into the mechanisms of gene regulation during development.

By integrating these datasets, we identify active promoters, regions of regulatory importance, and characterize the gene regulatory mechanisms involved. We have developed a method to classify active promoters using chromatin mark profiles with high specificity and sensitivity, enabling the characterization of promoter dynamics and identifying potential new promoters.

To identify regions of regulatory importance, we focus on CBP, a protein that enhances TF binding and has been used to successfully identify regulatory regions in mouse. We grouped CBP regions based on the set of TFs and chromatin remodeling factors also bound in these regions using a large set of >100 ChIP datasets from the modENCODE group and the literature using a finite mixture model of multivariate Bernoulli distributions. This approach indentified subsets of CBP regions bound by TFs known to physically interact with CBP, such as bicoid and dorsal, and these regions are highly enriched for known enhancers. We are currently using the CBP classifications to generate predictions that will be tested by the Posakony lab with *in vivo* transgenic reporter assays.

We have also developed algorithms to extract a robust representation of recurrent spatial gene expression patterns in a systematic and unbiased way. We clustered the extracted mRNA stain patterns based on a spatial similarity metric to assemble groups of genes that show coherent expression patterns. The inferred clusters show specific enrichment in known regulatory motifs, and binding data, associated with transcription factors involved in embryo development, and suggest specific regulatory connections to candidate regulators for these recurrent patterns.

The chromatin mark profiles and transcription factor binding data used in this analysis were generated by Kevin White and the cisreg modENCODE group. The *in situ* hybridization data was generated by Sue Celniker and the BDGP.

QUANTITATIVE ANALYSIS OF THE DROSOPHILA SEGMENTATION REGULATORY NETWORK USING PATTERN GENERATING POTENTIALS

Majid Kazemian[1], Charles Blatti[1], Sudhir Kumar[2], Scot Wolfe[3], Saurabh Sinha[1], Michael Brodsky[3]
[1]University of Illinois at Urbana-Champaign, Department of Computer Science, 201 N. Goodwin Avenue, Urbana, IL, 61801-2302, [2]Arizona State University, Center for Evolutionary Functional Genomics, 727 East Tyler Street, Tempe, AZ, 85287-5301, [3]University of Massachusetts Medical School, Program in Gene Function and Expression, 364 Plantation St, Worcester, MA, 01605-2324

Computational discovery of cis-regulatory modules commonly relies on identifying statistically significant clusters of putative binding sites for transcription factors that act in a common regulatory network. We describe a new computational strategy to use both the DNA binding specificity and the expression pattern of transcription factors to annotate genomic sequences based on "pattern generating potential". We use this approach to convert the extensive but qualitative understanding of interactions that regulate Drosophila anterior-posterior (A-P) patterning into a network model in which a confidence value is associated with each transcription factor-module interaction. In this approach, transcription factor binding motif profiles are first combined across multiple species using a statistical method that accommodates an irregular distribution of evolutionary distances. Next, a logistic regression model is used to combine these binding site profiles with transcription factor expression information to predict module activity patterns; interestingly, this simple model was as effective as a previous thermodynamics-based model in predicting activity patterns from sequence. This model is then used to identify modules by scanning genomic sequences for the potential to generate all or part of the expression pattern of a flanking gene. Finally, an in silico genetic analysis is used to infer edges in the transcriptional regulatory network; each edge in this network depicts the direct contribution of individual factors to individual modules and has an associated estimate of statistical significance.

We use pattern generating potential to systematically describe the location and function of cis-regulatory modules in the A-P patterning network. We identify known and new genomic regions with the potential to direct a specific part of a gene's expression pattern. We also identify many examples of two or more modules for a gene that are predicted to have closely related activity patterns. Surprisingly, predictions using conserved transcription factor binding site frequencies were more highly enriched in known factor-modules interactions than predictions using chromatin immunoprecipitation data. We propose that measurement of pattern generating potential provides a general solution to integrate genome sequence, expression data and factor binding specificities to describe the complex transcriptional regulatory networks that function during metazoan development.

IN VIVO QUANTIFICATION OF DYNAMIC GENE EXPRESSION IN THE ARABIDOPSIS ROOT

Wolfgang Busch[1], Richard W Twigg[1], Daniel L Mace[2], Bradley Martsberger[1], Uwe Ohler[2], Philip N Benfey[1]

[1]Duke University, Department of Biology and IGSP Center for Systems Biology, 124 Science Drive, Durham, NC, 27708, [2]Duke University, Institute for Genome Sciences and Policy, 101 Science Drive, Durham, NC, 27708

Gene expression is a dynamic continuous phenomenon. To describe the expression of a gene it is therefore important to quantitatively capture spatio-temporal patterns of gene expression under defined conditions. As for many other quantitative experiments, it is essential to acquire measurements from multiple individuals. As yet, it is almost impossible to perform live imaging with cellular resolution on developing organs in multiple replicates and under different environmental conditions.
We developed a microfluidics device, called the RootArray, which enables such studies. It permits more than 60 roots to be grown in parallel and to be imaged in short time intervals by confocal microsopy. The design of the root array allows for rapid exchange of growth media to alter environmental conditions and to observe subsequent alterations of gene expression. Our pipeline includes automated tracking and detection of growing roots, automated image acquisition and gene expression quantification. We currently use promotor:GFP as well as GFP fusion protein-based reporter lines to systematically capture expression patterns in the developing root. To assess the dynamics of gene expression, time courses and perturbations with changing media composition are conducted. The use of the root array can be extended to a variety of purposes that involve measurements at cell type resolution. For instance, we use molecular FRET sensors to conduct *in vivo* sensing of metabolites in the developing root in different growth media.

EFFECT OF DNA METHYLATION ON THE TRANSCRIPTION OF INTRONIC NONCODING RNAS IN CANCER CELL LINES

Lauren Camargo, Vinícius Maracajá-Coutinho, Sergio Verjovski-Almeida, Eduardo M Reis

Instituto de Química, Departamento de Bioquímica, Av. Profº Lineu Prestes, São Paulo, 05508-000, Brazil

DNA methylation is an epigenetic mechanism of control of gene expression in normal and pathological processes, such as cancer. There is mounting evidence in the literature correlating downregulation of several tumor suppressor genes with DNA methylation of CpG islands (CGIs) located on their promoter regions. Genome-wide tiling array experiments have shown the transcriptional output of eukaryotic organisms is largely composed of noncoding RNAs (ncRNAs), some of which were shown to play important regulatory functions in the cell gene expression program. Our group is characterizing a subclass of ncRNAs that are transcribed within intronic regions of protein-coding genes. We have shown that the expression levels of a subset of these correlate with the degree of prostate tumor malignancy (Reis et al., Oncogene 2004, 23:6684). In this work we examined the effect of chemical demethylation of DNA on the expression pattern of intronic ncRNAs in three cancer cell lines (prostate cell line DU-145, breast cell line MCF-7 and pancreas cell line Mia PaCa 2). RNA isolated from each cell line treated with the demethylating agent 5-Aza-deoxicytidine (5–AZA) or from untreated control cells were labeled and hybridized to custom designed exon-intron Agilent oligoarrays containing probes for 14,270 totally intronic ncRNAs (sense/antisense TINs), 4,439 antisense partially intronic ncRNAs (antisense PINs) (i.e. overlapping an exon of the corresponding protein-coding transcript) and 13,219 exons of protein-coding genes. Following data extraction, filtering and normalization, approximately 2%, 5% and 16% of the protein-coding exonic transcripts detected in all hybridizations were differentially expressed, in DU-145, MCF-7 and Mia PaCa 2, respectively, (False-Discovery Rate ≤ 1%) following 5-AZA treatment. The effect of DNA demethylation on the expression of intronic ncRNAs was more subtle; only 0.4%, 1.2% and 4.4% of detected intronic ncRNAs were deregulated after 5-AZA-treatment, in DU-145, MCF-7 and Mia PaCa 2 cancer cell lines, respectively. Most 5-AZA-responsive intronic ncRNAs were upregulated, suggesting a direct effect of demethylation of promoter regions in the expression of these transcripts. Inspection of the genomic regions upstream of intronic ncRNAs revealed the presence of CGIs in 31% (DU-145), 26% (MCF-7) and 29% (Mia PaCa 2) of 5-AZA regulated transcripts. We are currently mapping the promoter region and characterizing the CGI methylation status of selected 5-AZA-responsive intronic ncRNAs. In conclusion, our results indicate that CpG island methylation influences the transcription of a subset of intronic ncRNAs in a cell type specific manner.

COMBINATORIAL BINDING CODES OF TRANSCRIPTION FACTORS AND NUCLEOSOMES

Varodom Charoensawan, Sarath C Janga, Madan M Babu, Sarah A Teichmann

MRC Laboratory of Molecular Biology, Structural Studies, Hills road, Cambridge, CB2 0QH, United Kingdom

Transcription factors (TFs) are important to the establishment and maintenance of cellular phenotypes because they bind to specific DNA sequences and directly control the expression of their target genes. In eukaryotes, the likelihood that TFs bind to DNA and thus regulate transcription is not entirely determined by their intrinsic binding affinity (*i.e.* personal preference to DBD sequences). This is because different TFs have to compete with one another, as well as with other DNA-binding proteins such as nucleosomes. Here, we investigate the relationship between sites on DNA bound by TFs and nuclesomes in budding yeast. We integrate DNA-binding specificities of TFs from high-throughput experiments (*in vitro* and *in vivo*) with various genome-wide studies of nucleosome position. This allows us to estimate the extent to which the intrinsic DNA-binding affinities of TFs and nucleosomes play a role in regulating their *in vivo* binding. Based on the minimal intrinsic affinity of sites bound by TFs *in vivo*, we calculate that as small fraction as approximately 0.1% of putative sites in the yeast genome are actually bound by TFs. On naked DNA, nearly 70% of the yeast TFs with binding specificity data available would prefer to bind to sites similar to those preferred by nucleosomes. Although most TF binding sites are occluded by well-positioned nucleosomes in purified chromatin, we find evidence showing that these nucleosomes are repositioned to less favoured positions upon TF binding *in vivo*. The ultimate outcome is a greater proportion of accessible TF binding sites in promoters. We also find the cooperative binding of TFs in proximal sites to be one of the mechanisms that facilitate nucleosome occlusion *in vivo*.

THE LIMITED SHARING OF FACTOR OCCUPANCY BETWEEN SPECIES IS ENRICHED FOR INDUCTION OF GENE EXPRESSION

Yong Cheng, David C King, Quanbei Chen, Ross C Hardison

The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, 303 wartik, university park, PA, 16801

Recent genome-wide mapping of DNA segments occupied *in vivo* by transcription factors reveals many insights into gene regulation. However, we do not know how often factor occupancy is conserved between species or how strong the evolutionary constraint is on binding sites motif. Furthermore, efforts to correlate these evolutionary patterns with functional inference are in their infancy. To address these questions, we mapped occupancy by the hematopoietic regulatory factor GATA1 throughout the mouse erythroid genome and compared the GATA1 occupancy to that in human erythroid cells. Almost all the GATA1 occupied segments (GATA1 OSs) contain the specific binding sites motif WGATAR, and we evaluated the phylogenetic depth of preservation of those motifs. About 15000 GATA1 OSs are identified in each species but only around 1000 are bound in both mouse and human. Those GATA1 OSs shared by mouse and human have a higher level of GATA1 occupancy than the lineage specific GATA1 OSs. The WGATAR motifs preserved in multiple mammalian lineages are under strong evolutionary constraint. The constrained motifs are enriched in the shared GATA1 OSs. GO category results further show that genes close to shared GATA1 OSs and likely regulated by them are enriched in well-known erythroid function such as heme biosyntheisis, cytoskeleton, endosomes. The evolutionary history of most GATA1 OSs associated with gene induced during late erythroid maturation differs dramatically from that of the GATA1 OSs close to repressive genes. GATA1 OSs shared between human and mouse are enriched in induced genes whereas lineage specific GATA1OSs are distributed equally among up and down regulated genes. This suggests the regulatory mechanisms for up-regulated genes are under stronger purifying selection than those for down-regulated genes.

NFKB P50 RESTRICTS THE INTERFERON RESPONSE BY BINDING IRE SEQUENCES

Christine S Cheng[1,2,3], Kristyn Feldman[1,2], James Lee[1,2], Shilpi Verma[4], De-Bin Huang[2], Kim Huynh[2], Chris Benedict[4], Gourisankar Ghosh[2], Alexander Hoffman[1,2,3]

[1]University of California, San Diego, Chemistry & Biochemistry, Signaling Systems Laboratory, Bioinformatics & systems Biology, 9500 Gilman Drive, La Jolla, CA, 92093, [2]University of California, San Diego, Bioinformatics & systems Biology, 9500 Gilman Drive, La Jolla, CA, 92093, [3]University of California, San Diego, Bioinformatics & systems Biology, 9500 Gilman Drive, La Jolla, CA, 92093, [4] LIAI, Div. of Molecular Immunology, 9500 Gilman Drive, La Jolla, CA, 92037, [5]niversity of California, San Diego, Chemistry & Biochemistry, 9500 Gilman Drive, La Jolla, CA, 92093, [6]niversity of California, San Diego, Chemistry & Biochemistry, 9500 Gilman Drive, La Jolla, CA, 92093, [7]niversity of California, San Diego, Chemistry & Biochemistry, 9500 Gilman Drive, La Jolla, CA, 92093

B p50's critical function in repressing triple-G IREs is to enforce the requirement for transcriptional synergy on combinatorial promoter AND gates; this ensures pathogen-appropriate immune responses and suppresses unneeded, cytotoxic IFN signaling.κ. Our results from these top-down and bottom-up systems biology approaches demonstrate that NFβB sites; predictions were then tested experimentally by examining the combinatorially controlled anti-viral master regulator IFNκB p50 homodimer. Unbiased genome-wide expression and bioinformatic analysis revealed that p50 attenuates a subset of interferon-inducible genes. Biophysical studies explain how p50 binds newly defined composite triple-G IREs. To develop hypotheses about the functional consequences, we constructed mathematical models of alternate promoter architectures, including promoters synergistically controlled by IRE and κB-site and the interferon regulatory element (IRE), respectively. Here, we report unexpected cross-regulation of IREs by the abundant NFκB and IRF, which are known to bind the κSpecific binding by transcription factors to cognate sequence elements is thought to be critical for generating stimulus-specific gene expression programs. Cells produce pathogen-specific inflammatory and innate immune responses via NF

A METHOD FOR OPTIMIZING GENE COMBINATION TO INDUCE ADIPOCYTE DIFFERENTIATION FROM MESENCHYMAL STEM CELLS

Hirokazu Chiba[1], Taku Tanaka[2], Masato Miyake[2], Wataru Fujibuchi[1]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Computational Biology Research Center (CBRC), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan, [2]National Institute of Advanced Industrial Science and Technology (AIST), Research Institute for Cell Engineering (RICE), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Following the recent establishment of iPS (induced pluripotent stem) cells, efficient induction of specific tissues from stem cells has become increasingly important. Based on comparisons of gene expression data between stem and other differentiated cells, we have developed bioinformatics methods for predicting candidate genes that might be responsible for cell differentiation. With our recent analysis of transcription factor (TF) activities, we observe that the averaged expression levels of TFs tend to decrease as cell differentiation proceeds. Assuming the above TF reverse-activity model, we search for TFs that decrease their expression levels over the differentiation from human mesenchymal stem cells (hMSCs) into adipocytes. As a result of comparison of expression profiles taken from adipose tissues and hMSCs, we detect 55 TFs that might be involved in adipocyte differentiation. Among them, several TFs are reported as negative regulators of adipogenesis (fat generation). When we experimentally evaluate if adipocytes are induced by knock-down of other TFs, we observe that introduction of siRNAs against highly-ranked TFs can partially replace the function of a chemical inducer, 3-isobutyl-1-methyl-xanthine (IBMX). These results suggest that our model and approach be effective in finding novel TFs that play key roles in the stem cell differentiation.

QUANTITATIVE MODELS OF TRANSCRIPTION FACTOR SPECIFICITY USING HIGH THROUGHPUT SEQUENCING AND A BACTERIAL ONE HYBRID ASSAY

Ryan G Christensen, Gary D Stormo

Washington University in St Louis, Genetics, 4444 Forest Park Parkway; Camp Box 8510, St Louis, MO, 63108

The bacterial one hybrid (B1H) developed by Meng et al. (Nat Biotechnol, 2005) provides a simple and rapid way to characterize the DNA specificity of a transcription factor (TF) without the need for protein purification. The B1H assay was employed successfully in several studies to characterize the DNA binding preferences of more than a hundred different TFs. These studies all used low throughput Sanger sequencing to identity about 20 transcription factor binding sites (TFBSs) per TF. In this study, the high throughput Illumina Solexa sequencing platform was used, yielding thousands of TFBSs per TF. We present results for 96 homeodomain proteins and for the well studied C2H2 zinc finger, Zif268, and 14 Zif268 mutants. Having thousands of sites per TF allowed us to formulate a nonlinear, biophysically motivated model that relates a binding site's composition to the bacterial growth rate. The model includes terms for the relative energy contribution of each base at each position in the TFBS, as well as a term for the chemical potential of the TF and a term relating the probability of a TFBS being bound to the bacterial growth rate. We show through simulations and by comparison to existing position specific weight matrices (PWMs) that our model recovers quantitative PWMs. We also discuss how the growing catalog of high quality PWMs can be used to learn DNA recognition models for different transcription factor families.

NUTRIENT SIGNAL INTEGRATION THROUGH PKA IN *SACCHAROMYCES CEREVISIAE*

Michael K Conway, Warren Heideman

University of Wisconsin at Madison, Pharmaceutical Science, 777 Highland Ave., Madison, WI, 53705

Recent work from our lab has shown that *Saccharomyces cerevisiae* starved for glucose, nitrogen, or phosphorous, all conditions that produce growth arrest, generate a very similar transcriptional response when the missing nutrient is supplied. As expected, growth related genes are upregulated while stress and alternative metabolism genes are downregulated. Despite unique nutrient limitations and nutrient specific receptor mechanisms, the transcriptional response is highly similar with only very small subsets of nutrient specific differences. The common transcriptional response, which involves approximately one-third of the genome, is uniformly dependent upon cAMP/PKA and to a lesser extent TOR. Such similarity despite different nutrient sensing mechanisms led us to address how yeast integrate nutrient information before producing the characteristic transcriptional response described above. In order to trigger the large transcriptional response and commence growth, do specific nutrient conditions need to be met? Herein we report our early findings. To establish nutrient requirements for signaling through PKA, we measured the effect of multiple nutrient limitation followed by individual nutrient repletion on the transcriptome. We explored the connection between specific nutrient sensors such as Gpr1, Gap1, and Pho84 with PKA signaling and whether each is required to produce the large transcriptional response to nutrient repletion. We also examined the ability of other sugars to stimulate or to antagonize the characterized nutrient response.

GLOBAL ANALYSIS OF CHROMATIN STATE AND GENE EXPRESSION IN DEVELOPING MAMMALIAN TISSUES.

Justin Cotney[1], Albert Ayoub[2], Jing Leng[3], Sunghee Oh[1], Shrikant Mane[4], Pasko Rakic[2,5], James Noonan[1,5]
[1]Yale University School of Medicine, Department of Genetics, 333 Cedar St., New Haven, CT, 06520-8005, [2]Yale University School of Medicine, Department of Neuroscience, 333 Cedar St., New Haven, CT, 06520-8005, [3]Yale University School of Medicine, Graduate Program in Computational Biology and Bioinformatics, 333 Cedar St., New Haven, CT, 06520-8005, [4] Yale University School of Medicine, Yale Center for Genome Analysis, 300 Heffernan Drive, West Haven, CT, 06516, [5]Yale University School of Medicine, Yale Kavli Institute for Neuroscience, 333 Cedar St., New Haven, CT, 06520-8001

Numerous studies of global chromatin architecture and gene expression have focused on cultured cells. While these efforts provide insight into mechanisms of genome regulation in cells, they have not revealed how these mechanisms are utilized to form complex multicellular structures during mammalian embryonic development. To address this question, we have developed methods to examine the dynamics of chromatin architecture and global gene expression during early development of the mouse and primate limb and cortex using limiting amounts of input material.

We are combining ChIP-Seq using marks for enhancer and chromatin boundary elements, Sono-Seq to distinguish euchromatin and heterochromatin domains, and RNA-Seq to create an aggregate map of chromatin status and gene expression in the developing mouse and primate limb. In contrast to previous ChIP-Seq studies using milligram quantities of chromatin, we have identified over 25000 sites bound by the insulator protein CTCF with only 15 µg of chromatin from limb buds of E10.5 mouse embryos. Initial analysis shows that over 60% of CTCF binding events in forelimb are shared with hindlimb. Comparisons to ChIP-ChIP and ChIP-Seq studies in human and mouse cultured cells reveal a core set of CTCF binding events common across tissues and species. Despite significant site overlap there are distinct differences in the profiles of CTCF occupancy between forelimb, hindlimb and human tissues, that could contribute to general developmental mechanisms or tissue-specific biology. We are currently expanding these studies by ChIP-Seq analysis of the enhancer related protein p300 and RNA-Seq to identify global patterns of active chromatin demarcated by CTCF binding.

For mouse cerebral cortex we aim to identify differences in global gene expression necessary for the development of early stem cell populations in this structure. We have optimized RNA-Seq methods to conduct transcriptome profiling of small amounts of total RNA (50 ng) extracted from laser microdissected tissue from E14.5 mouse ventricular and subventricular zones and cortical plate. Our analyses have revealed genes that are differentially expressed in and characteristic of each tissue. Confirmation of tissue specific patterns with in situ hybridization has demonstrated the utility of this method in identifying unique expression profiles. Current analysis is focused on identifying genes exhibiting tissue specific expression patterns not previously known participate in cortex formation or function. We are applying these methods to early primate and human tissues to identify changes in gene expression underlying species-specific mechanisms of development.

GLOBALLY UNIFORM TRANSITIONS IN TRANSCRIPTION COMPLEX COMPOSITION

Andreas Mayer, Michael Lidschreiber, Matthias Siebert, Kristin Leike, Johannes Soeding, Patrick Cramer

Ludwig-Maximilians-Universität, Gene Center and Department of Biochemistry, Feodor-Lynen-Strasse 25, Munich, 81377, Germany

It has long been known that RNA polymerase II (Pol II) assembles with general initiation factors on promoter DNA to form an initiation complex, which we have now elucidated structurally (Kostrewa et al., Nature 19 November 2009). However it has not been investigated systematically whether elongation and chromatin transcription factors are recruited to Pol II on all transcribed genes, or only on selected genes, and whether there is a general order of factor recruitment. Here we present unpublished high-resolution genomic elongation factor profiles obtained by chromatin immunoprecipitation in S. cerevisiae coupled to tiling microarray analysis. The data indicate a general, dynamic Pol II elongation complex that undergoes uniform transitions in factor composition. Downstream of the transcription start site, "primary" elongation factors enter, and peak levels of phosphorylation of the Pol II C-terminal domain (CTD) at serines Ser5 and Ser7 of the consensus heptameric repeat Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 are reached ("early 5'-transition"). Further downstream, "secondary" elongation factors and the CTD Ser2 kinase Ctk1 enter ("late 5'-transition"). Immediately upstream of the transcript termination site, levels of a subset of elongation factors decrease sharply, and Ser2 phosphorylation peaks ("3'-transition"). The three transitions in transcription complex composition occur at all protein-coding genes in yeast, independent of their function, length, and expression level.

MODELING YEAST TRANSCRIPT DEFINITION

Carl G de Boer[1], Harm van Bakel[2], Kyle Tsui[3], Joyce Li[2], Quaid D Morris[1,2], Corey Nislow[1,2], Jack F Greenblatt[2], Timothy R Hughes[1,2]

[1]University of Toronto, Department of Molecular Genetics, 160 College St., Toronto, M5S 3E1, Canada, [2]University of Toronto, Banting and Best Department of Medical Research, 160 College St., Toronto, M5S 3E1, Canada, [3]University of Toronto, Department of Pharmaceutical Sciences, 160 College St., Toronto, M5S 3E1, Canada

Transcript definition is central to the expression of genomic sequence. Yet, in general, we do not know what signals eukaryotic cells use to define individual genes. Here, we have derived classifiers that can predict promoters, terminators, and transcript bodies which rely solely on nucleic acid sequence features detectable by the cell through DNA- and RNA-binding proteins. The classifiers reveal which inputs are most influential overall, and which are most critical for individual transcripts. Our predictions of promoter-defining cis-elements are supported by the results of tiling array RNA analyses of the corresponding trans-acting factor mutants. Similarly, mutation of RNA-binding proteins involved in termination results in dramatic alteration of the transcriptome. Surprisingly, genes display a remarkable strand-specific base content bias: for example, there is a 1.2-fold enrichment of As over Us in yeast transcripts, perhaps as an evolutionary response to the relative abundance of ATP in the cell. This difference results in strand biases for many RNA-binding proteins in transcripts; for example, the termination factor Nab3 shows a 2-fold reduction of binding motifs in the sense compared to the antisense strand. Our current goal is to unify these three classifiers into a single model that describes genome-wide transcription and can predict the outcome of genetic perturbations on a nucleotide level.

SPECIFIC TRANSCRIPTIONAL REGULATORY CIRCUITS AND NODES ARE AFFECTED IN FAMILIAL COMBINED HYPERLIPIDEMIA SYNDROME AND UPON STATIN TREATMENT.

Francesco Vuolo, Mario Coiro, Luisa De Magistris, Michele Oliviero, Francesca Bergantino, Vincenzo De Simone

University "Federico II", Biochemistry and Medical Biotechnology, Via Pansini 5, Napoli, IT-80131, Italy

We have employed “network biology” computational tools to identify regulatory “nodes” or “modules” that are altered in the familial combined hyperlipidemia (FCHL), the most frequent multifactorial and genetically complex dislipidemic syndrome in our population.
By microarrays transcriptome analysis, we find 1913 genes hyper- or hypo-expressed) in FCHL patients, and 688 genes significantly altered after statin (the elective drugs for FCHL) treatment. Out of the 97 genes present in the intersection of these two lists, the majority are hypo-expressed in FCHL patients and become normo- or hyper-expressed after statin treatment. These genes have been selected for promoter analysis.
MEME, Jaspar and CisRed analysis of the -400/+1 regions of these genes, reveal seven most significantly enriched motifs. Many promoters contain two or more motifs. Electrophoretic mobility shift assays (EMSA) show that all these motifs bind specific nuclear proteins. Moreover, in transient expression assays many of these sequences display a clear transcriptional activatory function when cloned upstream of a reporter gene, in a dose-dependent manner.
Network analysis reveals a transcriptional regulatory circuitry connecting these genes. This circuitry is a specific node of a wider transcriptional network of genes involved in FCHL syndrome, and is a promising target for a specific regulatory “interference” approach to FCHL.

GENOME-WIDE MAPPING OF THE PRECISE DNA BINDING LOCATIONS FOR ~120 TRANSCRIPTION FACTORS USING A SINGLE DNASE-SEQ ASSAY.

Roger Pique-Regi[1], Jacob F Degner[1,2], Athma A Pai[1], Daniel J Gaffney[1,3], Yoav Gilad[1], Jonathan K Pritchard[1,3]
[1]University of Chicago, Department of Human Genetic, 920 E. 58th St., Chicago, IL, 60637, [2]University of Chicago, Committee on genetics genomics and systems biology, 920 E. 58th St., Chicago, IL, 60637, [3]University of Chicago, Howard Hughes Medical Institute, 920 E. 58th St., Chicago, IL, 60637
*RPR & JFD contributed equally

A complete map of the DNA targets of transcription factors will be fundamental to our understanding of the eukaryotic regulatory code. Here, we describe an alternative to ChIP that uses the DNaseI footprint of each TF given by genomewide DNaseI cutsite data. In contrast to ChIP, our method can assay many TFs in a single experiment, does not require a specific antibody, and defines the coordinates of TF binding with basepair level precision.
We have developed a novel Bayesian hierarchical mixture model that captures differences between bound and unbound TF motifs in the distribution of DNaseI sensitivity with singlebase resolution. Additionally, our model utilizes annotation information that affects our prior probability a site is bound (e.g. match to the TF position weight matrix (PWM), sequence conservation, and distance to the transcription start site). An Expectation Maximization (EM) algorithm is used fit the parameters of our model and we make our inference from an estimate of the posterior probability that a specific motif instance is active (i.e. bound by a TF).
We apply our model to every occurrence of each of the PWMs in the TRANSFAC and JASPAR databases (~1000 PWMs) using data from HAPMAP cell lines. For four of these PWMs, we use publicly available ChIPseq data from the Encode Project to validate our model and find that our predictions are quite congruent with ChIPseq results (area under curve for ROC curves > 0.95). For PWMs where publicly available ChIP data is unavailable, we use a likelihood ratio test to determine if a 2-component mixture model (corresponding to bound and unbound states) outperforms a comparable single component model. We find that this test statistic is highly correlated with the sequence conservation of our bound predictions and together with sequence conservation can be used to select the PWMs for which our model indicates that the TF is active. Out of the ~1000 initial PWMs, our model estimates that 277 PWMs (~120 unique TF) are active and predicts ~0.5 million active TF binding sites (post. prob. >99.0%).
In addition, we have applied our model to enriched 10mers in DNase1 hypersensitivity sites with significant sequence conservation. Using this approach, we independently recover many of the PWMs in TRANSFAC and JASPAR. Furthermore, we estimate a PWM with no good match in these databases that corresponds to a recently discovered regulatory motif (consensus sequence TCTCGCGAGA Wyrwicz et al., 2007, Guo et al. 2008).

Given these results, we anticipate that the model we describe here, termed CENTIPEDE, will become a valuable tool for mapping TF binding.

PRIMER-INITIATED SEQUENCE SYNTHESIS TO IDENTIFY AND ASSEMBLE SEQUENCE VARIANTS USING NEXT-GENERATION SEQUENCING DATA

Andreas Massouras, Korneel Hens, Bart Deplancke

Swiss Federal Institute of Technology Lausanne (EPFL), School of Life Sciences/Institute of Bio-engineering, Station 15, Lausanne, 1015, Switzerland

We present a novel depth-of-coverage approach, Primer-Initiated Sequence Synthesis (PrInSeS), which uniquely maps and assembles sequence variants up to the nucleotide level independent of the read format (single- or paired-end). Contrary to existing depth-of-coverage methods which excel at pinpointing large structural variants, PrInSeS primarily detects smaller events (1 bp – ~10 kb) and has the capacity to assemble novel sequence insertions. We evaluated PrInSeS's performance using simulated data and show that our approach compares favorably to other recently developed structural variation mappers. In addition, we used experimental data 1) to sequence verify and assemble pooled clones derived from cDNA libraries in high-throughput fashion, and 2) to identify and assemble sequence variants in human chromosomal segments as well as in *Mycobacterium tuberculosis* and *Drosophila melanogaster* whole genomes. For the latter, we used data from the *Drosophila* Genetic Reference Panel (Ayroles et al., Nature Genetics, 2009), selecting one out of 40 inbred, phenotyped *D. melanogaster* lines for which next-generation genome sequencing data is available. In total, we found 36,312 non-SNP variants of which 38% were insertions (Longest: 4,816 bp), 44% deletions (longest: 1,043 bp), and 18% substitutions (longest: 4,299 bp), thereby obtaining an independent validation rate of >90%. Our data are consistent with those obtained using microarray comparative genome hybridization in that we found significantly fewer variants per Mb on the X compared to the autosomal chromosomes and in exons (including UTRs) compared to non-exonic regions. In addition, genes (including the 500 bp up- and downstream regions) harboring non-SNP structural variants exhibited significantly more gene expression variation than genes without variants consistent with the notion that genes tend to be closely associated with variants that impact their expression. We are now in the process of analyzing and comparing the genomes of the other 39 *Drosophila* lines and will report our latest findings.

PREDICTING GENE EXPRESSION: FROM A GRAPHICAL REPRESENTATION OF PROTEIN INTERACTIONS TO A FORMULA FOR GENE EXPRESSION

Jacqueline M Dresch

MSU, Mathematics, Wells Hall, Michigan State University, East Lansing, MI, 48824

Much is still unknown about transcriptional regulation. It is believed that there are two main components involved in transcription: transcriptional factors and DNA cis regulatory elements. In this study, we have explored these components, and in particular, the role of short-range repressors in the Drosophila melanogaster early-blastoderm embryo. There are many advantages of using the Drosophila embryo. Many of the DNA regulatory elements have been identified and many of the transcription factors have been studied extensively on a molecular level.

We began our study by looking at a data set, from the Arnosti lab, comprised of quantitative measurements of gene expression from synthetic enhancers with binding sites for endogenous activators and repressors. These enhancers were designed to test specific interactions between activators and short-range repressors. By looking at the architecture of the enhancer region, along with protein concentrations and mRNA output, we started to develop a clearer picture of how the transcriptional regulation process works inside a living organism.

From this set of data, using numerical techniques, we have created a mathematical model for DNA transcription. Our model incorporates the transcription factor concentrations, positioning along the DNA of binding sites, the range affected by the binding sites of short range repressors (in base pairs), and the cooperation between binding sites. Due to the importance of protein interactions, a graph representation of the proteins affecting a certain gene is the basis for our model. We have built such a graph by fitting parameter values in a corresponding rational function for expression. Due to the combinatorial nature of protein-DNA binding and a graphical approach, the corresponding formula for gene expression fits very well into the general framework of thermodynamic modeling.

We believe that designing a network of interactions between binding sites and a set of rules to take us from this graphical representation to a gene expression formula could be of great use. Despite the simplicity of this system, the idea may be expanded to predict gene expression in other, more complex systems. As we look at larger enhancers, the underlying rules should not change, new vertices and edges can easily be added to the graph and a new formula for gene expression will emerge.

LONG INTRONIC NONCODING RNA SIGNATURES OF MALIGNANCY AND SURVIVAL OUTCOME IN CLEAR CELL RENAL CELL CARCINOMA

Angela A Fachel[1], Ana C Tahira[1], Vinicius Maracaja-Coutinho[1], Etel R Gimba[2], Giselle M Vignal[2], Franz S Campos[2], Rodrigo Louro[1], Eduardo M Reis[1], Sergio Verjovski-Almeida[1]
[1]Universidade de Sao Paulo, Departamento de Bioquimica, Rua Professor Lineu Prestes, 748, 1200, Sao Paulo, 055800-000, Brazil, [2]Instituto Nacional de Cancer, Serviço de Pesquisa Clinica, Rua Andre Cavalcanti, 37, Rio de Janeiro, 20231-050, Brazil

Renal cell carcinoma (RCC) is the most common malignancy of the adult kidney, and the clear cell subtype is the most prevalent and lethal cancer of the urinary system. Large numbers of long (> 400 bp) noncoding RNAs (ncRNAs) are transcribed from intronic and intergenic regions of the human genome. There are examples of their involvement in the regulation of gene expression and of their correlation to cancer. However, the use of long intronic ncRNAs as predictor markers of RCC has not been studied so far. Herein we aimed to identify intronic ncRNA expression signatures correlated to malignancy and to disease outcome in RCC. We also looked for new protein-coding molecular markers, using a customized cDNA microarray. We identified a signature of 33 differentially expressed long intronic ncRNAs (FDR < 5 %) comparing 15 paired tumor and adjacent non-tumor samples from clear cell RCC patients. Simultaneously, a set of 268 significantly altered (FDR < 5 %) protein-coding genes was identified, pointing to possible new biomarkers of malignancy in RCC. An ncRNA expression signature correlated to the five-year follow-up survival status was identified by a supervised statistical analysis of 16 patients (15 described above plus one with only tumor sample available), comprised of 26 ncRNAs with significantly altered expression (FDR < 5 %, p-value ≤ 0.01). A set of 65 protein-coding genes significantly correlated to disease outcome was also identified (FDR < 5 %, p-value ≤ 0.01). The most significantly enriched network (p < 10-44) correlated to increased risk of renal cancer death involves HerbB4 and PSMC1 proteasome component, both showing increased expression. The ncRNA expression signature has stratified the 16, plus two new patients into two groups, a group with low risk of cancer-specific death within a five-year period after surgery (n = 9) and another with a poor prognosis (n = 9, 1 false-positive). A Kaplan-Meier survival analysis showed that these groups were significantly distinct (p = 0.0024; Hazard ratio = 9.0, 95 % CI = 2.2 to 37.2). Patients classified by the signature expression profile shown here as having a high risk of cancer-specific death within a five-year period after surgery might be candidates for testing two new drugs targeting HerbB4 gene product and the proteasome, which are currently under Phase I /Phase II clinical trials in other types of cancer. Our results suggest that long intronic ncRNAs are relevant to RCC tumor biology. The long ncRNAs and protein-coding RNA signatures may contribute to patient management improvements, impacting patient survival.

HARNESSING NATURAL SEQUENCE VARIATION TO DISSECT POST-TRANSCRIPTIONAL NETWORKS IN YEAST

Mina Fazlollahi[1], Eunjee Lee[2], Harmen Bussemaker[2,3]

[1]Columbia University, Physics, 538 West 120th Street, New York, NY, 10027, [2]Columbia University, Biological Sciences, 1212 Amsterdam Ave, New York, NY, 10027, [3]Columbia University, Center for Computational Biology and Bioinformatics, 1130 St Nicolas Ave, New York, NY, 10032

We recently developed a sequence-based method for mapping the loci whose inheritance modulates a given transcription factor. This approach explicitly treats the regulatory activity of the factor as a quantitative phenotype. As input, it takes parallel expression and genotyping data for segregants from an experimental cross between two parental strains/lines [1], along with prior information about the sequence specificity of the factor. We used the same approach to elucidate the role of post-transcriptional regulation of mRNA stability as a source of genetic variation in gene expression levels. This allowed us to dissect the network of proteins modulating or supporting a given RNA-binding protein (RBP). First, we analyzed a recent compendium of affinity-purification based in vivo mRNA binding data [2] to model the sequence specificity of RBPs in terms of a position-specific affinity matrix (PSAM) using MatrixREDUCE [3]. This suggested that the Pumilio homolog Puf3p is able to bind to RNA in two distinct conformations. Subsequently, we used each PSAM to predict the affinity of binding of the corresponding RBP to the 3'-UTR, 5'-UTR, or CDS of each transcript. Next, we performed regression of mRNA expression on these predicted affinities to infer RBP activity in each segregant. This allowed us to successfully map activity quantitative trait loci ("aQTLs") for about a dozen RBPs. Most interestingly, we identified three distinct aQTLs modulating Puf3p, one of which is specific to one of the identified RNA binding modes. Thus, it seems that the connectivity between signal transduction pathways and the trans-acting factors they control can depend on the precise sequence of the cis-regulatory element.

1. Brem RB, Kruglyak L: The landscape of genetic complexity across 5700 gene expression traits in yeast. PNAS 2005, 102(5): e1572-7
2. Hogan DJ, Riordon DP, et al: Diverse RNA-binding proteins interact with functionally related sets of RNAS, suggesting an extensive regulatory system. PLoS Biol, 2008, 6(10):e255
3. Foat BC, Morozov AV, Bussemaker HJ: Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics 2006, 22(14):e141-9.

PREDICTING ENHANCERS USING CHROMATIN MODIFICATIONS AND TIME-DELAY NEURAL NETWORK

Hiram A Firpi[1], Kai Tan[1,2]

[1]University of Iowa, Internal Medicine, 283 Newton Road, Iowa City, IA, 52242, [2]University of Iowa, Biomedical Engineering, 1402 Seamans Center, Iowa City, IA, 52242

Background: Identification of regulatory DNA elements is of utmost importance for understanding gene regulation in both healthy and diseased cells. Recent large-scale chromatin state mapping studies have revealed characteristic chromatin modification signatures for various types of functional DNA elements. In addition, it is now well established that many regulatory elements carry these epigenetic modifications only in specific cell/tissue types or according to environmental conditions, which cannot be determined by DNA sequence analysis alone. Given the important influence of chromatin states on gene regulation and the increasing amount of genome-wide chromatin state maps generated in recent years, there is a pressing need for computational methods to analyze these data in order to identify functional DNA elements. However, existing computational tools do not exploit data transformation and feature extraction as a mean to achieve a more accurate prediction.
Results: A Fisher discriminant analysis (FDA) and a time-delay neural network (TDNN), a configuration insensitive to genomic position-variant shifts, were implemented to classify enhancers using six histone modifications as inputs. The histone modification signals were processed through two features generating a twelve feature set. We fed those features through a FDA and the output signature is used to train a TDNN. The trained TDNN is used to predict enhancers in the ENCODE regions. The TDNN achieved a PPV of 60.93%, up by 6% when compared to previous approaches.
Conclusion: By implementing a well-established pattern recognition paradigm and evaluating this through the ENCODE regions for predicting enhancers, the TDNN achieved better results when compared to previous methods. Experiments demonstrated that feature extraction and integration of the histone modifications signals are of paramount relevance for a tool in order to exploit and extract the needed patterns to pinpoint the regulatory DNA elements.

INDUSTRIAL STRENGTH GENE EXPRESSION

Barrett Foat, Jaishree Chittoor, Gene Expression Technology Team

Monsanto Company, Gene Expression Technology, 800 N. Lindbergh Blvd., St. Louis, MO, 63167

The regulation of gene expression has been a focus of research for half of a century. However, only in the last decade has the advent of genomics technology generated sufficient data to enable reverse engineering of the genome-wide logic of gene expression regulation. While the systems biology community has developed many approaches for working backwards from observations of gene expression to the regulatory sequences that produced them, attempts to move from inferences about gene regulation to engineered gene regulation are much less common. Nevertheless, it is the forward engineering of protein expression quantities and patterns that provides the basis for transgenic products in biotechnology applications.

Here, we outline the design considerations for creating protein expression patterns for biotechnology traits in crop plants. Transgene design for an agricultural biotechnology product presents several challenges not common to academic settings, for example requiring consistent expression across many generations, environments, and genetic backgrounds. A top-level overview of our Gene Expression Technology Pipeline will also convey a sense of the broad range of research topics in the area of gene expression regulation that are relevant to our endeavor.

THE IMPACT OF MEASUREMENT ERRORS IN THE IDENTIFICATION OF GENE REGULATORY NETWORKS

Andre Fujita[1], Alexandre G Patriota[2], Joao R Sato[3], Satoru Miyano[1,4]

[1]RIKEN, Computational Science Research Program, 2-1 Hirosawa, Wako, 351-0198, Japan, [2]University of Sao Paulo, Institute of Mathematics and Statistics, Rua do Matao, 1010, Sao Paulo, 05508-090, Brazil, [3]Universidade Federal do ABC, Center of Mathematics, Computation and Cognition, Rua Santa Adelia, 166, Santo Andre, 09210-170, Brazil, [4] University of Tokyo, Human Genome Center, 4-6-1 Shirokanedai, Tokyo, 108-8639, Japan

This work investigates the effects of measurement error (present in microarrays) on the estimation of the parameters in regulatory networks. Simulation studies indicate that, in both time series (dependent) and non-time series (independent) data, the measurement error strongly affects the estimated parameters of the regulatory network models, biasing them as predicted by the theory. Moreover, when testing the parameters of the regulatory network models, p-values computed by ignoring the measurement error are not reliable, since the rate of false positives are not controlled under the null hypothesis. In order to overcome these problems, we present improved estimators to independent and dependent data when the variables are subject to noises. Moreover, measurement error estimation procedures for microarrays are also described. Simulation results also show that both corrected methods perform better than the standard ones (i.e., ignoring measurement error). The proposed methodologies are illustrated using actual microarray data from lung cancer patients and mouse liver time series data.

Measurement error dangerously affects the identification of regulatory network models, thus, they must be reduced or taken into account in order to avoid erroneous conclusions. This could be one of the reasons for high biological false positive rates identified in actual regulatory network models.

REGULATORY ELEMENTS THAT DEFINE BREAST CANCER PROGRESSION AND SUBTYPES

Paul G Giresi[1], Heather A Hirsch[3], Charles M Perou[1,2], Kevin Struhl[3], Jason D Lieb[1,2]

[1]University of North Carolina at Chapel Hill, Department of Biology, 407 Fordham Hall, CB#3280, Chapel Hill, NC, 27599, [2]University of North Carolina at Chapel Hill, Carolina Center for Genome Sciences, 5000 Genetic Medicine Building, CB#7265, Chapel Hill, NC, 27599, [3]Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, 240 Longwood Ave, Building C2, Boston, MA, 02155

Identification of open chromatin regions has been one of the most accurate and robust methods to identify functional promoters, enhancers, silencers, insulators, and locus control regions in mammalian cells. Here we have used FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) to survey the set of regulatory elements active during cellular transformation and between clinical tumor samples.

We were able to infer the function of active regulatory elements specific to tumor subtypes, regulatory factor binding sites, and gene expression profiles. FAIRE regions at sites distant from proximal promoter elements distinguished cancer subtypes and were highly associated with changes in transcriptional status. These sites reflected the activity of regulatory pathways specific to each subtype, including hormone receptor status. Thus, FAIRE is useful for characterizing human cancer and furthering our understanding of tumor biology.

In addition, we characterized the set of regulatory elements activated throughout the process of cellular transformation. In this model of oncogenesis, transformation is initiated by transient activation of Src, which induces a switch within untransformed MCF10A cells resulting in complete transformation within 36 hours. This switch entails the activation of the inflammatory response and other transcriptional regulatory pathways, including NF-kB and the Jak/Stat. We found that the majority of chromatin responses during cellular transformation were initiated within 4 hours. The set of regulatory elements were organized into distinct regions associated with differentially expressed genes, which we call COREs (Clusters of Open Regulatory Elements).

NOVEL METHOD FOR COMPUTING ENRICHMENT OF DNA BINDING MOTIFS PROVIDES CLUES TO *C. ALBICANS* PATHOGENICITY

Raluca Gordân[1], Saumyadipta Pyne[2], Martha L Bulyk[1,3,4]

[1]Brigham and Women's Hospital and Harvard Medical School, Department of Medicine, Division of Genetics, 77 Ave Louis Pasteur, Boston, MA, 02115, [2]Dana-Farber Cancer Institute and Harvard Medical School, Department of Medical Oncology, 44 Binney Str, Boston, MA, 02115, [3]Brigham and Women's Hospital and Harvard Medical School, Department of Pathology, 77 Ave Louis Pasteur, Boston, MA, 02115, [4] Harvard Medical School, Harvard-MIT Division of Health Sciences and Technology (HST), 77 Ave Louis Pasteur, Boston, MA, 02115

We have recently developed a novel method for computing the enrichment of a transcription factor (TF) DNA binding motif in a set of foreground DNA sequences compared to a set of background sequences. We have used this method to successfully distinguish between direct and indirect TF-DNA interactions in the yeast *S. cerevisiae* (Gordan *et al.*, Genome Res 2009, 19(11):2090-100). Here, we apply the same enrichment method to several sets of promoters of *Candida albicans* cell cycle regulated genes, as reported by Cote *et al.* (Mol Biol Cell 2009, 20(14):3363-73). Unlike *S. cerevisiae*, *C. albicans* is truly dimorphic: it has the ability to undergo morphological changes between the yeast pdeusohyphal and the true hyphal forms. Dimorphism has been shown to play a crucial role in pathogenesis. Connections between dimorphims, cell cycle checkpoints and cell cycle regulators have also been actively studied, although the absence of cell cycle expression data has limited these analyses. Recently, Cote *et al.* reported a set of 494 genes that are periodically expressed during the *C. albicans* cell cycle, 100 of which do not have homologs in *S. cerevisiae* (we henceforth refer to these genes as "*Candida*-specific"). We investigated the transcriptional regulation of these 100 genes, in an attempt to find possible clues about the pathogenicity of *C. albicans*, as compared to that of *S. cerevisiae*. We analyzed the promoter regions of periodically expressed genes that peak during transition between different phases of the cell cycle: G1/S, S/G2, G2/M, and M/G1. We found two TF DNA binding motifs that are significantly enriched upstream of *Candida*-specific genes that peak during the S/G2 transition, and not enriched in general upstream of genes that peak at this stage. The first motif is that of *S.cerevisiae* TF Tec1, which is conserved in many yeast species, including *C. albicans*. In *S. cerevisiae*, the factor is known to be required for haploid invasive and diploid pseudohyphal growth. The second enriched motif is that of *S. cerevisiae* TF Rox1, whose homolog in *C. albicans* is a hypothetical protein similar to Rgf1, a regulator of filamentous growth and virulence. We are currently investigating the *Candida*-specific genes that are regulated by the two TFs.

A GLOBAL VIEW OF PRE-MRNA PROCESSING SUGGESTS THAT SPLICING IS SLOW RELATIVE TO TRANSCRIPTION

Jesse M Gray[1,3], David A Harmin[1,3], Michael Springer[2], Michael E Greenberg[1]

[1]Harvard Medical School, Neurobiology, 200 Longwood Ave, Boston, MA, 02115, [2]Harvard Medical School, Systems Biology, 200 Longwood Ave, Boston, MA, 02115, [3]these authors contributed equally to this work, _, _, _, MA, 02115

The prevailing view of pre-mRNA processing is that splicing occurs co-transcriptionally, taking place soon after each intron is synthesized. However, while co-transcriptional splicing has been observed at individual genes, it remains unknown whether these findings generalize to whole metazoan genomes. Moreover, while pre-mRNA levels have been postulated to be very low, a genome-wide census of primary versus mature metazoan transcripts has yet to be undertaken. We have performed RNA sequencing (RNA-Seq) of total RNA from mouse and human cells to assess pre-mRNA processing globally. The density of RNA-Seq reads across annotated genes provides quantitative information about the expression levels of RNA corresponding to exons, introns, and junctions between them. Read densities across exons reflect both primary and mature mRNA transcript levels; densities within introns reflect both primary transcripts and introns that have been spliced out but not yet degraded; densities across exon-intron junctions reflect the expression levels of transcribed but as-yet unspliced primary transcripts; and densities across exon-exon splice junctions reflect levels of mature mRNAs. We find that transcript maturity, the relative abundance of pre-mRNA versus mature mRNA at an individual locus, varies across the genome by several orders of magnitude. Our data can also be used to infer relative rates of transcription, splicing, and RNA degradation, assuming steady-state conditions. For example, the relative numbers of reads spanning the junction at each end of an intron imply relative rates of transcription and splicing. The inferred relative rates differ depending on assumptions about the details of *in vivo* pre-mRNA processing, such as whether splicing of introns along a gene must occur sequentially. Our data preliminarily suggest that (1) in general, splicing may be significantly slower than transcription, and (2) splicing and intron degradation occur on a similar timescale.

MOTIFADJUSTER: A TOOL FOR COMPUTATIONAL REASSESSMENT OF TRANSCRIPTION FACTOR BINDING SITE ANNOTATIONS

Jens Keilwagen[1], Jan Baumbach[2], Thomas Kohl[3], Ivo Grosse[4]
[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Department of Molecular Genetics, Correnssstr. 3, Gatersleben, 06466, Germany, [2]International Computer Science Institute, Department of Bioinformatics, 1947 Center Street, Berkeley, CA, 94704, [3]Bielefeld University, Center for Biotechnology (CeBiTec), Universitaetsstr. 27, Bielefeld, 33615, Germany, [4]Martin Luther University Halle-Wittenberg, Institute of Computer Science, Von-Seckendorff-Platz 1, Halle, 06120, Germany

Valuable binding site annotations are stored in databases. However, several types of errors can, and do, occur in the process of manually incorporating annotations from the scientific literature into these databases. Three frequently occurring errors are (i) spuriously annotated binding sites, (ii) shifts of binding sites, and (iii) missing or wrong strand orientations of binding sites. Here, we introduce MotifAdjuster, a tool that helps to detect these errors.

For testing the efficacy of MotifAdjuster and improving the annotation of binding sites of *Escherichia coli*, we apply MotifAdjuster to the seven largest data sets of binding sites of CoryneRegNet 4.0 corresponding to the transcription factors CpxR, Crp, Fis, Fnr, Fur, Lrp, and NarL. Surprisingly, we find that all of these data sets are considered questionable. Specifically, 35% of the 536 binding sites annotations are proposed for removal or shifts, and the percentage of questionably annotated binding sites ranges from 9.3% for Fnr to 95.7% for Fur.

Based on a systematic analysis of the proposed re-annotations for the nitrate regulator NarL, we find that all proposals are in accordance with the originally published annotations as given in the literature. We apply the adjusted position weight matrix to promoter regions for which binding by NarL could be experimentally verified, and we find a novel binding site of NarL in the promoter region of the gene *torC*, for which no binding site of NarL could be detected in the past. MotifAdjuster is publicly available as part of Jstacs, an open source Java library for statistical analysis of biological sequences, and can be obtained at http://dig.ipk-gatersleben.de/MotifAdjuster.html.

GENETIC VARIATION OF THE AUXIN SIGNALING PATHWAY AND TRANSCRIPTIONAL AUXIN RESPONSE NETWORKS IN NATURAL *ARABIDOPSIS THALIANA* ACCESSIONS

Marcel Quint[1], Carolin Delker[1], Anja Raschke[1], Yvonne Poeschl[2], Ivo Grosse[2]

[1]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, Halle, 06120, Germany, [2]Martin Luther University, Institute of Computer Science, Von-Seckendroff-Platz 1, Halle, 06120, Germany

Natural genetic variation has been observed for several adaptational processes that underlie various patterns of natural selection in different environmental habitats. Phytohormone signaling cascades are known to be involved in the regulation of the responses to most environmental cues with auxin as a key player in many of them. We were therefore interested to find out if genetic variation in the *Arabidopsis* gene pool also exists on this level of regulation and investigated the population genetics of auxin signaling genes and transcriptional response networks between accessions from diverse habitats. While we identify extensive natural variation between accessions for physiological responses to various auxins, the signaling genes responsible for the transduction of the auxin stimulus seem to be rather conserved at the sequence level. We examined 20 *Arabidopsis* accessions for SNPs. Several population genetic parameters indicate some degree of purifying selection acting on the auxin receptors, the Aux/IAA repressors, and the ARF transcription factors. Hence, it seems unlikely that sequence divergence in signaling genes is responsible for the dramatic variation between accessions we identified in whole-genome expression analysis of auxin responses. However, co-expression network analysis using the Local Context Finder algorithm revealed highly significant differences on the level of transcriptional regulation of auxin signaling genes. These differences may cause the variation of downstream response gene expression. Cluster analysis coupled with co-expression networks reveal accession-specific response gene clusters that likely translate into the physiological variation we observed between accessions in response to auxin stimuli.

BRANCHING PROCESS DECONVOLUTION ALGORITHM REVEALS A DETAILED CELL-CYCLE TRANSCRIPTIONAL PROGRAM

Xin Guo[1], Allister Bernard[1], David Orlando[2,3], Steve Haase[2], Alexander J Hartemink[1]

[1]Duke University, Computer Science, 450 Research Drive, Durham, NC, 27708, [2]Duke University, Program in Computational Biology and Bioinformatics, 101 Science Drive, Durham, NC, 27708, [3]Duke University, Department of Biology, 125 Science Drive, Durham, NC, 27708

Time-series gene expression has been intensively studied to elucidate transcriptional events during the eukaryotic cell cycle. However, due to current microarray technology constraints, we can only accurately measure gene expression levels from a population of initially synchronized cells instead from single cells. Unfortunately, perfect cell synchrony is neither attainable at synchronization nor maintainable after release, so the measured gene expression levels at population level are actually a convolution of the expression levels of single cells distributed across cell cycle phases. Here, we present a novel deconvolution algorithm to accurately recover the underlying single-cell like transcriptional program. The algorithm is built on a branching process construction that accounts for synchrony loss. Coupled with this branching process construction, we use a wavelet-basis regularization approach to smooth our expression estimates, which also enables us to learn robust estimates jointly from multiple replicates of the same experiment. Utilizing transcript level data from cell-cycle synchrony experiments in the budding yeast, *S. cerevisiae*, we demonstrate that our algorithm, in contrast to previous approaches, can not only increase the resolution of expression profiles, but also explicitly learn distinct, semi-overlapping cell-cycle transcriptional programs for both mother and daughter cells. In addition, our deconvolved estimates indicate that nearly half of yeast genes exhibit substantial changes in expression amplitude during the cell cycle, suggesting that many previously unrecognized genes may play a role in this fundamental biological process.

FLEXIBLE AND ACCURATE DISCOVERY OF *CIS*-REGULATORY ELEMENTS IN INSECTS AND MAMMALS

Majid Kazemian[1], Jia-yu Chen[1], Miriam R Kantorovitz[1], Qiyun Zhu[2], Saurabh Sinha[1], Marc S Halfon[2]
[1]University of Illinois at Urbana-Champaign, Department of Computer Science, 201 N. Goodwin Ave., Urbana, IL, 61801, [2]University at Buffalo, State University of New York, NY State Center of Excellence in Bioinformatics & Life Sciences, 701 Ellicott St., Buffalo, NY, 14203

The vast majority of gene regulatory sequences in the genome remain uncharacterized. Despite recent advances in experimental approaches to identifying transcriptional *cis*-regulatory modules (CRMs, "enhancers"), direct empirical discovery of CRMs for all genes in all cell types under the full range of environmental conditions is likely to remain an infeasible and elusive goal. Effective, sensitive, and widely-applicable methods for computational CRM discovery are thus a critically needed complement to empirical approaches. This is particularly true for CRMs regulating genes involved in less fully described processes for which lack of knowledge of relevant transcription factor (TF) binding makes ChIP-based experimental approaches unavailable.

We are developing a flexible suite of methods for the prediction, identification and characterization of regulatory sequences when beginning with different degrees of knowledge of the underlying regulatory network—gene coexpression, known CRMs, characterized transcription factor binding sites, etc.—and compatible with a variety of post-processing steps to incorporate additional genomic information.

Our most successful efforts to date center on "motif blind" in silico CRM discovery methods that do not depend on knowledge or accurate prediction of TF binding sites and that are effective when limited knowledge of existing CRMs is available. A sliding genomic window is scored for similarity to a training set consisting of a small (as few as seven) training set of known CRMs that direct a common pattern of gene expression. Scoring is based on the statistics of short word (k-mer) frequencies under any of several novel measures. Empirical testing of our predictions using this strategy so far yields a 95% success rate in correctly identifying sequences with CRM activity in transgenic reporter assays in both Drosophila and mouse (n=20). We have also seen promising results with unsupervised methods that rely on gene coexpression data both with and without TF binding knowledge. We are continuing to refine all of these approaches. In particular we are working to increase the accuracy of our CRM discovery with respect to successful prediction of patterns of regulated gene expression and to extend our ability to perform CRM discovery in biomedically and agriculturally important, but experimentally less well-characterized, arthropod species such as *Anopheles gambiae* and *Apis mellifera*. Overall, our methods represent a flexible, accurate way of identifying regulatory sequences in both insect and mammalian genomes and provide an important in silico adjunct to and extension of empirical CRM discovery approaches.

CHARACTERIZATION OF GSE AND GSE-INTERACTING NOVEL GENE, GIAP, IN PRIMORDIAL GERM CELLS.

Yuki Hatanaka, Manabu Satou, Natsumi Shimizu, Mikiko Tokoro, Seungwook Shin, Satoshi Nishikawa, Hyangheun Lee, Tomoko Amano, Satoshi Kishigami, Kazuhiro Saeki, Yoshihiko Hosoi, Kazuya Matsumoto,

Kinki University, Biology-Oriented Science and Technology, Nishimitani, 930, Kinokawa, Wakayama, 649-6234, Japan

Germ cells are the only cells that are possible to communicate to next generation and having pluripotency. In the development of primordial germ cells (PGCs), specific molecular mechanisms such as epigenetic regulation and signal transduction are involved. Recently, several reports have showed that PGCs specific expression genes are essential for PGCs development and the mutants lack the normal PGCs development (Loveland et al., 1997; Kehler et al., 2004; Ohinata et al., 2005).
We have identified the novel gene named as GSE (gonad specific expression gene), which are expressed at the ovary and testis of adult mice (Zhang et al., 2002 ;Mizuno et al., 2006). Subsequently, we have isolated a novel gene named as GIAP (GSE-interacting protein), which are identified by the yeast two hybrid system. GSE are confirmed to interact with GIAP in both ovary and testis of adult mice by immunoprecipitation. To investigate the involvement of GSE-GIAP interaction in the development of PGCs, we performed the expression analysis of GSE and GIAP in the fetus stage. Additionally, we obtained presumed biological characters of GIAP by the public database.
As a result, we showed that GSE expresses from embryonic day 7.5 (E7.5) to E11.5 fetus and both of ovary and testis at E12.5 fetus. We also showed that GSE expresses in E11.5 PGCs by in situ hybridization. Furthermore, immunohistochemical analysis indicated that GSE and GIAP co-localized in E11.5 PGCs. Therefore, it suggested that GSE also interacts with GIAP in PGCs. The informatics analysis indicated that it is possible for GIAP to have PRMT activity. PRMT family is protein arginine methyltransferase and is known to play an important rule for the development of PGCs. BLIMP1, master regulator of PGCs development, interacts with PRMT5 and regulates the development related factors (Ancelin et al., 2006). We will address the function of GSE and GIAP in the development of PGCs.

DISTINCT EPIGENOMIC LANDSCAPES OF HUMAN PLURIPOTENT AND LINEAGE-COMMITTED CELLS

David Hawkins[1], Gary Hon[1], Ryan Lister[2], James Thomson[3], Joseph Ecker[2], Bing Ren[1]

[1]Ludwig Institute for Cancer Research, Gene Regulation, 9500 Gilman Dr, La Jolla, CA, 92093, [2]Salk Instititute, Genome Analysis, 10010 N. Torrey Pines, La Jolla, CA, 92037, [3]University of Wisconsin School of Medicine, Genome Center, 425 Henry Mall, Madison, WI, 53706

Human embryonic stem cells (hESCs) share an identical genome with lineage-committed cells, yet possess the remarkable properties of self-renewal and pluripotency. The diverse cellular properties in different cells have been attributed to their distinct epigenomes, but how much epigenomes differ remains unclear. Here, we report that epigenomic landscapes in hESC and lineage committed cells are drastically different. By comparing the chromatin modification profiles and DNA methylomes in hESCs and primary fibroblasts, we find that nearly one-third of the genome differs in chromatin structure. Most changes arise from dramatic redistributions of repressive H3K9me3 and H3K27me3 marks, which form blocks that significantly expand in fibroblasts. A large number of potential regulatory sequences also exhibit a high degree of dynamics in chromatin modifications and DNA methylation. Additionally, we observe novel, context-dependent relationships between DNA methylation and chromatin modifications. Our results provide new insights into epigenetic mechanisms underlying properties of pluripotency and cell-fate commitment.

APPLYING REACHABILITY PRIORS TO REGULATORY NETWORK INFERENCE

Brian C Haynes[1,2], Michael R Brent[1,2]

[1]Washington University, Department of Computer Science, 4444 Forest Park Ave, St Louis, MO, 63110, [2]Washington University, Center for Genome Sciences, 4444 Forest Park Ave, St Louis, MO, 63110

Motivation: Recently, there has been a major focus on gaining a mechanistic understanding of gene regulation by using gene expression data to infer networks of regulatory interactions. However, previous efforts have neglected a key source of information inherent in any data set that contains genetic perturbations; namely, which gene was perturbed and which genes were differentially expressed in response to the perturbation. Surprisingly, genetic perturbations have largely been treated as generic samples of the possible cellular states, without regard to the particular gene perturbed in each sample. We show how using the identity of the perturbed gene and the genes that respond to it leads to major improvements in the accuracy of network inference. Specifically, genes that respond to a perturbation of gene X must be reachable from gene X via a directed path in the network. This imposes a global constraint on the influence graph - a constraint whose satisfaction depends on the entire graph rather than a local piece of the graph.

Results: We present a new method, N-sieve, for inferring the structure of a transcriptional network from gene expression data. N-sieve combines a probabilistic measure of the fit between predictions and observations with a prior that favors structures consistent with the reachability constraints implied by perturbation data. We evaluate our approach and compare it to other methods on synthetic benchmarks and the transcriptional networks of *Escherichia coli* and *Saccharomyces cerevisiae*. Our method yields substantial accuracy improvements on all evaluations, with a 40% gain in accuracy over the next best method.

BIOINFORMATIC PREDICTIONS, EXPERIMENTAL VALIDATION AND ANALYSIS OF CIS-REGULATORY MODULES: APPLICATION TO D.MELANOGASTER'S CARDIOGENESIS

Delphine Potier[1,2], Stein Aerts[3], Carl Herrmann[2], Laurent Perrin[1]

[1]IBDML - UMR6216, CNRS & Université de la Méditerranée, Campus de Luminy, Marseille, 13288, France, [2]TAGC - U928, Inserm & Université de la Méditerranée, Campus de Luminy, Marseille, 13288, France, [3]LCB, Dep. of Human Genetics K.U. Leuven, Herestraat 49, Leuven, 3000, Belgium

Organogenesis and differentiation require the coordinated expression in time and space of different groups of genes. The accuracy of this process, governed by transcription factors (TFs) acting within a complex gene regulatory network, ensures the acquisition of specific organ shape and physiology. However, the logic of the cis-regulatory mechanisms is far from being understood so far. Bioinformatic approaches to predict CRMs from genomic sequences can greatly help to characterize new enhancers and the associated developmental regulatory network. Approaches based on combining expression data with comparative genomics are expected to allow predicting regions of DNA that regulate the expression of genes with greater accuracy.

We focus our interest on the development of the cardiovascular system in Drosophila in order to investigate the regulatory logic of this process. We recently reported a precise temporal map of gene expression of adult heart formation through the analysis of the temporal dynamics of heart-specific gene expression profiling.
Starting from clusters of co-expressed genes during cardiac tube remodeling during metamorphosis, we applied a new method that uses a comprehensive library of position weight matrices, combined with phylogenetic conservation, to identify potential cis regulatory modules common to a cluster of co-expressed genes.With this method we have been able to predict several CRMs involving a particular class of TFs for one of the clusters, in which gene expression is induced at 42h after pupation. Potential binding sites are evolutionary conserved and overrepresented in the surrounding non-coding sequences of co-expressed genes with a high statistical significance. The TF is likely to correspond to a nuclear receptor. A likely candidate, Hr46, is highly expressed during heart remodeling, just before the induction of the expression of the cluster of genes. In vivo validations, using transgenesis, suggest that the predicted CRMs reproduce the expected temporal expression pattern. Indeed, all tested CRMs drive a transitory expression in different tissues from 42h to 96h after pupation. Our approach thus seems successful in identifying the regulatory actors responsible for the specific temporal expression of the genes.

DISCOVERY OF PROMOTER MOTIFS IN *ARABIDOPSIS THALIANA* STRESS RESPONSE GENES

Richard Hickman[1], Vicky Buchanan-Wollaston[2,3], Laura Baxter[3], Sascha Ott[3]

[1]University of Warwick, Systems Biology DTC, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom, [2]University of Warwick, Warwick HRI, Wellesbourne, Warwick, CV35 9EF, United Kingdom, [3]University of Warwick, Systems Biology Centre, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom

Environmental stress is responsible for reduced crop productivity throughout the world. Plant adaptation to stress is dependent upon the initialisation of cascades of molecular networks involved in stress perception, signal transduction and the expression of specific stress-related genes. By examining large, high-resolution microarray datasets we can track gene expression changes over time in response to fungal pathogen infection, bacterial infection, drought, high-light and senescence in the model organism *Arabidopsis thaliana*. Dramatic variations in gene expression are observed at the onset of stress with different groups of genes showing different expression time-courses. This observation must, for a large part, be down to the action of different transcription factors (TFs) or combinations of TFs binding to the promoters of genes in each group.

We combined both expression and sequence data analysis in order to identify promoter motifs that may be responsible for recruiting specific TFs that orchestrate the activation of stress-response genes. We use position specific scoring matrices (PSSMs) retrieved from experimentally verified binding site databases and via *de novo* motif finding methods to assess whether particular patterns are statistically overrepresented within groups of genes that share similar expression profiles in response to a particular stress. We expand this analysis to groups of genes that exhibit similar expression behaviour across multiple experiments. We also examine whether there is a tendency for these patterns to exhibit a positional bias in relation to the transcription start site. In addition to testing single PSSMs, we also investigate the potential for specific pairs of PSSMs, with or without certain spacing rules, to be enriched in the promoters of stress response genes and to what extent they contribute to the observed expression patterns. This approach allows us to identify sequence patterns with statistical links to the microarray data, and in turn provides us with putative transcriptional modules, which may play important roles in the plant-stress response.

IDENTIFYING LARGE AND SMALL CHROMATIN DOMAINS FROM CHIP-SEQ DATA

Gary Hon, R. David Hawkins, Bing Ren

University of California, San Diego, Ludwig Institute for Cancer Research, 9500 Gilman Dr., La Jolla, CA, 92093

Histone modification maps provide global views of epigenetic status and allow us to investigate the influence of epigenetics in development and disease. However, while numerous methods exist to identify binding sites of transcription factors (TFs) from ChIP-Seq data, methods for analyzing histone modification profiles are still lacking due to unique challenges that have not been encountered with TF data. For example, while binding sites for TFs are generally discrete peaks and are sparsely scattered throughout the genome, histone modifications can also form large block-like structures enriched over many consecutive nucleosomes.

To systematically enumerate genomic regions enriched for chromatin modifications, we devised a computational method specifically tailored to identify both peak-like structures and large chromatin domains from ChIP-Seq data.

THE -13KB A/G POLYMORPHISM IN THE LMX1A UPSTREAM REGULATORY REGION THAT POTENTIALLY AFFECTS FOXF2/DEC2 BINDING IS ASSOCIATED WITH OSTEOPOROSIS

Qingyang Huang[1,2]

[1]The University of Hong Kong, Li Ka Shing Faculty of Medicine, 21 Sassoon Road, Kong Kong, 00000, China, [2]The University of Hong Kong, Research Centre of Heart, Brain, Hormone & Healthy Aging, 21 Sassoon Road, Kong Kong, 00000, China

Osteoporosis is a complex disease with both genetic and environmental risk factors. Bone mineral density (BMD) is one of the major determinants of risk for osteoporotic fracture. We previously reported linkage of chromosome 1q21-24 (LOD=2.36) with spine BMD in Southern Chinese pedigrees. In this study, we used a two-stage genotyping approach to identify genes in the linked region that contributed to the variation of spine BMD. In the first stage, 2000 simple nucleotide polymorphisms (SNPs) across the linked region were genotyped in 798 women with low (Z-scores $\leq$ -1.28) and high BMD (Z-score $\geq$ +1.0) at either the L1-4 lumbar spine or femoral neck. In the second stage, top 9 SNPs were genotyped in additional 720 women with extreme BMD. The most significant evidence of association for spine BMD ($P = 1.2 \times 10^{-6}$) was found with rs989335 (A/G) which is located -13 kb upstream of the LMX1A gene. Computational analysis showed that the A/G allele can bind FOXF2/DEC2 respectively. Interestingly, FOXF2 enhanced the rate of adipocyte differentiation (Gerlin et al. 2009), while DEC2 inhibits adipogenic differentiation (Gulbagci et al, 2009). In conclusion, the functional SNP rs989335 in the LMX1A upstream regulatory region that potentially affects FOXF2/DEC2 binding is associated with spinal BMD in Hong Kong Chinese population.

This study received funding from: Hong Kong General Research Fund and The University of Hong Kong.

PREDICTING TRANSCRIPTIONAL CONTROL PATTERNS IN COMPLEX GENOMIC DATA

Michael Hubank, Martino Barenco

University College London, Institute of Child Health, 30, Guilford Street, London, WC1N 1EH, United Kingdom

Mechanistic models of transcription can be used to identify the major activities controlling a transcriptional response. Using short time series of gene expression microarray data obtained from gamma-irradiated T-cells, we developed linear models that successfully predicted previously unknown targets of the transcription factor and tumour suppressor, p53. By next incorporating global transcript degradation rates, we found the model could be reformulated to identify the main transcriptional activities, and their targets, in the DNA damage response. This work was performed with data from a single activation dose (5 Gy). Here, we extend the models to allow quantitative prediction of transcriptional behaviour for any given dose. Separate time courses of gene expression were run on Affymetrix GeneArrayST microarrays at 0.5, 1 and 5 Gy. We found that simple linear models are unable to predict expression patterns across different doses. This is because the mechanisms governing gene transcription are inherently non-linear. We explain how we verified that this non-linearity is due to the transcription mechanism rather than to upstream causes and the process of model selection. We show that new non-linear models are able to accurately predict gene expression at intermediate doses. In addition, we found that the set of kinetic parameters governing the expression of single genes are surprisingly reproducible across different microarray platforms, implying that these parameters may correspond to biologically meaningful entities (such as affinity). Finally, given the rise of ultra-high content sequencing as an alternative to microarray use, we demonstrate the feasibility of using such data in our models. We discuss the avenues of research this opens towards a fuller understanding of the mechanisms governing gene transcription.

DISCOVERY AND CHARACTERIZATION OF NCRNAS INVOLVED IN CELL CYCLE REGULATION

Tiffany Hung[1,2], David Wong[1,2], Howard Chang[1,2]

[1]Stanford University, Epithelial Biology, 269 Campus Dr., Stanford, CA, 94305, [2]Howard Hughes Medical Institute, N/A, 4000 Jones Bridge Road, Chevy Chase, MD, 20815

Non-codingRNAs (ncRNAs) have recently emerged as critical regulators of chromatin states (Amaral, 2008). Ranging from 30 bases to 10kb, these diverse RNAs are implicated in epigenetic processes ranging from imprinting and dosage compensation to cell fate decisions. Divergent transcription initiation can lead to production of many promoter-associated ncRNAs, and some gene-proximal long ncRNAs have been to regulate gene silencing in cis. To further explore the roles of ncRNAs in cell cycle regulation, we utilized high density tiling arrays to profile transcription at all human cell cycle genes from 5kb upstream to 2kb downstream of their transcription start sites. We examined transcription under 40 conditions, including specific perturbation of oncogenic or tumor suppressor pathways, detailed characterization through cell cycle progression, human clinical tumor samples, and environmental changes such as growth factor deprivation and DNA damage induction. Preliminary data from this analysis has identified a network of ~200 novel transcripts that may regulate cell cycle processes.

ANALYSIS OF DEGRADED MATERNAL PROTEINS BY UBIQUITIN-PROTEASOME PATHWAY IN MOUSE PREIMPLANTATION EMBRYOS

LEE Hyang-Heun, SHIN Seung-Wook, TOKORO Mikiko, NISHIKAWA Satoshi, HATANAKA Yuki, KISHIGAMI Satoshi, SAEKI Kazuhiro, HOSOI Yoshihiko, MATSUMOTO Kazuya

Kinki University, Graduate School of BOST, Kinokawa, Wakayama, 649-6493, Japan

Development of mammalian embryos starts with fertilization. After fertilization, the developmental program is controlled by maternally inherited transcripts and proteins. In mice, the transition from maternal to embryonic control occurs during the late 1-cell and early 2-cell stages, and is an essential event during early development termed “zygotic gene activation (ZGA)”. Intracellular proteolysis occurs in the nonlysosomal ATP-dependent pathway (ubiquitin-proteasomal pathway: UPP). The degradation of maternal proteins is essential for remodeling the oocyte into a totipotent zygote. Zygote arrest 1 (Zar1), Argonaute 4 (Ago4) and Oogenesin 1 (Oog1) are three of many oocyte-specific maternal effect genes that function during the oocyte-to-embryo transition. Here we investigated the degradation mechanism of Zar1, Ago4 and Oog1 proteins by UPP.
We obtained MII oocytes from superovulated mice. After in vitro fertilization, embryos at 7hpi (hours post insemination) were cultured with or without proteasome inhibitor (MG-132) to confirm ubiquitinated maternal proteins in early stage embryos. Western blotting was performed by using anti-ub antibody. Immunoprecipitation was performed by using anti-maternal protein (Zar1, Ago4 and Oog1) antibodies. Then, western blotting was performed using anti-ub antibody.
In mouse preimplantation development, we observed gradual decrease of ubiquitinated proteins until 4-cell stage and could not detected at 8-cell or compacted morula stages. MG-132 treatment inhibited the development of fertilized embryos to 2-cell stage, in which we observed ubiquitinated proteins were accumulated. This result indicated that UPP may involve in mechanisms of protein degradation in mouse preimplantation development. Subsequently, we biochemically examined ubiquitination of the three maternal proteins in 1 cell mouse embryo by using IP. As a result, we confirmed that examined three maternal proteins were degenerated by UPP. These results suggest that UPP play an important role during maternal-embryonic transition.

THE ORIGIN OF VARIATION IN TRANSCRIPTION ELONGATION.

Ricardo Neves[1], Gupta Rajeev[1], Nick Jones[2], Tariq Enver[1], Francisco Iborra[1,3]

[1]University of Oxford, Molecular haematology, Jonh Radcliffe Hospital, Oxford, OX3 9DS, United Kingdom, [2]University of Oxford, Physics, South Parks, Oxford, OX1 3RE, United Kingdom, [3]CSIC, Centro Nacional de Biotecnologiua, Darwin, Madrid, 208049, Spain

Why populations of genetically identical eukaryotic cells show significant cell to cell variability remains unclear(Raser and O'Shea 2005; Raj and van Oudenaarden 2008). Most of this variability is due to differences in the transcription profile of these cells. We have discovered that cells elongate transcription at different speed, with all the RNA pol II molecules inside a given cell transcribing at similar speed.
Fusion studies suggest this variability can be attributed to small diffusible factors. In vitro studies suggest that transcription rate has a sensitive dependence on [ATP] but not other NTPs. Perturbing populations by changing nutrient levels and available [ATP] shows this connection holds in vivo. Since [ATP] relaxes quickly in the presence of mitochondrial poisons, variability in [ATP] might depend on variability in mitochondrial density. Selecting cells with differing mitochondrial densities shows that cells with higher [ATP] have a faster rate of transcription. Daughter studies show that a cause of variability in mitochondrial density is apparently stochastic segregation of mitochondria at division; we also find evidence that transcription rate variability is substantially modulated by the presence of anti(pro)-oxidants.
We conclude by noting that daughters with a lower mitochondrial density than their sisters have longer cell cycles, consistent with noise controlling cell cycle length(Di Talia et al. 2007). Our findings link energy metabolism to transcription elongation; providing a means for cells to generate variability.

CHARACTERIZATION OF GENOMIC REGIONS ASSOCIATED WITH THE NUCLEAR ENVELOPE IN *C. ELEGANS*

Kohta Ikegami, Jason Lieb

University of North Carolina at Chapel Hill, Department of Biology and Carolina Center for Genome Sciences, 408 Fordham Hall, Chapel Hill, NC, 27599-3280

The genome is highly compacted within the nucleus, but this organization must be coordinated with the regulation of gene expression. This coordination is mediated in part by the association of the genome with the nuclear lamina and the nuclear pore complex (NPC), both of which are major components of the nuclear envelope (NE). We seek to reveal rules governing NE-genome interaction. Using chromatin immunoprecipitation coupled with whole-genome tiling arrays or high-throughput sequencing in *C. elegans*, we identified genomic regions bound by LEM-2, a lamina protein, or by NPC component NPP-13.
LEM-2 associates with chromosomal patches over 10 kb in length. The intervals between the regions are often occupied by RNA polymerase II. Concordantly, genes outside of LEM-2 associated regions are more highly expressed than those within LEM-2 regions. These LEM-2 regions cluster and form several mega-basepair domains at the repeat-rich arm regions of the chromosomes and are excluded from the center of chromosomes. A very different pattern of association is seen between chromosomal loci and nuclear pore component NPP-13. Sites of interaction are dispersed, occurring both at the arm and center of the chromosomes. NPP-13 binding sites are short and mostly (83%) found at non-coding RNA genes, 70% of which encode tRNA. Our preliminary result indicates that this binding does not correlate with levels of tRNA gene expression. These results suggest very different determinants for genome interaction between the NPC and the nuclear lamina. Our data suggests that lamina-genome binding is associated with lower gene expression, whereas at least one component of the NPC, NPP-13, associates primarily with non-coding RNA loci.

DYNAMICS OF CHROMATIN FOLDING, LOOPING AND REPOSITIONING IN A FRACTAL GLOBULE

Maxim Imakaev[1], Erez Lieberman-Aiden [2,3], Leonid Mirny[1,3]

[1]MIT, Physics, 77 Massachusetts ave, Cambridge, MA, 01239, [2]Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, [3]MIT, Harvard-MIT Division of Health Sciences and Technology, 77 Massachusetts ave, Cambridge, MA, 02139

In a recent collaboration we demonstrated that the structure of interphase chromatin is consistent with a fractal globule conformation. The fractal globule is a dense and unknotted conformation into which a polymer (e.g., a chromatin fiber) can spontaneously fold.

Here, we use computer simulations to examine structural and dynamic properties of chromatin folded into a fractal globule and their connection to such processes as gene expression, DNA replication, and silencing.

First, we study mechanisms of chromatin condensation, and examine criteria required for formation of the fractal globule. We explore how crosslinking interactions guide the formation of distinct chromatin domains, the formation of high-density heterochromatic domains, and the spatial segregation of these domains. Next, we examine dynamical properties of the fractal globule that are essential for gene regulation: opening of chromatin domains, formation of loops between distant regions of the genome, and mobility of individual loci. We demonstrate that the fractal globule allows not only folding and unfolding, but also rapid locus repositioning – a process essential for gene expression. Furthermore, we demonstrate that the organization of chromatin into sectors, where genomically proximal regions are located close to one another in space, facilitates access of transcription factors to proximal genes. We suggest how such an organization can lead to coordinated chromatin modification and gene expression. Biophysical modeling and theoretical analysis of chromatin folding will provide us with new insights into the nature of genomic organization and gene regulation.

COMPUTER ANALYSIS ON THE CORRELATION BETWEEN SURFACE STRUCTURE AND GENE TRANSCRIPTION EFFICIENCY OF THE HYALURONIC ACID-COATED DNA/POLYCATION COMPLEX

Tomoko Ito[1], Yoshiyuki Koyama[2], Makoto Otsuka[1]

[1]Musashino University, Research Institute of Pharmaceutical Science, 1-1-20 Nishitokyo-shi, Tokyo, 202-8585, Japan, [2]Otsuma Women's University, Department of Textile Science, Sanbancho 12 Chiyoda-ku, Tokyo, 102-8357, Japan

To attain the efficient gene expression, non-viral vectors should overcome not only the barriers in delivery to the nucleus of the target cells, but a problem with transcriptional activity of the tightly compacted plasmid/polycation complex. We have found that hyaluronic acid (HA) deposited onto the positively charged DNA/polycation complex, and formed ternary complexes with negative surface charge, which exhibited an improved transcription efficiency. HA were thought to moderately loosen the DNA complex to allow the access of transcription factors. However, excess HA rather hindered the transcription.
To analyze the influence of the surface structure of the DNA/polycation/HA ternary complex on the transcription efficiency, ζ-potential and the Raman spectrum were measured, and their correlation was investigated with chemoinformatics analysis. At the mixing ratio of HA/PEI (COOH/NH) = 0.5, all the amino groups were just protonated, and the surface was completely covered by HA. At this ratio, the DNA complex showed the highest gene-expression level, probably attributed to the reduced interaction between the DNA and PEI in which all the amino groups have already interacted with carboxyls of HA. More amount of HA caused further decrease of the ζ-potential, though the Raman spectrum showed no significant change. Outer HA layer around the complex would be formed, and caused the suppression of transcription.

SYSTEMATIC DISCOVERY OF NOVEL MOTIFS WHICH MODULATE MICRORNA REGULATION

Anders Jacobsen[1,2], Jiayu Wen[1], Debora S Marks[3], Anders Krogh[1]

[1]University of Copenhagen, Bioinformatics Centre, Ole Maaloes Vej, Copenhagen, 2200, Denmark, [2]Memorial Sloan Kettering Cancer Center, Computational Biology Center, York Avenue, New York, NY, 10065, [3]Harvard Medical School, Systems Biology Department, Longwood Ave, Boston, MA, 02115

MicroRNAs and siRNAs, bound to Argonaute proteins (RISC), destabilize mRNAs through base-pairing with the mRNA. However, the gene expression changes after perturbation by small RNAs are only partially explained by the predicted microRNA targeting and hence the change in expression of hundreds of genes in each experiment remains unexplained. MicroRNA/siRNA targeting may be modulated by other mRNA sequence elements such as binding sites for the hundreds of RNA binding proteins expressed in any cell. This aspect of small RNA regulation has not yet been systematically explored.

Across a panel of published experiments using rigorous computational methods, we investigated to what extent sequence motifs in 3'UTRs correlates with expression changes following perturbation by small RNAs. We discover hundreds of motifs, in addition to the microRNA target sites, that are significantly correlated with up or down-regulation in all transfection experiments. Among the most significantly overrepresented motifs in down-regulated mRNAs were two novel binding motifs recently discovered for the RNA binding protein HuD. The most significantly overrepresented motif in up-regulated mRNAs is the heptanucleotide AU-rich element (ARE), which is known to affect mRNA stability via at least twenty different ARE binding proteins. Genes with predicted ARE motif over-representation were significantly up-regulated after all microRNA transfections ($P < 2.5e-60$). We confirmed this perturbed ARE-stability signal a range of other types of published experiments and provide genome-wide evidence that the endogenous microRNA sites are cooperative with the ARE stability sites when occurring in the same 3'UTR.

This is the first global assessment of candidate co-regulatory 3'UTR motifs that modulate regulation by microRNAs. Our results suggest that microRNA and siRNA binding sites should not be considered in isolation when interpreting and predicting effects of these small RNAs in-vivo. We find that the existence of AU-rich elements, HuD binding sites and other 3' UTR motifs have a significant concomitant impact on mRNA levels. Overall, these discoveries have implications for the analysis of microRNA targeting, siRNA design and the design of small RNA therapeutics.

IDENTIFICATION OF ALTERED REGULATORY PATHWAYS IN CANCER USING INTOGEN SYSTEM

Alba Jene-Sanz, Gunes Gundem, Nuria Lopez-Bigas

Universitat Pompeu Fabra, Research Unit on Biomedical Informatics, Dr. Aiguader 88, Barcelona, 08003, Spain

Cancer is a complex disease that can be triggered through a myriad of dysregulated pathways and mechanisms. The scientific community is applying novel genome-wide approaches that will lead to a more integrated knowledge of this disease. Large consortia, such as ENCODE, are generating highly valuable data that can be exploited to gain insights into the regulatory pathways altered in cancer.

Previously, we have developed IntOGen, a resource that integrates multidimensional OncoGenomics data for the identification of genes and groups of genes (biological modules) involved in cancer development. IntOGen system includes a web resource (www.intogen.org) designed to be a discovery tool for cancer researchers. Now we have used IntOGen system to identify the regulatory pathways misregulated in different types of cancer.

Recently it is becoming clear that a combination of different epigenetic-related factors have a role in the transcriptional dysregulation that leads to transformation of cancer cells. We have collated, annotated and analyzed hight-throughput regulatory data across different cell and cancer types, mainly in the form of ChIP-seq human whole-genomic experiments. Its thorough annotation is crucial to understand the analyses derived from the integration with cancer data, and we thus use appropriate ontologies for classification purposes. Our focus is on histone modifications and polycomb group (PcG) proteins, which regulate silencing in genes poised for activation through chromatin remodeling, in some selected cell types. Some PcG proteins are over-expressed in several human cancers, so aberrations in this system can potentially give rise to global alterations in gene silencing in cancer. Also, there is some evidence for an oncogenic role of PcG, which has traditionally been associated to tumor suppressor activity.

We have analyzed the data at the level of Regulatory Modules (sets of genes that share a common regulatory property) in order to study the mechanisms that regulate the cancer process. We will present IntOGen system and the identification of the main regulatory pathways altered in cancer using this system.

REGULATION OF ALTERNATIVE MRNA POLYADENYLATION IN CELL REPROGRAMMING

Zhe Ji, Bin Tian

UMDNJ-New Jersey Medical School, Department of Biochemistry and Molecular Biology, 185 South Orange Avenue, Newark, NJ, 07101

Over half of mammalian genes contain multiple mRNA polyadenylation sites that lead to transcript variants with different open reading frames and/or 3' untranslated regions (3'UTRs). The alternative polyadenylation (APA) pattern varies across tissues, and is dynamically regulated in proliferating or differentiating cells. Both transcriptional and epigenetic mechanisms are involved in generation of induced pluripotent stem (iPS) cells, in which cells are developmentally reprogrammed. Here we show that regulation of 3′UTRs by APA is also an integral part of iPS cell generation. In general, 3'UTRs shorten during reprogramming of somatic cells, but lengthen during reprogramming of spermatogonial cells. Polyadenylation sites responsible for 3'UTR changes in cell reprogramming show various characteristics, which are consistent with concomitant regulation of the mRNAs encoding polyadenylation factors. We also found that APA regulation is coordinated with other steps of gene regulation, such as microRNA targeting. Taken together, our findings indicate that reprogramming of 3′UTRs by APA, which result from regulation of both general polyadenylation activity and cell type-specific factors, is an important part of the gene regulatory program in iPS cell generation, and the APA pattern can be a good biomarker for cell type and state, useful for sample classification.

A GENOMIC APPROACH TO MAP TRANSCRIPTION PATHWAYS IN *SACCHAROMYCES CEREVISIAE*

Pinay Kainth[1], Jeffrey Fillingham[2], Christoph Kurat[1], Kyle Tsui[1], Corey Nislow[1], Timothy Hughes[1], Jack Greenblatt[1], Brenda Andrews[1]

[1]University of Toronto, The Donnelly Centre, 160 College Street, Toronto, M5S 3E1, Canada, [2]Ryerson University, Department of Chemistry and Biology, Kerr Hall East 122, Toronto, M5B 2K3, Canada

We are using tools of yeast functional genomics to systematically discover new regulators controlling gene expression in Saccharomyces cerevisiae. To identify such proteins, we combined a promoter-reporter system with our capacity for high-throughput genetics to determine the effect of gene deletions on a promoter of interest. Our general strategy utilizes the synthetic genetic array (SGA) platform to systematically introduce a test promoter-GFP reporter construct along with a control promoter-RFP gene into the array of ~5000 yeast deletion mutants. Fluorescence intensities from each reporter are assayed from individual colonies arrayed on solid agar plates using a scanning fluorimager and the ratio of GFP to RFP intensity reveals deletion mutants that cause differential GFP expression.

We carried out a proof-of-principle project by screening the promoter that normally drives transcription of histone H2A (*HTA1*) to identify the regulators responsible for its S-phase specific expression. We discovered that the histone chaperone protein Rtt106, along with the HIR complex, is required for proper repression of histone transcription. Genome-wide nucleosome occupancy assays revealed that these histone chaperones are required for proper chromatin organization at the promoters of histone genes. Furthermore, our genomic screen and follow-up assays revealed a previously unappreciated role for the chromatin boundary protein Yta7 in properly localizing Rtt106 to the promoter region of the *HTA1* gene. Our data suggest that the Yta7-Rtt106/HIR pathway may regulate gene repression broadly throughout the genome and ongoing research involves ChIP-seq experiments to characterize HIR-Rtt106-Yta7 binding genome-wide. We have also expanded our Reporter-SGA screens to include a panel of cell cycle reporters and yeast arrays that allow assessment of other genetic perturbations, including gene overxpression.

DISCOVERING NOVEL INTERACTIONS BETWEEN TRANSCRIPTION FACTORS: THE ALLELE BINDING COOPERATIVITY TEST

Konrad J Karczewski[1], Maya Kasowski[2], Fabian Grubert[1], Manoj Hariharan[1], Mark Gerstein[3], Jan Korbel[4], Michael Snyder[1]

[1]Stanford University, Genetics, Pasteur Drive, Stanford, CA, 94305, [2]Yale University, Molecular, Cellular and Developmental Biology, Whitney Avenue, New Haven, CT, 06520, [3]Yale University, Molecular Biophysics and Biochemistry, Whitney Avenue, New Haven, CA, 06520, [4] EMBL, Genome Biology Research Unit, Meyerhofstrasse, Heidelberg, 69117, Germany

Regulation of gene expression at the transcriptional level is achieved by binding of transcription factors to their target promoter sequences. While there is often a single most important factor for the regulation of a certain gene in a certain condition (such as NFκB in inflammatory response), in reality, transcription is controlled by a combination of factors, each contributing to switching transcription on or off in a given condition. Here, we develop a method to discover potential interactors of a given transcription factor by analysis of differences in ChIP-Seq binding data in 10 individuals. Using this new test, the Allele Binding Cooperativity (ABC) test, we find a number of potential transcription factors associating with NFκB in a search of the JASPAR motif database. Although some factors have been suggested previously to work with NFκB, many have not.

TCAP REVEALS TRANSCRIPTIONAL MODULES IN ARABIDOPSIS

Steven J Kiddle[1,2], Katherine J Denby[1,2], Sach Mukherjee[3,4]

[1]Warwick University, Warwick Systems Biology Centre, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom, [2]Warwick University, Warwick HRI, Wellesbourne, Warwick, CV35 9EF, United Kingdom, [3]Warwick University, Statistics, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom, [4] Warwick University, Centre for Complexity Science, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom

Identifying genes that may be regulating each other from gene expression time series is an important task. Clustering algorithms are often used for this purpose. However, gene regulatory events may induce complex temporal features in a gene expression profile, including time delays, inversions and transient correlations, which are not well accounted for by current clustering methods. As the cost of microarray experiments continues to fall, the temporal resolution of time course studies is increasing. This has led to a need to take account of detailed temporal features of this kind. Thus, while standard clustering methods are both widely used and much studied, their shared shortcomings with respect to such temporal features motivates the work presented here.

Here, we present a temporal clustering approach for high-dimensional gene expression data which takes account of time delays, inversions and transient correlations (Kiddle et al. 2009). It exploits a recently introduced, message-passing-based algorithm called Affinity Propagation (AP). It takes account of temporal features of interest following an approximate but efficient dynamic programming approach due to Qian et al. (2001). The resulting approach is demonstrably effective in its ability to discern non-obvious temporal features, yet efficient and robust enough for routine use as an exploratory tool. We show results on validated transcription factor-target pairs in yeast and on gene expression data from a study of *Arabidopsis thaliana* under pathogen infection. The latter reveals a number of biologically striking findings.

THE BIOPHYSICAL BASIS OF TRANSCRIPTIONAL REGULATION REVEALED THROUGH DEEP SEQUENCING

Justin B Kinney[1,2,5], Anand Murugan[1], Curtis G Callan[1,3], Edward C Cox[1,4]

[1]Princeton University, Department of Physics, Princeton University, Princeton, NJ, 08544, [2]Princeton University, Lewis-Sigler Institute, Princeton University, Princeton, NJ, 08544, [3]Princeton University, Princeton Center for Theoretical Science, Princeton University, Princeton, NJ, 08544, [4] Princeton University, Department of Molecular Biology, Princeton University, Princeton, NJ, 08544, [5]Cold Spring Harbor Laboratory, Koch Building, 1 Bungtown Rd., Cold Spring Harbor, NY, 11724

Cells use protein-DNA and protein-protein interactions to regulate transcription. A biophysical understanding of this process at individual promoters and enhancers, however, has been hampered by a lack of sufficiently powerful experimental methods. Here we show how deep sequencing can be used to quantitatively characterize, *in vivo*, the protein-DNA and protein-protein interactions that allow a targeted regulatory sequence to modulate transcription. To extract such biophysical information from sequence data, we use a fundamental relationship [1] between likelihood and mutual information. As a result, our analysis method is highly robust to both experimental noise and variations in experimental design. Applying our experimental and theoretical tools to the *E. coli lac* promoter, we identified regulatory protein binding sites *de novo*, determined the sequence-specificity of the proteins that bind these sites, and even measured the *in vivo* interaction energy between a transcription factor and RNA polymerase in their native DNA-bound configuration. The principles of our method can be readily applied to a wide range of problems in molecular biology, opening the door to many new applications for ultra-high-throughput DNA sequencing.

[1] Kinney, J.B., Tkačik, G., and Callan, C. (2007). Precise physical models of protein-DNA interaction from high-throughput data. Proc. Natl. Acad. Sci. USA *104*, 501-506.

ICLIP REVEALS THE FUNCTION OF HNRNP PARTICLES IN SPLICING AT INDIVIDUAL NUCLEOTIDE RESOLUTION

Julian Konig[1], Kathi Zarnack[2], Gregor Rot[3], Tomaz Curk[3], Nick M Luscombe[2], Jernej Ule[1]

[1]MRC Laboratory of Molecular Biology, Structural Studies, Hills Road, Cambridge, CB2 0QH, United Kingdom, [2]European Bioinformatics Institute, EBI, Welcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom, [3]University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, Ljubljana, SI-1000, Slovenia

In the nucleus of eukaryotic cells, nascent transcripts undergo multiple processing steps. Nascent transcripts are generally associated with heterogeneous nuclear ribonucleoprotein (hnRNP) particles that are nucleated by hnRNP C. Despite their abundance, it remained unclear whether these particles control pre-mRNA processing. Here, we developed individual-nucleotide resolution UV-crosslinking and immunoprecipitation (iCLIP) to study the role of hnRNP C in splicing regulation. iCLIP data demonstrate that hnRNP C recognizes uridine tracts with a defined long-range spacing in pre-mRNAs consistent with hnRNP particle organization. hnRNP particles assemble on both introns and exons, but remain generally excluded from splice sites. Integration of transcriptome-wide iCLIP data and alternative splicing profiles into an 'RNA map' indicates how the positioning of hnRNP particles determines their effect on inclusion of alternative exons. The ability of high-resolution iCLIP data to provide insights into the mechanism of this regulation holds promise for studies of other higher-order ribonucleoprotein complexes.

AN ALIGNMENT-FREE SEQUENCE COMPARISON MODEL FOR DETECTION OF FUNCTIONAL CONSERVATION OF REGULATORY SEQUENCES

Hashem H Koohy[1], Sascha S Ott[2], Georgy G Koentges[3]

[1]MOAC, Warwick University, Coventry, CV4 7AL, United Kingdom, [2]Systems Biology Centre, Warwick University, Coventry, CV4 7AL, United Kingdom, [3]Biological Sciences, Warwick University, Coventry, CV4 7AL, United Kingdom

Recent studies show that cis-regulatory modules (CRMs) may function similarly despite their substantial sequence divergence (M. Ludwig et al. PLosBIOLOGY 2005 and Hare et al. PLosGENETICS 2008). These establish that some of the CRMs must share some common patterns that drive almost identical regulatory outputs but possibly with different arrangements and hence it may be too hard to detect their functional similarity by alignment-based sequence comparison tools.
We have developed an alignment-free sequence comparison model for the detection of functional conservation of regulatory sequences based on the distribution of transcription factor binding sites. It has been aimed to detect:
Other enhancers in the same genome that are likely to have similar function to the enhancer of interest.
Functionally conserved regions in the orthologous genomes, even if the enhancer regions do not align.
As input, our model takes a template sequence, a test sequence and a set of transcription factor motifs. The mathematical framework of the model has two main components. In the first model component we compute the probability of each of the possible valid configurations, then we associate each motif with its expected number of occurrences in the given sequence over all valid configurations. The output of the first model component is therefore a mathematical vector representing the strength and multiplicity of each of the motifs with respect to the sequence . This model component is in fact a modification of an elegant thermodynamic model developed by T. Raveh-SAdka et al. (Nature 2008). The second model component is to establish a pairwise similarity score. This takes place from the comparison of the expected vectors associated to both template and test sequences by defining a probabilistic measure function.
We apply our model to some fly sequences to detect orthologues of some known enhancers from D.melanogaster in other drosophila species.

A MACHINE LEARNING FRAMEWORK FOR INTEGRATIVE ANALYSIS OF ENCODE II DATA

Anshul Kundaje[1], Max Libbrecht[1], Serafim Batzoglou[1], Arend Sidow[2]
[1]Stanford University, Computer Science Dept., James H. Clark Center, 318 Campus Drive, RM S266, Stanford, CA, 94305-5428, [2]Stanford University, Pathology Dept., SUMC L235 (Edwards Bldg), Stanford, CA, 94305-5324

The ENCODE consortium is generating a variety of genome-wide, functional genomics datasets in several cell lines using deep sequencing technologies. These include transcription factor (TF) binding profiles, histone modification profiles, open-chromatin assays, transcriptome data and high quality annotations (GENCODE). We present a supervised machine learning framework to integrate and automatically learn relationships between these diverse datasets. We define several biologically motivated classification problems. We use boosting algorithms as well as support vector machines to learn high-accuracy, discriminative models. We restrict analysis to ENCODE datasets from the K562 and Gm12878 cell lines.

First, we present a discriminative analysis of transcription start sites (TSSs). We train binary classifiers using sequence-based and epigenomic features defined in windows around various genomic locations labeled as TSS or non-TSS. The features capture the signal 'shape' and 'strength' of the ENCODE data tracks. The learned models exhibit high classification accuracy (auROC=0.8). High DNase-I hypersensitivity and CpG content are strongly associated with TSS locations whereas H3k4me1, H3k36me3 and H4k20me1 signals are anti-correlated with TSS locations. We then define a multi-class classification problem to predict cell-line specific TSS activity. We use RNA-seq and CAGE data along with Pol2 ChIP-seq data to divide TSSs into 3 classes – active (Pol2+, RNA/CAGE+), poised (Pol2+, RNA/CAGE-) and inactive (Pol2-, RNA/CAGE-). We are able to discriminate inactive TSSs from Pol2-bound TSSs with high accuracy (auROC=0.95). Inactive TSSs are characterized by a lack of H3k4me3, H3k4me2 and H3k27ac marks and the occasional presence of the repressive H3k27me3 mark. Active TSSs have higher levels of H3k9ac as compared to poised TSSs.
We also use our boosting framework to learn models of in-vivo binding of the c-Jun and c-Myc TFs. Peaks in TF ChIP-seq experiments enable us to label genomic locations as being bound or unbound by a TF in a particular cell-line. We use this data as our training labels in a binary classification setting to learn sequence-motifs as well as epigenomic marks that can accurately (auROC=0.86) discriminate bound and unbound loci. The algorithm learns several variants of each TF's DNA binding motif. However, the inclusion of epigenomic features is essential to obtain high prediction accuracy. DNase-I hypersensitivity and several histone-3 methylation marks contribute significantly to improving the classification accuracy by >20% over using sequence features alone.
This research is supported by NIH grant U54 HG004576.

NUCLEOSOME SLIDING AND REMODELING: LIMITS AND PATTERNS

Yaakov Belch[1], Jingyi Yang[2], Sridhar A Malkaram[3], Rong Liu[4], Jean-Jack M Riethoven[5], Istvan Ladunga[6]
[1]University of Nebraska-Lincoln, Statistics, telecommuting, Chativat Harel 3/4, Karmiel, 21613, Israel, [2]Microsoft, R & D, 1 Microsoft Way, Redmond, WA, 98052, [3]University of Nebraska-Lincoln, Center for Biotechnology, 1901 Vine St., Lincoln, NE, 68588, [4] University of California Los Angeles, Chemical and Biological Engineering, 5409 Boelter Hall, Los Angeles, CA, 90095, [5]University of Nebraska-Lincoln, School for Biological Sciences, E204 Beadle, 1901 Vine St., Lincoln, NE, 68588-0665, [6]University of Nebraska-Lincoln, Statistics, E145 Beadle, 1901 Vine St., Lincoln, NE, 68588-0665

Nucleosome sliding, remodeling and eviction are poorly understood mechanisms for balancing between DNA accessibility and genome compaction. To examine the enigmatic co-existence of nucleosomes with the progressing transcriptional machinery, we designed and implemented minimally biased computational methods to analyze 17 published *in vivo* and *in vitro* genome-wide nucleosome maps in yeast and 2 *in vitro* maps of the ovine β-lactoglobulin gene. In the yeast experiments, peaks of sequencing reads represent alternative nucleosome loci in populations of millions of cells. We mapped the majority of H2A.Z variant nucleosomes to reproducible loci of ~140-153 base pairs on the genome. This confirms not only their static nature but also the fidelity of the experiments, including low sequencing, amplification and micrococcal nuclease cleavage bias. In contrast to the H2A.Z variants, most nucleosomes containing the canonical H2A subunits form peaks that span significantly wider than the general ~147 base pair footprint of a single histone octamer on the DNA. Peaks as wide as multiple footprints are likely artifacts caused by the limited accessibility of DNA for nuclease digestion. However, thousands of nucleosome peaks that are either narrower than a double footprint or sterically isolated from others clearly show extensive histone remodeling and/or sliding.

We found significantly higher mobility in coding sequences, CpG islands and evolutionary conserved domains than in promoters and introns. Sliding and eviction are prominent at the mid-third of actively transcribed genes. To reveal the limits to remodeling, we searched for large-scale remodeling events that would span over 2000 base pairs but found no one. Typically, the fuzziest nucleosomes are surrounded by gradually narrowing peaks, possibly indicating a series of weak momentum absorber and/or relatively eviction-resistant nucleosomes. The results indicate the highly statistical nature of sequence-specific nucleosome positioning with implications on transcription, regulation and chromatin structure.

This work was supported by NSF Grant EPS-0701892.

C_2H_2 ZINC FINGER MODULES TYPICALLY RETAIN SEQUENCE SPECIFICITY IN MODULAR ASSEMBLIES

Kathy N Lam[1], Timothy R Hughes[1,2]

[1]University of Toronto, Department of Molecular Genetics, 160 College Street, Toronto, M5S 3E1, Canada, [2]University of Toronto, Banting and Best Department of Medical Research, 160 College Street, Toronto, M5S 3E1, Canada

C_2H_2 zinc fingers are the most prevalent type of vertebrate transcription factor DNA-binding domain. They generally contain multiple fingers, which can behave in a modular manner, and this quality may be responsible for the motif's evolutionary success. Taking advantage of their modularity, novel zinc finger proteins have been engineered using a modular assembly approach: for example, three individual C_2H_2 zinc fingers, each with a known 3-base specificity, can be combined to generate a novel zinc finger array (ZFA) that has a desired 9-base binding site. However, a recent survey of a large number of ZFAs constructed by modular assembly reported unexpectedly high failure rates (~70%) (Ramirez et al., Nature Methods, 2009), casting doubt on the general modularity of C_2H_2 zinc fingers and complicating analysis of their sequence specificity and its determinants.

Here, we used Protein Binding Microarrays to analyze several dozen ZFAs that failed in the aforementioned study. We find that most of these engineered ZFAs do in fact bind specifically to DNA containing sequences resembling their intended targets. Similarly, most individual zinc fingers retain their known DNA-binding specificity when modularly assembled into functional ZFAs, including cases in which we tested multiple contexts for the same zinc finger. Our findings are in agreement with previous work describing C_2H_2 modularity, and support the validity of efforts to understand the determinants of individual zinc finger sequence specificity as well as computational prediction of DNA-binding sites for zinc finger proteins with unknown specificities.

UNDERSTANDING LYSINE ACETYLATION SIGNALING

Benjamin Lang, M. Madan Babu

MRC Laboratory of Molecular Biology, Structural Studies Division, Hills Road, Cambridge, CB2 0QH, United Kingdom

Post-translational modifications occur in the majority of proteins. A portion of dynamic, reversible modifications have been studied in depth in the context of the unstructured histone tails in chromatin. Their potential for combinatorial logic in signaling is vast, and it is likely that these essential signals of the elaborate eukaryotic transcriptional regulatory system are also used extensively outside the nucleus.

Very recently, evidence that lysine acetylation of non-histone proteins is widespread and conserved from bacteria to humans has been acquired by mass spectrometry. Our functional understanding of such modifications, however, is still limited to a small number of exemplary cases, where one modification may exclude another, create or disrupt an interaction surface or affect enzyme activity. Detailed examples of this kind have been found among chromatin factors and in receptor tyrosine kinase endocytosis. By applying a data integration approach, we aim to understand the overall dynamics and higher-order characteristics of modification networks, as well as uncovering new interactions.

Here, we present a theoretical framework for the inference of functional relationships within modification networks, taking into account protein-protein interactions, domains, localization, conservation and disordered regions. We apply this framework to lysine acetylation signaling in humans.

DIRECT OBSERVATION OF TRANSCRIPTION INITIATION AND ELONGATION CONTROL IN LIVING CELLS

Daniel R Larson[1,2], Daniel Zenklusen[1], Robert H Singer[1,2]

[1]Albert Einstein College of Medicine, Anatomy and Structural Biology, 1300 Morris Park Ave., Bronx, NY, 10031, [2]Albert Einstein College of Medicine, Gruss-Lipper Biophotonics Center, 1300 Morris Park Ave., Bronx, NY, 10031

The level of expression of individual genes is determined by the balance of production and degradation of mRNA and is subject to stochastic fluctuations. The synthesis of mRNA is mediated by a single class of enzymes, the RNA polymerases, which exhibit stochastic behavior, including variable elongation velocity and pausing, when observed at the single molecule level *in vitro*. Whether such variation in RNA polymerase dynamics exists *in vivo* and what the downstream consequences are for gene expression is unknown. Previously, we showed that transcript levels are 3-fold higher than previously predicted. Here we show how transcription initiation and elongation of individual RNA polymerases are independently controlled on a single endogenous gene in a single living cell. We are able to directly observe the enzymatic activity of single RNA polymerase II molecules in *Saccharomyces cerevisiae* by observing the production of fluorescently-labeled RNA at an active locus. We have developed a method of single molecule fluctuation analysis for transcription which allows us to separate the effects of initiation and elongation. Using this single molecule live-cell approach, we find that initiation of the housekeeping gene *MDN1* occurs via a single rate-limiting step, but elongation kinetics are the result of many sequential steps, none of which are rate-limiting or pause-inducing. Using a reporter gene with the complete *POL1* upstream activating sequence, we directly observe diffusion-limited initiation rates when the gene is active at the G1/S transition. This *in vivo* transcription rate can be explained by a model of facilitated diffusion for a transcription factor to find its target sequence in the nucleus, a model that synthesizes information on nucleosome occupancy and identity of upstream factors determined by protein binding microarrays. These measurements of transcription initiation kinetics, independent of downstream processes such as mRNA decay, translation rates, and protein decay provide a template for a quantitative description of in vivo genetic networks.

MULTI-GENE CHROMATIN DOMAINS FOUND IN MOUSE GENOME VIA HIDDEN MARKOV MODELS

Jessica L Larson[1], Guo-Cheng Yuan[1,2]

[1]Harvard School of Public Health, Biostatistics, 677 Huntington Avenue, Boston, MA, 02115, [2]Dana-Farber Cancer Institute, Biostatistics and Computational Biology, 44 Binney Street, Boston, MA, 02115

Eukaryotic chromatin is not uniformly packaged but is organized into higher-order (i.e., multi-gene) regulatory domains. There is evidence that neighboring genes, although not always involved in the same pathways, are still similarly regulated at the level of transcription via various histone modifications. We discovered and characterized 76 of these largest domains through a novel genome-wide analysis of chromatin immunoprecipitation following by DNA sequencing (ChIP-seq) histone modification data in mouse embryonic stem cells. We examined the activity of five of these modifications (H3K4me2, H3K4me3, H3K27me3, H3K9me3, H3K36me3) at all known mouse genes. To determine each domain we first obtained a 5-dimensional score for each gene based on average modification activity in select gene regions (e.g. near the transcription start site). Then, with hidden Markov models and corresponding algorithms, we were able to determine the most probable domain status of each gene. To verify our findings biologically, we found that our domain boundaries are often in agreement with prior literature concerning gene clusters. Our method located the known olfactory receptor (OR) and *Hoxa* gene clusters. Moreover, certain domains contain genes only found in select gene ontology (GO) groups. We also noted less gene expression variability within each of our domains when compared to randomly selected boundaries (p-value<0.0001). We thus have evidence of multi-gene domains in mouse stem cells, which are characterized by similar patterns in five histone modifications. As we continue to apply our method to other cell lines, we will provide important insight into the general structure, organization, and regulation of the mammalian genome.

IDENTIFYING THE GENETIC DETERMINANTS OF TRANSCRIPTION FACTOR ACTIVITY

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

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

Understanding how phenotype relates to genotype is one of the central goals of biology. Analysis of parallel genotyping and expression profiling data has shown that mRNA expression level is highly heritable. The molecular mechanisms underlying the heritability of gene expression levels are poorly understood, but they are expected to often involve mediation by transcription factors (TFs). We here present a transcription-factor-centric and sequence-based method for dissecting the transcriptional response to genetic perturbations. In our approach, we first predict the affinity with which each TF binds to the promoter region of each gene using quantitative prior information about the DNA binding specificity. Next, we perform genomewide linear regression of differential mRNA expression on predicted promoter affinity to estimate segregant-specific TF activity as a quantitative phenotype. Genetic mapping of the TF activity trait allows us to identify the activity quantitative trait loci ("aQTLs") whose inheritance modulates the regulatory activity of each specific TF. Our method has a greatly improved statistical power to detect regulatory mechanisms underlying the heritability of genomewide mRNA expression. Specifically, it identifies seven times as many locus-TF associations and more than twice as many trans-acting loci from a genetic cross between two haploid yeast strains [2] as all existing methods combined. We validated our ability to predict locus-TF associations in yeast using gene expression profiles for allele replacement strains. Furthermore, application to mouse data from an F2 intercross [3] identified an aQTL on chromosome VII modulating the activity of Zscan4 in liver cells, demonstrating that our method also works in higher eukaryotes.

1. Foat, B.C., R.G. Tepper, and H.J. Bussemaker, TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. Nucleic Acids Res, 2008. 36(Database issue): p. D125-31.
2. Brem, R.B. and L. Kruglyak, The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A, 2005. 102(5): p. 1572-7.
3. Schadt, E.E., et al., Genetics of gene expression surveyed in maize, mouse and man. Nature, 2003. 422(6929): p. 297-302.

TRANSCRIPTION REGULATORY NETWORK OF ALZHEIMER'S DISEASE

Jiya Sun, Hongxing Lei

Beijing Institute of Genomics, CAS, Bioinformatics, No.7 Beitucheng West Road, Beijing, 100029, China

The AlzGene database contains hundreds of candidate AD (Alzheimer's disease) genes from genome wide association studies (GWAS). It is one of the best resources for genetic/genomic studies of Alzheimer's disease. However, the relationship among those candidate genes is still unclear. In mammalian cells, mRNAs are transcribed starting from the transcription start sites (TSS) by the RNA polymerase□with the help of transcription factors (TF) and usually there is a 5'-UTR ahead of start codon. TSSs are DNA sequence motifs where the RNA polymerase□binds and begins to synthesize the first ribonucleotide of mRNA. It has been shown in previous studies that a single gene may have more than one TSS, thus can have multiple gene products. It has also been demonstrated that TFs tend to function in the form of cis-regulation module (CRM) rather than regulating the gene expression alone by itself.

In this study, we focused on how the candidate AD genes interact on the level of gene transcription regulation. First, we retrieved upstream sequences relative to the precise TSSs. We used the DBTSS database where TSS sequences obtained from eight tissues and six cell lines can be downloaded. These TSS sequences can be mapped to human genome and the distribution of TSS sequences on human genome can be displayed. Thus, we can retrieve the reliable regulatory sequences referring to the precise TSS sequences. Next, we used the tool matrix-scan, a component of RSAT which is available on the website http://rsat.ulb.ac.be/rsat/, to scan the prepared regulatory sequence with the TF position weight matrix (PWM) and to analyze the distribution of known TF binding motifs in the target regulatory sequences. The TF PWMs, the total being 62, were downloaded from the Jaspar database at the website http://jaspar.cgb.ki.se/. In this step, CRMs of regulatory sequences can be displayed statistically, according to which we can find out whether genes are co-regulated by some common TFs. Finally, after we obtained the relationship between TFs and targets, we studied the regulatory strength of a particular TF based on gene expression chips. Hereby, we applied the network component analysis (NCA), which has been proved to be effective, to the quantitative analysis of the specific TFs. In summary, from the above-described investigation, a clearer picture of the transcription regulatory network of AD genes has emerged.

GENOME-WIDE ANALYSIS OF FOXO1 OCCUPANCY IN MOUSE NEURAL STEM CELLS USING NEXT-GENERATION SEQUENCING

Willey Liao[1,3], Ji-hye Paik[2], Yifan Mo[1,3], Justin B Kinney[1], Ronald A Depinho[2], Michael Q Zhang[1,4]
[1]Cold Spring Harbor Laboratory, Computational Biology, 1 Bungtown Road, Cold Spring Harbor, NY, 11724, [2]Dana Farber Cancer Institute, Medical Oncology, 44 Binney Street, Boston, MA, 02115, [3]Stony Brook University, Applied Mathematics and Statistics, Math Tower, Stony Brook, NY, 11790, [4] Tsinghua University, Automation, Central Main Building, Beijing, 100084, China

The FOXO transcription factors are linked, via the PI3K-AKT pathway, to regulation of several cellular functions such as apoptosis, glucose metabolism, oxidative stress response, and cell cycle control. Dysregulation of the FOXOs can lead to cancer, glaucoma, diabetes, and other disease phenotypes. Expression patterns of FOXOs can vary significantly between multiple tissue types and developmental stages, suggesting a temporal and spatial mechanism that contributes to its context-specific functionality. In mouse neural stem cells (NSCs), in particular, FOXOs appear to be involved in stem cell homeostasis through maintenance of stem cell quiescence and regulation of self-renewal properties related to neurogenesis. Understanding the NSC-specific FOXO landscape may provide a glimpse into its context-dependent role in the central nervous system.

To address the NSC-specific functions of FOXO transcription factors, we characterized FOXO1 transcription factor binding sites using chromatin immunoprecipitation coupled with high-throughput, massively-parallel, next-generation sequencing (ChIP-seq). This technology provides a high-resolution, genome-wide approach that enables a global analysis of FOXO1 occupancy. We used a peak-calling algorithm to infer putative FOXO1 binding sites, which we subsequently searched for enrichment of binding motifs to gain a more comprehensive understanding of potential FOXO1 target genes in mouse NSCs. Furthermore, it has been suggested that FOXO target specificity is determined by distinct cofactor interactions. We therefore searched regions flanking putative binding sites for additional proximal cofactor binding motifs perhaps involved in the modulation of FOXO1's functional specificity. We report here the results of our genome-wide profiling of FOXO1 occupancy in NSCs, which we hope will help in clarifying the FOXO1-mediated regulatory network and the role it plays in stem cell homeostasis.

Acknowledgement:
Grant support comes from the STARR foundation award (MQZ), NIH grants, ES017166 (MQZ), the Claudia Adam Barr Foundation (J-H.P), NIH grants, 5P01CA95616 (R.A.D). R.A.D. is an American Cancer Society Research Professor and supported by the Robert A. and Renee E. Belfer Foundation Institute for Innovative Cancer Science.

PREDICTING MEMBERSHIP OF REGULATORY PROTEIN COMPLEXES – INTEGRATING PROTEIN INTERACTION DATA WITH TRANSCRIPTIONAL REGULATION.

Howard B Lightfoot[1], Sascha Ott[2], Gynanendra Tripathi[3]

[1]University of Warwick, MOAC Doctoral Training Centre, Coventry House, Coventry, CV4 7AL, United Kingdom, [2]University of Warwick, Warwick Systems Biology Centre, Coventry House, Coventry, CV4 7AL, United Kingdom, [3]University of Warwick Hospital, Clinical Sciences Research Institute, Clifford Bridge Road, Coventry, CV2 2DX, United Kingdom

Most bioinformatics models of DNA-protein interaction examine only the likelihood of a single transcription factor binding a given sequence, but proteins rarely act alone but in concert. Here we present a method to combine DNA binding data with protein-protein interaction data to identify putative protein members of DNA-regulatory complexes.

First protein interaction networks are extracted from public databases based on one or more seed proteins. Probabilities are then assigned to the network edges (the likelihood of a protein-protein interaction) and nodes (the likelihood of an interaction with DNA). Random sampling of both the nodes and edges provides a collection of sub-networks that represent hypothetical complexes. By using a flexible sampling approach the probability model is easily adapted to the biological system of interest. A wide variety of information can be integrated to score both edges and nodes, ranging from experimental type for a given binary interaction to nuclear localization (GO annotation) for individual proteins.

Complex membership can then be tested using experimental methods e.g. ChIP. This method/tool offers biologists a way to sift large protein-protein interaction (interactome) datasets for relationships of particular interest. As a test case we examine interactions between proteins involved in the regulation of adipogenesis and endoplasmic reticulum stress in humans.

YEAST AXIAL ELEMENT PROTEIN RED1 BINDS SUMO CHAINS TO PROMOTE MEIOTIC INTERHOMOLOG RECOMBINATION AND CHROMOSOME SYNAPSIS

Feng-Ming Lin[1,2], Ting-Fang Wang[1,2]

[1]Graduate Institute of Life Sciences, National Defense Medical Center, Min-Chuan East Road, Taipei, 114, Taiwan, [2]Institute of Molecular Biology, Academia Sinica, Academia Road, Taipei, 115, Taiwan

The synaptonemal complex (SC) is a tripartite protein structure consisting of two parallel axial elements and a central region. During meiosis, the SC connects paired homologous chromosomes, promoting interhomolog recombination. Here, we report that, like the CE component Zip1, Saccharomyces cerevisiae axial element structural protein Red1 can bind small ubiquitin-like modifier (SUMO) polymeric chains. The Red1-SUMO chain interaction is dispensable for the initiation of meiotic DNA recombination, but it is essential for Tel1- and Mec1-dependent Hop1 phosphorylation, which ensures interhomolog recombination by preventing the inter-sister chromatid DNA repair pathway. Our results also indicate that Red1 and Zip1 may directly sandwich the SUMO chains to mediate SC assembly. We suggest that Red1 and SUMO chains function together to couple homologous recombination and Mec1/Tel1 kinase activation with chromosome synapsis during yeast meiosis

INFER THE TRANSCRIPTION REGULATORY NETWORK IN MOUSE INTESTINE DEVELOPMENT FROM HISTONE MARK DYNAMICS

Hyunjin Shin[1,2], Michael Verzi[1,3], Cliff Meyer[1,2], Ramesh Shivdasani[1,3], X. Shirley Liu[1,2]

[1]Dana-Farber Cancer Institute, Biostatistics & Computational Biology, 44 Binney Street, Boston, MA, 02115, [2]Harvard School of Public Health, Biostatistics, 677 Huntington Avenue, Boston, MA, 02115, [3]Harvard Medical School, Medical Oncology & Cancer Biology, 25 Shattuck Street, Boston, MA, 02115

Nucleosome-resolution ChIP-seq of H3K4me2 is a cost-effective approach to obtain the genome-wide nucleosome occupancy information at functional promoters and enhancers. Between two related biological conditions, differential transcription factor binding events are often accompanied by a distinct H3K4me2 profile, where the nucleosomes at the transcription factor binding sites are displaced showing reduced H3K4me2 signals, while the two nucleosomes flanking the binding sites are better positioned with increased H3K4me2 signal. By selecting genome-wide regions with these dynamic H3K4me2 dynamic profiles and conducting transcription factor motif discoveries within the two flanking nucleosomes, we could identify driving transcription factors in a biological process, and infer their genome-wide binding locations and regulated genes. In addition, the regulators of these driving transcription factors can also be predicted from the differential H3K4me2 regions near the factors' genes. Therefore, using histone mark ChIP-seq, gene expression profiling, and bioinformatics motif analysis, we could detect stimulus dependent transcription factor activities and deduce the transcription regulatory network in a biological process. We demonstrate the application of this technique to study mouse intestine development, and infer the regulatory network involving Cdx2, HNF4, GATA6, as well as Tcf, Sox, Fox, and Hox factors.

SIGNIFICANCE-BASED CLUSTERING OF GENE EXPRESSION DATA

Marta Łuksza[1], Michael Lässig[2], Johannes Berg[2]

[1]Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Ihnestraße 63-73, Berlin, 14195, Germany, [2]Universität zu Köln, Institut für Theoretische Physik, Zülpicher Straße 77, Köln, 50937, Germany

The identification of groups of co-expressed genes is a key step in gene expression analysis. Clustering is used widely to infer putative functional relationships between genes, with gene expression clusters arising for instance through common biological pathways or shared modes of regulation. Clustering has also been applied to classify tumors based on their expression profiles.

But could such clusters have arisen simply by chance, without an underlying biological mechanism? Even unrelated objects can form cluster-like structures simply due to density fluctuations. To distinguish such random clusters from biological signals, we compute the cluster *p*-value: the probability that a random data set contains a cluster with similarity score equal or higher than a given score *S*. Based on a probabilistic model and using methods from statistical mechanics, we obtain an analytical solution to the cluster significance problem. This result is analogous to the Karlin-Altschul theorem in sequence alignment. As the cluster *p*-value is a natural measure to choose between alternative clusterings, we propose a new, parameter-free clustering method. This significance-based clustering identifies clusters with minimal combined *p*-value.

We apply significance-based clustering to genome-wide expression data in yeast with samples corresponding to various environmental stress conditions. Tracking the dependence of co-expression clusters on cluster scoring parameters and cluster sizes, we find a striking correspondence between cluster *p*-value and the biological significance of a cluster (quantified by enrichment in Gene Ontology terms). Moreover, our significance-based clustering method finds more distinct and functionally enriched clusters than standard methods, such as k-means clustering and Gaussian mixture-models, agglomerative approaches or superparamagnetic clustering. Similarly, our method outperforms other approaches in the tumor classification problem, as shown on several datasets.

AN EPIGENETIC SWITCH THAT LINKS INFLAMMATION TO CANCER: REGULATORY CIRCUITS AND CANCER TREATMENT

Kevin Struhl, Dimitrios Iliopoulos, Heather A Hirsch, Marianne Lindahl-Allen

Harvard Medical School, Biological Chemistry & Molecular Pharmacology, 240 Longwood Ave., Boston, MA, 02115

Transient activation of Src oncoprotein can mediate an epigenetic switch from immortalized breast cells to a stably transformed line that forms self-renewing mammospheres that contain cancer stem cells. Src activation triggers an inflammatory response mediated by NF-κB that directly activates Lin28 transcription and rapidly reduces the level of let-7 microRNAs. Let-7 inhibits IL6 expression through a direct interaction with the 3'UTR of IL6 mRNA, thus resulting in higher levels of IL6 than achieved by direct activation by NF-κB. IL6 activates NF-κB, thereby completing a positive feedback loop. In addition, IL6 activates the STAT3 transcription factor, which directly activates the miR-21 and miR-181b microRNAs. These microRNAs are sufficient to induce transformation, and this occurs by their respective inhibition of the PTEN and CYLD tumor suppressors, which results in NF-κB activation and induction of the positive feedback loop. This complex regulatory circuit is operates in other cancer cells lines and in a variety of human cancers. Thus, inflammation activates a positive feedback loop that maintains the epigenetic transformed state for many generations in the absence of the inducing signal.

AN ATLAS OF OPEN CHROMATIN SPANNING DIVERSE HUMAN CELL TYPES IN HEALTH AND DISEASE

Jorge Ferrer[3], Marie P Fogarty[2], Kyle J Gaulton[2], Paul G Giresi[1], Linda L Grasfeder[1], Jason D Lieb[1], Karen L Mohlke[2], Takao Nammo[2], Tami M Panhuis[2], Lorenzo Pasquali[2], Charles M Perou[2], Jeremy M Simon[1,2]

[1]University of North Carolina at Chapel Hill, Department of Biology, 408 Fordham Hall, Chapel Hill, NC, 27599, [2]University of North Carolina at Chapel Hill, Department of Genetics, 5096 Genetic Medicine Building, Chapel Hill, NC, 27599, [3]Hospital Clínic de Barcelona, Department of Endocrinology, Institut d'Investigacions Biomediques August Pi i Sunyer, Barcelona, 08036, Spain

FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) is a simple, low-cost genomic method for the isolation and identification of nucleosome-depleted regions in eukaryotic cells. FAIRE-seq data from many human cell lines, clinical breast tumor samples, and pancreatic islet cells will be presented. The typical cell type appears to contain about 100,000 open regions, with sites of open chromatin varying widely between cell types. Clinical breast tumors exhibit characteristic chromatin profiles that can readily distinguish tumor subtype and hormone responsiveness. In human Islets, rs7903146, a *TCF7L2* intronic variant strongly associated with type 2 diabetes, is located in islet-selective open chromatin. rs7903146 heterozygotes exhibit allelic imbalance in islet FAIRE signal, and the variant alters enhancer activity, indicating that genetic variation at this locus acts in cis with local chromatin and regulatory changes. Thus, FAIRE data can guide identification of functional regulatory variants important for human disease.

Authors are listed alphabetically.

ULTRA-HIGH RESOLUTION NUCLEOSOME ORGANIZATION MAPS AND GENE EXPRESSION ANALYSIS IN PURIFIED PRIMARY HUMAN CELLS

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

[1]Stanford University School of Medicine, Pathology and Genetics, 300 Pasteur Drive, Stanford, CA, 94305, [2]Brigham Young University, Microbiology and Molecular Biology, 775 WIDB, Provo, UT, 84602

Chromatin organization of a cell is defined by interactions between nucleosomes packaging the DNA, transcription factors binding to their response elements, and chromatin remodeling complexes changing the chromatin landscape according to internal and external signals. What forces affect positioning of nucleosomes, to which degree positioning is reproducible across different tissues, what determines nucleosome occupancy, how histone modifications reflect or govern gene regulation, and how nucleosomes affect expression, are currently actively debated topics in chromatin systems biology. Using SOLiD sequencing we genereated several billion reads from nucleosome core fragments from three blood cell types, from in vitro reconstituted chromatin, and from micrococcal nuclease-digested naked DNA. DNA sequence has influence on, but does not determine, nucleosome positioning. Certain functional features of chromatin such as promoters exhibit strong signals of nucleosome positioning, enrichment, or depletion, that are not dictated by sequence. Other functional features, such as exons, exhibit sequence-correlated chromatin structure. We furthermore investigate how nucleosome organization and histone modifications interact with gene expression, revealing striking global and local differences in chromatin organization among the cell types.

GENOME-WIDE LONG RANGE CHROMATIN INTERACTIONS AND TRANSCRIPTION REGULATION NETWORKS

Yijun Ruan

Genome Institute of Singapore, Genome Technology and Biology, 60 Biopolis Street #02-01, Genome, Singapore, 138672, Singapore

Genomes are thought to be organized into high-level 3-dimensional structures, and DNA elements separated by long genomic distances are expected to functionally interact. This view has been furthered emphasized by the abundant observations that many transcription factor binding sites are remote to gene promoters. To study transcription regulation mediated by potential long range chromatin interactions in genome-wide manner, we have developed the Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) for de novo detection of global chromatin interactions, and have comprehensively mapped the chromatin interaction network bound by oestrogen receptor α (ERα) in the human genome (Nature 2009 462: 58-64). We now apply the ChIA-PET strategy to map other specific chromatin interactions mediated by a number of specific transcription factors and general chromatin interactions involved in all transcription regulation using general factors such as RNAPII and histone markers as analysis targets. Our results suggest that long range chromatin interaction is a primary mechanism for transcription regulation networks in mammalian genomes.

LONG-RANGE GENE REGULATORY ARCHITECTURE OF HUMAN CHROMOSOME 21

Nynke L van Berkum[1], Richard Humbert[2], Bryan R Lajoie[1], Erez Lieberman-Aiden[3,4], Tobias Ragoczy[5,6], Robert Thurman[2], Louise Williams[3], M. A Bender[7], Mark Groudine[5,6], Eric S Lander[3], John Stamatoyannopoulos[2], Job Dekker[1] [1]University of Massachusetss Medical School, Program in Gene Function and Expression and Department of Biochemistry and Molecular Pharmacology, 364 Plantation Street, Worcester, MA, 01605, [2]University of Washington, Department of Genome Sciences, Seattle, WA, 98195, [3]Broad Institute of Harvard and MIT, Cambridge, MA, 02139, [4] MIT, Division of Health Sciences and Technology, Cambridge, MA, 02139, [5]Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, [6]University of Washington School of Medicine, Department of Radiation Oncology, Seattle, WA, 98195, [7]University of Washington School of Medicine, Department of Pediatrics, Seattle, WA, 98195

The spatial organization of the genome plays a critical role in its regulation, including the control of gene expression. Enhancers, insulators, and repressors can act over large genomic distances. This often involves direct looping interactions between regulatory elements and their target genes, giving rise to complex spatial organization of chromosomes.

To probe the spatial arrangement of genomes we developed Hi-C, a method that combines 3C and high-throughput sequencing to map chromatin interactions in an unbiased, genome-wide fashion. Application of Hi-C to the human genome revealed a novel layer of genome organization in which open and closed chromatin are spatially segregated, forming two genome-wide compartments. The contents of the compartments are dynamic: changes in chromatin state and/or expression correlate with movement from one compartment to the other.

To explore the properties of three-dimensional chromatin interaction networks at higher resolution, we employed 5C technology. We generated a comprehensive long-range interaction map between 188 gene promoters and 1,252 loci distributed evenly along human chromosome 21 and identified approximately 4000 specific long-range looping interactions. Analysis of this set of interactions provides new insights into the architecture of long-range control in the human genome. First, promoters are found to interact with a surprisingly large number of distant elements. Second, many distant elements can also loop to multiple promoters. Third, the interacting elements frequently contain DNAse I hypersensitive sites, predicted enhancer elements, and/or CTCF-bound elements. This suggests that our analysis identified bona fide regulatory elements interacting with promoters. Fourth, only a small fraction of the observed interactions are very frequent and span a relatively small genomic distance, whereas the large majority of interactions are infrequent and long-range (>2 Mb). Finally, promoters preferentially interact with elements that belong to the same compartment (as determined by Hi-C), though elements belonging to the other compartment may be closer in the linear genome.

Combined, our Hi-C and 5C data provide a first view of the architecture and specificity of gene-element associations and of the potential role of higher order folding of chromosomes in facilitating gene regulation.

EPIGENOMIC LANDSCAPES OF PLURIPOTENT AND LINEAGE-COMMITTED HUMAN CELLS

Bing Ren[1], David Hawkins[1], Gary Hon[1], Ryan Lister[2], Mattia Pelizolla[2], Ron Stewart[3], Wei Wang[4], James Thomson[3], Joe Ecker[2]

[1]Ludwig Institute for Cancer Research, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, [2]The Salk Institute for Biological Studies, Genomic Analysis Laboratory, na, La Jolla, CA, 92037, [3]Morgridge Institute for Research, University of Wisconsin, na, Madison, WI, 53706, [4] University of California, San Diego, Department of Biochemistry, 9500 Gilman Drive, La Jolla, CA, 92093

Human embryonic stem cells (hESCs) share an identical genome with lineage-committed cells, yet possess the remarkable properties of self-renewal and pluripotency. The diverse cellular properties in different cells have been attributed to their distinct epigenomes, but how much epigenomes differ remains unclear. Here, we report that epigenomic landscapes in hESC and lineage committed cells are drastically different. By comparing the chromatin modification profiles and DNA methylomes in hESCs and primary fibroblasts, we find that nearly one-third of the genome differs in chromatin structure. Most changes arise from dramatic redistributions of repressive H3K9me3 and H3K27me3 marks, which form blocks that significantly expand in fibroblasts. A large number of potential regulatory sequences also exhibit a high degree of dynamics in chromatin modifications and DNA methylation. Additionally, we observe novel, context-dependent relationships between DNA methylation and chromatin modifications. Our results provide new insights into epigenetic mechanisms underlying properties of pluripotency and cell-fate commitment.

*This work was performed within the NIH Roadmap Epigenomics Project (http://nihroadmap.nih.gov/epigenomics/).

PROXIMAL PROMOTER ELEMENTS DETERMINE DNA METHYLATION DURING SOMATIC DIFFERENTIATION

Florian Lienert[1], Fabio Mohn[1], Ann Dean[2], Dirk Schuebeler[1]

[1]Friedrich Miescher Institute for Biomedical Research, Laboratory of propagation and dynamics of epigenetic states, Maulbeerstr. 66, Basel, 4058, Switzerland, [2]National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Laboratory of Cellular and Developmental Biology, 50 South Dr., Bethesda, MD, 20892

DNA methylation is an epigenetic modification which can lead to gene silencing probably by mediating and stabilizing a local repressive chromatin state. Importantly however, DNA methylation is not a default hallmark of silent genes and therefore dynamic changes of promoter methylation appear tightly regulated. Yet, it is unclear how DNA methylation targets are specified. A recent genome-wide study of our laboratory determined distribution patterns of epigenetic modifications during differentiation of embryonic stem cells (ES cells) into pyramidal neurons (Mohn et al., Mol Cell, 2008). This analysis revealed a comprehensive set of dynamically methylated gene promoters and it further suggested substantial crosstalk between different histone modifications and DNA methylation. Based on the genome-wide data, we here systematically tested the contribution of promoter DNA sequence to target specification in stem cells and *de novo* methylation during neuronal differentiation. By repeatedly placing partial and complete promoter sequences into the same genomic site in stem cells we tested the autonomy of *cis*-acting elements in mediating a defined DNA methylation state. We show that ectopically inserted sequences closely recapitulate DNA methylation patterns of the respective endogenous promoters in ES cells and that they get correctly reprogrammed during differentiation. This result suggests that promoter DNA methylation and its developmental dynamics are not epigenetically but genetically determined by proximal promoter elements.

WHOLE-GENOME BISULFITE SEQUENCING OF HUMAN PLURIPOTENT AND DIFFERENTIATED CELLS REVEALS DYNAMIC CHANGES OCCURING DURING DIFFERENTIATION

Louise C Laurent[1,2], Eleanor Wong[3], Guoliang Li[3], Tien Huynh[4], Aristotelis Tsirigos[4], Isidore Rigoutsos[4], Jeanne F Loring[2], Chia-Lin Wei[3]
[1]UCSD, Department of Reproductive Medicine, 200 West Arbor Drive, San Diego, CA, 92103, [2]The Scripps Research Institute, Center for Regenerative Medicine and Department of Chemical Physiology, 10550 North Torrey Pines Rd., La Jolla, CA, 92037, [3]Genome Institute of Singapore, Genome Technology & Biology Group, 60 Biopolis Street, Singapore, N/A, Singapore, [4] IBM Thomas J Watson Research Center, Bioinformatics & Pattern Discovery Group, 1101 Kitchawan Road, Yorktown Heights, NY, 10598

Pluripotent stem cells (PSCs), including embryonic stem cells (ESCs), have the capability to indefinitely self-renew, as well as to differentiate into a wide array of cell types. DNA methylation is an epigenetic mechanism that plays important roles in mammalian development and disease. In order to understand how DNA methylation contributes to pluripotence, we performed whole-genome bisulfite sequencing on human ESCs (hESCs), fibroblastic cells differentiated from hESCs, and primary neonatal fibroblasts, representing three different stages of differentiation. Analysis of these data revealed that some patterns of DNA methylation were highly cell-type specific, while others were common to all three cells types.

By using a whole-genome bisulfite sequencing approach, we were able to avoid the biases present in many previous studies, which focused on specific areas of the genome, such as CpG-rich regions or promoter regions. We found that global DNA methylation progressively decreased with differentiation status. While virtually all methylated cytosines in the differentiated hESCs and primary fibroblasts were found in CpG dinucleotides, we saw a surprisingly high level of CpA methylation in hESCs. Differentially methylated regions that demonstrated increasing DNA methylation with differentiation included developmentally regulated genes, such as the HOX genes, and pluripotence-associated genes such as POU5F1 and TCF3. We observed progressive demethylation of repeat elements, such as LINEs, with differentiation.

In all three cell types, low levels of promoter methylation and high levels of gene body methylation were correlated with transcription. A possible role for DNA methylation in regulation of transcript splicing was indicated by our findings that exons generally had higher levels of DNA methylation than introns, and that there were sharp spikes of DNA methylation at exon-intron boundaries.

These results point to the value of whole-genome DNA methylation maps for uncovering previously unrecognized roles for DNA methylation in the regulation of splicing and transcription of both coding and noncoding genetic elements, as well as tracking changes in DNA methylation that occur during differentiation.

STATIC AND DYNAMIC GENOME-WIDE VIEWS OF YEAST CHROMATIN

Oliver J Rando[1], Marta Radman-Livaja[1], Chih Long Liu[1], Nir Friedman[2]

[1]University of Massachusetts Medical School, Biochemistry and Molecular Pharmacology, 364 Plantation St., Worcester, MA, 01760, [2]Hebrew University of Jerusalem, School of Computer Science and Engineering, XX, Jerusalem, 91904, Israel

Genomic approaches to chromatin structure have provided a great deal of insight into mechanisms governing the establishment of chromatin states, and constrain hypotheses regarding the functional roles of chromatin in gene expression and other processes. For the past several years, we have gathered whole-genome data on nucleosome positions, histone modifications, histone variant incorporation, and histone dynamics in the budding yeast S. cerevisiae growing under a number of conditions. Here, I will present some of our data on chromatin dynamics during the cell cycle, with a focus on 1) the role of cell cycle timing in erasure of active chromatin states, and 2) the dynamics of ancestral nucleosome behavior during replication. These results have implications for the role of chromatin in epigenetic inheritance.

COMBINED COMPUTATIONAL AND EXPERIMENTAL APPROACHES PINPOINT A GC-RICH ELEMENT SUFFICIENT FOR THE RECRUITMENT OF POLYCOMB COMPLEXES IN ES CELLS.

Eric M Mendenhall[1,2,3], Richard P Koche[1,3], Thanh T Truong[1,3], Vicky W Zhou[1,3], Bradley E Bernstein[1,2,3]

[1]Massachusetts General Hospital and Harvard Medical School, Department of Pathology, Cambridge St., Boston, MA, 02114, [2]Howard Hughes Medical Institute, ., Jones Bridge Road, Chevy Chase, MD, 20815, [3]Broad Institute, Genome Biology, Cambridge Center, Cambridge, MA, 01239

The Polycomb Repressive Complexes (PRC1 and PRC2) and their associated histone mark, H3 lysine 27 (H3K27) tri-methylation, localize to many silent developmental genes in embryonic stem (ES) cells. Though well studied in *Drosophila*, the signals that underlie the recruitment of Polycomb to these sites in vertebrates are just beginning to be elucidated. Here we combine computational analysis with a transgenic cell assay to define the DNA sequence elements sufficient for recruitment of Polycomb. Genome-wide Polycomb and H3K27 tri-methylation mapping experiments enabled the identification of a series of elements that correlate with PRC2 localization. We have devised a BAC cloning approach to assess the sufficiency of such elements. We have determined through stable incorporation of BACs representing a series of deletions that a 1.7 Kb CpG island is the crucial element for PRC2 localization. Additionally when this CpG island is placed exogenously into a gene desert region it recruits PRC2 and shows robust H3K27 tri-methylation. This data supports a functional role for GC-richness in PRC2 localization. Additionally, the stable integration of candidate DNA sequence elements into a gene desert region provides a robust functional readout of Polycomb recruitment. This supports a model where GC-rich sequences, in the absence of active transcription, serve as functional cis elements which drive the recruitment of PRC2. Thus, in addition to roles at housekeeping genes, and as targets of DNA methylation in cancer, CpG islands also appear to be critical regulators of gene silencing during development.

BIMODALITY IN GENE EXPRESSION LEVELS CORRELATES WITH AN EPIGENETIC MODULE

Daniel Hebenstreit[1], Miaoqing Fang[2], Varodom Charoensawan[1], Daniel J Turner[3], Alexander van Oudenaarden[2], Sarah A Teichmann[1]

[1]MRC Laboratory of Molecular Biology, Structural Studies Department, Hills Rd, Cambridge, CB20QH, United Kingdom, [2]MIT, Department of Physics, 77 Massachusetts Ave, Cambridge, MA, 02139, [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB101HH, United Kingdom

Gene expression levels are believed to be continuously distributed from very low to very high levels, with most genes at an intermediate level. We studied the transcriptome-wide distribution of expression levels in T helper cells based on microarray and RNA-seq technology and integrated this with RNA-fish experiments for well-characterized genes, allowing the quantification of mRNA transcripts at single-molecule resolution. The results of these analyses and experiments show that virtually all genes are expressed at $> \sim 1$ molecule per cell and that two major expression levels can be identified which vary by roughly one to two orders of magnitude, giving rise to strongly bimodal expression distributions. Analysis of histone modifications by ChIP-seq indicates that activating modifications such as H3K9/14ac and H3K4me3 are involved in this 'digital' expression switch. Our findings have broad implications for the analysis of microarray, RNA-seq and ChIP-seq data and for the understanding of the regulation of gene expression.

PROGRAMMING CHROMATIN STATES BY LONG NONCODING RNAS

Howard Y Chang

HHMI, Program in Epithelial Biology, Stanford University School of Medicine, 269 Campus Drive, Stanford, CA, 94305

An organizing principle of the diverse cell types in multicellular organisms is their anatomic location. In turn, anatomic location is patterned by the positional identities of cells along developmental axes. Recent progress in functional genomics and chromatin biology illustrates how cells use specific gene expression programs to encode location. Dynamic chromatin states of key genes, notably the Hox loci, serve as the internal representation in cells of their positional identity within the animal. Here I will present new findings on long noncoding RNAs (lincRNAs) that are involved in long-range gene activation. Such lincRNAs can physically interact with the MLL/trithorax complexes and position MLL binding on chromatin, thereby influencing histone H3 lysine 4 methylation and transcription. Analysis of genomic location, noncoding transcription, and chromatin states provides an integrated view of how the genome is configured to encode diverse positional identities.

A PROTEIN-DNA INTERACTION NETWORK REGULATING RESPONSES TO REACTIVE OXYGEN SPECIES

Lesley T MacNeil, A. J. Marian Walhout

University of Massachusetts Medical School, Program in Gene Function and Expression, 364 Plantation St., Worcester, MA, 01605

Reactive oxygen species (ROS), including peroxides, superoxides and hydroxyl radicals, are highly reactive species that can cause cellular damage by oxidizing proteins and DNA. ROS are detoxified in the cell by a number of enzymes, and additional mechanisms exist to reverse the damage caused by ROS. Oxidative stress, defined as the buildup of ROS, has been observed in a number of human diseases including juvenile diabetes and neurodegenerative diseases. Although the role of oxidative stress in disease is not completely understood, the abnormal accumulation of these species early in disease progression suggests that they may contribute to cellular degeneration.

In order to understand the cellular processes that result in oxidative stress, we are studying the transcriptional regulation of genes involved in ROS production and detoxification. Using *C. elegans* as a model system, we have cloned promoters for genes implicated in these processes. In order to identify transcriptional regulators of these genes, we have employed high-throughput yeast one-hybrid assays. In this assay over 800 predicted transcription factors are assayed for interaction with each promoter. Thus far, we have identified over one hundred protein-DNA interactions and identified transcription factors that are highly connected in our network. In addition, we have identified two transcription factors that interact with the promoters of both *skn-1*/NTF2 and *daf-16*/FOXO, two major regulators of the oxidative stress response. Elucidation of the interconnectivity of these genes will allow us a more comprehensive understanding of the response to ROS and the events that lead to oxidative stress and subsequent cell death.

THE IMPACT OF GENETIC POLYMORPHISM ON THE TRANSCRIPTIONAL PATHWAY UNDERYLING GLUCOCORTICOID RESPONSE

Joseph C Maranville, Francesca Luca, Allison Richards, Matthew Stephens, Anna Di Rienzo

University of Chicago, Department of Human Genetics, 920 E. 58th Street, Chicago, IL, 60637

Glucocorticoids (GCs) are endogenous steroid hormones that help mediate physiological responses to the environment. GCs regulate physiological processes relevant to the etiology of many common diseases (e.g. autoimmune disorders), and synthetic GCs are widely used as pharmaceutical interventions (e.g. in the treatment of a variety of inflammatory diseases and some types of cancer). Cellular response to GC treatment is largely mediated through changes in target gene transcription. The activation, and subsequent transcriptional regulatory activity, of the glucocorticoid receptor (GR) is thought to initiate a cascade of transcriptional response where some direct targets of the GR proceed to regulate transcription of additional genes. We used expression microarrays to profile transcriptional response in HapMap EBV-transformed lymphoblastoid cell lines (LCL) from 8 individuals sampled from two populations (Yoruba from Nigeria and Tuscans from Italy) over a 24-hour time course, revealing 1,666 GC-responsive genes (FDR<=0.01). Substantial inter-individual variation in many GC-related phenotypes has been observed clinically and, in some cases, has been shown to be heritable. Genetic polymorphisms in regulatory elements controlling target gene transcriptional response are likely to contribute to this phenotypic variation. Consequently, we performed quantitative trait loci (QTL) mapping, with the magnitude of transcriptional response to GC treatment at each gene as a quantitative trait, in a sample of HapMap LCLs corresponding to 116 individuals (58 Yoruba and 58 Tuscans). Transcriptional response to GCs for the QTL mapping study was measured after 8 hours of treatment, as the greatest number of responsive genes was observed at this time point during the 24 hour time course. Mapping revealed cis response QTLs, likely representing polymorphisms in cis-regulatory elements, for 193 GC target genes ($p<10^{-4}$). Many GC-related phenotypes are unevenly distributed across geographic populations, consistent with varying selective pressures across human populations in response to local environmental conditions. Through population genetic analyses of the response QTLs, we are currently directly testing various evolutionary models regarding the impact of natural selection on the GC response pathways.

GENOME-WIDE CHARACTERIZATION OF THE TRANSCRIPTOME IN ENCODE CELL LINES

Georgi K Marinov[1], Brian A Williams[1], Ali Mortazavi[1], Timothy E Reddy[2], Jason Gertz[2], Richard M Myers[2], Barbara Wold[1]

[1]California Institute of Technology, Biology Division, MC 156-29, Pasadena, CA, 91125, [2]HudsonAlpha Institute for Biotechnology, HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL, 35806

Transcription is the outcome of the integration by the cell of different gene regulatory inputs. Major goals of the ENCODE project are the annotation of all transcripts in the genome and the characterization of the functional DNA elements that govern genome output. This entails mapping and quantifying the transcriptome of all ENCODE experimental cell lines. To contribute to this effort, we sequenced deeply polyA-selected RNA from the seven Tier1 and Tier2 ENCODE cell lines as paired-end 75mers. We also generated directional RNA-Seq for the same samples as single-end 75mers, which allows us to assign strandedness to transcripts and to resolve RNA molecules emanating from different strands at the same locus. We are using these data to identify novel transcript isoforms, to position and orient novel transcripts, and to quantify the transcriptome at the level of individual transcripts. We examine how properties of the different kinds of RNA-Seq data contribute to transcript isoform assembly and quantification. In addition, RNA-Seq measurements are combined with ChIP-Seq measurements of RNA polymerase II and components of the preinitiation-complex in order to examine the correlation between the association of those factors with DNA and transcripts levels. We are further extending these comparisons with genome-wide measurements of transcriptionally active polymerase molecules to understand better the relationship between polymerase loading, immediate RNA synthesis, and subsequent RNA prevalence. Together, these data provide a detailed annotation of the transcriptome and reveal the extent and dynamics of transcription in the human genome across a diverse sample of cell types.

EVIDENCE OF BET-HEDGING IN THE SPATIAL VARIATION OF MUTATION RATE ALONG BACTERIAL GENOMES

Iñigo Martincorena, Aswin Seshasayee, Nicholas M Luscombe

EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom

A well-accepted assumption of current evolutionary theory is that mutations occur randomly with respect to their fitness effect. Whether a particular mutation occurs or not is believed to be unrelated to how useful that mutation is. However, organisms live in varying environments where changes are often predictable to some extent. For example, in the long-term selection consistently tends to favor the fixation of mutations in antigenic sequences while keeping essential genes more invariable. Organisms able to modulate the mutation rate of different genes to reflect the consistently different selection pressures would have a great evolutionary advantage. This idea is well supported by bet-hedging (a risk spreading gambling strategy recently related to biological noise) and other theoretical approaches, but there is virtually no evidence supporting its existence. Notable exceptions are somatic hypermutation of immunoglobulin genes in many vertebrates and hypermutable contingency loci in bacteria, but background mutation rate (which accounts for the vast majority of genomic changes) is believed to be random.

We have performed an intensive and rigorous computational analysis of many bacterial genomes to estimate the mutation rate at gene resolution for several species. Our results strongly suggest that mutation rate varies very significantly among different genes, with essential genes showing a lower mutation rate and antigenic genes having a higher background mutation rate. Globally, we observe a coupling between mutation rate and selection pressure, as predicted by bet-hedging. We also found an association of mutation rate with DNA topology and with gene expression level (unrelated to selection on codon usage). We finally demonstrate that both selection pressure and mutation rate of orthologous genes are highly conserved between related bacterial species, even in the lack of synteny.

PROVIRAL SILENCING IN ES CELLS REQUIRES THE HISTONE METHYLTRANSFERASE ESET

Toshiyuki Matsui[1,2], Yoichi Shinkai[1,2]

[1]Institute for Virus Research, Kyoto University, Experimental Research Center for Infectious Diseases, 53 Shogoin Kawara-cho Sakyo-ku, Kyoto, 606-8507, Japan, [2]Graduate School of Biostudies, Kyoto University, Department of Molecular and Cellular Biology, 53 Shogoin Kawara-cho Sakyo-ku, Kyoto, 606-8507, Japan

DNA methylation plays an important role in transcriptional repression of transposable elements. While the existence of a "stem cell-specific" silencing pathway for murine leukemia virus (MLV) retroviruses has been recognized for more than 30 years, the molecular basis of this pathway was reported only recently. ZFP809, a member of the Kruppel-associated box (KRAB)-zinc finger protein (ZFP) superfamily, was shown to bind to the primer binding site (PBS) of an MLV provirus and to induce silencing of this exogenous retrovirus (XRV) in ES cells by recruiting the co-repressor protein KAP-1. Intriguingly, the histone H3 lysine 9 (H3K9) methyltransferase (HMTase) ESET/Setdb1 is known to interact biochemically with KAP-1. However, the role of histone methylation in silencing of XRVs, and the importance of this pathway in the repression of endogenous retroviruses (ERVs) remain unexplored. Employing *ESET* conditional knockout (CKO) ES cells, we show that ESET plays a critical role in silencing of representative Class I (MLV) and class II (Intracisternal A particle (IAP) and MusD) ERVs and MLV-based XRVs. ESET HMTase activity is required for this silencing pathway. While ESET, Kap-1, H3K9 trimethylation (H3K9me3) and H4K20me3 are enriched at ERVs in ESET+ ES cells, ESET depletion results in the reduction of H3K9me3 and H4K20me3 at these elements. In contrast, while LINE-1 retrotransposons are reactivated in *Dnmt1*, *Dnmt3a* and *Dnmt3b* triple knockout (*Dnmt* TKO) ES cells, Class I and II ERVs are not or only weakly derepressed. Furthermore, enrichment of ESET, H3K9me3 and H4K20me3 at these ERVs is maintained in *Dnmt* TKO ES cells. We propose that ESET plays a crucial role in silencing of ERVs early in embryogenesis, in a pathway that functions independent of DNA methylation.

HIGH-RESOLUTION CHIP-CHIP PROFILING REVEALS UNIFORM TRANSITIONS IN TRANSCRIPTION COMPLEX COMPOSITION

Andreas Mayer, Michael Lidschreiber, Matthias Siebert, Kristin Leike, Johannes Soeding, Patrick Cramer

Gene Center, Department of Biochemistry, and Center for Integrated Protein Science Munich, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, Munich, 81377, Germany

RNA polymerase (Pol) II synthesizes all eukaryotic mRNA in the course of the gene transcription cycle that consists of initiation, elongation and termination/re-initiation. During initiation of transcription, Pol II and general initiation factors assemble on promoter DNA to form an initiation complex, which was recently elucidated structurally (Kostrewa et al., Nature 19 November 2009). Transcription elongation represents another distinct stage within the transcription cycle in which nascent RNA is synthesized by a processive Pol II elongation complex on a chromatin DNA template. Many proteins are involved in this process including Spt4/Spt5 (human DSIF), Spt16/Pob3 (human FACT), Spt6, TFIIF, TFIIS and the Paf1 complex in yeast *Saccharomyces cerevisiae* (*S. cerevisiae*). It has not been investigated systematically whether elongation and chromatin transcription factors are recruited to Pol II on all transcribed genes, or only on a subset of genes, and whether there is a general order of factor recruitment. Here we present unpublished high-resolution genomic elongation factor profiles obtained by chromatin immunoprecipitation in *S. cerevisiae* coupled to tiling microarray analysis (ChIP-chip). To this end we established a highly reproducible ChIP-chip protocol including quality controls after each key step of this modular technique. The data indicate a general, dynamic elongation complex that undergoes uniform transitions in factor composition. Downstream of the transcription start site, "primary" elongation factors enter, and peak levels of phosphorylation of the Pol II C-terminal domain (CTD) at serines Ser_5 and Ser_7 of the consensus heptameric repeat Tyr_1-Ser_2-Pro_3-Thr_4-Ser_5-Pro_6-Ser_7 are reached ("early 5'-transition"). Further downstream, "secondary" elongation factors and the CTD Ser_2 kinase Ctk1 enter ("late 5'-transition"). Immediately upstream of the transcript termination site, levels of a subset of elongation factors decrease sharply, and Ser_2 phosphorylation peaks ("3'-transition"). The three transitions in transcription complex composition occur at all protein-coding genes in yeast, independent of their function, length, and expression level.

RPB2 ORTHOLOG JHD2 MEDIATES GLOBAL HISTONE DEMETHYLATION AND GENE EXPRESSION TO CONTROL SPORE DIFFERENTIATION TIMING DURING BUDDING YEAST GAMETOGENESIS.

Marc D Meneghini[1], Mathieu Ranger[1], Mike Schertzberg[1], Zarna Shah[1], Hiten D Madhani[2]

[1]University of Toronto, Dept. of Molecular Genetics, 1 King's College Circle, Toronto, M5S 1A8, Canada, [2]University of California, San Francisco, Dept. of Biochemistry and Biophysics, 600 16th St, San Francisco, CA, 94158

The ancient and expansive Jumonji family of histone demethylases controls stem cell differentiation and tumorigenesis in mammals. While *bona fide* histone demethylases of this family are found in budding yeast, their physiological functions have yet to be identified. We have discovered that *JHD2*, an ortholog of mammalian JARID1A/Rbp2, functions to accomplish developmentally programmed global demethylation of histone H3 lysine-4 (H3-K4) during yeast sporulation, with the bulk of demethylation occurring post-meiotically. Using a sensitive cell-biological assay, we determined that Jhd2-demethylation of H3-K4 controls the timing of spore differentiation: Amazingly, both *jhd2*Δ and an enzymatically-dead allele of *jhd2* cause precocious terminal differentiation. To confirm and extend these findings, we performed extensive genome-wide microarray experiments and found that *jhd2* mutants exhibit a gene expression defect that expressly corroborates our differentiation timing model: 1. Transcripts that are specifically induced penultimate to terminal spore differentiation are greatly reduced, and 2. Concomitantly, transcripts that are normally induced at terminal differentiation are expressed precociously. Thus, *JHD2* may function both positively and negatively to regulate gene expression during spore differentiation. We think this dual role likely reflects differing mechanistic contributions of H3-1/2meK4 compared with H3-3meK4. To test this idea, we performed ChIP experiments to monitor H3meK4 dynamics during sporulation in WT and *jhd2*Δ. Interestingly, our results show that Jhd2 demethylation of H3-3meK4, an epigenetic mark canonically regarded as activating, is required for gene targets to evade widespread gene repression occurring during spore differentiation (*i.e.* H3-3meK4 mediates gene repression in this developmental context). Finally, the classes of genes controlled by both Jhd2 during yeast spore differentiation and JARID1A/Rbp2 during mammalian stem cell differentiation suggest that inter-organelle signaling between the mitochondria and the nucleus is prominent during eukaryotic differentiation. As Jumonji demethylase activity is dependent on alpha-ketoglutarate, a metabolite produced by the TCA cycle, this phenomenon seems intriguingly fundamental.

DYNAMICS AND EVOLUTION OF GENOME - NUCLEAR LAMINA INTERACTIONS

Wouter Meuleman[1,2,3,4], Daan Peric Hupkes[1,4], Marcel Reinders[3], Lodewyk Wessels[2,3], Bas van Steensel[1]

[1]Netherlands Cancer Institute, Gene Regulation, Plesmanlaan 121, Amsterdam, 1066CX, Netherlands, [2]Netherlands Cancer Institute, Molecular Biology, Plesmanlaan 121, Amsterdam, 1066CX, Netherlands, [3]Delft University of Technology, Delft Bioinformatics Lab, Mekelweg 4, Delft, 2628CD, Netherlands, [4] These authors contributed equally

The three-dimensional organization of chromosomes within the nucleus and its dynamics during differentiation are largely unknown. To visualize this process in molecular detail, we generated high-resolution molecular maps of genome – nuclear lamina interactions during subsequent differentiation of mouse embryonic stem cells via lineage-committed neural precursor cells into terminally differentiated astrocytes. This reveals that a basal chromosome architecture present in embryonic stem cells is cumulatively altered at hundreds of sites during lineage commitment and subsequent terminal differentiation. This remodeling involves both individual transcription units and multi-gene regions, and affects many genes that determine cellular identity. Often, genes that move away from the lamina are concomitantly activated; many others however remain inactive yet become unlocked for activation in a next differentiation step. These results suggest that lamina-genome interactions are widely involved in the control of gene expression programs during lineage commitment and terminal differentiation.

We are currently creating an atlas of genome – nuclear lamina interactions in a larger number of cell types, covering all three germ layers. Using this resource, we are defining regions that show static and dynamic behavior in terms of lamina-genome interactions. We find that a substantial portion of the genome behaves identical in all cell types in terms of lamina-genome interactions. Interestingly, comparison of mouse and human data reveal a high degree of evolutionary conservation of lamina-genome interactions. Initial results show that given the mouse lamina interaction map and the primary genome sequence, we can accurately predict lamina status in a similar human cell type. This is an indication that the organization of the genome is, in part, hard-coded in the primary sequence.

NUCLEOSOME-MEDIATED COOPERATIVITY BETWEEN TRANSCRIPTION FACTORS

Leonid Mirny

MIT, Health Sciences and Technology, and Physics Dept, 77 Mass ave, Cambridge, MA, 02139

Cooperative binding of transcription factors (TFs) to promoter and other regulatory regions is essential for precise gene expression. The classical model of cooperativity requires direct interactions between TFs, thus constraining the arrangement of TFs sites in regulatory regions. Recent genomic and functional studies, however, demonstrate a great deal of flexibility in such arrangements with variable distances, numbers of sites, and identities of the involved TFs. Such flexibility is inconsistent with cooperativity by direct interactions between TFs. Here we demonstrate that strong cooperativity among non-interacting TFs can be achieved by their competition with nucleosomes. We find that the mechanism of nucleosome-mediated cooperativity is analogous to cooperativity in another multimolecular complex of hemoglobin. This surprising analogy provides deep insights, with parallels between heterotropic regulation of hemoglobin (e.g. Bohr effect) and roles of nucleosome-positioning sequences and chromatin modifications in gene expression. Nucleosome-mediated cooperativity is consistent with several experimental studies, allows substantial evolutionary flexibility in and modularity of regulatory regions, and provides a rationale for a broad range of genomic and evolutionary observations. Striking parallel between cooperativity in hemoglobin and in transcription regulation points to a general mechanism that may be used in various biological systems.

A NOVAL BAYESIAN SEGMENTATION MODEL FOR CHIP-SEQ DATA ANALYSES

Yifan Mo[1,2], HaiPeng Xing[2], Will Liao[1,2], Michael Q Zhang[1,3,4]

[1]Cold Spring Harbor Laboratory, Michael Zhang's Lab, Bungtown road, Cold Spring Harbor, NY, 11724, [2]Stony Brook University, Applied Mathematics and Statistics, Stony Brook, NY, 11794, [3]Tsinghua University, Bei Jing, 100084, China, [4] University of Texas at Dallas, Department of Molecular Cell Biology, Richardson, Dallas, TX, 75080

The Chromatin Immunoprecipitation (ChIP) followed by high throughput massively parallel next-generation sequencing technology has being widely used to analyze the interaction between protein and chromatin DNA interaction, as a result, large amounts of ChIP sequencing (ChIP-seq) data consisting of tens of millions short reads (25bp-50bp) become the focus of intense interest in studies of various transcription factor binding sites (TFBSs) and/or histone modifications.

To meet such an analysis challenge, we present a novel stochastic Bayesian segmentation model and the associated estimation procedure to analyze such ChIP-seq data. The main advantage of our model is that it could use explicit formula for posterior means, which is the criterion to estimate the "peak" signals without performing segmentations first and hence can avoid the uncertainty of choosing a "window size" as done in many other methods. We also develop a fast approximation algorithm that can be implemented parallelly. Since its computation time is linear in sequence length, it is practically efficient for all ChIP-seq data analyses without sacrificing accuracy comparing to the common Bayes estimator.

We will show the results of the comparison of our method against MACS, QuEST, SISSRs on both the real data (CTCF, Foxa2) and the simulation dataset. The results show the good performance on both data saturation and false discover rate (FDR).

Acknowledgement:
This work is supported by STARR foundation award (MQZ), NIH grant ES017166 (MQZ) and NSF grant DMS0906593 (HPX).

TATA IS A MODULAR COMPONENT OF SYNTHETIC PROMOTERS

Ilaria Mogno, Barak A Cohen

Washington University in St. Louis, Genetics, 4444 Forest Park Ave., St. Louis, MO, 63108

The expression of most genes is regulated by multiple transcription factors. The interactions between regulators produce complex patterns of gene expression that are not always obvious from the arrangement of *cis*-regulatory elements in a promoter. One critical element of all promoters is the docking site for the RNA polymerase holoenzyme, referred to as the TATA box. We studied the effect of the strength of the TATA box on combinatorial *cis*-regulation. We constructed three synthetic promoter libraries with three different TATA boxes in *S. cerevisiae*. Using this synthetic promoter system coupled to a thermodynamic model of combinatorial regulation, we analyzed the effects of different strength TATA boxes on various aspects of combinatorial *cis*-regulation, in the effort to decipher the relationship between the strength of the TATA box and the combinatorial interactions that occur on promoters.

The thermodynamic model was able to explain 75% of the variance in gene expression in synthetic promoter libraries with different strength TATA boxes. Thus, many of the salient aspects of *cis*-regulation in our system were captured by the model. Further analysis revealed several properties of the TATA box with respect to *cis*-regulation. Our results suggested that the effect of changing the TATA box on gene expression is the same for all promoters regardless of their *cis*-regulatory content. Our analysis also showed that the strength of the RNA polymerase-TATA interaction does not alter the combinatorial interactions between transcription factors, or between transcription factors and RNA polymerase. Finally, we found that although stronger TATA boxes increase expression in a predictable fashion, stronger TATA boxes do not add any additional noise to promoters, regardless of the *cis*-regulatory content of promoters.

Our results support a modular model of promoter function, where *cis*-regulatory elements can be mixed and matched (programmed) with outcomes on expression that are predictable based on the rules of simple protein-protein and protein-DNA interactions.

NUCLEOTIDE RESOLUTION PROTEIN-DNA BINDING CHARACTERISTICS BY DOUBLE FRAGMENTATION CHIP-SEQ

Michal Mokry[1], Pantelis Hatzis[1], Jurian Schuijers[1], Ewart de Bruijn[1], Jan Koster[2], Rogier Versteeg[2], Marc van de Wetering[1], Victor Guryev[1], Hans Clevers[1], Edwin Cuppen[1,3]

[1]Hubrecht Institute and University Medical Center Utrecht, KNAW, Cancer Genomics Center, Uppsalalaan 8, Utrecht, 3584CT, Netherlands, [2]Academic Medical Center, University of Amsterdam, Department of Human Genetics, P.O. Box 22700, Amsterdam, 1100DE, Netherlands, [3]University Medical Center Utrecht, Department of Medical Genetics, Universiteitsweg 100, Utrecht, 3584CG, Netherlands

ChIP-seq has become the method of choice for studying functional DNA-protein interactions on a genome-wide scale. In current ChIP-seq protocols DNA fragments smaller than ~300 bp are size-selected and converted into sequencing libraries followed by next-generation sequencing of the fragment ends. However, this approach ignores larger DNA fragments covered by big protein complexes and requires computational deconvolution of + and – strand reads to identify individual binding sites. In addition it also discards large amounts of specifically immunoprecipitated material in the larger size range. The latter increases demands on the amount of starting material and makes genome wide DNA-protein interaction profiling complicated in case of lower affinity antibodies, indirect DNA-protein interaction or difficult sources of starting material (e.g. primary tissue material or sorted cells).
To address these limitations we applied a strategy with a first gentle shearing step with the objective to preserve biological relevant complexes. After immunoprecipitation a second shearing step was introduced to sub-shear the DNA into small fragments (75-125 nt) followed by AB/SOLiD sequencing.
Using Tcf4 ChIP-Seq we show that this approach requires only relatively small amounts of input material and results in quantitative sequence coverage over the complete binding region. The presence and size of the Tcf4-binding regions were found to correlate with Wnt-dependent gene expression changes and revealed putative direct regulation of specific miRNAs. We were able to reliably detect substructures in the peaks, providing high-resolution information regarding true Tcf4-binding sites and potential novel co-factors. In addition we present evidence that long-distance chromatin interactions by co-immunoprecipitation of promoter regions of Tcf4-regulated genes can be detected. Finally, taking advantage from presented technique, we were able to produce genome wide ChIP-seq data for various proteins including those indirectly associated with DNA like Argonautes protein family members.

AN INCOHERENT FEED FORWARD LOOP DEFINES THE PATTERN OF *AGO4* EXPRESSION DURING EARLY *ARABIDOPSIS THALIANA* TRICHOME DEVELOPMENT

Kengo Morohashi[1], Alper Yilmaz[1], Dan Siegal-Gaskins [1,2,3], Erich Grotewold[1,2,3]

[1]The Ohio State University, Plant Cellular and Molecular Biology, 1060 Carmack Dr., Columbus, OH, 43210, [2]The Ohio State University, Plant Biotechnology Center, 1060 Carmack Dr., Columbus, OH, 43210, [3]The Ohio State University, Mathematical Biosciences Institute, 1735 Neil Ave., Columbus, OH, 43210

The development of trichomes from pluripotent epidermal cells in *Arabidopsis thaliana* provides a powerful model for the study of gene regulatory networks involved in plant cell differentiation. We have previously shown that the R2R3-MYB protein GL1 and bHLH transcription factor GL3 together directly regulate approximately 20 genes involved in trichome initiation, including the WRKY transcription factor gene *TTG2* [1, 2]. To better establish the trichome initiation network, we combined literature-based ChIP analyses with ChIP-chip experiments using TTG2-GFP transgenic plants and GFP antibodies. Among the 372 genes identified as TTG2 direct targets is *AGO4*, a gene involved in siRNA mediated DNA methylation that was also found to be a target for GL3. Further expression analyses conducted on mutant strains suggest that *GL3*, *TTG2* and *AGO4* constitute an incoherent feed forward loop in which *AGO4* is up-regulated by TTG2 and down-regulated by GL3, *TTG2* is itself activated by GL3. We propose that it is this network architecture that is responsible for the narrow window of *AGO4* expression that occurs at early stages of trichome development. We are currently using a combination of experimental assays and mathematical modeling to probe the functional relationships between AGO4, GL3 and TTG2, and further elucidate the mechanisms underlying epidermal cell fate determination.

[1] Morohashi, et al., (2009). PLoS Genet 5(2): e1000396.
[2] Morohashi, et al., (2007). Plant Phys 145: 736.

EXTENSIVE ROLE OF THE GENERAL REGULATORY FACTORS, ABF1 AND RAP1, IN DETERMINING GENOME-WIDE CHROMATIN STRUCTURE IN BUDDING YEAST

Mythily Ganapathi[1], Michael J Palumbo[2], Suraiya A Ansari[1], Qiye He[1,3], Kyle Tsui[4], Corey Nislow[5], Randall H Morse[1,3]

[1]Wadsworth Center, Molecular Genetics, 120 New Scotland Ave, Albany, NY, 12201-0509, [2]Wadsworth Center, Bioinformatics, 120 New Scotland Ave, Albany, NY, 12201-0509, [3]University at Albany School of Public Health, Biomedical Sciences, 120 New Scotland Ave, Albany, NY, 12201-0509, [4] University of Toronto, Pharmaceutical Sciences, 112 College St, Toronto, ON M5S 3E1, Canada, [5]University of Toronto, Molecular Genetics, 112 College St, Toronto, ON M5S 3E1, Canada

The packaging of eukaryotic DNA into chromatin has profound consequences for gene regulation, as well as for other DNA transactions such as recombination, replication, and repair. Understanding how this packaging is determined is consequently a pressing problem in molecular genetics. DNA sequence, chromatin remodelers, and transcription factors affect chromatin structure, but the scope of these influences on genome-wide nucleosome occupancy patterns remains uncertain. Here, we report results from using high resolution tiling arrays to examine the contributions of two general regulatory factors, Abf1 and Rap1, to nucleosome occupancy in *Saccharomyces cerevisiae*. Abf1 and Rap1 are abundant, essential DNA-binding proteins that function in transcriptional activation at hundreds of promoters, and both are known to create local regions of open chromatin surrounding their binding sites. Correspondingly, we find that these factors have a widespread influence on chromatin structure. In fact, the role that Abf1 and Rap1 play in governing nucleosome occupancy is considerably greater than anticipated based on the number of predicted binding sites for these factors, with thousands of loci showing localized regions of altered nucleosome occupancy within one hr of loss of Abf1 or Rap1 binding. We present evidence that Abf1 and Rap1 can affect chromatin structure via binding sites having a wide range of affinities. These results indicate that DNA-binding transcription factors affect chromatin structure, and probably dynamics, throughout the genome to a much greater extent than previously appreciated.

INTEGRATIVE ANALYSIS OF CHIP-SEQ AND RNA-SEQ DATA USING SELF-ORGANIZING MAPS

Ali Mortazavi[1], Shirley Pepke[2], Georgi Marinov[1], Barbara Wold[1]

[1]California Institute of Technology, Division of Biology, 1201 East California Blvd, Pasadena, CA, 91125, [2]California Institute of Biology, Center for Advanced Computing Research, 1201 East California Blvd, Pasadena, CA, 91125

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. When we use ChIP-seq for multiple factors, cofactors and marks, plus RNA-seq to measure define inputs and outputs of these physical networks for multiple time-points and cell states, the integrative analysis of the resulting high-dimensional data matrix becomes limiting for 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 ChIP-seq and RNA-seq datasets to cluster the genome into thousands of coherent units and then identifying units representing co-regulated regions across multiple cell types and states. 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.

A CELLULAR RESOLUTION ATLAS OF EMBRYONIC GENE EXPRESSION IDENTIFIES DYNAMIC TEMPORAL CONTROL OF TEMPORAL IDENTITY AND FATE

John I Murray[1], Elicia Preston[2], Dionne Vafeados[2], Thomas Boyle[2], Mihail Sarov[2], Robert Waterston[2]

[1]University of Pennsylvania, Genetics, 433 South University Avenue, Philadelphia, PA, 19104, [2]University of Washington, Genome Sciences, 1705 NE Pacific St, Seattle, WA, 98195

We recently developed methods to quantify gene expression dynamics with cellular resolution in *C. elegans* embryos. We have described embryonic gene expression with cellular resolution for about 150 genes, mostly encoding transcription factors. The resulting data set is a useful resource for *C. elegans* biologists interested in the development of particular cell types because it enumerates which genes are specific to each cell's lineage.

Unexpectedly, many of the gene expression patterns are well correlated with lineage identity but not with fate. The variety of these patterns suggest that an initial lineage-based pattern of combinatorial transcription factor expression may be important for directing cells to adopt the correct terminal fate. We have identified reporters that differentiate the two daughter lineages of 65% of the first 200 cell divisions in the *C. elegans* embryo. We have identified many genes whose expression is correlated with anterior-posterior position or is expressed asymmetrically in the left and right sides of the animal.

To investigate temporal dynamics of regulation, we compared expression of promoter-fusion reporters to that of protein-fusion reporters. In general the two types of reporters produced expression in similar sets of cells, however the protein fusions exhibited striking dynamic patterns of expression and localization. These patterns appeared to be tightly and reproducibly coordinated with events such as cell division and embryo stage, suggesting that novel mechanisms exist to coordinate cell identity and temporal identity to achieve correct spatiotemporal regulation of gene expression.

RHYTHMIC PROTEIN-DNA INTERACTOMES AND CIRCADIAN TRANSCRIPTION REGULATORY NETWORKS

Guillaume Rey[1,2], Felix Naef[1,2]

[1]Ecole polytechnique Fédérale de Lausanne (EPFL), The Institute of Bioengineering (IBI), School of Life Sciences, Station 15, Lausanne, CH-1015, Switzerland, [2]Swiss Institute of Bioinformatics (SIB), Computational systems biology group, Genopode, Lausanne, CH-1015, Switzerland

Cells measure time using circadian clocks, which has profound implications for cell biology and physiology. In mammals, disrupting these clocks leads to reduced fitness, increased susceptibility to cancer, and metabolic syndrome. Here, we combine targeted and systems-wide approaches to dissect circadian transcriptional regulatory networks. We ask the key question of how circadian transcription regulatory networks achieve phase specific gene expression, and how these drive tissue specific output programs. The circadian clocks uses interlocked negative feedback mechanisms that rely on key transcriptional regulators. Among those, the principal bHLH activator is the heterodimer CLOCK/BMAL1 in mammals. Using comparative genomics we found a highly conserved CLOCK/BMAL1 enhancer, which revealed two highly conserved tandem E-box like (E1-E2) motifs that predict known CLOCK/BMAL1 targets in mouse. *In vitro* binding assays reveal that the tandem element is bound by two BMAL1/CLOCK dimers with a strong cooperativity. To complement this targeted approach, we predict a systems-wide cartography of active enhancers by using regression methods that combine time series profiling with cis-regulatory bioinformatics. Analysis of data from mouse liver shows that we can reliably identify phase-specific circadian cis-elements, but also points to novel regulatory elements showing circadian activity patterns in liver. These predictions are validated *in vivo* using time resolved ChIP-seq experiments in mouse liver, revealing for the first time the dynamic nature of circadian protein-DNA interactomes on a genome-wide scale. To investigate the functionality of these interactions, we are investigating how rhythmic CLOCK/BMAL1 protein-DNA interactions drives phase-specific circadian expression by integrating the binding data with high resolution time-series transcription profiling and mRNA stability measurements. In summary, these combined approaches allow us to understand how a molecular clock circuit regulates rhythmic and phase-specific gene expression in mouse liver.

USING C-MYC-GFP KNOCK-IN MICE TO STUDY THE REGULATION AND FUNCTION OF C- MYC IN NORMAL CELLS, IN VIVO

Zuqin Nie[1], Gang Wei[2], Gangqing Hu[2], Keji Zhao[2], Dave Levens[1]

[1]NIH, NCI, 10 Center Dr., Bethesda, MD, 20892, [2]NIH, NHLBI, 10 Center Dr., Bethesd, MD, 20892

The function and the regulation of the expression of the c-Myc proto-oncogene are very complicated. Many transcription factors, cis-regulatory elements and signaling pathways are involved in both processes.
In order to understand the regulation of c-Myc expression in vivo, we generated c-Myc-GFP knock-in mice, in which the stop codon of c-Myc was replaced by pd4-GFP using homologous recombination. Homozygous knock-in mice are viable and encode the c-Myc- GFP fusion protein, as confirmed by Western-blot, fluorescence microscopic and flow-cytometric analysis. The downstream targets of c-Myc were compared between normal B-, T-cells and mitogen activated B-, T-cells from splenocytes using Chip-Seq and expression arrays. The MAX independent targets of c-Myc are being studied by knock-down of the MAX gene in the MEF cells. Studies of the networks that regulate of c-Myc by serum, growth factors, and transcription factors are ongoing.

OCT4/POU5F1-DEPENDENT TRANSCRIPTIONAL NETWORKS IN TEMPORAL CONTROL OF EARLY DEVELOPMENT

Daria Onichtchouk[1], Florian Geier[2,3], Rebecca Moessner[1], Bjoern Wendik[1], Jens Timmer[3], Wolfgang Driever[1]

[1]University of Freiburg, Developmental Biology, Hauptstrasse 1, Freiburg, D-79104, Germany, [2]ETH Zürich, Department of Biosystems Science and Engineering (D-BSSE, Mattenstrasse 26, Basel, CH-4058, Switzerland, [3]University of Freiburg, Physics Department, Albertstrasse 19, Freiburg, D-79104, Germany

Control of stemness and differentiation programs from pluripotent cell populations is of central interest in current biomedical research. To maintain pluripotency, at least two programs have to be activated, the core self-maintenance circuits, and programs repressing cell differentiation. The transcription factor Oct4 (also called Pou5f1) is a key component of the core pluripotency circuit in mammalian embryonic stem (ES) cells. While the ES cell maintenance circuits are well studied, a mechanistic understanding of differentiation control is rudimentary so far, because loss of embryonic cell lineages in mutant mice precludes functional analysis of Pou5f1 in the embryo.

In the presented work, we utilize zebrafish as an experimental system to identify and analyze the functions of Pou5f1 in regulating the progression of cell differentiation. We used Pou5f1 mutant zebrafish embryos (MZspg) to analyze temporospatial aspects of the transcriptional network controlled by Pou5f1 by microarray analysis. We identified direct Pou5f1 targets during development, which show an unexpected level of evolutionary conservation with mammals. We demonstrate that components of various tissue-specific differentiation programs are prematurely activated in MZspg mutant embryos.

We identify two Pou5f1-dependent mechanisms to control developmental timing. First, Pou5f1 directly activates transcriptional repressors, mediating repression of differentiation genes in distinct embryonic compartments. Second, the extent of cooperation of Pou5f1 with transcriptional coactivator SoxB1 controls temporal dynamics of direct target activation. We establish a quantitative mathematical model of the early Pou5f1 and SoxB1 gene network to illustrate regulatory characteristics important for developmental timing. The significant overlap between zebrafish and mammalian Pou5f1 targets together with the ability of mouse Oct4/Pou5f1 to functionally replace the zebrafish Pou5f1 suggests that the mammalian network may have evolved from a basal situation similar to what is observed in teleosts.

GENOME WIDE NUCLEOSOME OCCUPANCY AND AFFINITY OF TRANSCRIPTION FACTORS BINDING: YEAST STUDY

Yuriy L ORLOV, Wee Siong GOH, Jingmei LI, Jin-Quan RUN, Mikael HUSS, Neil D CLARKE

Genome Institute of Singapore, Computational Systems Biology, 60 Biopolis St., Singapore, 138672, Singapore

Study of high-resolution nucleosome position data that has recently become available in yeast together with transcription factor binding data provides a model for global regulation of gene expression. Chromatin immunoprecipitation microarray experiments (ChIP-chip) years ago [Lee et al., 2004] revealed an under-representation of nucleosomes in promoter regions, relative to transcribed regions. In contrast, TFs are underrepresented in transcribed regions and enriched in promoter regions. We revisit the data of Kaplan et al. [2009], analyzing it in the context of existing ChIP-chip data [Harbison et al., 2004] as well as new, more quantitative ChIP-qPCR experiments [Goh et al., in press]. The maps have been obtained from both crosslinked and uncrosslinked chromatin in vivo, and from chromatin assembled from genomic DNA and nucleosomes in vitro. Additionally we used new in vivo nucleosome sequencing data from yeast in different environmental conditions. We confirm a role for intrinsic nucleosome binding preferences in the binding of transcription factors. To assess quantitatively the correlation between low nucleosome occupancy and TF binding, we asked how well nucleosome tag counts correctly distinguish TF-bound sites from random sites selected from yeast promoters. We use the area under the ROC curve as a measure of this association. The analysis was done to the 41 yeast TFs for which there are at least 50 binding motifs bound according to the ChIP-chip data of Harbison et al. [2005] ($p<1e-3$) and the subsequent motif analysis of MacIsaac et al. [2006]. Bootstrap analysis of ROC curve areas shows a significant association between TF occupancy and nucleosome depletion for most of the 41 TFs. The ROC values are higher with in vivo chromatin than with in vitro reconstituted chromatin for most TFs (33 of 41). To assess more directly how much of an effect on TF binding is encoded by the intrinsic DNA binding specificity of nucleosomes, we determined the apparent binding occupancies of 107 perfect consensus binding sites in the genome using ChIP-qPCR. ChIP enrichment values are inversely correlated with nucleosome occupancy. Kaplan et al. [2009] used two different methods in their nucleosome mapping experiments, one involving formaldehyde crosslinking and the other a more traditional non-crosslinking protocol. We have shown that TF binding sites tend to be associated with excess nucleosome counts in crosslinked chromatin vs. uncrosslinked (nucleosome occupancy difference). In conclusion, we argue that regional nucleosomal density in yeast promoters is generally more relevant to TF binding than precise nucleosome position.

COMBINATORIAL REGULATION OF TRANSCRIPTION FACTORS AND HISTONE MODIFICATIONS REVEALED FROM INTEGRATED MODELING OF GENOME-WIDE CHIP-SEQ AND GENE EXPRESSION DATA

Zhengqing Ouyang[1], Wing H Wong[2]

[1]Stanford University, Department of Biology, Gilbert Hall, Stanford, CA, 94305, [2]Stanford University, Departments of Statistics, Health Research and Policy, and Biology, 390 Serra Mall, Stanford, CA, 94305

The genome-wide binding locations of transcription factors (TFs) and histone modifications are now routinely generated by chromatin-immunoprecipitation combined with next-generation sequencing technologies (ChIP-Seq). A prominent example is the mouse embryonic stem cells (ESCs), in which more than a dozen ChIP-Seq profiles have been reported. Although these rich data sets have provided strong implications of individual regulators on the regulation of gene expression, the interplay between TFs and histone modifications has not been fully illustrated. Here we report the joint analysis of ChIP-Seq profiles for histone marks and TFs in mouse ESCs together with gene expression data provided by RNA-Seq. Using a quantitative modeling framework, we determined the association strength of 5 histone modifications (H3K4me3, H3K27me3, H3K36me3, H3K9me3, and H3K20me3) and 12 sequence-specific factors (Oct4, Sox2, Nanog, Stat3, Smad1, Esrrb, Tcfcp2l1, Klf4, Zfx, E2f1, Myc, and Mycn) on 19,319 genes. We performed the PC-regression analysis, which first extracts principal combinations of TFs and histone modifications and then selects those that most affect gene expression. In addition to confirming the high predictive power of ChIP-Seq on gene expression, our analysis revealed intriguing functional redundancy of both types of regulators. TFs and histone modifications, while each explaining ~65% of gene expression variation in mouse ESCs, only differ by 3% of additional prediction power to each other. The results suggest that H3K4me3 and H3K36me3 cooperate with all the 12 TFs to activate gene expression, while H3K27me3 interacts with only a subset of the TFs (including Oct4, Nanog and Sox2) to play inhibition roles in mouse ESCs. Furthermore, we found that an interesting linear combination of the histone modification marks can be accurately predicted from the 12 TFs (R^2=0.82).

GENOME-WIDE ORGANIZATION OF OPEN CHROMATIN IN HUMAN PANCREATIC ISLETS

Lorenzo Pasquali[1,2], Takao Nammo[1,2], Kyle J Gaulton[3], Jeremy M Simon[3,4], Paul G Giresi[4], Piotr Mieczkowski[3], Antonio Secchi[5], Thierry Berney[6], Eduard Montanya[2], Karen L Mohlke[3], Jason Lieb[4], Jorge Ferrer[1,7]

[1]Genomic Programming of Beta Cells, Institut d'Investigacions Biomediques August Pi i Sunyer, Casanova, 143, Barcelona, 08036, Spain, [2]CIBER de Diabetes y Enfermedades Metabólicas Asociadas, CIBERDEM, NA, Barcelona, 08036, Spain, [3]University of North Carolina at Chapel Hill, Department of Genetics, 5096 Genetic Medicine Building, Chapel Hill, NC, 27599, [4] University of North Carolina at Chapel Hill, Department of Biology, 408 Fordham Hall, Chapel Hill, NC, 27599, [5]Clinical Transplant Unit, Department of Internal Medicine, Via Olgettina 60, Milano, 20132, Italy, [6]Cell Isolation and Transplantation Center, Department of Surgery, rue Michel-Servet 1, Geneva, 1211, Switzerland, [7]Hospital Clínic de Barcelona, Department of Endocrinology, Villarroel,170, Barcelona, 08036, Spain

Pancreatic islets are composed of endocrine cells that secrete insulin, glucagon, and other polypeptide hormones. Islet cells are essential for glucose homeostasis, and therefore elucidating the transcriptional control of islet-cells has implications for understanding diabetes pathogenesis and treatment. We have now examined how islet-cell selective active regulatory elements are organized in the human genome. To identify active regulatory elements in islet cells, we performed FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) coupled with high-throughput sequencing (FAIRE-Seq) in three human pancreatic islet samples. We then compared islet FAIRE-seq to five ENCODE non-islet cell lines, and thus identified islet-selective open chromatin sites. We found that islet-selective open chromatin sites are highly clustered, forming unexpectedly broad domains that we term Clusters of Open Regulatory Elements (CORE). We identified □3300 islet-selective COREs, which had a median size of 25 kb, with the largest containing 148 FAIRE sites spanning 602 kb. Islet-selective COREs typically encompassed single genes exhibiting islet-specific expression, and provided rich maps of putative cis-regulatory elements surrounding many known and putative novel islet cell regulators. The unbiased maps generated by FAIRE-seq reveal new insight into the higher order organization of tissue-specific cis-regulatory elements, and provide a foundation for mechanistic understanding of transcriptional regulation of genes important for pancreatic islet function.

GENOME-WIDE MAPPING OF PPARΓ/RXR BINDING IN MACROPHAGES CONNECTS THE TRANSCRIPTION FACTOR PU.1 TO THE SELECTION OF TISSUE-SPECIFIC TARGET GENES.

Sebastian Pott[1,2], Nima K Kamrani[1,2], Jolanta Mazurek[2,3], Martin L Hibbert[4], Sven Pettersson[2], Edison T Liu[1]

[1]Genome Institute of Singapore, Cancer Biology and Pharmacology 2, 60 Biopolis Street, Singapore, 138672, Singapore, [2]Karolinska Institute, Department of Microbiology, Tumor and Cell Biology, Nobels vaeg 16, Stockholm, 17177, Sweden, [3]Swedish Institute for Infectious Disease Control, SMI, Nobels vaeg 18, Solna, 17182, Sweden, [4] Genome Institute of Singapore, Infectious Diseases, 60 Biopolis Street, Singapore, 138672, Singapore

While the nuclear receptor PPARγ regulates adipogenesis and reduces insulin resistance in adipose tissue and in myeloid cells, the relationship between PPARγ activity, its target genes and the systemic consequences remains poorly understood. To study mechanisms that establish tissue-specific activity of PPARγ, we set up a functional genomics approach to identify direct PPARγ targets in human PMA induced THP-1 cells. Here we report a complete map of binding sites for PPARγ and its heterodimerization partner, the retinoid x receptor (RXR) by use of a ChIP-Seq approach. Our genomic binding data were complemented by expression profiles from the ligand induced transcriptional response to a well-defined PPARγ ligand. As expected, PPARγ/RXR binding sites are enriched around regulated genes. Interestingly, we also noticed a preferential enrichment of sites at loci implicated in the control of triglyceride and low-density lipoprotein levels. This further connects PPARγ to the control of cholesterol metabolism in macrophages. Surprisingly, we found that in general PPARγ/RXR binding sites were associated with genes preferentially expressed in monocytes. Subsequent in silico analysis identified an ETS motif as significantly enriched in the vicinity of PPARγ/RXR binding regions. A complementary ChIP-Seq study confirmed the physical presence of the ETS transcription factor PU.1 at a large proportion of PPARγ/RXR binding sites. In conclusion, our results link the PPARγ/RXR target genes to loci important in the myeloid lineage and implicate PU.1 to be a possible partner that contributes to the activity of the nuclear receptor PPARγ in macrophages.

GLOBAL ANALYSIS OF PHOSPHORYLATION NETWORKS IN HUMANS

Rob Newman[1], Hee Sool Rho[1], Zhi Xie[2], Jianfei Hu[2], Crystal Woodard[2], Shaohui Hu[2], Lizhi Jiang[1], Sheng-ce Tao[1], Eric Cox[1], Jin Zhang[1], Heng Zhu[1], Jiang Qian[2]

[1]Johns Hopkins, Pharmacology and Molecular Sciences, 733 N Broadway, Baltimore, MD, 21287, [2]Johns Hopkins, Ophthalmology, 400 N Broadway, Baltimore, MD, 21287

Protein phosphorylation is one of the most important posttranslational modifications that regulate different cellular processes. However, a comprehensive understanding of the phosphorylation networks and pathways is still lacking in higher eukaryotes. Here we used a combined bioinformatics and protein microarray-based strategy to determine the kinase-substrate interactions in humans. We performed kinase reactions with 450 human kinases (~130 kinase families) on protein microarrays that each carries 4,191 human proteins. Among them, we identified over 29,296 in vitro kinase-substrate interactions. To improve physiological relevance of these interactions, we developed a Bayesian model to filter the interactions that are likely to occur in vivo. The information used for model includes protein-protein interaction, cellular localization, and gene expression patterns across tissues. We then obtained 2,096 highly confident kinase-substrate interactions. Following in vivo validation suggests this is a reliable data set to construct the human phosphorylation networks. Furthermore, we explored the specificity of kinase-substrate interaction in terms of phosphorylation motif, docking motif and scaffolding proteins. Our results provide the first global insights into the structure of phosphorylation networks and pathways in humans.

DETECTION AND REFINEMENT OF TRANSCRIPTION FACTOR BINDING SITES USING HYBRID MONTE CARLO METHOD

Ming Hu[1], Jindan Yu[2], Jeremy M Taylor[1], Arul M Chinnaiyan[3,4,5,6], Zhaohui S Qin[1]

[1]University of Michigan, Biostatistics, 1420 Washington Heights, Ann Arbor, MI, 48109, [2]Northwestern University Feinberg Medical school, Hematology/Oncology, 303 E. Superior St., Lurie 5-117, Chicago, IL, 60611, [3]University of Michigan, Michigan Center of Translational Pathology, 1400 E. Medical Center Drive, Ann Arbor, MI, 48109, [4] University of Michigan, Pathology, 1301 Catherine, Ann Arbor, MI, 48109, [5]University of Michigan, Urology, 500 E. Medical Center Dr., Room 3875, Ann Arbor, MI, 48109, [6]Howard Hughes Medical Institute, 1400 E. Medical Center Drive, Ann Arbor, MI, 48109

Coupling chromatin immunoprecipitation (ChIP) with recently developed massively parallel sequencing technologies has enabled genome-wide detection of protein-DNA interactions with unprecedented sensitivity and specificity. This new technology, ChIP-Seq, presents opportunities for in-depth analysis of transcription regulation. In this study, we explore the value of using ChIP-Seq data to better detect and refine transcription factor binding sites (TFBS). We introduce a novel computational algorithm named Hybrid Motif Sampler (HMS), specifically designed for TFBS motif discovery in ChIP-Seq data. We propose a Bayesian model that incorporates sequencing depth information to aid motif identification. Our model also allows intra-motif dependency to describe more accurately the underlying motif pattern. Our algorithm combines stochastic sampling and deterministic "greedy" search steps into a novel hybrid iterative scheme. This combination accelerates the computation process. Simulation studies demonstrate favorable performance of HMS compared to other existing methods. When applying HMS to real ChIP-Seq datasets, we find that (i) the accuracy of existing TFBS motif patterns can be significantly improved; and (ii) there is significant intra-motif dependency inside all the TFBS motifs we tested; modeling these dependencies further improves the accuracy of these TFBS motif patterns. These findings may offer new biological insights into the mechanisms of transcription factor regulation.

H3.3 LEVELS AT ENHANCERS AND THE 3' ENDS OF GENES CORRELATE WITH GENE EXPRESSION

Haruhiko Ishii[1], Nisha Rajagopal[2], Bing Ren[2], Zhanwen Du[3], Zhenghe Wang[3]
[1]University of California at San Diego, Division of Biological Sciences, 9500 Gilman Drive, La Jolla, CA, 92093, [2]University of California at San Diego, Bioinformatics and Systems Biology Program, 9500 Gilman Drive, La Jolla, CA, 92093, [3]Case Western Reserve University, Department of Genetics and Case Comprehensive Cancer Center, 10900 Euclid Avenue, Cleveland, OH, 44106

The histone variant H3.3 is known to be incorporated along the body of active genes in a replication-independent manner. Mass spectrometry analysis showed that H3.3 was enriched for certain active modifications such as K4 di- and tri-methylations and K9 and K14 acetylations, and under-enriched for markers of inactive chromatin such as K9 dimethylation. Most of the previous studies focused on the distribution of H3.3 relative to the body of active genes, but did not look at their abundance near enhancers. Enhancers are known to have a distinct chromatin signature characterized by the presence of H3K4me1 and the absence of H3K4me3. An algorithm was previously developed in our group to predict enhancers based on this property.

In this study, we used human cell lines derived from HCT116 with Flag epitope tagged to H3.3 or H3.1. ChIP-seq was performed using antibodies against Flag, H3, H3K4me1, H3K4me3, and H3K27ac. Bioinformatics analyses were performed to correlate the levels of histone modifications with the histone variants.

As seen before, H3.3 was more enriched for H3K4me3 and H3K27ac than H3 and H3.1. In addition, it was also seen to be more enriched for H3K4me1. However when the levels of these modifications along H3.3 enriched domains were studied, the profile of H3.3 deposition was seen to be almost identical to that of H3K4me1, but opposite to that of H3K4me3. This strongly suggests that H3.3 is an enhancer mark.

Enhancers were predicted and then associated with their corresponding promoters using CTCF-regulated domains. An interesting discovery that followed was that the levels of H3.3 at the enhancers were much better correlated with the expression of the associated genes than other enhancer marks such as H3K4me1 and H3K27ac.

In keeping with previous studies, the levels of H3.3 along the gene bodies increased gradually and had a correlation with the expression level. Also, the levels of H3.3 at the 3' ends were directly correlated with gene expression. This suggests that H3.3 may have a role to play in the 3' end processing of the mRNA.

Future studies will be carried out to find the correlation between p300 binding and H3.3 deposition, to further validate the role of H3.3 in enhancer action. Also, association of H3.3 with the transcription termination machinery will be closely examined, to understand the role of H3.3 at the 3' ends of genes.

STATISTICAL ANALYSIS REVEALS FINE SCALE PATTERNS IN METHYLATION AND EXPRESSION OF GENES IN THE NCI 60 PANEL

Mark A Reimers[1], William C Reinhold[2], Matthias Ehrich[3]

[1]Virginia Commonwealth University, Biostatistics, 730 E Broad St, Richmond, VA, 23298, [2]National Cancer Institute, Laboratory of Molecular Pharmacology, Bldg 37, 9000 Rockville Pike, Bethesda, MD, 20892, [3]Sequenom Inc, Research, 3595 John Hopkins Court, San Diego, CA, 92121-1331

Aberrant DNA methylation changes play an important role in the development of cancer, but the specific processes and mechanisms of methylation in cancer are poorly understood. This study performs a detailed multivariate analysis of *de novo* methylation profiles of 112 promoters, at almost single methylation site resolution, across the NCI 60 cell lines, and six normal samples (data originally described in Ehrich et al, ***PNAS***, 2008). We find that patterns of *de novo* methylation are distinctive in the leukemias, but that no pattern reflected tissue of origin among cell lines derived from epithelial tumors. Furthermore there is evidence for considerable change in methylation during cell culture. *De novo* methylation seems to occur in clusters. Finally we find that methylation seems be correlated with splice variation in an unexpected way. We speculate that methylation may influence splicing machinery.

BINDING-SITE AFFINITY MODELING OF POSITIONAL DEPENDENCIES AND CONTEXT-SENSITIVE NUCLEOTIDE INSERTIONS AND DELETIONS

Todd R Riley[1,2], Harmen J Bussemaker[1,2]

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

The regulation of gene expression by transcription factors (TFs) is of paramount importance to the overall control of cell function. However, our current understanding of the sequence specificity of TFs is limited. Current models typically assume that each nucleotide position in a putative binding site contributes independently to the overall binding affinity of the TF for the site [1,2]. In addition, these models assume that the residue-nucleotide binding interaction geometry is static for all possible nucleotide sequences, and that consequently all the binding-sites are of equal length [1,2].

However, analysis has shown that for some TF-DNA interactions the positional-independence assumption is not valid [3,4], and that some TF-DNA interactions tolerate nucleotide insertions and deletions relative to the consensus motif [5]. Subtle differences in binding specificity between TFs can lead to qualitative differences in the downstream processes they control. [6]. It is therefore crucial to develop accurate quantitative models for predicting TF binding affinity landscapes from genome sequence.

We have developed an extension to the biophysical model underlying the MatrixREDUCE algorithm [2] that detects deviations from the position specific affinity matrix (PSAM) model [2] due to dinucleotide dependencies and tolerated context-sensitive nucleotide insertions and deletions. Our new model pinpoints exactly where in the binding site the positional independence assumption breaks down. In addition, it estimates the energetic costs of context-sensitive nucleotide insertions and deletions within a half-site and within variable-length spacers between half-sites.

[1] Stormo, GD, Fields DS (1998). Trends Biochem. Sci. 23, 109-113.
[2] Foat, B.C., et al. (2006). Bioinformatics 22:14: e141-9.
[3] Roulet E, et al. (2002). Nature Biotech. 20, 831-835.
[4] Berger MF, et al. (2006). Nature Biotech. 11, 1429-1435.
[5] Riley TR, et al. (2009). Bioinformatics 2009, 10:111.
[6] Maerkl SJ, Quake SR (2007). Science 315, 233-237.

COMPARATIVE GENOMICS OF HSF-1 REGULATED ACTIVITY

Ana Rodrigues, Gerard Manning

Salk Institute, Razavi Newman Center for Bioinformatics, 10010 N Torrey Pines Rd, La Jolla, CA, 92037

In eukaryotes, heat shock (HS) induces an emergency stress response that involves a complex transcriptional program mediated by HS Factor-1 (HSF-1). HSF-1 also mediates several other external and internal stresses, and has recently been found to influence carcinogenesis, lifespan and promote disaggregation of amyloids in Alzheimer's disease models. A large body of research on HSF-1's canonical target, the hsp-70 promoter, has unveiled much detailed information on its binding site and binding mechanism. However, attempts to understand the full genomic scale and logic of its action have had limited success. We are leveraging available expression profiling data with large scale comparative promoter analysis to describe the HS response in four diverse eukaryotic phyla, characterize the variety of cis-response elements and track the evolution of both response elements and expression program within and between these phyla.

We analyzed all publicly available whole-genome expression profiles involving HS and/or HSF-1 manipulations in S. cerevisae (177 chips), C. elegans (41 chips), D. melanogaster (77 chips), and H. sapiens (110 chips) to establish a set of core consistent HS responding genes, beyond its canonical targets. Comparison of such large numbers of relatively noisy experiments, combined with enhanced orthology detection, enabled us to identify both high confidence genes across the four model systems, as well as phylum specific genes. We found that the proteins encoded by these genes are enriched for expected stress-response functions, but also include many unexpected metabolic functions, frequently in key pathway positions. We integrated this information with the limited publicly available HSF-1 ChIP-chip data to investigate how HSF-1 interacts with this larger set of putative targets. Our methodology for binding site identification exploits the cooperative nature of HSF-1 DNA binding to suggest degenerate binding sites, and uses the comparative information distilled from genome sequences to support them. This enabled us to expand the canonical HSF-1 binding site definition, and identify groups of genes enriched for novel site architectures, including various gapped multimeric sites. In addition, we found that despite the limited overlap between experimentally determined DNA-binding and transcriptional response, this overlap is substantially enriched in promoters with conserved binding sites, indicating that evolution can help distinguish productive from non-productive binding.

Through experimental collaborations, we are also studying the targets and timing of HSF-1 activity in proteostasis, protection against amyloid formation and aging.

REGULATION OF INTERSECTIN FAMILY GENE EXPRESSION

Lyudmyla Tsyba, Inessa Skrypkina, Sergii Kropyvko, Olga Novokhatska, Mykola Dergai, Oleksandr Dergai, Dmytro Gerasymchuk, Tetyana Gryaznova, Dmytro Morderer, Alla Rynditch

Institute of Molecular Biology and Genetics, Department of Functional Genomics, 150 Zabolotnogo str, Kyiv, 03680, Ukraine

Clathrin-dependent endocytosis is the major pathway for the uptake of nutrients and signaling molecules in higher eukaryotic cells. It is mediated by machinery of proteins organized into functional complexes on multimodular scaffolds. One of such undermembrane scaffolds are proteins of intersectin (ITSN) family. Intersectins form complexes with many regulatory proteins involved in apoptosis, mitogenic signaling pathways and cytoskeleton rearrangements. Structural and functional diversity of intersectins in vertebrates is generated by alternative splicing, the use of alternative promoters and presence of paralogous gene ITSN2. We identified alternative transcription initiation site in the fifth intron of ITSN1 gene and revealed 17 alternative splicing events affecting ITSN1 pre-mRNA. ITSN1 isoforms differ in their domain organization, interaction with protein partners, localization in different tissues and stages of development, which could affect the pattern and specific organization of the complexes of ITSN1 with protein partners. Moreover, analysis of 3'UTR of ITSN1 mRNAs revealed several target sites for miRNAs and AU-rich binding protein TTP. Some of these sites were common for ITSN1 and its interacting proteins suggesting their posttranscriptional regulation in different cellular processes. The results demonstrate the complex regulation of ITSN proteins that provide interfaces for interaction between basic endocytic machinery and different cell processes such as signaling, sorting and cytoskeleton rearrangements.

ASSOCIATIONS BETWEEN GENOMIC ALTERATIONS IN CANCER

Felix Sanchez Garcia[1], Uri David Akavia[1,2], Dylan Kotliar[1], Dana Pe'er[1,2]

[1]Columbia University, Department of Biological Sciences, 2960 Broadway, New York, NY, 10027, [2]Columbia University, Center for Computational Biology and Bioinformatics, 2960 Broadway, New York, NY, 10027

Cancer is a disease of genomic alterations in which a group of cells show uncontrolled growth, invasion, and sometimes metastasis. A combination of independent mutations produces the deregulation of key biological processes such as cell growth, programmed cell death and DNA repair. Tumor samples harbor a vast number and variety of alterations and most computational methods focus on distinguishing alterations that contribute to oncogenesis from passenger alterations. However, a single mutation by itself is not sufficient to turn a normal cell into a cancer cell and indeed different combinations of alterations can result in cancer formation. It is vital to understand how driver alterations interact to develop the disease.

We developed a computational framework to systematically identify interactions between mutations. Specifically, we consider relationships with AND and OR logic between genetic alterations via co-occurrence and mutual exclusion. We defined a statistical score that assesses the significance of the interaction and an efficient search algorithm that detects high scoring combination. For the initial prototype of the system we consider chromosomal copy number aberrations and point mutations as candidate alterations.

While the study of co-occurring alterations provides insight into epistatic interactions that drive the disease, the analysis of associations of mutual exclusion allows us to move from a gene-centric model to a pathway-centric model. Recent studies show that pathways and not specific genes are ultimately targeted by cancer. A large scale study performed by the TCGA (Nature, 2008) discovered that alterations within targeted pathways tend to be mutually exclusive. Our approach is to go beyond annotated pathways and use this property to identify and define targeted pathways.

We applied our method to the TCGA datasets generating a genetic mapping that goes beyond the study of isolated alterations, analyzing synergistic interactions and targeted pathways without any previous knowledge. We extended previous analysis of this data and discovered additional pathways involved in Glioblastoma and their interactions.

IDENTIFICATION OF ALLELIC EXPRESSION DIFFERENCES IN RETINAL EXPRESSED DISEASE GENES

Simone Schimpf-Linzenbold, Sukirthini Balendran, Bernd Wissinger

Molecular Genetics Laboratory, Institute for Ophthalmic Research, Roentgenweg 11, Tübingen, 72076, Germany

Reduced penetrance and variability in disease expression with respect to onset, course and severity is common in retinal dystrophies and can be observed even between and within families with the same primary gene defect. This feature hampers solid and reliable genotype/phenotype correlations as well as individual disease prognosis. It has been suggested that modifications in gene regulation are responsible for most of the observed phenotypic variations. Expression quantitative trait loci (eQTL) studies have become a widely used tool for identifying genetic variants that affect gene regulation.
For identifying such eQTLs we crossed five inbred mice (C57BL/6, BALB/c, CAST, CBA and LP) to form an identical F1 generation with a high genetic variability and isolated DNA and RNA from ear and retina, respectively. We screened 20 different retinal disease genes for heterozygous cSNPs applying PCR and sequencing. To determine allelic expression differences based on the identified cSNPs, we applied Pyrosequencing assays on RT-PCR amplified retinal cDNAs. The results were calibrated for equimolar ratios by used genomic DNA as a control. Using the Pyrosequencing technology, a highly accurate method to detect allele-specific expression differences, we have detected an allelic imbalance in four different genes. In the *Opa3* gene we have seen the allelic imbalance already on the genomic level suggesting a copy number variation. In *Pde6c* we were able to detect a mutation within this gene that results in a premature termination codon leading, due to the nonsense mediated mRNA decay, to a downregulation of the mutant transcript. For the other two genes we will determine the promoter regions and identify variants functionally assessed applying reporter gene assays.

COORDINATE TRANSCRIPTIONAL REGULATION OF THE RESPONSE TO ENVIRONMENTAL EXTREMES

Amy K Schmid[1], Min Pan[2], Nitin S Baliga[2]

[1]Duke University, Biology Department and Center for Systems Biology, 124 Science Dr., Durham, NC, 27708, [2]Institute for Systems Biology, Institute for Systems Biology, 1441 N 34th St., Seattle, WA, 98103

How do transcriptional regulatory networks precisely orchestrate physiological events even when confronted with environmental stress? In archaeal systems, which live in regimes ranging from boiling temperatures to saturated salt, extreme stress is a constant threat. We hypothesize that the heartiness required in such conditions has evolved through cross-regulated transcriptional network modules to ensure the appropriate execution of dynamic changes in a core set of stress response genes. For example, our recent results suggest that the response to iron starvation in organisms dwelling in saturated salt is coordinately regulated by two transcription factors acting through interconnected feed forward regulatory network motifs. This network enables the expression of core metabolic genes, such as those associated with growth and chemotaxis, to be calibrated to the availability of iron. We discuss the evolutionary implications of the conservation of feed forward motifs across all three domains of life.

STUDYING GENE STRUCTURE, EXPRESSION AND REGULATION USING THE ILLUMINA HISEQ2000 SYSTEM

Gary P Schroth, Shujun Luo, Lu Zhang, Raymond McCauley, Robin Li, Irina Khrebtukova

Illumina, Expression Applications, 25861 Industrial Blvd, Hayward, CA, 94545

The Illumina HiSeq2000 is a new high-throughput DNA sequencing platform that produces more than 1 billion reads and over 200 Gb of high quality data per run. We have used the throughput of the HiSeq2000 to create four enormous new data sets: a Gene Expression Body Map that includes very deep coverage of transcripts from 16 individual human tissues, a complete map of all coding and non-coding RNAs in the human transcriptome, a RNA-Seq profiling study of more than 200 cancer samples, and a whole genome comparison of DNA methylation in a cancer/normal pair. Each of these very large sequencing projects can now be done in a single instrument run using the HiSeq2000. In this presentation we will review the data from these four projects, look in detail at the system performance for each application, and discuss how this new system enables new genome-wide studies in functional genomics.

HISTONE H3 METHYLATION AND H2B UBIQUITINATION SIGNATURES IN THE YEAST GENOME AND THEIR ROLES IN CELL CYCLE CONTROL, TRANSCRIPTION, AND MRNA SPLICING.

Julia M Schulze[1], Thomas Hentrich[2], Shima Nakanishi[3], Ali Shilatifard[3], Michael S Kobor[1]

[1]Centre for Molecular Medicine and Therapeutics, University of British Columbia, Medical Genetics, 950 West 28th Ave, Vancouver, V5Z4H4, Canada, [2]School of Computing Science, Simon Fraser University, Computer Science, 8888 Burnaby Drive, Burnaby, V5A1S6, Canada, [3]Stowers Institute for Medical Research, Stowers Institute, 1000 East 50th Street, Kansas City, MO, 64110

In eukaryotic cells, genomic DNA is packaged with histone proteins into a dynamic chromatin structure composed of distinct neighborhoods. These neighborhoods are specified by histone variants, ATP-dependent chromatin remodeling, and various posttranslational modifications of histones such as acetylation, ubiquitination, and methylation.
In yeast, there are three histone H3 methyl-transferases: Set1 methylates lysine H3K4, Set2 methylates lysine H3K36, and Dot1 methylates lysine H3K79. All these enzymes can mono-, di- and tri-methylate their target, thereby adding an extra layer of complexity. Using high-resolution ChIP-on-Chip, we have shown that H3K79 di- and trimethylation were located at distinct and mutually exclusive regions of chromatin. In addition, M/G1 cell cycle-regulated genes were significantly and specifically enriched for H3K79 dimethylation. Swi4 and Swi6, components of the cell cycle-regulated SBF transcription complex, were required for normal levels of dimethylation, but not trimethylation, of H3K79.
Importantly, histone modifications do not occur independently but rather exist in combination thereby defining complex patterns. We hypothesized that loci of similar regulatory function share similar chromatin signatures. Therefore, we systematically examined histone H3K4, K36, and K79 methylation signatures for the promoter, 5'end, middle, and 3'end of all genes to search for frequently occurring patterns. Interestingly, we were able to identify over-represented methylation signatures that were associated with promoter binding of specific transcription factors and specific gene expression profiles.
Since methylation of H3K4 and H3K79 requires prior monoubiquitination of H2BK123, we also profiled this modification and its ubiquitin-ligase Bre1 genome-wide. Cluster analysis of these profiles and their downstream methylation marks revealed that specific signatures were associated with intron-containing genes. Subsequent functional studies indicated a possible role for Bre1 in splicing.
Taken together, our studies established a framework for the circuitry between chromatin structure and processes like cell cycle regulation, transcription, and mRNA processing.

GENOME INSTABILITY IN AGING AND LONGEVITY

Bjoern Schumacher
University of Cologne, CECAD, Zuelpicher Str. 47a, Cologne, 50674, Germany

Genome Instability has been recognized as causal factor of cancer and recently also as a major contributing factor of aging. A number of skin cancer susceptibility and progeroid (premature aging-like) syndromes are linked to defects in nucleotide excision repair (NER). NER thus provides a highly relevant experimental system to study the role of genome integrity both in cancer and in aging. Using the NER system we recently uncovered a novel link between DNA damage accumulation and the regulation of longevity assurance programs and tumor suppressor mechanisms. Based on genome-wide comparative correlation analysis we uncovered similarities between progeroid mouse models and mice with extended longevity. Furthermore, we demonstrated the validity of genome-wide correlation analysis for assessing biological aging. Mechanistically, we identified a response program to persistent DNA damage that is triggered amid increasingly damaged genomes with aging. DNA damage treatment in vitro led to similar gene expression changes as observed in various tissues with aging. Low amounts of persistent lesions that interfere with RNA polymerase II elongation led to attenuation of somatotropic genes that is linked to extended longevity. Functionally, this response program evoked enhanced stress resistance and antagonized hyperplasia. We propose that sensing of low levels of persistent DNA damage by RNAPII comprises a mechanistic basis for hormesis and shifts the organism's resources from growth to somatic maintenance in aging.

PREDICTING ENHANCER REGIONS AND TRANSCRIPTION FACTOR BINDING SITES IN D. MELANOGASTER USING SUPERVISED LEARNING APPROACHES

Rachel S Sealfon[1,2], Christopher A Bristow[1,2], Pouya Kheradpour[1,2], Manolis Kellis[1,2]
[1]Computer Science and Artificial Intelligence Laboratory, EECS, 32 Vassar St., Cambridge, MA, 02139, [2]Broad Institute, 7 Cambridge Center, Cambridge, MA, 02139

New experimental datasets, such as genome-wide profiling of chromatin marks and transcription factor binding, offer the potential for gaining insight into the combinatorial code of gene regulation and understanding the relative roles played by enhancers, transcription factors, chromatin marks, and individual motifs in driving gene expression. Towards this goal, we used a machine learning approach to study the predictive power of experimental and sequence-based combinations of features in the context of both enhancer prediction and prediction of transcription factor binding sites.

We developed an integrative approach to enhancer prediction that leverages the wealth of available experimental data on chromatin marks and transcription factor binding. Using a supervised learning framework, we identified combinations of ChIP-chip protein binding data, chromatin marks, chromatin-associated factors, and sequence conservation features that are characteristic of the experimentally validated enhancers in the REDFly database. We observe that the most confidently predicted enhancers are enriched near genes patterned in the early embryo, and include recently validated enhancers while excluding recently tested regions for which experimental validation of enhancer function failed. We also found that while chromatin marks alone had low predictive power, including chromatin mark features as well as transcription factor binding features dramatically improved the power of our classifier. The improvement in classifier performance using combinations of types of features relative to any individual feature type suggests that each class of functional elements plays distinct yet necessary roles in defining enhancer regions in the cell.

We have also applied supervised learning methods for predicting transcription factor binding locations based on combinations of regulatory motifs. For each experiment in a compendium of ChIP-chip studies, we constructed a classifier to distinguish between regions bound by the given factor and regions bound by any other factor. For each factor, we compared the performance of subsets of enriched and depleted motifs, and examined the improvement in classifier performance as individual motifs are added to the feature set. While the results differ across factors, we found that combinations of features typically outperformed individual motifs, and predictive power increased when depleted motifs were included as features. This result suggests that binding of an individual transcription factor at a given site may be highly dependent on the local combination of bound factors, which provide both synergistic and antagonistic influences.

LOCALIZED DNA DEMETHYLATION MARKS RECOMBINATION INTERMEDIATES DURING IMMUNOGLOBULIN HEAVY CHAIN GENE ASSEMBLY

Roza Selimyan[1], Rachel M Gerstein[2], Irina Ivanova[1], Thomas Perlot[3,4], Frederick W Alt[3], Ranjan Sen[1]
[1]National Institutes of Health, National Institute on Aging, Laboratory of Cellular and Molecular Biology, 251 Bayview Boulevard, Baltimore, MD, 21224, [2]University of Massachusetts Medical School, Department of Molecular Genetics and Microbiology, 55 Lake Avenue, Worcester, MA, 01655, [3]The Howard Hughes Medical Institute, Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA, 02115, [4] Austrian Academy of Sciences, Institute of Molecular Biotechnology, Dr. Bohr-Gasse 3, Vienna, 1030, Austria

The highly diverse antigen receptor repertoire of B- and T-cells is generated by gene rearrangements. Recombination of seven antigen receptor loci is governed by the same recombinase machinery and is initiated by regulated access of lymphocyte-restricted RAG-1 and RAG-2 proteins. Numerous studies indicate that chromatin structure is the key factor that regulates accessibility. Accessibility generally correlates with DNA demethylation, H3K9 acetylation, H3K4 methylation. There is considerable evidence that developmentally regulated histone modifications accompany sequential rearrangement of the immunoglobulin heavy chain (IgH) gene locus, which occurs in two well coordinated recombination reactions occur. In early pro-B cells one of several diversity (DH) gene segments rearranges to one of four joining (JH) gene segments to generate a DJH recombined allele. This is followed by variable region (VH) gene recombination to the DJH junction. DH to JH recombination also occurs in T lymphocytes, but VH recombination is B cell specific. DNA methylation has been implicated in control of Ig light chain recombination; however, little information exists on DNA methylation of the IgH locus. We therefore undertook a systematic study of DNA methylation of the IgH locus during B-cell development.

We found that prior to rearrangements the IgH locus was methylated in the DH-Cμ region with the exception of two regulatory regions, PDQ52 and the intronic enhancer Eμ that are marked by tissue-specific DNase I hypersensitive sites. No correlation with histone modifications was found in the germline locus. However, rearranged DJH junctions were hypomethylated in pro-, pre- and mature B-cells. We found that hypomethylation was highly localized to recombined DJH junctions and did not extend to unrearranged DH gene segments. In contrast to wild type pro-B cells, DJH junctions were hypermethylated in Eμ-/- pro-B cells and CD4+CD8+ double-positive thymocytes. In both cell types D to JH recombination is the last event that happens to the IgH locus. Because VH gene rearrangement is blocked in both cell types, we suggest that DNA demethylation at DJH junctions may be required for the second step of VDJ recombination demonstrating that junctional demethylation is Eμ-dependent.

MICRORNA-MEDIATED REGULATION OF GENE EXPRESSION AND MOLECULAR NETWORKS IN ORAL SQUAMOUS CELL CARCINOMA

Patricia Severino[1], Holger Brüggemann[2], Flavia M Andreghetto[1], Renata M Soares[1], Marcelo A Bezerra-Paparelli[3], Flavio Borges[4], Raquel A Moyses[5], Fabio D Nunes[6], Ariane Machado-Lima[3], Eloiza H Tajara[7], GENCAPO Head and Neck Genome Project[7]
[1]Albert Einstein Research and Education Institute, Center for Experimental Research, Av. Albert Einstein, Sao Paulo, 06561-901, Brazil, [2]Max Planck Institute for Infection Biology, Department of Molecular Biology, Charitéplatz, Berlin, D-10117, Germany, [3]University of Sao Paulo, School of Arts, Science and Humanities, R. Arlindo Béttio, Sao Paulo, 03828-000, Brazil, [4] University of Sao Paulo, Department of Public Health, Av. Dr. Arnaldo, Sao Paulo, 03828-000, Brazil, [5]University of Sao Paulo, Faculty of Medicine, Av. Dr. Eneas de Carvalho, Sao Paulo, 05403-000, Brazil, [6]University of Sao Paulo, Department of Oral Pathology, Av. Prof. Lineu Prestes, Sao Paulo, 05403-000, Brazil, [7]Faculty of Medicine, Department of Molecular Biology, Av. Brigadeiro Faria Lima, Sao Jose do Rio Preto, 01311-300, Brazil

MicroRNAs (miRNA) are endogenous, noncoding RNAs that regulate gene expression by degrading or destabilizing the RNA message or by inhibiting protein translation. They are believed to integrate networks wherein variation in the expression level of one miRNA could set up broad changes involving distinct pathways. High levels of miRNA-196 have been recently associated with oncogenic phenotypes in different cancer types. We detected statistically significant over-expression of both miRNA-196a and miRNA-196b in oral squamous cell carcinoma (OSCC) samples compared to cancer-free surgical margins by means of quantitative real time-PCR. In order to evaluate the impact of miRNA-196a/b in OSCC, we selected possible direct targets using two algorithms - PicTar and TargetScan – and integrated data related to these targets using Ingenuity Pathway Analysis. Cancer-related networks linked predicted direct or indirect miRNA-196a/b targets to tumor suppressor genes and oncogenes. Squamous cell carcinoma was the highest enriched cancer type in this analysis (p-value 3.2 $x10^{-7}$), indicating that miRNA-196a/b deregulation could be a crucial event in OSCC. Additionally, for all genes presenting a miRNA-196a/b target predicted by at least one tool (266 genes), we searched for miRNA-196a/b target sites in their promoter region. This search was performed using miRanda software with default cutoffs and options --shuffle -s 100000. Alternative transcription start positions were considered for all genes. We found 195 genes with predicted target sites at the promoter region, several of which possessed more than one predicted target site at the promoter region per transcript isoform. Our results suggest that alternative transcription regulation by miRNAs should be considered and could help in the comprehension of their systemic effect. Ultimately, an in depth understanding of the cross talk between deregulated miRNAs and related molecular networks should contribute for the development of novel strategies for targeting molecular components and, thus, for the effective treatment of this disease.

Financial Support: FAPESP (04/12054-9 and 09/04166-5) and Instituto Israelita de Ensino e Pesquisa (IEP.PE.08-0125)

DIFFRNCES IN p53 BINDING UPON ACETYLATION

Idit Shiff, Itamar Simon

Hebrew University of Jerusalem, Department of Microbiology and Molecular Genetics, Hebrew University Medical School, Jerusalem, 91120, Israel

Activation of p53 leads the cells to apoptosis or cell cycle arrest, and the mechanism p53 distinguishes between the different outcomes is not well understood. The tumor suppressor p53 becomes acetylated upon its activation, thus acetylation may play an important role in the p53 regulation mechanism. We study differences in p53 binding by ChIP on chip using cells containing p53 with mutations in the extreme C terminal region mimicking either p53 acetylated state or unacetylated state. Surprisingly we found more targets bound in the wild-type state than in the mutants. Also in the mutant mimicking p53 acetylation we got less targets than in the unacetylated mimicking mutant. Further analysis is needed to determine which form binds the different sets of target sites and how the differential binding is acheived.

ANALYSIS OF TRANSCRIPTIONAL REGULATORS GOVERNING ESTABLISHMENT OF CELL-TO-CELL REPROGRAMMING NETWORK

Jay W Shin, Harukazu Suzuki, Yoshihide Hayashizaki

RIKEN Yokohama, Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan

Trans-differentiation describes a switch in one cell-fate to that of another and the inference of transcriptional networks that regulate this process presents valuable information for cell-replacement therapy. However, transcriptional regulatory modulators (TMs) responsible for implementing cell reprogramming are largely unknown. In this report, we demonstrate the creation and analysis of a database of global gene expression profiles of over 20 different human primary cell types derived from different cell-lineages including induced-pluripotent stem (iPS) cells. Further using genome-wide predictions of transcription factor binding sites in proximal promoters, we identified key motifs and TMs that activate, in part, transcriptional regulatory network of each cell type. We also showed that cells derived from the same cell-lineage clustered together, but most interestingly, ectopic expression of important TMs in human fibroblasts deviated towards the associated target cells based on unsupervised clustering methods. These findings demonstrate apt extraction of key transcriptional modulators and plasticity, in part, to trans-differentiate into multiple targets with defined factors. Using further bioinformatic analysis, we uncovered a protein-protein network that is shared by the lineage-related cell types and also between fibroblasts to various target cells. Using such integrative approach, we are able to reverse-engineer transcriptional regulatory network governing cell-to-cell reprogramming process.

INFERRING CIS-REGULATORY NETWORK VIA DIFFERENTLY PERTURBED MULTIPLE TIME-COURSE GENE EXPRESSION DATA

Yuichi Shiraishi, Takeshi Nagashim, Mariko Okada

RIKEN, Research Center for Allergy and Immunology, W518 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan

Distinct extracellular stimulations lead to different cell fates in many cell lines. For example, epidermal growth factor (EGF) induces proliferation whereas heregulin (HRG) induces differentiation in MCF7 human breast cancer cells. Unraveling the mechanism of different cell decisions elicited by distinct stimulations is one of the most important problems of molecular biology.

In this talk, we will introduce our statistical approaches for understanding the difference of cell decision programs in terms of transcriptional sequence motifs. From multiple time-course gene expression data collected under several stimulation (EGF, HRG and several inhibitors) conditions of MCF7 cells, we extract several regulatory motifs which correlate significantly with temporal profiles under specific conditions. Joining inferred regulatory motifs above and statistical modeling, we estimate a regulatory network which takes multiple conditions into account. Furthremore, we briefly discuss biological aspects of the inferred network.

PRECISE TEMPORAL CONTROL OF THE EYE REGULATORY GENE PAX6 VIA ENHANCER BINDING SITE AFFINITY

Sheldon Rowan[1], Trevor W Siggers[1], Salil A Lachke[1], Yingzi Yue[1], Martha L Bulyk[1,2,3], Richard L Maas[1]

[1]Brigham & Women's Hospital and Harvard Medical School, Genetics, 77 Ave Louis Pasteur, Boston, MA, 02115, [2]Brigham & Women's Hospital and Harvard Medical School, Pathology, 77 Ave Louis Pasteur, Boston, MA, 02115, [3]Harvard Medical School, Harvard-MIT Division of Health Sciences and Technology (HST), 77 Ave Louis Pasteur, Boston, MA, 02115

How transcription factors interpret the *cis* regulatory logic encoded within enhancers to mediate quantitative changes in spatiotemporally-restricted expression patterns during animal development is not well understood. *Pax6* is a dosage-sensitive gene essential for eye development. We have identified the Prep1 (pKnox1) transcription factor as a critical dose-dependent upstream regulator of *Pax6* expression during lens formation. We identify a pair of phylogenetically conserved, lower-affinity Prep1 DNA binding sites, each uniquely required for *Pax6* lens enhancer activity. Using mathematical modeling and *in vivo* experiments, we show that Prep1 binds non-cooperatively to these sites but functions synergistically to regulate *Pax6* expression in response to changing Prep1 levels, and that it is the affinity of Prep1 for these two sites that dictates the timing of *Pax6* lens enhancer activation. Vast numbers of lower-affinity transcription factor binding sites present in vertebrate genomes may likewise play crucial roles in controlling the precise timing of gene expression through such affinity-dependent mechanisms.

MODEL- AND KNOWLEDGE-BASED ANALYSIS OF GENE DEREGULATION

Ewa M Szczurek[1,2], Florian Markowetz[3], Irit Gat-Viks[4], Przemyslaw Biecek[2], Jerzy Tiuryn[2], Martin Vingron[1]

[1]Computational Biology Department, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, Berlin, 14195, Germany, [2]University of Warsaw, Faculty of Mathematics, Informatics and Mechanics, Banacha 2, Warszawa, 02-097, Poland, [3]Cancer Research UK Cambridge Research Institute, Computational Biology, Robinson Way, Cambridge, CB2 0RE, United Kingdom, [4] Broad Institute of MIT and Harvard, Aviv Regev Lab, 7 Cambridge Center, Cambridge, MA, 02142

Understanding changes in gene regulation under different cellular conditions is an important step in elucidating disease mechanisms, including cancer. For example, one condition could refer to healthy cells, whereas the other condition to tumor cells. In the tumor cells, one or more pathway member is often mutated and thus not able to interact with other proteins, affecting the responsive genes. Such a drastic switch in cellular conditions leads to deregulation — differently activated signaling pathways and differently expressed genes. Different activation of the pathway and its downstream targets implies changes in regulatory control that governs the targets. There is a need for a systematic method reconciling the reconstruction of regulatory networks with an analysis of changes in gene expression between two cellular conditions.

Here, we propose a knowledge-based approach to identify how gene regulation changes between two conditions. Our approach utilizes prior models of pathways active in both conditions, as well as perturbation data. Data analysis yields probabilities of differential expression for the genes under each perturbation, and also benefits from prior knowledge, e.g., of the genes expected to be regulated. We dissect differences in gene regulation by explicitly modeling changes in the signaling pathway and in expression profiles of its downstream genes measured after perturbing pathway members under both cellular conditions. With the model and data, we recover the system-specific deregulated transcriptional control relations and the deregulated genes.

Our approach is exemplified in a case study on DNA damage response network in human, which mediates cellular response to genomic alterations and thus functions as one of the key protections against cancer development. We find major transcriptional deregulation after exposure of cells to a damaging agent and evaluate the results both statistically and biologically.

ADIPOGENESIS INDUCED BY SHORT INTERFERENCE RNAS PREDICTED BY INFORMATICS

Taku Tanaka[1], Hirokazu Chiba[2], Wataru Fujibuchi[2], Masato Miyake [1]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Research Institute for Cell Engineering (RICE), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan, [2]National Institute of Advanced Industrial Science and Technology (AIST), Computational Biology Research Center (CBRC), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Emerging iPS cells induction suggests that mammalian cell phenotypes are artificially reversible. One of the next challenges would be informatics-driven prediction of molecules to control cell differentiation. Here, we report adipogenesis induction by siRNAs predicted by informatics. Human mesenchymal stem cells (hMSCs) differentiate into adipocytes by at least three compounds, indomethacin, dexamethasone and 3-isobutyl-1-methyl-xanthine (IBMX). PPARgamma expression, which is a well-known marker of commitment to adipocytes, increases in hMSCs within 7 days after induction by the three compounds. We found that IBMX function was partially replaced by a set of siRNAs predicted through comparison of genes expression specificity and level in hMSCs and the derived adipocytes. The set of siRNAs enhanced PPARgamma expression in hMSCs in the presence of indomethacin and dexamethasone. The results show that informatics give us fruitful hints towards a complete prediction of the adipogenesis pathways.

NUCLEAR RECEPTOR-INDUCED DNA BREAKS THAT CAN UNDERLIE SPECIFIC TRANSLOCATIONS IN CANCERS

Bogdan Tanasa[1,2], Chunru Lin[1], Liuqing Yang[1], Bong-gun Ju[4], Michael G Rosenfeld[1,3]

[1]Howard Hughes Medical Institute, (HHMI), 9500 Gilman Drive, La Jolla, CA, 92093, [2]The Scripps Research Institute, (TSRI), 10550 North Torrey Pines, La Jolla, CA, 92037, [3]University of California, San Diego, (UCSD), 9500 Gilman Drive, La Jolla, CA, 92093, [4] Sogang University, Life Sciences, Life Sciences, Seoul, 121-742, South Korea

Chromosomal translocations are a hallmark of leukemias and lymphomas and also appear in solid tumors, but the underlying mechanism remains elusive. By establishing a cell model that mimics the relative frequency of translocation events without proliferative selection, we report mechanisms of nuclear receptor-dependent tumor translocations. We show that intronic binding of liganded androgen receptor (AR) first juxtaposes translocation loci by triggering intra- and interchromosomal interactions. Secondly, AR promotes site-specific DNA double-stranded breaks (DSBs) at translocation loci by recruiting two types of enzymatic activities induced by genotoxic stress and liganded AR, including AID and LINE1 ORF2. These enzymes synergistically generate site-selective DSBs at juxtaposed translocation loci that are ligated by nonhomologous end joining pathway for specific translocations. Employing a protocol of BrdU-labeling by terminal deoxynucleotide transferase (TdT) coupled with ChIP-seq analysis suggested potential DNA breaks in response to hormonal treatments in the intronic regions of the translocation partners, that were confirmed by specific DNA sequencing.

TRANSCRIPTIONAL REGULATORY SUBNETWORKS IN THE MOUSE BRAIN AS DERIVED FROM DATA SETS IN THE ALLEN MOUSE BRAIN ATLAS

Ronald C Taylor[1], George Acquaah-Mensah[2], Jason McDermott[1]
[1]Pacific Northwest National Laboratory (U.S. Dept. of Energy, operated by Battelle), Computational Biology & Bioinformatics Group, 902 Battelle Blvd., Richland, WA, 99352, [2]Massachusetts College of Phamacy and Health Sciences, Pharmaceutical Sciences, 19 Foster Street, Worchester, MA, 01608

The gene expression measurements generated by high-throughput in situ hybridization and stored in the mouse brain atlas (http://mouse.brain-map.org/ welcome.do) at the Allen Institute for Brain Science form a unique resource. For our work, we have employed the data in this repository in the inference of mRNA transcriptional regulatory (sub)networks, said reconstruction being based on algorithms that employ correlations in gene state to infer gene-to-gene regulation. In this poster we present preliminary results using such methods on Allen data for the determination of high-confidence transcriptional regulatory subnetworks in the mouse brain. We used several state-of-the-art algorithms for such network inference: the Context Likelihood of Relatedness (CLR) algorithm (mutual information), the Inferelator algorithm (standard regression and model shrinkage), and the Supervised Inference of Regulatory Networks (SIRENE) algorithm (Support Vector Machine based supervised machine learning). We identified mouse transcription factors (TFs) and known TF targets for use in training SIRENE using GeneInfoViz and Biobase. In addition, we used the gene sets in the high-confidence connections inferred by the algorithms denoted above as input to Cross-Ontological Analytics (XOA, Sanfilippo et al., 2007), a Gene Ontology (GO) based algorithm recently developed at Pacific Northwest National Lab that can link genes into networks using aggregated semantic similarities between GO annotations found for those genes in the GO database.

We combined the inferred networks from these algorithms in the Cytoscape environment. We compared their overlap, and performed further topological analyses, including identification of "bottleneck" genes and the subnetworks that surrounded them. In brief: “betweenness”, which measures the number of shortest paths going through a given node (gene), is one of the most important topological properties of a network. Nodes having the highest betweenness we call “bottlenecks”. Such genes control most of the information flow in a network, and can be viewed as critical points which have been found to correlate well with gene essentiality. We have earlier shown the usefulness of such bottleneck analysis for biological insight (McDermott et al, 2009). Our focus here was on finding important (e.g., bottleneck-involved) novel regulatory connections, and confirming known connections, around transcription factors known to be involved in neurodegenerative processes. Finally, we discuss the biological relevance of several of the novel connections found, using functional relationships found by XOA and using background information from the literature.

A NOVEL META-ANALYSIS IDENTIFIES DAF-16/FOXO TRANSCRIPTION FACTOR TARGET GENES

Ronald G Tepper[2,3], Harmen J Bussemaker[1,2]

[1]Columbia University, Biological Sciences, 1212 Amsterdam Ave., New York, NY, 10027, [2]Columbia University, Center for Computational Biology and Bioinformatics, 1130 St Nicholas Ave., New York, NY, 10032, [3]Columbia University, Integrated Program in Cellular, Molecular, and Biophysical Studies, 630 W 168 St, New York, NY, 10032

The conserved DAF-16/FOXO transcription factor mediates aging and longevity in C. elegans and other species. While a number of studies have been done to identify genes acting downstream of DAF-16 to regulate lifespan, to date agreement on the set of targets has been limited. Here we have used a novel method of meta-analysis to combine seven experimental datasets consisting of 46 genome-wide expression microarrays from different protocols, genotypes, time points, and platforms. This has yielded a more complete target list, with increased sensitivity and selectivity. Following array-specific standardization, normalization, and re-mapping, a fold-change value was calculated for each transcript on each array, considering all within-array features targeting that transcript. A standard deviation was estimated for each fold-change value based on pixel data, within-array replicates, and an overall error model for each individual array. Two related voting schemes were then used to determine the response of each transcript. One scheme used only the direction of the fold-change on each array, regardless of magnitude. The binomial distribution was used to evaluate the probability that the transcript was consistently up- or down-regulated. The second voting method combined each transcript's fold-change with its standard deviation into a score related to the net probability of the transcript actually being up- or down-regulated in that array. These scores were summed across arrays, and compared to an empirical null distribution based on permuting all genes within each array. As a result of this meta-analysis, approximately 1,200 up-regulated and 1,000 down-regulated targets were identified at a 1% false discovery rate (FDR) level, the majority of which were not previously identified. In addition, a significant number of prior positives inferred from each individual dataset were not supported by the combined data.

HIGH NUCLEOSOME OCCUPANCY IS ENCODED AT HUMAN REGULATORY SEQUENCES

Desiree Tillo[1], Noam Kaplan[2], Irene K Moore[3], Yvonne Fondufe-Mittendorf[3], Andrea J Gossett[4], Yair Field[2], Jason D Lieb[4], Jonathan Widom[3], Eran Segal[2], Timothy R Hughes[1]

[1]University of Toronto, Department of Molecular Genetics, 160 College Street, Room 1350, Toronto, M5S1E3, Canada, [2]Weizmann Institute of Science, Department of Computer Science and Applied Mathematics, Ziskind bldg., Room 149, Rehovot, 76100, Israel, [3]Northwestern University, Department of Biochemistry, Molecular Biology, and Cell Biology, 2145 Sheridan Road, Evanston, IL, 60208, [4] University of North Carolina at Chapel Hill, Department of Biology, 408 Fordham Hall, Chapel Hill, NC, 27599-3280

Active eukaryotic regulatory sites are characterized by open chromatin, and yeast promoters and transcription factor binding sites (TFBSs) typically have low intrinsic nucleosome occupancy - i.e. these sequences are disfavoured when naked DNA and histone octamers are assembled *in vitro*. Here, using a computational model of intrinsic nucleosome sequence preference derived from assembly of yeast genomic DNA on chicken nucleosomes, we show that in contrast to yeast, DNA at human promoters, enhancers, and TFBSs generally encodes high intrinsic nucleosome occupancy. These elements also have a strong tendency towards high experimentally measured nucleosome occupancy *in vivo*. These regions typically have high G+C content, which correlates positively with intrinsic nucleosome occupancy, as well as the reduced probability of rigid, nucleosome-excluding polyA-like sequences. The only systematic deviation between our model and measured nucleosome occupancy *in vivo* is that CpG-containing promoters, while having high overall predicted intrinsic occupancy, are depleted of nucleosomes *in vivo*. This observation has previously been attributed to the presence of the RNA polymerase preinitiation complex; however, we find that poly-A-like sequences are also particularly enriched at this location, and may also contribute to relative nucleosome depletion. We propose that high nucleosome affinity is directly encoded at regulatory sequences in the human genome to restrict access to regulatory information that will ultimately be utilized in only a subset of differentiated cells. Our findings present a functional consequence of variation in base content that is observed at diverse scales in eukaryotic genomes.

SINGLE MOLECULE ANALYSIS IN YEAST CELLS REVEALS CELL-CYCLE DEPENDENT mRNA DEGRADATION

Tatjana Trcek, Daniel R Larson, Robert H Singer

Albert Einstein College of Medicine, Anatomy and Structural Biology, 1300 Morris Park Avenue, The Bronx, NY, 10461

Messenger RNA decay is a major determinant of cellular mRNA abundance. By regulating stability a cell can adjust its mRNA levels rapidly in response to an array of cellular and environmental stimuli. Conventional methods that use population ensemble measurements for measuring mRNA decay rates are intrinsically inaccurate because of population heterogeneity, normalization of mRNA signal and the use of transcriptional inhibitors. Average decay rates reported in the literature vary by more than 50% from each other and misrepresent how cyclically expressed transcripts decay. In this work we demonstrate how Fluorescent In Situ Hybridization (FISH) can be used to measure endogenous mRNA decay rates with single molecule resolution in individual, unperturbed yeast cells. We used morphological indicators to determine cell cycle timing. We detected both cytoplasmic transcripts and nascent transcripts in order to calculate how frequently cells transcribed a gene during cell cycle phases and determined how many transcripts were made per transcriptional event. Decay rates were then calculated using a mathematical model.
We focused on the expression of *SWI5* and *CLB2* genes, which peaks during G2/M transition. Prior to anaphase onset SWI5 and CLB2 mRNAs decay with a 110 min and 40 min half-life, but with a much shorter half-lives afterwards (1.5 min and 1.2 min respectively). Pre-anaphase stabilization acts as counterbalance to infrequent initiations of transcription during early stages of their expression providing for an increase in mRNA abundance. During post-anaphase the rapid decay prevents the carry-over of mRNAs from one cell cycle into another. Two redundant kinases, Dbf2p and Dbf20p, regulate this decay; Dbf20p preferentially stabilizes *SWI5* and *CLB2* mRNAs during S phase while Dbf2p preferentially stabilizes *SWI5* and *CLB2* mRNAs during G2 phase and mitosis. Therefore, in a single cell, two types of *SWI5* and *CLB2* mRNPs must coexist during this transition. Their diversity in composition in turn implies diversity in their function, which could aid in regulation of the cell cycle transition. This approach provides a platform for the analysis of genes that are responsible for regulating cell cycle mediated RNA turnover. Supported by NIH GM57071 to RHS.

THE ROLE OF NUCLEOSOME POSITIONING IN THE EVOLUTION OF GENE REGULATION

Alex M Tsankov[1,2], Dawn A Thompson[2], Amanda Socha[2], Aviv Regev[2,3], Oliver J Rando[4]

[1]MIT, Electrical Engineering and Computer Science, 77 Massachusetts Ave, Cambridge, MA, 02140, [2]Broad Institute, MIT and Harvard, 7 Cambridge Center, Cambridge, MA, 02142, [3]MIT, Biology, 77 Massachusetts Ave, Cambridge, MA, 02140, [4]UMass Medical, Biochemistry and Molecular Pharmacology, 55 Lake Avenue North, Worcester, MA, 01605

Chromatin organization plays a major role in gene regulation and can affect the function and evolution of new transcriptional programs. However, it can be difficult to decipher the basis of changes in chromatin organization and their functional effect on gene expression. Here, we present the first large-scale comparative genomic analysis of the relationship between chromatin organization and gene expression, by measuring mRNA abundance and nucleosome positions genome-wide in 12 Hemiascomycota yeast species. We found substantial conservation of global and functional chromatin organization in all species, including prominent nucleosome-free regions (NFRs) at gene promoters, and distinct chromatin architecture in growth and stress genes. Chromatin organization has also substantially diverged in both global quantitative features such as spacing between adjacent nucleosomes, and in functional groups of genes. Expression levels, intrinsic anti-nucleosomal sequences and trans-acting chromatin modifiers all play important, complementary, and evolvable roles in determining NFRs. We identify four mechanisms that couple chromatin organization to evolution of gene regulation and have contributed to the evolution of respiro-fermentation and other key systems, including (1) compensatory evolution of alternative modifiers associated with conserved chromatin organization; (2) a gradual transition from constitutive to trans-regulated NFRs; (3) a loss of intrinsic anti-nucleosomal sequences accompanying changes in chromatin organization and gene expression, and (4) re-positioning of motifs from NFRs to nucleosome-occluded regions. Our study sheds light on the molecular basis of chromatin organization, and on the role of chromatin organization in the evolution of gene regulation.

NOVEL INSIGHTS INTO THE GENE REGULATORY NETWORKS REGULATING HUMAN T HELPER (TH) CELL DIFFERENTIATION

Laura L. Elo[1,2,*], Henna Järvenpää[2,3,*], Soile Tuomela[2,3,*], Sunil Raghav[2,*], Helena Ahlfors[2,4], Kirsti Laurila[5], Bhawna Gupta[2], Riikka J. Lund[2,6], Johanna Tahvanainen[2,7], R. David Hawkins[8], Matej Orešic[9], Harri Lähdesmäki[5,10], Omid Rasool[2], Kanury V. Rao[11], Tero Aittokallio[1,2], Riitta Lahesmaa[2] [1]Biomathematics Research Group, Department of Mathematics, University of Turku, FI-20014 Turku, Finland. [2]Turku Centre for Biotechnology, University of Turku and Åbo Akademi, P.O. Box 123, FI-20521 Turku, Finland. [3]Turku Graduate School of Biomedical Sciences, Kiinamyllynkatu 13, FI-20520 Turku, Finland. [4]The National Graduate School in Informational and Structural Biology, Åbo Akademi University, FI-20520 Turku, Finland. [5]Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland. [6]Department of Biological Science, University of Sheffield, S10 2TN, Sheffield, UK. [7]Drug Discovery Graduate School, University of Turku, FI-20014 Turku, Finland. [8]Ludwig Institute for Cancer Research, University of California, San Diego, California 92037, USA.[9]VTT Technical Research Centre of Finland, P.O. Box 1000, FI-02044 Espoo, Finland. [10]Deparment of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland. [11]International Centre for Genetic Engineering and Biotechnology, P.O. Box 10504, 110067 New Delhi, India.
*Equal contribution

In response to signals from pathogen activated innate immune cells, naïve CD4+ T cells can differentiate into different T helper cell (Th) lineages including effector Th2 cells which are involved in many immunological diseases, such as asthma and allergy. We have used systematically the state-of-the art genome-scale measurement technologies together with effective computational methods to investigate the complicated networks and the molecular mechanisms involved in human Th2 differentiation.

Our study describes a whole spectrum of cellular regulation during the early stages of Interleukin-4 (IL-4) induced Th2 cell differentiation as a dynamic network that regulates and gives rise to the complex Th2 phenotype. Our data shows differential regulation of both known genes coding for components of certain pathways and novel genes with unknown functions. In addition, we have exploited siRNA-mediated knockdown experiments followed by genome-wide transcriptome analysis to reveal potential upstream key factors involved in the regulation of Th2 differentiation. The data provides a comprehensive starting point for generating novel hypotheses, to be experimentally tested, on the genes and pathways involved in the process.

ICLIP RNA MAPS ELUCIDATE TIA1 AND TIAL1 AS MASTER REGULATORS OF RNA SPLICING.

Zhen Wang[1], Melis Kayikci[1], Kathi Zarnack[2], Gregor Rot[3], Tomaž Curk[3], Jernej Ule[1]

[1]MRC Laboratory of Molecular Biology, Structural Studies, Hills Rd, Cambridge, CB2 0QH, United Kingdom, [2]EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom, [3]University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, Ljubljana, SI-1000, Slovenia

RNA splicing regulation involves interactions between the splice sites and snRNAs, and between pre-mRNA positions close to the splice sites and RNA-binding proteins. We used individual-nucleotide resolution UV-crosslinking and immunoprecipitation (iCLIP) to quantitatively profile the precise binding of TIA1 and TIAL1 RNA-binding proteins on RNAs. We show that the two proteins bind to intronic uridine-rich motifs downstream of 5' splice sites of both alternative and constitutive exons. iCLIP RNA maps demonstrated how TIA1 and TIAL1 regulate alternative exons, splice sites or introns in a highly position-dependent manner, such that a local TIA function as splicing enhancer can lead to either a positive or negative effect on exon inclusion. Finally, we find that TIA-regulated exons are overrepresented in genes with functions in RNA splicing, further indicating the importance of TIA1 and TIAL1 as master regulators of RNA splicing.

SELECTIVE DEREPRESSION POLYCOMB TARGETS DURING PANCREATIC ORGANOGENESIS CREATES A NEURAL PROGRAM IN INSULIN-PRODUCING B-CELLS.

Joris van Arensbergen[1,2], Xavi Garcia[1,2], Ignasi Moran[1], Miguel Maestro[1,2], Xiaobo Xu[3], Mark Van de Casteele[3], Anouchka L Skoudy[4], Matteo Palassini[5], Harry Heimberg[3], Jorge Ferrer[1,2,6]

[1]Genomic Programming of Beta Cells, IDIBAPS, Casanova, 143, Barcelona, 08036, Spain, [2]CIBER de Diabetes y Enfermedades Metabolicas, CIBERDEM, c. Mallorca, 183, Barcelona, 08036, Spain, [3]Diabetes Research Center, Vrije Universiteit Brussel, Laarbeeklaan, 103, Brussels, B1090, Belgium, [4] Cancer Research Program, Institut Municipal d'Investigació Mèdica, Dr Aiguader, 88, Barcelona, 08003, Spain, [5]Departament de Física Fonamental, Universitat de Barcelona, Diagonal 647, Barcelona, 08028, Spain, [6]Department of Endocrinology, Hospital Clínic de Barcelona, Casanova 143, Barcelona, 08036, Spain

The epigenome changes that underlie cellular differentiation in developing organisms are poorly understood. To gain insights into how pancreatic β-cells are programmed, we profiled key histone methylations in embryonic stem cells, multipotent progenitors of the nascent embryonic pancreas, purified β-cells, and 10 differentiated tissues. We report that despite their endodermal origin, β-cells show a transcriptional and active chromatin signature that is most similar to ectoderm-derived neural tissues. In contrast, the β-cell signature of trimethylated H3K27 (a mark of Polycomb-mediated repression) clusters with pancreatic progenitors, acinar cells and liver, consistent with the transmission of this mark from endoderm progenitors to their differentiated cellular progeny. In addition, we identified two H3K27 methylation events that arise in the β-cell lineage after the pancreatic progenitor stage. One is a wave of cell-selective de novo H3K27 trimethylation in non-CpG island genes. Another is the selective loss of H3K27me3-repressed chromatin in a core program of neural developmental regulators that enables a convergence of the gene activity state of β-cells with that of neural cells. The data reveals a dynamic regulation of Polycomb repression programs that shape the identity of differentiated β-cells.

MOST “DARK MATTER” TRANSCRIPTS ARE ASSOCIATED WITH KNOWN GENES

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J Blencowe[1,2], Timothy R Hughes[1,2]

[1]University of Toronto, Banting and Best Department of Medical Research, 160 College St., Toronto, M5S 3E1, Canada, [2]University of Toronto, Department of Molecular Genetics, 160 College St., Toronto, M5S 3E1, Canada

A series of reports over the last few years have indicated that a much larger portion of the mammalian genome is transcribed than can be accounted for by currently annotated genes, but the quantity and nature of these additional transcripts remains unclear. Here, we have used RNA-seq and tiling arrays to assess the quantity and composition of transcripts in PolyA+ RNA from human and mouse tissues. Relative to tiling arrays, RNA-seq identifies many fewer transcribed regions (“seqfrags”) outside known exons and ncRNAs. Most nonexonic seqfrags are in introns, raising the possibility that they are fragments of pre-mRNAs. The chromosomal locations of the majority of intergenic seqfrags in RNA-seq data are near known genes, consistent with termination read-through, promoter-associated transcripts, or new alternative exons; indeed, reads that bridge introns identified 4,544 new exons, affecting 3,554 genes. Most of the remaining seqfrags correspond to either single reads that display characteristics of random sampling from a low-level background, or several thousand small transcripts (median length = 111bp) present at higher levels, which also tend to display sequence conservation. We conclude that, while there are bona fide new intergenic transcripts, their number and abundance is low in comparison to known exons, and the extent of steady-state transcripts in these regions is far lower than previously reported in other studies.

CHROMATIN SIGNATURE OF EMBYONIC PLURIPOTENCY IS ESTABLISHED DURING ZYGOTIC GENOME ACTIVATION

Nadine L Vastenhouw[1], Yong Zhang[2], Ian G Woods[1], Farhad Imam[1], Aviv Regev[3], X. Shirley Liu[4], John Rinn[3,5], Alex Schier[1,3]

[1]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA, 02138, [2]School of Life Science and Technology, Tongji University, 1239 Siping Road, Shanghai, -, China, [3]Broad Institute, MIT and Harvard, 7 Cambridge Center, Cambridge, MA, 02142, [4] Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street, Boston, MA, 02115, [5]Department of Pathology, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA, 02215

The genome is inactive after fertilization until embryonic transcription is initiated during the maternal-zygotic transition (MZT). This universal process coincides with the formation of pluripotent cells, which in mammals can be used to generate embryonic stem (ES) cells. To study the changes in chromatin structure that accompany zygotic genome activation and pluripotency, we mapped the genomic locations of histone H3 modifications before and after MZT in zebrafish embryos. Repressive H3 lysine 27 trimethylation (H3K27me3) and activating H3 lysine 4 trimethylation (H3K4me3) are only detected after MZT. We find that more than 80% of genes are marked by H3K4me3. These genes include many developmental regulatory genes that are occupied by H3K4me3 and H3K27me3. Sequential chromatin immunoprecipitation demonstrates that both methylation marks occupy the same promoter regions, providing the first evidence that the bivalent chromatin domains found in cultured ES cells also exist in embryos. Interestingly, we also find a large group of genes that are monovalently marked by H3K4me3 but not H3K27me3. These H3K4me3 monovalent gene loci are neither transcribed nor stably bound by RNA polymerase II. Closer inspection of in vitro data sets reveals similar monovalent H3K4me3 domains in ES cells. Moreover, the analysis of an inducible transgene indicates that H3K4me3 domains can form in the absence of sequence-specific transcriptional activators or stable association with RNA pol II. These results suggest that both bivalent and monovalent domains poise embryonic genes for activation and that the chromatin profile associated with pluripotency is established during MZT.

GENOME-WIDE PROBING OF RNA STRUCTURE IN YEAST

Yue Wan[1], Michael Kertesz[2], Elad Mazor[2], John Rinn[1], Howard Y Chang[3], Eran Segal[2]

[1]Stanford, Epithelial Biology, 269 Campus Drive, Stanford, CA, 94305, [2]Weizmann Institute of Science, Computer Science and Applied Mathematics, Ziskind Building, Rehovot, 76100, Israel, [3]The Broad Institute, Pathology, 7 Cambridge Center, Cambridge, MA, 02142

RNA structures play important roles in the function of RNAs in the cell. Structure in the 5' and 3' UTRs directly affect RNA subcellular localization, stability, translation and ability to respond to intracellular stress. However, despite our understanding on a few RNAs, such as tRNAs, rRNAs, the solution structures of most mRNAs in the cell are unknown. Here we devised a method called PARS (Parallel Analysis of RNA Structure) that probes RNA structures in a genome-wide manner. PARS couple RNA footprinting, using single and double strand specific enzymes, with high throughput sequencing to obtain structural information of many RNAs simultaneously. This information provides insights to structural organization of mRNAs as well as how these structures may affect cellular biology.

GENOME-WIDE PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES USING AN INTEGRATED MODEL

Kyoung-Jae Won[1], Bing Ren[2], Wei Wang[1]

[1]University of California, San Diego, Chemistry and Biochemistry, 9500 Gilman Drive, La Jolla, CA, 92093-0359, [2]Ludwig Institute for Cancer Research and UCSD, Cellular and Molecular Medicine, 9500 Gilman Drive, La Jolla, CA, 92093-0653

Identification of target loci of transcription factors (TFs) in a specific tissue or at a specific developmental stage is crucial for understanding transcriptional regulation in eukaryotes. To date, genome-wide prediction of TF binding sites in mammalians has suffered from limited information available to find functional elements in a genome.

We present here an integrated approach, called Chromia, that integrates data from multiple sources including discrete sequence motif information and continuous chromatin modification data to tackle this task. Chromia is composed of a hidden Markov model (HMM) that captures the characteristic patterns of a TF binding motif occurrences and the histone modification signature associated with regulatory elements such as promoters and enhancers. We demonstrated its usefulness on genome-wide predictions of target loci of 13 TFs in the mouse embryonic stem (mES) cell. Using the independent ChIP-seq analyses of these TF bindings as the gold standard, we showed that the performance of our HMM model was significantly better than many other computational methods. In addition, we observed that target genes of a TF identified by Chromia had a larger overlap with the genes affected by knocking down the TF using RNAi than the target genes identified by the ChIP-seq binding peaks. This observation implies that Chromia is useful to locate functional TFBSs.

The encouraging results of this study suggest Chromia as a novel approach of predicting condition specific TF binding sites at a genomic scale. Because of its capability to integrate different types of information, when more epigenomic data become available, Chromia is expected to become more useful in identifying functional TF binding sites in a genome, which can guide further experimental investigation.

CO-EVOLUTION OF TRANSCRIPTION FACTOR AND ITS BINDING SITES

Shu Yang, Xinran Li, Hari K Yalamanchili, Junwen Wang

Department of Biochemistry, Faculty of Medicine, the University of Hong Kong, Laboratory Block, 21 Sassoon Rd., Hong Kong, 00000, China

The interaction between a transcription factor (TF) and its DNA binding sites (TFBS) is essential to gene regulation. Mutation in either TF or TFBS elements may weaken the interactions and cause genetic disorders in animals. During evolution, many TF families, such as Fox, Sox and Hox, have expanded to dozens of members through gene duplication. These members have protein sequences with various levels of sequence identity; similarly, their TFBS are different in various degrees. We used Pearson's correlation coefficient to measure the co-evolution between TFs and TFBS for the families with TFBS annotated in JASPAR database. The evolution of TFs was measured by protein sequence alignment, and the evolution of TFBS was measured by the similarities of Position Weight Matrices (PWMs). We found that for many TF families, significant co-evolutionary relationships exist between TFs and TFBS, either at whole protein sequence level, or protein domain level, depending on the contact surface of protein-DNA interaction. We further developed an algorithm to discover the co-evolved protein residues and DNA bases based on mutual information. In summary, our study discovered the co-evolution relationship between TF and TFBS. The method we developed can be used to indentify interacting residue-base pairs that are important to TF-TFBS interaction.

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

Ting Wang[1], Alika Maunakea[2], Raman Nagarajan[2], Steve Jones[3], Tracy Ballinger[4], David Haussler[4], Marco Marra[3], Martin Hirst[3], Shaun Fouse[2], Brett Johnson[2], Chibo Hong[2], Joseph Costello[2]
[1]Washington University, Genetics, 4444 Forest Park ave, St. Louis, MO, 63108, [2]UC San Francisco, Neurosurgery, 1450 3rd Street, San Francisco, CA, 94158, [3]BC Cancer Agency, Genome Sciences Centre, 675 W. 10th Avenue, Vancouver, V5Z 1L3, Canada, [4]UC Santa Cruz, Center for Biomolecular Science and Engineering, 1150 High Street, Santa Cruz, CA, 95064

We present two complementary next-generation sequencing-based approaches to detect methylated and unmethylated genomic DNA. The first, methyl DNA immunoprecipitation and sequencing (MeDIP-seq), uses antibody-based immunoprecipitation of 5-methylcytosine and sequencing to map the methylated fraction of the genome. In the second method, unmethylated CpG sites are identified at single CpG site resolution by sequencing size-selected fragments from parallel DNA digestions with the methyl-sensitive restriction enzymes (MREs) HpaII, Hin6I, and AciI (MRE-seq). Combining these two methods, we generated a genome-wide, high-resolution methylome map of human brain tissue, frontal cortex gray matter, and a second map of human embryonic stem cell H1. These maps on average interrogate close to 90% of all CpGs (25 million of 28 million total) and 98% of CpG islands in the human genome, at the modest expense of relatively a small amount of initial specimen and just a few lanes of Illumina flowcell.

We investigated the role of DNA methylation in gene bodies with these methylome maps. While the methylation of DNA in 5' promoters suppresses gene expression, the role of DNA methylation in gene bodies is unclear. From the dense, high-resolution coverage of CpG islands, the majority of methylated CpG islands were revealed to be in intragenic and intergenic regions, while less than 3% of CpG islands in 5' promoters were methylated. The CpG islands in all three locations overlapped with RNA markers of transcription initiation, and unmethylated CpG islands also overlapped significantly with trimethylation of H3K4, a histone modification enriched at active promoters. The general and CpG-island-specific patterns of methylation are conserved in mouse tissues.

An in-depth investigation of the human SHANK3 locus and its mouse homologue demonstrated that this tissue-specific DNA methylation does indeed regulate intragenic promoter activity in vitro and in vivo. These methylation-regulated, alternative transcripts are expressed in a tissue and cell type-specific manner, and are expressed differentially within a single cell type from distinct brain regions. These results support a major role for intragenic methylation in regulating cell context-specific alternative promoters in gene bodies.

PROMISCUOUS, STRESS-RESPONSIVE BINDING OF TRANSCRIPTION FACTORS TO YEAST NCRNA GENES AND NUCLEOSOME-DEPLETED REGIONS

Lucas D Ward[1], Junbai Wang[1], Harmen J Bussemaker[1,2]
[1]Columbia University, Biological Sciences, 1212 Amsterdam Ave., New York, NY, 10027, [2]Columbia University, Center for Computational Biology and Bioinformatics, 1130 St. Nicholas Ave., New York, NY, 10032

Recent genomic studies of chromatin in fly and mouse have revealed "hotspots" at which many DNA-binding proteins colocalize independently of their sequence specificities[1]. The most commonly-analyzed compendium of yeast transcription factor (TF) ChIP-chip data[2] was originally analyzed in a manner that normalized enrichment values across experiments, which would have eliminated the evidence for any analagous colocalization phenomenon in this organism. We have reanalyzed the raw data and discovered that many probes are indeed occupied by many TFs. These colocalization hotspots are associated with (a) nucleosome depletion and (b) proximity to genes encoding noncoding RNAs such as tRNAs and snoRNAs. Significantly, this phenomenon is generally seen only in rich media (YPD) conditions and is abrogated in most stress conditions.

We found that for many TFs, the occupancy at ncRNA genes was on the same order of magnitude as the occupancy at annotated targets. In rich media, a total of 47 TFs significantly occupied both ncRNA genes and their annotated targets, another 28 TFs significantly occupied ncRNA genes and not their annotated targets, and only three TFs significantly occupied their annotated targets but not ncRNA genes.

Taken together with the recent discovery that the genomically dispersed tRNA genes in yeast colocalize to the nucleolus, our results suggest that this organelle may act as a sink for transcription factors in yeast. Several proteins in higher eukaryotes, including p53[3], are regulated through nucleolar sequestration; however, this mechanism has never been observed for such a broad range of transcription factors.

This newly-discovered landscape of promiscuous binding dominates our re-analyzed, non-normalized data: the median occupancy by all other transcription factors explains 14% of the variance in any given rich-media ChIP-chip experiment, far more than the 1% typically explained by known *in vitro* sequence specificity. Our results underscore the importance of chromatin accessibility and nuclear organization in driving genomic protein occupancy, even in the relatively compact yeast genome.

1. Moorman, C., et al. (2006) Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster. PNAS* **103**(32).
2. Harbison, C.T., et al. (2004) Transcriptional regulatory code of a eukaryotic organism. *Nature* **431**.
3. Rubbi CP & Milner J (2003) Disruption of the nucleolus mediates stabilization of p53 in response to DNA damage and other stresses. *EMBO J* **22**(22).

"PFAM-WIDE" DETERMINATION AND INFERENCE OF TRANSCRIPTION FACTOR DNA SEQUENCE SPECIFICITIES

Matthew T Weirauch[1], Ally Yang[1], Anriy Vorobyob[1], Atina Cote[1], Ishminder K Mann[1], Timothy R Hughes[1,2]

[1]University of Toronto, Banting and Best Department of Medical Research, 160 College St. Room 1350, Toronto, M5S 3E1, Canada, [2]University of Toronto, Department of Molecular Genetics, 160 College St. Room 1350, Toronto, M5S 3E1, Canada

Knowledge of the sequence binding preferences of transcription factors (TFs) is central to the understanding of gene regulation and genome function. However, information is lacking regarding the sequence specificities of many TFs, even in well-studied organisms. Similar DNA binding domains (DBDs) often have similar sequence preferences, and in some cases rules have been derived for inferring sequence specificities within TF families.

Our aims are three-fold. First, we are using high-resolution DNA-binding data (e.g. from Protein Binding Microarrays) to refine and test rules for inference of TF sequence specificity. Second, we are generating the data needed to produce accurate "Pfam-wide" inferences of sequence specificity for as many eukaryotic DBD classes as possible. Third, we have constructed a database (Catalog of Inferred Sequence Binding Preferences, or Cis-BP) to house both known and inferred sequence preferences in the form of 8-mer binding scores, position weight matrices, and IUPAC consensus motifs.

Using our current dataset, we provide a summary of knowledge of eukaryotic TF binding preferences. We also show that, for multiple DBD classes, simple sequence rules (e.g. percent amino acid identity) can readily identify TFs with new and different DNA-binding activities. Using these sequence rules to guide selection of proteins for further analysis, we have experimentally obtained distinctive sequence specificities for DBDs from a wide variety of eukaryotes.

GEOMETRIC PERSPECTIVES ON UNCERTAINTY MINIMIZATION FOR FUZZY SPECTRAL CLUSTERING

Brian White[1], Daniel Korenblum[2], David Shalloway[1]

[1]Cornell University, Molecular Biology and Genetics, 267 Biotechnology Building, Ithaca, NY, 14850, [2]PAIRE, Palo Alto Institute for Research and Education, 3801 Miranda Avenue, Palo Alto, CA, 94304

Spectral clustering uses the global information embedded in eigenvectors of an inter-item similarity matrix to correctly identify clusters of irregular shape, an ability lacking in commonly used approaches such as k-means and agglomerative clustering. However, traditional spectral clustering partitions items into hard clusters, and the ability to instead generate fuzzy item assignments would be advantageous for the growing class of biological problems for which cluster overlap and uncertainty are important, including the separation of subpopulations by fluorescence activated cell sorting (FACS), automated biological database curation, and gene expression analysis. Korenblum and Shalloway [Phys. Rev. E 67, 056704 (2003)] extended spectral clustering to fuzzy clustering by introducing the principle of uncertainty minimization. However, this posed a challenging non-convex global optimization problem that they solved by a brute-force technique unlikely to scale to data sets having more than $O(10^2)$ items. Here we develop a new method for solving the minimization problem, which can handle data sets at least two orders of magnitude larger. In doing so, we elucidate the underlying structure of uncertainty minimization using multiple geometric representations. Uncertainty minimization can be applied to a wide variety of existing hard spectral clustering approaches, thus transforming them to fuzzy methods.

UNDERSTANDING TRANSCRIPTIONAL REGULATION DURING DEVELOPMENT USING A PROBABILISTIC MODEL

Bartek Wilczynski, Zhen Xuan Yeo, Eileen E Furlong

European Molecular Biology Laboratory, Genome Biology, Meyerhofstrasse 1, Heidelberg, 69117, Germany

Development of multi-cellular organisms depends on tight regulation of spatial and temporal patterns of gene expression. This process is mediated through a complex network of regulatory dependencies between transcription factors (TFs) and their respective targets. These networks are often described on the gene level, but in fact they depend on patterns of sequence specific binding of transcription factors to cis-regulatory modules (CRMs). Although there has been significant progress in predicting the activity of CRMs in a well defined developmental context [1], it remains a key challenge to translate individual CRM activity to the complex gene expression patterns that arise through the activity of multiple enhancers acting in the context of dynamic chromatin states. Mesoderm development in *Drosophila* is a perfect model system to tackle this complex question, as we can combine the genome-wide database of mesodermal cis-regulatory elements [1] with genome-wide data on dynamic chromatin state from ModENCODE project [2] and tissue-specific gene expression data from BDGP database [3].

We approach this problem by constructing a probabilistic model of tissue-specific transcription initiation including all the aforementioned components. Using the available experimental data we can identify optimal model parameters for different tissues related to mesoderm development, which can be then used to predict new genes expressed specifically in mesoderm and muscle tissues with high accuracy (AUC>80%). In addition, the Bayesian framework [4] underlying our model allows us to assess not only the probability of a given gene to be expressed in a given tissue, but also the relative importance of different datasets (e.g. chromatin modifications, TF binding) for prediction accuracy.

References:
1. Zinzen, R.P., et al., Combinatorial transcription factor binding predicts spatio temporal *cis*-regulatory activity. Nature, 2009. 462(7269): p. 65-70.
2. Celniker, S.E., et al., Unlocking the secrets of the genome. Nature, 2009. 459(7249): p. 927-930.
3. Tomancak, P., et al., Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biology, 2002. 3(12): p. research00881–8814.
4. Wilczynski, B. and N. Dojer, BNFinder: Exact and efficient method for learning Baesian networks. Bioinformatics, 2009, 25(2): p.286-287

EPIGENOMIC LANDSCAPE OF ERYTHROID MATURATION

Weisheng Wu*[1], Yong Cheng*[1], Swathi Kumar[1], Kuan-Bei Chen[1,2], Chris Morrissey[1], Cheryl K Capone[1], Yoichiro Shibata[3], Christine Dorman[1], Francesca Chiaromonte[1,4], Greg Crawford[3], Mitchell J Weiss[5], and Ross C Hardison[1]
[1]The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics of the Huck Institutes of Life Sciences, Curtin Road, University Park, PA, 16802, [2]The Pennsylvania State University, Department of Computer Science and Engineering, Curtin Road, University Park, PA, 16802, [3]Duke University, Institute for Genome Sciences & Policy, Science Drive, Durham, NC, 27708, [4] The Pennsylvania State University, Department of Statistics, Curtin Road, University Park, PA, 16802, [5]Children's Hospital of Philadelphia, Division of Hematology, Civic Center Blvd, Philadelphia, PA, 19104
* These authors have contributed equally.

The commitment and differentiation of multipotent cells result from induction of lineage-specific genes and repression of other genes that are not part of the lineage. This regulation of gene expression involves occupancy of *cis*-regulatory modules (CRMs) by transcription factors, recruitment of co-activators and co-repressors and modifications of the chromatin structure. However, a full picture of these epigenomic features and how they result in induction and repression of batteries of genes are not known. We explore these questions in a mouse cell line model of late erythroid maturation. G1E cells are derived from mouse ES cells with a knockout of the gene *Gata1*, which encodes a transcription factor required for erythroid maturation. We restore GATA1 in an estradiol-inducible manner by expressing a GATA1-ER hybrid protein in the G1E-ER4 subline. After activation of GATA1-ER, the cells progress from proerythroblast to mature erythroblasts, making abundant hemoglobin and changing morphology dramatically. We have measured comprehensively changes in gene expression during this GATA1-dependent maturation and concomitantly, genome-wide occupancy by the transcription factors GATA1, GATA2, TAL1, and CTCF, as well as chromatin accessibility (DNase hypersensitive sites) and histone modifications in the chromatin (activating marks H3K4me1 and H3K4me3 and the repressing Polycomb mark H3K27me3), using Illumina sequencing technology for ChIP-seq. The data are being analyzed to determine globally the changes of protein occupancy and histone modification levels after restoration of active GATA1. Further, we are seeking the features that can account for the expression patterns of the genes and their differential response to GATA1. Surprisingly, we find that only limited changes in histone modification status accompany the substantial changes in gene expression, and rather than altering chromatin structure upon restoration, GATA1 binds to genomic sites that have already activated histone modifications. The occupancy of CRMs by transcription factors does correlate in many respects with responses in gene expression. In particular, induced genes tend to have GATA1 binding relatively close to the transcription start site (within 10kb), many induced genes show co-occupancy with GATA1 and TAL1, and some repressed genes show a loss of TAL1 upon restoration of GATA1. These global data are a valuable resource for further studies of gene expression during differentiation and maturation, and analyses may lead to more comprehensive understanding of regulatory mechanisms.

GENOME-WIDE REDUCTION OF NFR SIZE AND SUPPRESSION OF CRYPTIC RNA TRANSCRIPTION BY AN ATP-DEPENDENT CHROMATIN REMODELING ENZYME

Adam N Yadon[1,2], Daniel Van De Mark[1], Ryan Basom[3], Jeffrey Delrow[3], Iestyn Whitehouse[1], Toshio Tsukiyama[1]

[1]Fred Hutchinson Cancer Research Center, Basic Sciences, 1100 Fairview Ave N, Seattle, WA, 98109, [2]University of Washington, Molecular and Cellular Biology, University of Washington Box 357275, Seattle, WA, 98195, [3]Fred Hutchinson Cancer Research Center, Genomics Resources, 1100 Fairview Ave N, Seattle, CA, 98109

The compaction of eukaryotic genomes into chromatin presents challenges to all DNA-dependent processes. Studies have revealed many cryptic antisense RNA transcripts initiate at the edges of nucleosome free regions (NFRs) at the 5'- and 3'-end of genes. This study investigated the role of the ATP-dependent chromatin remodeling enzyme Isw2 in the regulation of cryptic RNA transcription. We systematically annotated NFRs across the yeast genome from multiple nucleosome maps and discovered four distinct classes of NFRs located at the 5'-end of genes, the 3'-end of genes, within ORFs, and far from ORFs. Isw2 targets were statistically enriched at all classes of NFRs. Additionally, we found Isw2-dependent chromatin remodeling decreases the size of NFRs. To our knowledge, this is the first chromatin regulator required to decrease NFR size *in vivo*. We also find that Isw2 is globally required to suppress cryptic sense and antisense RNA transcription at its targets. Our results provide a unified model in which a major function of Isw2 is to restrict the size of NFRs and suppress cryptic RNA transcription. In addition, we show that suppression of cryptic RNA alleviates transcriptional interference of coding RNA at a large number of loci.

FROM TMPRSS2-ERG GENE FUSION TO PROSTATE CANCER

Jindan Yu[1,2], Jianjun Yu[2], Qi Cao[2], Ming Hu[2], Zhaohui S Qin[2], Arul M Chinnaiyan[2]

[1]Northwestern University, Medicine, 303 E. Superior St., Chicago, IL, 60540, [2]University of Michigan, Pathology, 1500 E. Medical Center Drive, Ann Arbor, MI, 48109

While chromosomal rearrangements fusing the androgen-regulated gene TMPRSS2 to the oncogenic ETS transcription factor ERG occur in approximately 50% of prostate cancers, how the fusion products regulate prostate cancer remains unclear. TMPRSS2-ERG gene fusions were thought to merely represent one of many downstream mutations emanating from androgen receptor (AR) signaling.

Using chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-Seq), we mapped the genomic landscape of AR, ERG and key histone modifications in prostate cancer cells and tissues. While AR was found to activate genetic programs involved in normal prostate differentiation, interestingly, we found that ERG disrupts AR signaling by inhibiting AR expression, binding to and inhibiting AR activity at gene-specific loci. Serial titration experiments using cells expressing different amount of ERG revealed gradual recruitment of ERG directly to AR target genes thus repressing their expression. In addition, ERG-mediated tumorigenesis are independent of AR expression, thus being functionally consistent with ERG repression of AR in prostate cancer. Bioinformatics analysis revealed an interconnected network of androgen signaling, ERG regulation and dedifferentiated epigenetic programs.

TMPRSS2-ERG, an early-onset lesion in prostate cancer, has pleotropic effects in guiding cells towards malignant pathways. Our results provide a working model in which TMPRSS2-ERG plays a critical role in cancer progression by disrupting lineage-specific differentiation of the prostate and potentiating cellular de-differentiation program. Instead of simply being a downstream mutation of AR signaling, TMPRSS2-ERG gene fusions may be fundmental to prostate cancer progression and the development of hormonal resistance.

SYSTEMATIC DISCOVERY OF CIS-REGULATORY ELEMENTS IN THE MOUSE GENOME

Feng Yue, Yin Shen, David McCleary, Lee Edsall, Bing Ren

Ludwig Institute for Cancer Research, UCSD, 9500 Gilman Dr., San Diego, CA, 92122

Temporal and tissue-specific gene expression in mammals depends on the cis-regulatory elements in the genome. These non-coding sequences can be divided into many classes depending on their regulatory functions. Among the better-characterized elements are promoters, enhancers, and insulators. With Chip-Seq technology, we identified these elements in 10 mouse tissues and cell types, including bone marrow, cerebellum, cortex, heart, kidney, liver, lung, embryonic stem cells and embryonic fibroblast cells (male and female). Overall, we predicted ~25,000 active promoters, ~100,000 potential enhancers and ~30,000 insulators in the mouse genome. We developed an entropy-based algorithm to identify tissue-specific elements and found that CTCF binding sites are highly invariant across multiple tissues/cell lines. The chromatin marks at promoters are also very similar, with ~20% of them being tissue specific. In contrast, enhancers are more dynamic and most of them are tissue-specific. These observations are in accordance with our previous reports for human regulatory elements. Next, we investigated how conserved these regulatory elements are and found that promoters are the most conserved, followed by insulators and enhancers. We also noticed that insulators contain a short and strong conserved region (~50 bp) around CTCF binding motifs, while the conserved regions for promoters and enhancers are broader. Finally, we designed a computational pipeline to search for the potential sequence motifs in the tissue specific enhancers, and found most of them are enriched with tissue-specific transcription factors motifs.

IDENTIFICATION OF LARGE-SCALE CHROMATIN DOMAINS FROM CHIP-SEQ DATA: A COARSE-GRAINING APPROACH

Chongzhi Zang, Weiqun Peng

The George Washington University, Department of Physics, 725 21st St. NW, RM 105, Washington, DC, 20052

Chromatin states are critical in regulating the gene expression and maintaining the cell identity. Many chromatin marks (e.g. histone modifications H3K27me3 and H3K9me3) are known to form large-scale domains that are functionally important in the genome. Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-Seq) technique has become widely used in profiling chromatin marks in the whole genome. Bioinformatic methods for systematically detecting large-scale domains have been lacking. We present a novel method that can identify enriched domains on a large range of length scales with proper statistics. This algorithm is based on a coarse-graining approach, which uses recursive block transformations to determine spatial clustering of enriched regions across multiple length scales. We confirm and optimize this method with synthetic data. We apply this method to identify H3K27me3 domains in several human cell types. Our results show that H3K27me3 form domains on different length scales, ranging from Kbps to Mbps. A significant number of domains are found to be bounded by CTCF binding sites, suggesting the validity and functional integrity of the identified domains. More generally, this method is applicable to the dissection of the high-order organization of chromosomes on multiple length scales.

INTEGRATIVE DISCOVERY AND ANALYSIS OF A GLOBAL SPLICING-REGULATORY NETWORK IN MOUSE BRAIN

Chaolin Zhang, Maria A Frias, Matteo Ruggiu, Taesun Eom, Christina Marney, Huidong Wang, Donny Licatalosi, Aldo Mele, John Fak, Robert B Darnell

Howard Hughes Medical Institute, The Rockefeller University, Laboratory of Molecular Neuro-Oncology, 1230 York Avenue, New York, NY, 10065

In mammals, alternative pre-mRNA splicing regulation through specific protein-RNA interactions is critical for diversifying gene expression products and their spatial and temporal distribution. However, identification and characterization of target transcripts and splicing-regulatory network for specific splicing factors in a global scale have been of limited success. Here we report a global, high-quality splicing-regulatory network of the neuron-specific splicing factor Nova in mouse brain. We performed deep genome-wide surveys of Nova-RNA interactions using high-throughput sequencing crosslinking immunoprecipitation (HITS-CLIP), and analyzed Nova-dependent splicing using splicing microarrays. We developed highly predictive statistical models for Nova binding sites and target transcripts by integrating CLIP and microarray data with motif, RNA structure, genomic, and evolutionary information using Hidden Markov Models (HMM) and Bayesian networks. As a result, we present a comprehensive Nova-regulated neuron-specific splicing–regulatory network, comprised of ~700 alternative splicing events from ~360 genes with both high specificity and sensitivity. Analysis of these targets not only confirmed our previous observation on the "RNA map" and coherent synaptic functions of Nova targets in a much larger scale, but also yielded several unsuspected insights into Nova function. First, this analysis predicted other neuron-specific splicing factors would be required for combinatorial regulation with Nova, and we were able to validate this prediction experimentally. Second, we found a strong coupling between splicing and post-translational regulatory networks that provide a mechanistic explanation how Nova contributes to signal transductions and other pathways important for synaptic functions. This study provides a global view of organization and functional impact of a splicing-regulatory network, as well as a roadmap for the use of bioinformatics to generate novel and validatable predictions about RNA-binding protein function.

FUNCTIONAL ANALYSIS OF THE C-TERMINAL BINDING PROTEIN (CTBP) IN *DROSOPHILA*

Yang Zhang, David Arnosti

Michigan State University, Department of Biochemistry and Molecular Biology, Wilson RD, East Lansing, MI, 48824

Transcriptional repression plays fundamental roles in development and disease. Gene repression is largely achieved by recruitment of transcriptional corepressor complexes by sequence-specific repressors to DNA and modification of chromatin structures. The C-terminal binding protein (CtBP) is a conserved corepressor in higher eukaryotes that functions in development and oncogenesis, associating with multiple transcription factors, and regulating multiple cellular processes. CtBP is homologous to NAD-dependent D-2 dehydrogenases and contains an in vitro dehydrogenase activity, but the function of such enzymatic activity in CtBP-mediated gene regulation is elusive. In fact, most cell-based assays suggest that this activity is dispensable for repression. In addition, CtBP has multiple isoforms and all vertebrate counterparts contain a C-terminal extension. Drosophila has two major CtBP splicing isoforms, termed CtBPL and CtBPS, due to the presence or absence of the C-terminal domain. Interestingly, however, the short isoform, CtBPS, is expressed throughout development and is the most abundant isoform in Drosophila, raising the question of what functions this C-terminal tail may possess.

To address these questions, we studied functions of the dehydrogenase activity and the C-terminal extension of CtBP in a developmental context. Various CtBP isoforms or mutants were introduced as genetic rescue constructs into a CtBP knockout background to test their abilities to rescue the lethality and restore development. Expression of only CtBPL or CtBPS was sufficient to rescue CtBP lethality, however CtBPL rescued flies exhibited significantly stronger wing phenotypes, suggesting that the C-terminal tail may play roles in wing development, but is not essential for other functions. Significantly, the CtBP catalytic mutant, lacking the dehydrogenase activity, rescued while the NAD(H) binding mutant did not. The catalytic mutant exhibited an even higher penetrance of wing phenotypes than did the CtBPL and CtBPS lines, revealing that the dehydrogenase activity is biologically relevant for CtBP function. The failure of NAD(H) mutant to rescue CtBP lethality demonstrates that binding of NAD(H) to CtBP is essential for its function. We propose that NAD(H) binding may affect the stability of CtBP and/or CtBP complexes and association with transcriptional factors. The dehydrogenase activity may be involved in certain functions of CtBP and regulate specific gene targets. We are performing genome-wide expression profiling analysis of CtBP catalytic mutant to test this idea and to identify potential dehydrogenase-specific targets.

ESTIMATING BINDING ENERGIES FROM PROTEIN BINDING MICROARRAY EXPERIMENTS

Yue Zhao, Gary D Stormo

Washington University, Department of Genetics, Campus Box 8510, St. Louis, MO, 63108

Protein Binding Microarray (PBM) is an important technology that provides high-throughput measurements of binding specificity for transcription factor (TF) proteins. Currently, PBM experiments are analyzed using non-parametric enrichment scores based on ranks of probe intensities.

We have performed quantitative analysis that directly models probe intensity as a function of probe sequence, using a modified version of our nonlinear regression method BEEML (Binding Energy Estimation by Maximum Likelihood). By using a biophysical model of TF-DNA interaction, BEEML is able to obtain quantitative and reproducible binding energy models from PBM data. When tested on independent replicate arrays, BEEML energy models outperform models of comparable complexity obtained by existing methods.

Analysis of existing PBM data shows that the specificities of most TFs are well described by a simple Position Weight Matrix (PWM). BEEML can also be used to fit more complicated models that include parameters for phenomena such as non-additive interactions as well as multiple modes of binding, where a single PWM is not adequate to describe TF specificity.

DETERMINANTS OF TRANSCRIPTION FACTOR BINDING AND REGULATION

Xu Zhou[1], Erin K O'Shea[1,2,3,4]
[1]Harvard University, Molecular and Cellular Biology, 52 Oxford, Cambridge, MA, 02138, [2]Harvard University, Howard Hughes Medical Institute, 52 Oxford, Cambridge, MA, 02138, [3]Harvard University, Faculty of Arts and Sciences Center for Systems Biology, 52 Oxford, Cambridge, MA, 02138, [4] Harvard University, Chemistry and Chemical Biology, 52 Oxford, Cambridge, MA, 02138

Gene expression programs play an important role in many biological processes. Regulation of these programs is carried out by transcription factors (TFs), whose binding typically requires recognition of a specific DNA binding motif. However, such sequence motifs by themselves are not sufficient to predict *in vivo* binding events. To complicate matters, binding of a TF is not sufficient for transcriptional regulation. Despite many studies on transcriptional mechanisms, it is unclear what determines TF binding and what determines whether this binding is functional. Answering these questions will help us to understand how specific gene transcriptional regulation is achieved. The phosphate responsive pathway (PHO) in *S. cerevisiae* has been widely used as a model system to study transcriptional activation. Its transcriptional activation is regulated by two TFs, Pho2 and Pho4. When cells are grown in Pi-limited media, the expression of a set of genes is specifically up-regulated, and appear to have diverse dependence on Pho2 and Pho4. Here we use the PHO pathway as a model system to explore the factors that determine TF binding and its regulation. We hypothesize that the DNA binding motif, the presence or absence of cofactors, and chromatin structure influence TF binding and subsequent gene expression. We first identified genome-wide *in vivo* binding sites for both Pho2 and Pho4 using ChIP-seq technique under both Pi-rich and Pi-limited conditions. In addition, we determined genome-wide nucleosome occupancy using the same sequencing technique. We found that Pho4 binding in Pi-limited conditions is influenced by both local nucleosome occupancy and a competitive binding factor Cbf1. High nucleosome occupancy prevents Pho4 from binding to its high affinity binding sites. At sites where nucleosomes are depleted, Cbf1 can block Pho4 binding through flanking sequence preference. In Pi-rich conditions, Pho4 is located primarily in the cytoplasm; Cbf1 occupies Pho4 binding sites to prevent spurious transcriptional activation by low levels of nuclear Pho4. Using expression epistasis analysis, we identified a set of genes that are regulated by both Pho2 and Pho4 in Pi-limited conditions, but they account for only a small portion of the genes bound by both factors. To explain these, we found that the binding events regulating transcription are correlated with the increase of both Pho2 and Pho4 occupancy after Pi limitation. The proximity between Pho2 and Pho4 binding events and the spatial organization of the Pho2 and Pho4 binding motif suggest that Pho4 transcriptional functionality is correlated with the cooperative binding between Pho2 and Pho4.

DIFFERENTIAL GENE EXPRESSION IN TISSUES WITH DIFFERENT TROPISM FOR FOOT-AND-MOUTH DISEASE VIRUS

James Zhu, Jonathan Arzt, Luis Rodriguez

FADRU, Plum Island Animal Disease Research Center, USDA-ARS, P.O. Box 848, Greenport, NY, 11944

Foot and mouth disease (FMD) is a highly contagious viral animal disease. Natural infection in cattle generally occurs via the respiratory route. After infection, virus replicates locally in primary infection sites such as the soft palate, pharynx, and lung. The virus then spreads throughout the body via viremia and causes vesicular lesions only at very defined targeted sites (i.e. mouth, tongue, and interdigital skin, etc.). The determinants and molecular mechanisms mediating this tissue tropism remain largely undefined. To understand the mechanisms, we used whole genome expression microarray analysis to identify genes differentially expressed between virus-targeted and non-targeted tissues. Thirteen genes were detected as differentially expressed at 0.05 of false discover rate. Interestingly, two of the differentially expressed genes (DEG) are involved in acetylcholine receptor signaling pathway and urokinase-type plasminogen activator system, which have been known to play important roles in cell-cell and cell-extracellular matrix attachment. Three DEG participate in the regulation of membrane trafficking via catalyzing the hydrolysis of phosphatidylcholine to produce phosphatidic acid and choline. Based on the known pathology of the virus and the functions of the DEG, the results of the microarray analysis suggest that these DEG probably play important roles in the pathogenesis of FMD. The identification of these genes provides unique opportunities to explore host determinants and mechanisms of FMDV pathogenesis.

THE REGULATORY PROGRAM OF EGF-INDUCED S-PHASE ENTRY

Yaara Zwang[1], Tal Shay[2], Yotam Drier[2], Eytan Domany[2], Yosef Yarden[1]

[1]Department of Biological Regulation, The Weizmann Institute of Science, Herzel, Rehovot, 76100, Israel, [2]Department of Physics of Complex Systems, The Weizmann Institute of Science, Herzel, Rehovot, 76100, Israel

Normal cells will pass through the G1-phase restriction point and enter the S-phase of the cell cycle only upon two short pulses of EGF exposure separated by an interval of a specific length. However, cells will not proliferate in the absence of a second pulse. Using gene expression microarrays, we profiled the global transcriptional program of the two pulses, and the inter-pulse interval. This revealed two interesting patterns of gene expression; one pattern, which includes the immediate-early-induced transcription factor EGR1, was characterized by stronger induction upon the second pulse compared to the first pulse. This was preceded by a stronger activation of the MAPK pathway upon the second pulse of EGF, as measured by reverse-phase phospho-protein arrays. Partial inhibition of MAPK activation during the second pulse suppressed the enhanced EGR1 induction, and abolished proliferation, suggesting that the stronger activation of MAPK and EGR1 during the second pulse is essential for cell proliferation. The second pattern of expression consisted of genes that were highly expressed at the end of the interval and were then suppressed only upon the second pulse of EGF. This group includes several anti-proliferative genes, such as CDKN1B, implying a mechanism which restrains proliferation unless a second pulse is provided at the correct timing. According to our model, the first pulse of EGF initiates a priming process. Once priming is completed, at the end of the interval, the second pulse of EGF licenses S-phase entry by suppressing anti-proliferative genes, while enhancing the activation of MAPK and EGR1. Conceivably, the uncovered bimodal regulation prevents proliferation on sporadic and transient exposures to growth factors, and ensures robust commitment to S-phase entry.

GEOMETRY, GENETICS, AND EVOLUTION

Eric D Siggia, Paul Francois

The Rockefeller University, Center for Studies in Physics and Biology, 1230 York Avenue, New York, NY, 10065

Darwin understood that elaborate organs such as the eye could evolve by a sequence of small steps if each step conferred increased fitness. We have applied this insight to developmental gene networks by doing in-silico evolution. A suitable fitness to evolve anterior-posterior (AP) patterning in short germ insects or vertebrates is the mutual information between position and concentration profiles of selector genes. When supplemented with the condition that patterning occurs during growth, gene networks evolve with properties similar to Hox genes, namely temporal colinearity (the temporal order of expression in a cell matches the AP order), and posterior prevalence (posterior Hox genes impose their fates on anterior genes). The networks we have evolved for this and other problems are interpretable by the geometric approach to differential equations which is a tool for visualizing the minimum number of parameters and variables necessary to realize a given space and time dependent process. We propose that networks amenable to incremental evolution are those that will be seen in Biology.

MAKING GLOBAL PREDICTIONS OF *CIS*-REGULATORY ACTIVITY

Eileen E Furlong

EMBL, Gene Expression Unit, Meyerhofstrate 1, Heidelberg, 69117, Germany

The precise regulation of gene expression is crucial for almost all biological processes. In development, spatio-temporal patterns of gene expression are controlled by extensive regulatory networks, where the activity of transcription factors converge on cis-regulatory modules (CRMs). Recent advances enable the genome wide identification of CRMs by assaying transcription factor (TF) binding via techniques such as ChIP-on-chip and ChIP-seq. Therefore, the location and even combinatorial occupancy of CRMs can be experimentally measured as specific stages of development, at high-resolution. A current major challenge is to interpret TF binding data in terms of the resulting spatio-temporal CRM activity. We have addressed this question in the context of Drosophila myogenesis by establishing a high-resolution TF binding atlas for 15 developmental conditions covering multiple stages of mesoderm specification. This binding atlas identified thousands of mesoderm and muscle CRMs and was used to train a machine-learning algorithm to predict CRM spatio-temporal activity in five exclusive expression patterns. In vivo transgenic reporter assays demonstrates the high accuracy of this approach, with >77% of the CRMs spatio-temporal activity matching their predicted expression domains. This data-driven approach is based solely on in vivo TF occupancy and in vivo enhancer activity and therefore should be widely applicable.

ELUCIDATING THE INTRINSIC SEQUENCE SPECIFICITY OF DNASE I USING HIGH-THROUGHPUT SEQUENCING

Allan Lazarovici[1], Richard Sandstrom[2], Peter J Sabo[2], Todd R Riley[3], John Stamatoyannopoulos[2], Harmen J Bussemaker[3]

[1]Columbia University, Electrical Engineering, 1212 Amsterdam Ave, New York, NY, 10027, [2]University of Washington, Genome Sciences and Medicine, 1705 NE Pacific St, Seattle, WA, 98195, [3]Columbia University, Biological Sciences, 1212 Amsterdam Ave, New York, NY, 10027

The enzyme DNaseI is widely used to probe interactions between proteins and DNA both in vitro and in vivo. DNaseI is widely believed to lack significant sequence specificity; however, the intrinsic sequence specificity of DNAse has never been accurately characterized. To address this, we coupled next-generation sequencing with a maximum-likelihood framework based on the Poisson distribution to model the cleavage specificity of DNaseI on purified yeast and human genomic DNA as a function of local sequence context. At least three base pairs up- and downstream of the cleavage site contribute to the cleavage rate, and the unprecedented depth of information provided by high-throughput sequencing allowed us accurately and comprehensively to model interactions between base positions within this range. We find that the rate at which DNA is cleaved is strand-specific and varies by more than two orders of magnitude between different sequence combinations. Our analysis also reveals a marked dependency between the first and second nucleotide positions downstream of the cleavage site, which are likely related to the local geometry of the DNA minor groove. Finally, comparison of the predicted specificities of DNaseI inferred independently from yeast and human genomic DNA cleavage patterns exposed species-specific differences in cleavage rates that were isolated to positions flanking CpG dinucleotides, indicating that the enzymatic efficiency of DNaseI may be significantly modulated by DNA methylation status.

WHAT SHAPES THE LANDSCAPE OF TRANSCRIPTION FACTOR BINDING DURING EARLY DROSOPHILA DEVELOPMENT?

Tommy Kaplan[1], Xiao-Yong Li[2], Peter J Sabo[3], Sean Thomas[3], John A Stamatoyannopoulos[3], Mark D Biggin[4], Michael B Eisen[1,2,4]

[1]Department of Molecular and Cell Biology, California Institute of Quantitative Biosciences, University of California, 387 Stanley Hall #3220, Berkeley, CA, 94720, [2]Howard Hughes Medical Institute, University of California, 387 Stanley Hall #3220, Berkeley, CA, 94720, [3]Department of Genome Sciences, University of Washington, 1705 NE Pacific St, Seatle, WA, 98195, [4] Genomics Division, Lawrence Berkeley National Laboratory, Cyclotron Road MS 84-181, Berkeley, CA, 94720

Highly specific spatial patterns of gene expression are formed in early development by differential binding of transcription factors to regulatory regions in DNA. We now have tools to characterize the DNA affinities of these proteins and to precisely measure their genome-wide distribution *in vivo*. But our understanding of the forces that control where and when factors bind remains primitive.

Here we describe a thermodynamic model of transcription factor binding that incorporates DNA sequence, 3D protein concentrations at single-nucleus resolution, competitive binding among transcription factors and nucleosomes, and cooperativity between factors, to predict binding of five regulators of anterior-posterior patterning in early stages of *Drosophila melanogaster* development.

After training, our predictions achieve a significant correlation of ~0.45 with ChIP-seq measurements on a set of held-our test loci. Yet, qualitatively, there are many false positive predictions that highlight the limitations of our current understanding of transcription factor binding and its relation to gene expression.

To improve our predictions, we used DNaseI-hypersensitivity assays to measure the genomic accessibility of DNA, and incorporated it into our model as a non-uniform positional prior on protein binding. This boosted our predictions by over 50% to a correlation of 0.7, enabling us to infer the 3D genomic landscape of in vivo binding across the embryo with significant precision.

Finally, we used our method to quantify the regulatory roles of DNA sequence, accessibility, binding competition, and binding cooperativity. Our results suggest that binding is controlled almost exclusively by sequence and accessibility, with only a minimal role for interaction with nucleosomes or other regulators.

INTEGRATING AND TESTING CHIP-SEQ AND RNA-SEQ DATA

Ali Mortazavi[1], Shirley Pepke[1], Georgi Marinov[1], Brian Williams[1], Anthony Kirilusha[1], Katherine Fisher[1], Gilberto DeSalvo[1], Richard M Myers[2], Barbara Wold[1]
[1]California Institute of Technology, Dept. of Biology, Pasadena, CA, 91125, [2]HudsonAlpha Institute for Biotechnology, R&D, Huntsville, AL, 35806

High resolution genomewide maps of transcription factor and cofactor occupancy, chromatin marks, polymerase occupancy and transcriptome structure have been made to measure the physical networks in multiple cell types and states. We are focusing on extracting information from these data and testing elements for function. To begin to identify relationships and build specific hypotheses from high dimensional data matrices, we have been investigating self-organizing maps (SOMs). SOMs are an unsupervised machine learning-method for clustering and visualizing high-dimensional data in a two dimensional map. A useful property of SOMs for this application is that additional datasets can be readily mapped onto a trained SOM to identify further relationships. We are using large, fine-grained self-organizing maps constructed from ChIP-seq datasets to cluster the genome into thousands of coherent units based on, for example, chromatin marks.

We then identify clusters of units that concentrate promoters, enhancers and other potential functional groups that differ - or remain constant - from one cell type or state to another by using transcription factor data, polymerase occupancy, binding motif and RNA output data. We think that mining such SOMs, combined with perturbation experiments and functional tests, suggests a path forward for probing genome-scale network structure and function

GLOBAL ANALYSIS OF RNA PROCESSING IN MOUSE MODELS OF MYOTONIC DYSTROPHY

Christopher B Burge[1], Eric T Wang[1], Sonali Jog[2], Gary P Schroth[3], Sita Reddy[2], Thomas Cooper[4]

[1]MIT, Biology, 31 Ames St., 68-271, Cambridge, MA, 02142, [2]University of Southern Califor, Biochemistry & Molecular Biology, CSC 240, 9075, Los Angeles, CA, 90089, [3]Illumina, Inc., Research & Development, 25861 Industrial Blvd, Hayward, CA, 94545, [4] Baylor College of Medicine, Pathology, 1 Baylor Plaza, Houston, TX, 77030

Myotonic dystrophy (DM) is the most common form of adult-onset muscular dystrophy, affecting about 1 in 8,000 adults, and is characterized by muscle wasting, cataracts, heart conduction defects, myotonia (difficulty relaxing muscles), and cognitive problems. The most common form of the disease, DM1, is caused by an expanded CTG repeat (typically 50 to 2000 repeats) in the 3' UTR of the DMPK gene. CUG repeat RNAs from this gene fold into hairpins that accumulate in nuclear foci, resulting in effective depletion of the alternative splicing factor Muscleblind (MBNL1) and hyperactivation of the splicing factor CUG Binding Protein 1 (CUGBP1). Mis-regulation of splicing by these factors is central in the disease, as supported by several mouse models. To obtain a global picture of DM-associated changes in RNA processing, we have conducted RNA-Seq analysis of heart, skeletal muscle and brain from MBNL1 knockout and wild-type mice (5 of each). These analyses have been complemented by UV cross-linking/immunoprecipitation/sequencing (CLIP-Seq) analysis of the binding targets of MBNL1 in mouse tissue, and by RNA-Seq analyses of systems in which CUGBP1 levels have been manipulated. We are working to paint a comprehensive picture of the roles of these factors in regulation of pre-mRNA processing and expression, and of their contributions to pathogenesis in DM.

TRANSCRIPTIONAL SILENCING BY MICRO-RNAS THAT TARGET GENE PROMOTERS

Scott T Younger, David R Corey

UT Southwestern Medical Center, Departments of Pharmacology and Biochemistry, 6001 Forest Park, Dallas, TX, 75390

Our laboratory has demonstrated that small duplex RNAs targeted to gene promoters or regions downstream of gene termini can silence or activate gene transcription in mammalian cells. The potency and robustness of transcriptional modulation suggests that endogenous small RNAs will also target non-coding regions of the genome. We have developed a computational method for identifying potential microRNA (miRNA) targets and have found that regions upstream and downstream of annotated genes are enriched for potential miRNA recognition sites. We designed 10 miRNA mimics predicted to target sequences upstream of the human Progesterone Receptor (PR) gene and identified 4 mimics that inhibit PR transcription. Transcriptional silencing is associated with recruitment of Argonaute 2 (AGO2) to a non-coding RNA (ncRNA) transcript that overlaps the PR gene promoter. This recognition results in a reduced occupancy of RNA polymerase II on the PR promoter that is transient and independent of DNA methylation. In addition, we designed 5 miRNA mimics predicted to target sequences downstream of the PR gene terminus and identified a mimic that inhibits PR transcription. Our findings implicate a direct role for miRNAs in the regulation of gene transcription.

MICRORNAS AND THEIR REGULATORY TARGETS

David Bartel

HHMI, MIT, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA, 02142-1479

MicroRNAs (miRNAs) are endogenous ~23 nt RNAs that can play important gene-regulatory roles by pairing to the mRNAs of protein-coding genes to direct their repression. Metazoan miRNAs dampen the expression of many mRNA targets to optimize expression, reinforce cell identity, and sharpen developmental transitions. This talk will describe recent insights regarding miRNAs and their regulatory targets. Potential topics include 1) computational results showing that most human mRNAs are conserved targets of microRNAs, 2) the identification of new types of regulatory sites, with experimental demonstration of their efficacy, and 3) proteomic and ribosome-profiling results showing the molecular consequences of miRNA-directed repression, including the relative contributions of translational inhibition and mRNA destabilization.

EUKARYOTIC TRANSCRIPTOMES: COMPLEX, MULTIFUNCTIONAL, COMPARTMENTALIZED AND ELEGANT

Thomas R Gingeras

Cold Spring Harbor Laboratory, Gingeras Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY, 11724

Deep sequencing and tiling array analyses of transcriptomes obtained from various cells and tissues of multiple organisms reveal that information stored in DNA sequences is complex, layered and compartmentalized. It is complex based the number of types of novel RNA continuing to be found that serve specified functions. The information is also layered because within a given length of genomic sequence multiple RNAs can be produced having distinct functional roles. Finally, it is compartmentalized because cells shuttle specific RNAs to designated compartments within cells after they are made. This compartmentalization provides an added level of control of expression and function. Three examples of RNAs that illustrate these characteristics of complexity and multifunctional layering will be described. The first example is the short RNAs found mapping at the 5' ends of genes. Based on length, 5' end modifications, distribution around the transcriptional start sites and standedness, at least 3 types of short RNAs have been identified. One of these types called Promoter Associated Short RNAs, (PASRs) exhibit differences between human and flies. The second example is chimaeric RNAs that utilize exonic sequences of genes separated by large distances across the genome (at times mapping on different chromosomes) to form transcripts. These chimaeric RNAs can serve to increase the protein coding potential of a genome and to provide a record of co-transcribed genes. Multiple characteristics of these RNAs underscore their biological importance. One of these characteristics is the network of interconnectivity of genes involved in contributing to the formation of each of the chimaeric RNAs. The third example involves RNAs enriched within various sub-compartments within the nucleus of cells. These three examples highlight not only the complexity observable in transcriptomes but illustrate the elegance of non-co-linear informational organization found at the RNA level compared to that found in DNA.

PROTEOME-WIDE SEARCH FOR NOVEL RBP-RNA INTERACTIONS IN S.CEREVISIAE USING PROTEIN MICROARRAYS

Nikoleta G Tsvetanova, Daniel M Klass, Patrick O Brown

Stanford University, Biochemistry, 259 Campus Drive, Stanford, CA, 94305

As a step toward developing a comprehensive map of the system that regulates the events in the lives of each mRNA in yeast, we have developed a method for comparative binding of in vitro synthesized mRNAs to a protein microarray. We printed microarrays with >4,700 proteins (representing >80% of the saccharomyces cerevisiae proteome). This approach successfully identified known specific RNA-protein interactions (eg. of Ash1 mRNA with Khd1 and She2) and picked out annotated RBPs with high confidence (p-value = 1.43x10¬-10 by Wilcoxon test). Using the protein array data and a validation rate for candidate RBPs from follow-up in vivo IP-microarray experiments, we estimate the total number of RBPs in S. cerevisiae to be at least 890, 200 of which are not currently annotated as RNA-binding. Therefore, the new method provides a useful unbiased global approach to look for proteins interacting with total as well as specific mRNAs in vitro. Using these protein microarrays to identify proteins that bound specifically to an individual mRNA uncovered evidence for an unexpected role for the yeast Rab homologs (the Ypt proteins) in post-transcriptional control. We are currently focusing on characterizing the putative Ypt1-mRNA interactions in greater detail and attempting to identify regulatory elements in RNA that enable interactions with Ypt1.

SYSTEMATIC IDENTIFICATION OF RNA-BINDING PROTEINS IN YEAST PROPOSES DUAL FUNCTIONS FOR ENZYMES

Tanja Scherrer[1], Nitish Mittal[2], Sarath Chandra Janga[2], André P Gerber[1]

[1]ETH Zurich, Chemistry and Applied Biosciences, Wolfgang-Pauli-Strasse 10, Zurich, 8093, Switzerland, [2]MRC, laboratory of Molecular Biology, Hills Road, Cambridge, CB2 0QH, United Kingdom

RNA-binding proteins (RBPs) control diverse aspects of post-transcriptional gene regulation. In the budding yeast, Saccharomyces cerevisiae, almost 600 proteins are predicted to act as RBPs, many of them bearing one or several characteristic RNA-binding motifs. To identify "novel" and unconventional RBPs, we have probed high-density protein microarrays with fluorescently labeled RNA, and identified almost 200 proteins that reproducibly interacted with different types of RNA. Surprisingly, more than 40% of these proteins represent previously known enzymes, one third of them acting in metabolism and hence, may have dual functions. Furthermore, several “novel” protein domains were overrepresented among the RBPs, including ubiquitin, PC_rep, WW and TPR1 domains adding potential to represent novel RNA-binding motifs. We mapped the RNA targets for 13 ”novel” RBPs, which are expressed to different levels in cells, by the affinity-purification of tagged proteins and DNA microarray analysis of associated RNAs. Particularly highly expressed "novel" RBPs were associated with distinct groups of mRNAs that code for functionally related proteins, which is reminiscent to "post-transcriptional operons"; and three of them bound to their own mRNA adding potential for auto-regulation. We further confirmed RNA-binding capacity for Map1p, a methionine aminopeptidase, by demonstrating that the transcript levels of Map1 mRNA targets were selectively decreased after short MAP1 overexpression. Based on these results, we propose a dense post-transcriptional regulatory network with roles for enzymes that potentially control mRNAs coding for proteins acting in related pathways, or establish auto-regulatory feedback loops.

CHROMATIN ASSOCIATED LARGE INTERGENIC NON-CODING RNAS (LINCRNAS) IN CANCER AND STEM CELLS

Maite Huarte[1,2,3], Mitchell Guttman[1], Manuel Garber[1], Ahmad Khalil[1,2,3], Aviv Regev[1], Eric S Lander[1], John L Rinn[1,2,3]

[1]Broad Institute, Genome Biology, 7 Cambridge Center, Cambridge, MA, 02142, [2]Beth Israel Deaconess Medical Center, Pathology, 3 Blackfan Circle, Boston, MA, 02215, [3]Harvard Medical School, Pathology, 3 Blackfan Circle, Boston, MA, 02215

One of the most fundamental and unsolved problems in biology is: how does the same genome present in every cell encode a multitude of different cellular states? It is well established that epigenetic regulation plays a key role in this process, yet the array of these epigenetic landscapes are established by ubiquitously expressed chromatin remodeling complexes. It has long been suspected that non-coding RNA molecules to bring these complexes to their sites of action. Indeed, we and others have recently discovered a few examples of large non-coding RNA molecules that 'guide' chromatin formation. Interestingly, all three of these known examples (XIST, HOTAIR, KCNQ1OT1 and AIR) confer distinctive epigenetic states, yet share a common mechanism: they physically associate with chromatin remodeling complexes and impart specificity at distinct genomic loci. Here we show that large ncRNAs may be a general mechanism for the establishment and maintenance of epigenetic states in development and disease.

We recently discovered a new class of highly conserved large intergenic non-coding RNAs (lincRNAs) and a method to identify their functions. This "guilt by association method" pointed to a clear association of lincRNAs with chromatin remodeling complexes, particularly in the context of cancer and embryonic stem cells. Here, we present a systematic and comprehensive approach that demonstrates a majority of lincRNAs associate with various chromatin-remodeling complexes and regulate specific genomic sites. Moreover, we have profiled these lincRNAs across multiple cancer types and discovered numerous lincRNAs that are both misregulated in cancer and bound to chromatin remodeling complexes. For example, we show that p53 directly and temporally induces several chromatin-associated lincRNAs in response to DNA damage. Remarkably, these lincRNAs serve to regulate many key genes in the p53 pathway. Together, these results point to a general mechanism of lincRNA mediated regulation in key cancer processes, via the guidance of chromatin remodeling complexes.

HIGH-THROUGHPUT EXPERIMENTAL IDENTIFICATION OF TISSUE/CELL-TYPE-SPECIFIC CIS REGULATORY MODULES IN DROSOPHILA

Stephen S Gisselbrecht[1], Yongsok Kim[4], Anastasia Vedenko[1], Brian W Busser[4], Anton Aboukhalil[1,5], Xianmin Zhu[4], Alan M Michelson[4], Martha L Bulyk[1,2,3]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, HMS NRB 466d, 77 Ave. Louis Pasteur, Boston, MA, 02115, [2]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 77 Ave. Louis Pasteur, Boston, MA, 02115, [3]Harvard-MIT Division of Health Sciences & Technology (HST), Harvard Medical School, 77 Ave. Louis Pasteur, Boston, MA, 02115, [4] Laboratory of Developmental Systems Biology, Division of Intramural Research, National Heart, Lung and Blood Institute, National Institutes of Health, 10 Center Dr, Bethesda, MD, 20892, [5]Department of Aeronautics and Astronautics, MIT, 77 Massachusetts Ave., Cambridge, MA, 02139

In metazoans, gene expression is regulated dynamically in a tissue/cell-type specific manner predominantly via stretches of noncoding sequence referred to as cis regulatory modules (CRMs). CRMs usually contain 1 or more DNA binding sites for 1 or more sequence-specific, regulatory transcription factors (TFs) that function to activate or repress the target gene(s); CRMs that activate gene expression are frequently referred to as "transcriptional enhancers", and have been the focus of many computational and experimental studies. Large-scale, experimental identification of tissue/cell-type-specific enhancers in metazoans remains a significant challenge. Many algorithms have been proposed for prediction of enhancers from sequence and/or TF in vivo genomic occupancy data. However, despite recent technological advances, a major, rate-limiting bottleneck that is impeding rapid progress in the field is the still quite low-throughput experimental testing of computationally predicted candidate enhancers.

We have developed a novel technology we call 'enhancer-FACS-Seq' for high-throughput experimental identification of tissue/cell-type-specific transcriptional enhancers. In this project we are focusing on the developing embryonic mesoderm in Drosophila melanogaster as a model system. We are using enhancer-FACS-Seq to identify transcriptional enhancers active in particular subsets of mesodermal cells, including somatic mesoderm (SM) founder cells (FCs) and fusion competent myoblasts (FCMs), and cardial cells (CCs) and pericardial cells (PCs) of the embryonic heart. We are combining the experimental data on the tested enhancers with computational sequence analysis in order to identify cis regulatory codes that control gene expression in a tissue- and cell-type-specific manner in these cell types.

TRANSCRIPTIONAL LEGO - TUNING EXPRESSION LEVELS IN A PREDICTABLE MANNER BY MANIPULATING PROMOTER BUILDING BLOCKS

Michal Levo*[1,2], Tali Raveh-Sadka*[1,2], Uri Shabi[1,3], Leeat Yankielowicz-Keren[1,2], Maya Lotan-Pompan[1,2], Danny Zeevi[1,2], Adina Weinberger[1,2], Eran Segal[1,2]
[1]Weizmann Institute of Science, Department of Computer Science and Applied Mathematics, Herzl Street, Rehovot, 76100, Israel, [2]Weizmann Institute of Science, Department of Molecular Cell Biology, Herzl Street, Rehovot, 76100, Israel, [3]Weizmann Institute of Science, Department of Biological Chemistry, Herzl Street, Rehovot, 76100, Israel

The ability to control the timing and levels at which genes are expressed is key to most biological processes. Although we largely understand the sequence preferences of key players such as transcription factors (TFs) and nucleosomes, we are still far from understanding how these elements combine within regulatory sequences to encode the transcriptional outcome. For the past several years we have developed an experimental system in yeast, targeted at deciphering the mapping between DNA sequence and expression. In our system, different promoters are integrated into the same genomic location and in front of the same fluorescent reporter, and the reporter activity is measured accurately and robustly in living cells. Thus, any differences in the measured reporter expression are attributable only to differences in the input promoter sequences, making our system ideal for pinpointing the sequence elements responsible for shaping the measured expression and characterizing the nature of their effect. Here, we present unpublished results employing our system to explore the effect of various combinations of promoter elements, mediating both TF and nucleosome binding, on the resulting expression. Based on current knowledge and our theoretical analyses, we devised hypotheses regarding the effect that different nucleosome disfavoring sequences that vary in length, composition, and distance from TF sites, will have on transcription. To systematically test these hypotheses, we then designed and synthesized ~80 promoter sequences, resulting in the largest library of designed promoter variants to date. Our results show that by manipulating either or both TF sites and nucleosome disfavoring sequences, we can tune expression levels in a predictable manner. Importantly, sequence manipulations that only alter nucleosome disfavoring sequences result in changes to expression comparable in magnitude to those that result from manipulations to TF sites. In fact, compared to TF site changes, alterations of nucleosome disfavoring sequences likely yield more gradual changes in expression levels, and thus offer means to fine-tune gene expression with high resolution. These results have intriguing implications for gene expression evolution, suggesting that sequence changes that alter the DNA-encoded nucleosome organization may provide an efficient genetic mechanism by which genomes may evolve and fine-tune expression. Overall, our results bring us closer towards understanding the role of various promoter elements and their combined effects on transcription, and suggest that directed design of promoter sequences that yield pre-specified expression patterns may be within reach.

WIDESPREAD TRANSCRIPTION AT THOUSANDS OF ENHANCERS DURING ACTIVITY-DEPENDENT GENE EXPRESSION IN NEURONS

Tae-Kyung Kim[1], Martin Hemberg[2], Jesse M Gray[1], David A Harmin[1,3], Scott Kuersten[4], Allen M Costa[1], Kellie Barbara-Haley[5], Eirene Markenscoff-Papadimitriou [6], Gabriel Kreiman[2], Michael E Greenberg[1]

[1]Harvard Medical School, Neurobiology, 300 Longwood Ave, Boston, MA, 02215, [2]Children's Hospital, Ophthalmology, 300 Longwood Ave, Boston, MA, 02215, [3]Children's Hospital, Informatics Program, Harvard-MIT Division of Health Sciences and Technology, Boston, MA, 02215, [4] Life Technologies, 2130 Woodward Street, Austin, TX, 78744, [5]Children's Hospital, Molecular Genetics Core Facility, 300 Longwood Ave, Boston, MA, 02215, [6]University of California, San Francisco, Graduate Program in Neuroscience, Parnassus Ave, San Francisco, CA, 94122

During development and in mature organisms cells respond to changes in their environment in part through changes in gene expression. Extracellular factors including growth factors, hormones, and neurotransmitters activate programs of new gene expression in a manner that is temporally and spatially controlled by the coordinated action of trans-acting transcription factors (TFs) that bind to cis-acting DNA regulatory elements including enhancers, insulators, and promoters. Most studies of the mechanisms by which gene expression is induced in response to extracellular stimuli have focused on promoters, which lie adjacent to the site at which mRNA synthesis is initiated. In contrast, the mechanisms by which enhancers, which lie far away from the start site of mRNA synthesis, contribute to stimulus-dependent gene expression are not well characterized. We used genome-wide sequencing methods to study stimulus-dependent enhancer function in neurons. We identified over 20,000 neuronal activity-regulated enhancers in neurons that are bound by the general transcriptional co-activator CBP in an activity-dependent manner. A function of CBP at enhancers may be to recruit RNA polymerase II (RNAPII), as we also observed activity-regulated RNAPII binding to thousands of enhancers. Remarkably, RNAPII at enhancers transcribes bi-directionally a novel class of non-polyadenylated enhancer RNAs (eRNAs) within enhancer domains defined by the presence of histone H3 that is mono-methylated at lysine 4 (H3K4me1). The level of eRNA expression at neuronal enhancers positively correlates with the level of mRNA synthesis at nearby genes, suggesting that eRNA synthesis
occurs specifically at enhancers that are actively engaged in promoting mRNA synthesis. These findings reveal that a widespread mechanism of enhancer activation involves RNAPII binding and eRNA synthesis.

STUDYING TRANSCRIPTION DYNAMICS IN YEAST – A SINGLE MOLECULE APPROACH

Daniel Zenklusen[1,2], Daniel R Larson[2], Saumil Gandhi[2], Sami Hocine[2], Robert H Singer[2]

[1]Université de Montréal, Department of Biochemistry, Pavillon Roger-Gaudry, Montréal, H3C 3J7, Canada, [2]Albert Einstein College of Medicine, Department of Anatomy & Structural Biology, 1300 Morris Park Avenue, Bronx, NY, 10461

The development of imaging tools to study transcription at the single cell level has led to a review of basic concepts of transcription regulation and revealed surprising new insights about the processes by which cells achieve regulated mRNA expression. Most of the approaches to study gene expression kinetics in single cells use GFP reporters or depend on the use of gene arrays to create a sufficient signal that permits real time kinetic measurements. We have established single molecule approaches in the yeast S. cerevisiae that permit the study of transcription kinetics of endogenous genes in fixed and in living cells. Yeast presents an ideal system for analyzing transcription dynamics, as the small size and low copy numbers for a large fraction of the transcriptome facilitates acquisition of 'whole cell' expression data by counting mRNA molecules within cells.

Using a fluorescence in situ hybridization approach that allows the detection of single mRNA molecules, we measured mRNA abundance and transcriptional activity for endogenous genes without the need for genetic modifications. Combining single molecule counting with mathematical modeling enabled us to reveal different modes of expression in yeast, showing that most constitutively active genes are expressed by Poisson distributed initiation events, whereas others are transcribed by transcription bursts. Expanding this approach to determine expression levels of multiple genes within a single cell, furthermore revealed a surprisingly low level of expression coordination for genes encoding multi-protein complexes, even when regulated by common upstream activators. Correlations between these genes were as low as the correlation between any random pair of constitutively expressed genes. These results underline the challenges cells encounter to achieve coordinated assembly of multi-protein complexes within the limits of expression noise.

To obtain higher temporal resolution of the transcription kinetics, we established a live cell imaging approach based on a modified version of the MS2 system, specifically designed to study transcription coordination and kinetics in real time. This approach permits to monitor the synthesis of individual mRNAs in living cells, simultaneously determining initiation timing and frequency and providing a unique tool to study the regulation of gene expression networks in real time within the boundaries of a single cell.

A COMBINATORIAL PROTEIN CODE DEFINES NEW PRINCIPAL CHROMATIN TYPES IN *DROSOPHILA*

Ulrich Braunschweig *[1], Guillaume J Filion *[1], Joke G van Bemmel *[1], Wendy Talhout *[1], Jop Kind[1], Aranzazu Rosado[1], Ines de Castro[1], Wim Brugman[2], Ron Kerkhoven[2], Bas van Steensel[1]
[1]Netherlands Cancer Institute, Division of Gene Regulation, Plesmanlaan 121, Amsterdam, 1066CX, Netherlands, [2]Netherlands Cancer Institute, Central Microarray Facility, Plesmanlaan 121, Amsterdam, 1066CX, Netherlands

The local protein composition of chromatin is important for the regulation of transcription and other functions. However, a global view is still lacking of how many types of chromatin exist, how these types differ in protein composition, how they are distributed along the chromosomes, and how they are linked to regulatory functions. We present here a comprehensive analysis of the chromatin states in *Drosophila* cells.

We generated high-resolution genome-wide binding maps of 53 chromatin proteins, including histone-modifying enzymes, modified histone binding proteins, transcription machinery components, nucleosome remodelers, heterochromatin components, insulators and structural components of chromatin. For the majority of these proteins, high-resolution maps were not previously available.

Using a combination of Principal Component Analysis and Hidden Markov Models, we show that the genome is segmented in five major chromatin types that are formed by unique yet overlapping combinations of proteins. For clarity, we refer to each type by a color: *red*, *yellow*, *green*, *blue*, and *black*.

The *blue* and *green* chromatin types correspond to previously reported Polycomb and HP1-bound types. *Black* chromatin covers about half of the genome and contains almost exclusively inactive genes. Remarkably, this chromatin type is devoid of the classic 'heterochromatin' proteins of either the Polycomb or the HP1 type. Instead, it is defined by histone H1, Lamin, and a few mostly uncharacterized proteins. Thus, our data reveal the existence of a prominent repressive chromatin type that has largely been overlooked.

The bulk of the active genes, encompassing one quarter of the genome, surprisingly segregates into two distinct chromatin types. *Yellow* chromatin is marked by (among others) the chromodomain protein MRG15 and comprises many genes involved in cell proliferation and growth. *Red* chromatin consists of two thirds of all tested proteins, including the remodeler Brahma. Despite this unusually high protein density, *red* chromatin is more accessible than *yellow* chromatin. *Red* regions carry a higher density of conserved noncoding sequences, including known transcription factor binding sites. This suggests two different mechanisms for activation of genes in *yellow* and *red* regions. Experiments are ongoing to further elucidate these mechanisms.

Taken together, our integrative approach identified novel principal chromatin types, which are defined by unique combinations of proteins and have distinct functional properties.

GLOBAL CORRELATION OF TRANSCRIPTION AND DNA CONFORMATION REVEALS NEW MODES OF GENE REGULATION

Fedor Kouzine[1], Ashutosh Gupta[1,2], Damian Wojtowicz[3], Matthew Vogt[1], Arito Yamane[4], Rafael Casellas[4], Teresa Przytycka[3], Craig J Benham[5], David L Levens[1]
[1]CCR, NCI, NIH, Laboratory of Pathology, 10 Center Dr, Bethesda, MD, 20892, [2]University of Maryland, Department of Physics, 1201 Turner Hall, College Park, MD, 20742, [3]NCBI, NLM, NIH, Computational Biology Branch, 8600 Rockville Pike, Bethesda, MD, 20894, [4] NIAMSD, NIH, Genomic Integrity and Immunity, 1 AMS Circle, Bethesda, MD, 20892, [5]UC Davis, Genome Center, One Shields Ave, Davis, CA, 95616

A general perception is that biological information is encoded directly by DNA through a sequence of nucleotides, and indirectly through epigenetic phenomena. Accordingly, specific sequences or chromatin features have been presumed to act locally with little direct communication between genes. In this view, the coordination of genetic processes is conducted by trans-acting factors with DNA participating only passively. But, the molecular structure of DNA is dynamic and could function as an active participant in the interpretation of biological information. The role and biological relevance of DNA structural variation has long been controversial, despite *in silico* prediction of genome-wide associations between DNA structures and genetic activities. The biological importance of these structures is predicated on their *in vivo* existence. Using psoralen as an *in vivo* probe, we have developed an approach to chart a genome-scale map of DNA supercoiling. The map suggests that waves of negative supercoiling diffuse at least 2 kb upstream of the active promoters. Thus, the torsional stress produced as result of transcription elongation leads to DNA supercoiling. To investigate the possibility that non-B-DNA structures induced by supercoiling may be involved in control mechanism of transcription, we have developed an approach to map DNA conformation across the genome using permanganate as a probe of base unpairing. The map reveals the existence of non-canonical DNA structures across the genome. Cross-correlating regions of permanganate reactivity with *in silico* predicted alternative DNA structures indicates the widespread and sequence-dependent rearrangement of DNA structures (from B-DNA to non-B, e.g. quadruplex and Z-DNA). Many such regions lie outside the bodies of transcription units and are sensitive to the inhibition of transcription, suggesting a long-distance coupling between transcription and DNA conformation. It was recently demonstrated that transcription-induced DNA supercoiling and the transition from B-DNA to non-B-DNA are the basis of a mechanism for tight control of the human *c-myc* promoter. Therefore, we are investigating whether non-B-DNA structures induced by transcription-generated supercoiling frequently serve as regulatory signals for transcription factors. As such, *in vivo* maps of supercoiling, DNA conformation, and the *in vivo* binding sites for DNA structure-sensitive transcription factors may reveal very new modes of transcriptional regulation on a global scale.

CHROMATIN ON THE MEGABASE SCALE: THE FRACTAL GLOBULE ARCHITECTURE AND ITS PHYSICAL PROPERTIES

Erez Lieberman-Aiden[1,2], Maxim Imakaev[3], Nynke van Berkum[4], Louise Williams[1], Andreas Gnirke[1], Najeeb Tarazi[5], Job Dekker[4], Eric S Lander[1], Leonid A Mirny[2,3]

[1]Harvard-MIT, Broad Institute, 7 Cambridge Circle, Cambridge, MA, 02139, [2]Harvard-MIT, Division of Health Sciences & Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, [3]Massachusetts Institute of Technology, Department of Physics, 77 Massachusetts Avenue, Cambridge, MA, 02139, [4] University of Massachusetts Medical School, Biochemistry & Molecular Pharmacology, 364 Plantation St, Worcester, MA, 01605, [5]Harvard University, Harvard College, Massachusetts Hall, Cambridge, MA, 02138

The human genome folds into a structure that is dense enough to fit inside the nucleus, but compatible with cellular processes such as gene expression and the cell cycle. Understanding how the genome folds can elucidate the role of chromatin structure in enabling these crucial cellular processes. We recently developed Hi-C, a method that couples proximity ligation of genomic loci with massively parallel sequencing in order to enable genome-wide discovery of long-range chromatin contacts (Lieberman-Aiden & Van Berkum et al., Science, 2009).

Combining theoretical arguments and original Monte Carlo simulations we demonstrate that on the scale of several megabases, the experimentally obtained distribution of chromatin contacts is consistent with a fractal globule: a dense, highly organized, and knot-free structure. First proposed on theoretical grounds in 1988, the fractal globule has never before been observed, either *in silico* or experimentally.

We use our simulations to explore the physical properties of the fractal globule. We show that unentangled polymers (e.g. chromatin fibers) tend to form fractal globules by condensation or relaxation in a constrained volume, and show how easily the fractal globule can unravel in comparison to other condensed polymer conformations. These effects could play a role in regulating chromatin architecture during the cell cycle.

The fractal globule organizes the genome into discrete spatial sectors, which correspond to contiguous chromosomal subregions. We demonstrate the capacity of individual sectors to rapidly unfold and re-fold without disturbing the global chromatin conformation, a mechanism that could play a role in gene activation. Furthermore, we show that the presence of a fractal globule structure does not interfere with the ability of individual loci to diffuse across the globule over time.

THE THREE-DIMENSIONAL FOLDING OF THE ALPHA–GLOBIN GENE DOMAIN REVEALS FORMATION OF CHROMATIN GLOBULES

Davide Bau*[1], Amartya Sanyal*[2], Bryan R Lajoie*[2], Emidio Capriotti[1], Job Dekker[2], Marc A Marti-Renom[1]

[1]Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigacion Principe Felipe, Valencia, 46012, Spain, [2]Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation St., Worcester, MA, 01605

We developed a general approach that combines comprehensive chromatin interaction mapping using 5C with the Integrated Modeling Platform to generate high-resolution three-dimensional models of chromatin at the Mb scale. We applied this approach to analyze a 500 Kb gene dense domain (ENm008) on human chromosome 16 that includes the α-globin locus. We obtained 3D models of this domain in cells that express the α-globin locus (K562) as well as in lymphoblastoid cells that do not (GM12878). The models accurately reproduce the known looping interactions between the α-globin genes and their distal regulatory elements. Further, we find that the domain folds into a single globular conformation in GM12878 cells, whereas two globules are formed in K562 cells. The central cores of these globules are enriched for actively transcribed genes, whereas non-transcribed chromatin is more peripheral. We propose that globule formation represents a higher order folding state that is related to clustering of actively transcribed genes around shared transcription machineries.

REGULATORY GENOMICS AND EPIGENOMICS OF MULTIPLE HUMAN CELL LINES

Jason Ernst[1], Pouya Kheradpour[1], Noam Shoresh[2], Chuck Epstein[2], Bradley Bernstein[2], Manolis Kellis[1,2]
[1]MIT, Computer Science and Artificial Intelligence Lab, 32 Vassar St., Cambridge, MA, 02139, [2]Broad Institute, MIT and Harvard, 7 Cambridge Center, Cambridge, MA, 02141

Recent large-scale experimental efforts such as the NIH ENCODE and Epigenome Roadmap projects have enabled the systematic study of epigenetic information across both multiple chromatin marks and multiple cell types across the complete human genome. The systematic study of these datasets requires new computational methods to enable the joint modeling of combinatorial patterns of both chromatin marks and gain/loss patterns across cell types.

To address this challenge, we developed a graphical probabilistic model for learning chromatin states from combinations of chromatin marks within each cell type, in a systematic and unbiased way, without prior information on genome annotation. The resulting states reveal surprisingly rich epigenomic information, in several distinct classes of enhancer, promoter, transcribed, active intergenic, repressed, and repetitive states. We have extended these methods to enable joint learning of chromatin states across multiple cell types. In pairwise cell type comparisons, this has enabled us to reveal chromatin state changes associated with enhancer and promoter activity consistent with the functional roles of corresponding target genes in the surveyed cell types. In multi-cell comparisons of nine ENCODE cell types, this has enabled us to define activity profiles for both chromatin states and gene expression patterns, and correlate these with active and repressed chromatin states, leading to several applications including the linking of enhancer regions to likely target genes based on the co-variation between gene expression and chromatin mark intensity associated with enhancer states in multiple cell types.

We also used maps of 9 chromatin marks across 9 human cell types to identify active regulators, based on the correlation between TF gene expression and enrichment of the corresponding motif in active and repressive chromatin states. We used these correlations to define activator and repressor signatures by distinguishing positive and negative correlations. We found that specificity can be further improved by more closely examining the intensity profile of each chromatin modification at the region of binding. We find that a dip in the chromatin signal within regions of high intensity helps identify precise locations of binding at higher resolution, likely due to nucleosome exclusion in the ‘footprint’ of bound factors. These chromatin dips improved our ability to identify functional motif instances across the whole genome. We also found that dips specific to a cell type were enriched in the motifs of factors specifically expressed in that cell type.

TRANSPOSABLE ELEMENTS HAVE REWIRED THE CORE REGULATORY NETWORK OF HUMAN EMBRYONIC STEM CELLS

Guillaume Bourque[1], Galih Kunarso[1], Justin Jeyakani[1], Catalina Hwang[1], Na-Yu Chia[2], Winston Chan[2], Huck-Hui Ng[2]

[1]Computational & Mathematical Biology, Genome Institute of Singapore, Biopolis Street, Singapore, 138672, Singapore, [2]Stem Cell & Developmental Biology, Genome Institute of Singapore, Biopolis Street, Singapore, 138672, Singapore

Detection of novelty in genomic control elements is critical for understanding the wiring of transcriptional regulatory networks in their entirety. To systematically investigate the evolution of response elements in a mammalian gene regulatory network, we study the role of three key regulators (OCT4, NANOG and CTCF) in Embryonic Stem (ES) cells by performing occupancy and knockdown experiments in human and mouse. Interestingly, and in contrast to CTCF, we find that the occupancy profiles of OCT4 and NANOG are drastically different with only ~5% of the regions bound in human also bound in mouse. Moreover, we demonstrate that species-specific transposable elements have been an important source of regulatory sites in both lineages and have contributed close to 25% of the regions occupied by OCT4 in human. We also show that regulatory elements embedded in specific classes of repeats are highly enriched in proximity of regulated genes. Finally, because we have matching datasets in the two species, we are also able to identify 65 human-specific OCT4 target genes that have been added to the core regulatory network of human ES cells via the insertion of transposable elements. Together our findings reveal a high rate of functional turnover in the core regulatory network of mammalian ES cells and implicate transposable elements as a major contributor to this evolutionary plasticity.

STOCHASTIC EPIGENETIC VARIATION AS A DRIVING FORCE OF DEVELOPMENT, EVOLUTIONARY ADAPTATION, AND DISEASE

Rafael A Irizarry[1], Andrew P Feinberg[2]

[1]Johns Hopkins University, Biostatistics, 615 N. Wolfe St, Baltimore, MD, 21205, [2]Johns Hopkins University, Center for Epigenetics, 855 N. Wolfe St., Rangos 570, Baltimore, MD, 21205

Neo-Darwinian evolutionary theory is based on selection of phenotypes caused by small genetic variations, which is the basis of quantitative trait contribution to phenotype and disease. Epigenetics is the study of nonsequence-based changes, such as DNA methylation, heritable during cell division. Previous attempts to incorporate epigenetics into evolutionary thinking have focused on Lamarckian inheritance, that is, environmentally directed epigenetic changes. Here, we propose a new non-Lamarckian theory for a role of epigenetics in evolution. We suggest that genetic variants that do not change the mean phenotype could change the variability of phenotype; and this could be mediated epigenetically. This inherited stochastic variation model would provide a mechanism to explain an epigenetic role of developmental biology in selectable phenotypic variation, as well as the largely unexplained heritable genetic variation underlying common complex disease. We will show three experimental results related to methylation, gene expression, and common disease data. I will also describe some of the data analysis challenges.

HUMAN-SPECIFIC LOSS OF REGULATORY DNA AND THE EVOLUTION OF HUMAN-SPECIFIC TRAITS

Gill Bejerano[1,2], Cory Y McLean [2], Philip L Reno[1,3], Alex A Pollen[1], Abraham I Bassan[1], Terence D Capellini[1], Catherine Guenther[1,3], Vahan B Indjeian[1,3], Douglas B Menke[1,4], Bruce Schaar[1], Aaron M Wenger[2], David M Kingsley[1,3]

[1]Stanford University, Department of Developmental Biology, Stanford, CA, 94305, [2]Stanford University, Department of Computer Science, Stanford, CA, 94305, [3]Howard Hughes Medical Institute, MD, [4] Current address: University of Georgia, Department of Genetics, GA

The availability of several primate whole genome sequences has spurred great excitement for the prospect of understanding the molecular basis of what makes us human. Recent investigations have discovered dozens of conserved non protein coding genomic loci which have experienced accelerated basepair changes in the human lineage, as well as some proteins that show evidence of positive selection. Many previous genome surveys have focused on small base pair changes in otherwise well aligned sequences.

Here we expand these studies to look for a type of event particularly likely to produce functional effects: complete deletion in humans of sequences that are otherwise highly conserved in other organisms. By searching for regions of the chimpanzee genome highly conserved over mammalian evolution that are clearly missing in humans, we discover several hundreds human-specific losses of putatively functional ancestral DNA. PCR and computational validation show that roughly 80% of the deletions are fixed in human populations, while others are polymorphic in different individuals. Most of the deletions removed conserved non-coding sequences rather than protein-coding regions, and many lie in proximity to genes involved in development, morphogenesis, neural function, and steroid hormone signaling.

We have functionally tested a subset of human specific deletions in transgenic mice, and have found intriguing examples of regulatory alterations in humans that appear to be associated with evolution of specific anatomical differences between humans and other animals.

SINGLE BASE MATTERS: A SYSTEMATIC APPROACH FOR DETECTING EXPRESSION VARIATION AT SINGLE NUCLEOTIDE RESOLUTION

Yue Yun, Ayodele Adesanya , Gary Stormo, Rob Mitra
Washington University School of Medicine, Genetics Department, 4444 Forest Park Parkview, St. Louis, MO, 63108

Regulatory single nucleotide polymorphisms (rSNPs) contribute to phenotypic variation between individuals by causing changes in gene expression. The identification of functional rSNPs remains a significant challenge. The most commonly used approach for identifying rSNPs focuses on variants in conserved regions, especially those in transcription factor binding sites; the fraction of functional polymorphism missed by this approach however is unknown.

To address this question, we performed a systematic analysis of gene expression variation at the single nucleotide level in the yeast S. cerevisiae. Our approach involved two steps: first, using synthetic biology techniques, we created a nearly exhaustive single nucleotide variation library for the GAL1 promoter. Next, we measured the corresponding changes in gene expression for each variant using a sensitive dual-color reporter assay in living yeast. This approach is able to detect expression variation as small as 10%. For 7% (43/582) of the bases in the GAL1 promoter, we observed an expression change greater than 10%. Among these, the majority (81%; 35/43) resided in conserved region, and 63% (27/43) resided in known transcription factor binding sites. We also identified that expression variation caused by single nucleotide variations within a transcription factor binding sites (TFBS) is directly correlated with the position weight matrix of the TF. The strategy of identifying rSNPs by focusing on conserved regions and TFBS will likely capture large fractions of functional rSNPs.

Our work also suggests an important role for ATG trinucleotides in 5' UTRs. Among mutations that caused expression variation, three of them created frame-shifted ATGs in the 5'UTR and completely abolished reporter gene expression, demonstrating a strong impact of ATG on gene regulation. We also observed that, on a whole genome scale, both gene transcript levels and protein levels anti-correlate with the occurrences of ATG in the 5' UTR of genes. This explains why ATG trinucleotides are statistically less represented in 5' UTR of species from yeast to human. More intriguingly, many genes still maintain ATGs in their 5' UTR. Compared to the S. paradoxus genome, we found that this phenomenon is evolutionarily conserved, indicating a functional role for ATG in gene regulation. We propose that selection against neomorphic mutations may explain the surprisingly high degree of conservation observed in regulatory regions.

This study conducts a quantitative analysis of the relationship between sequence conservation and regulatory function at the single nucleotide level, providing experimental evidence to predict rSNPs.

COMPARATIVE REGULATORY GENOMICS IN *DROSOPHILA*

Qiye He[1], Anaïs Bardet[2], Brianne Patton[1], Julia Zeitlinger[1], Alexander Stark[2]
[1]Stowers Institute for Medical Research, -, 1000 East 50th Street, Kansas City, MO, 64110, [2]Research Institute of Molecular Pathology (IMP), -, Dr. Bohr-Gasse 7, Vienna, 1050, Austria

During animal development, the transcription of genes is tightly controlled in a temporal and spatial fashion. Regulation is mediated by transcription factors, which bind to *cis*-regulatory sequences (enhancers) in the vicinity of genes. Key transcription factors and their target genes are often highly conserved during evolution, while transcription factor binding sites or enhancers have also been reported to diverge quite rapidly and are thought to be an important driving force for evolutionary changes. However, the degree and mechanisms by which binding site turnover occurs is largely unknown since most studies have been focused on individual enhancers, and large-scale comparative sequence analyses lack functional information.

Here, we used ChIP-Seq to determine the binding sites of the mesodermal transcription factor Twist in embryos of six Drosophila species, chosen at evolutionary distances comparable to human-primate, human-mouse, and human-chicken (*D. melanogaster, D. simulans, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura*).

Our study suggests that Twist binding sites show a surprisingly high degree of conservation. Based on strong enrichment near mesodermal genes and altered transcript levels in twist mutants, we first determined that Twist has around one thousand *bona fide* binding sites. These Twist binding sites are highly conserved in orthologously aligned sequences, even at large evolutionary distances, and their conservation pattern follows the phylogenetic tree. Conservation of binding is highest in the vicinity of mesodermal genes that likely constitute functional Twist target genes and lowest around housekeeping genes that are likely regulated independently of Twist. Across all genes, conservation of binding is lowest for peaks in coding regions and highest for peaks in intergenic regions, independent of their distance to the nearest gene.

To understand the mechanisms by which binding sites diverge, we analyzed the sequences around conserved and diverged binding sites. Interestingly, regions with conserved Twist binding sites do not stand out by elevated overall sequence similarity across species. Instead, these regions show preferential conservation of sequences that match the known Twist motif. In contrast, regions with loss of binding correlate with motif losses, which occurred either by point mutations, insertions or deletions. Taken together, these results suggest that key developmental transcription factors such as Twist have a large number of functional binding sites and that the maintenance of the corresponding binding sequences is under direct evolutionary pressure.

TRANSCRIPTION BINDING VARIATION IN EUCARYOTES

M Snyder, M Kasowski, W Zheng, F Grubert, C Heffelfinger, M Hariharan, A Asabere, S Waszak, L Habegger, J Rozowsky, M Shi, A Urban, K Karczewski, H Zhao, E Mancera, L Steinmetz, M Gerstein, J Korbel
Stanford University, Dept. of Genetics, Stanford, CA, 94305

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

Participant List

Dr. Sheera Adar
UNC Chapel Hill
sheera@email.unc.edu

Dr. Phaedra Agius
MSKCC
phaedragius@gmail.com

Ms. Ilaria Albieri
San Raffaele Scientific Institute
albieri.ilaria@hsr.it

Ms. KHOLOUD ALSHAFAI
Qatar Foundation
khokah00@hotmail.com

Dr. Brenda Andrews
University of Toronto
brenda.andrews@utoronto.ca

Mr. Aaron Arvey
Sloan Kettering Institute
aarvey@cbio.mskcc.org

Mr. Roi Avraham
Weizmann Institute of Science
roi.avraham@weizmann.ac.il

Dr. Songjoon Baek
National Cancer Institute, NIH
baeks@mail.nih.gov

Dr. Yuchen Bai
Pfizer
baiy@wyeth.com

Dr. Martino Barenco
University College London
m.barenco@ucl.ac.uk

Dr. David Bartel
Whitehead Institute/MIT/HHMI
dbartel@wi.mit.edu

Mr. Felipe Beckedorff
University of São Paulo
fbecke@iq.usp.br

Dr. Gill Bejerano
Stanford University
bejerano@stanford.edu

Dr. George Bell
Whitehead Institute for Biomedical Research
gbell@wi.mit.edu

Dr. Shifra Ben-Dor
Weizmann Institute of Science
shifra.ben-dor@weizmann.ac.il

Dr. Joris Benschop
University Medical Center Utrecht
j.j.benschop@umcutrecht.nl

Ms. Anne Bergstrom Lucas
Agilent Technologies
anne_lucas@agilent.com

Dr. Giorgio Bernardi
Università degli Studi Roma Tre
bernardi@accademiaxl.it

Dr. Bradley Bernstein
Mass Gen Hosp, Harvard Med School, Broad Institute
bernstein.bradley@mgh.harvard.edu

Dr. Andreas Beutler
Mayo Clinic
beutler.andreas@mayo.edu

Dr. Guillaume Bourque
Genome Insitute of Singapore
bourque@gis.a-star.edu.sg

Dr. Alan Boyle
Duke University
alan.boyle@duke.edu

Dr. Marjorie Brand
OHRI
mbrand@ohri.ca

Mr. Ulrich Braunschweig
Netherlands Cancer Institute
u.braunschweig@nki.nl

Prof. Michael Brent
Washington University
brent@cse.wustl.edu

Dr. Christopher Bristow
MIT
bristow@mit.edu

Prof. Michael Brodsky
University of Massachusetts Medical School
michael.brodsky@umassmed.edu

Dr. Laurakay Bruhn
Agilent Technologies
laurakay_bruhn@agilent.com

Dr. Martha Bulyk
Brigham & Womens Hospital and Harvard Medical Sch
mlbulyk@receptor.med.harvard.edu

Dr. Christopher Burge
Massachusetts Institute of Technology
cburge@mit.edu

Dr. Wolfgang Busch
Duke University
wolfgang.busch@duke.edu

Dr. Harmen Bussemaker
Columbia University
hjb2004@columbia.edu

Ms. Lauren Camargo
Universidade de São Paulo
lauren@iq.usp.br

Dr. Howard Chang
Stanford University
howchang@stanford.edu

Mr. Varodom Charoensawan
MRC Laboratory of Molecular Biology
varodom@mrc-lmb.cam.ac.uk

Dr. Feng-Chi Chen
National Health Research Institutes
fcchen@nhri.org.tw

Dr. Muhan Chen
Cold Spring Harbor Laboratory
chenm@cshl.edu

Ms. Christine Cheng
University of California, San Diego
c3cheng@ucsd.edu

Ms. Bi-Hua Cheng
Institute for
bihuacheng@hrca.harvard.edu

Dr. Yong Cheng
Penn state university
yzc109@psu.edu

Mr. Hirokazu Chiba
Nat'l Institute of Advanced Industrial Science & Technology
h.chiba@aist.go.jp

Dr. Jaishree Chittoor
Monsanto
jmchit@monsanto.com

Mr. Ryan Christensen
Washington University in St Louis
christensen@wustl.edu

Mr. Michael Conway
University of Wisconsin Madison
mconway@wisc.edu

Ms. Kate Cook
University of Toronto
kathleen.cook@utoronto.ca

Dr. Justin Cotney
Yale University School of Medicine
justin.cotney@yale.edu

Prof. Patrick Cramer
LMU University of Munich
cramer@lmb.uni-muenchen.de

Dr. Maya Dadiani
Weizmann Institute of Science
maya.chetrit@weizmann.ac.il

Mr. Carl de Boer
University of Toronto
carlgdeboer@gmail.com

Dr. Vincenzo De Simone
University of Naples "Federico II" (Naples-Italy)
vincenzo.desimone@unina.it

Mr. Jacob Degner
University of Chicago
jdegner@uchicago.edu

Dr. Job Dekker
University of Massachusetts Medical School
job.dekker@umassmed.edu

Prof. Bart Deplancke
Swiss Federal Institute of Technology Lausanne
bart.deplancke@epfl.ch

Ms. Jacqueline Dresch
MSU
dreschja@msu.edu

Ms. Elif Duman
Biopsychology PhD Program, Stony Brook University.
eduman@notes.cc.sunysb.edu

Dr. Noriko Esumi
Johns Hopkins University School of Medicine
nesumi@jhmi.edu

Dr. Angela Fachel
Universidade de Sao Paulo
afachel@iq.usp.br

Ms. Mina Fazlollahi
Columbia University
mina_f@phys.columbia.edu

Dr. Vladimir Filkov
UC Davis
filkov@cs.ucdavis.edu

Dr. Hiram Firpi
University of Iowa
hiram-firpi@uiowa.edu

Dr. Barrett Foat
Monsanto Company
barrett.foat@monsanto.com

Prof. Nir Friedman
Hebrew University
nir@cs.huji.ac.il

Dr. Erwin Frise
Lawrence Berkeley National Labs
erwin@fruitfly.org

Ms. Katja Fruehauf
LMU Munich
fruehauf@lmb.uni-muenchen.de

Dr. Wataru Fujibuchi
Advanced Industrial Science and Technology
w.fujibuchi@aist.go.jp

Dr. Andre Fujita
RIKEN
andrefujita@riken.jp

Dr. Eileen Furlong
EMBL
eileen.furlong@embl-heidelberg.de

Dr. Mythily Ganapathi
Wadsworth Center, Dept of Health
mythilyg@gmail.com

Dr. Clare Garvey
Genome Biology
clare.garvey@genomebiology.com

Dr. Andre Gerber
ETH Zurich
andre.gerber@pharma.ethz.ch

Dr. Thomas Gingeras
Cold Spring Harbor Laboratory
gingeras@cshl.edu

Mr. Paul Giresi
University of North Carolina at Chapel Hill
paulg@email.unc.edu

Dr. Raluca Gordan
Brigham&Women's Hospital/Harvard Medical School
rgordan@rics.bwh.harvard.edu

Dr. Jesse Gray
Harvard Medical School
jesse.gray@gmail.com

Dr. Ivo Grosse
Martin Luther University
grosse@informatik.uni-halle.de

Mr. Xin Guo
Duke University
xinguo@cs.duke.edu

Dr. Marc Halfon
State University of New York at Buffalo
mshalfon@buffalo.edu

Prof. Ross Hardison
Penn State University
rch8@psu.edu

Dr. David Harmin
Harvard Medical School
David_Harmin@hms.harvard.edu

Mr. Yuki Hatanaka
Kinki University
bm8012hy@waka.kindai.ac.jp

Dr. David Hawkins
Ludwig Institute for Cancer Research
d1hawkin@ucsd.edu

Mr. Brian Haynes
Washington University in St. Louis
brian.c.haynes@gmail.com

Mr. Daniel Hebenstreit
MRC Laboratory of Molecular Biology
danielh@mrc-lmb.cam.ac.uk

Dr. Martin Hemberg
Children's Hospital Boston
martin.hemberg@childrens.harvard.edu

Dr. Steven Henikoff
Fred Hutchinson Cancer Research Center
steveh@fhcrc.org

Dr. Carl Herrmann
TAGC - Inserm
herrmann@tagc.univ-mrs.fr

Mr. Richard Hickman
University of Warwick
richard.j.hickman@warwick.ac.uk

Dr. Gary Hon
University of California, San Diego
ghon@ucsd.edu

Dr. Kun Huang
The Ohio State University
khuang@bmi.osu.edu

Dr. Qingyang Huang
The University of Hong Kong
huangqy@hkucc.hku.hk

Dr. Mike Hubank
Institute of Child Health, UCL
m.hubank@ich.ucl.ac.uk

Prof. Timothy Hughes
University of Toronto
t.hughes@utoronto.ca

Ms. Tiffany Hung
Stanford U.
tiffany.hung@stanford.edu

Ms. Lee Hyang Heun
Kinki University
hyangheun@hanmail.net

Dr. Francisco Iborra
Centro Nacional de Biotecnologia (CSIC)
fjiborra@cnb.csic.es

Dr. Trey Ideker
University of California, San Diego
trey@bioeng.ucsd.edu

Dr. Kohta Ikegami
University of North Carolina at Chapel Hill
ikgmk@email.unc.edu

Mr. Maksim Imakaev
MIT
imakaev@mit.edu

Dr. Rafael Irizarry
Johns Hopkins Bloomberg School of Public Health
ririzarr@jhsph.edu

Dr. Tomoko Ito
Musashino University
t_ito@musashino-u.ac.jp

Dr. Sharoni Jacobs
Agilent
sharoni_jacobs@agilent.com

Dr. Anders Jacobsen
University of Copenhagen
andersbj@binf.ku.dk

Dr. James Jarvis
Univ of Okla Health Sciences Ctr
james-jarvis@ouhsc.edu

Ms. Alba Jene Sanz
Pompeu Fabra University
alba.jene@upf.edu

Mr. Zhe Ji
UMDNJ-New Jersey Medical School
jizh@umdnj.edu

Dr. Pinay Kainth
University of Toronto
pinay.kainth@gmail.com

Dr. Tommy Kaplan
UC Berkeley
tomkap@berkeley.edu

Mr. Konrad Karczewski
Stanford University
konradjk@stanford.edu

Dr. Kerstin Kaufmann
Wageningen University and Research Centre
kerstin.kaufmann@wur.nl

Prof. Manolis Kellis
MIT
manoli@mit.edu

Mr. Steven Kiddle
University of Warwick
s.j.kiddle@warwick.ac.uk

Dr. Jonghwan Kim
HMS/CHB
jkim@bloodgroup.tch.harvard.edu

Dr. Justin Kinney
Cold Spring Harbor Laboratory
jkinney@cshl.edu

Dr. Julian Konig
Medical Research Council
jkonig@mrc-lmb.cam.ac.uk

Dr. Hashem Koohy
Warwick University
hashem.koohy@warwick.ac.uk

Dr. Fedor Kouzine
CCR, NCI, NIH
kouzinef@mail.nih.gov

Dr. Sunita Kumari
Cold Spring Harbor Lab
kumari@cshl.edu

Dr. Anshul Kundaje
Stanford University
akundaje@stanford.edu

Dr. Christoph Kurat
CCBR
cf.kurat@utoronto.ca

Prof. Istvan (Steve) Ladunga
University of Nebraska-Lincoln
sladunga@unl.edu

Mr. Bryan Lajoie
University of Massachusetts Medical School
bryan.lajoie@umassmed.edu

Ms. Kathy Lam
University of Toronto
itskathylam@gmail.com

Mr. Benjamin Lang
MRC Laboratory of Molecular Biology
blang@mrc-lmb.cam.ac.uk

Dr. Daniel Larson
Albert Einstein College of Medicine
dlarson@aecom.yu.edu

Ms. Jessica Larson
Harvard University
jllarson@fas.harvard.edu

Dr. Louise Laurent
UCSD
llaurent@ucsd.edu

Dr. Eunjee Lee
Columbia University
el2380@columbia.edu

Dr. Hongxing Lei
Beijing Institute of Genomics, CAS
leihx2008@yahoo.com

Dr. David Levens
National Cancer Institute
levens@helix.nih.gov

Ms. Michal Levo
Weizmann Institute of Science
michal.levo@weizmann.ac.il

Ms. Lili Li
McGill University
lili.li@mail.mcgill.ca

Dr. Wei Li
Baylor College of Medicine
WL1@bcm.edu

Dr. XiaoChing Li
Louisiana State University
xli4@lsuhsc.edu

Mr. Willey Liao
Cold Spring Harbor Laboratory
will.liao@gmail.com

Mr. Michael Lidschreiber
LMU Munich
lidschreiber@lmb.uni-muenchen.de

Dr. Jason Lieb
University of North Carolina, Chapel Hill
jlieb@bio.unc.edu

Mr. Erez Lieberman-Aiden
Harvard/MIT
erez@mit.edu

Mr. Florian Lienert
Friedrich Miescher Institute
florian.lienert@fmi.ch

Mr. Howard Lightfoot
University of Warwick
h.b.lightfoot@warwick.ac.uk

Mr. Feng-Ming Lin
Academic Sinica
rsmrlin@yahoo.com.tw

Dr. X. Shirley Liu
Harvard / Dana-Farber
xsliu@jimmy.harvard.edu

Dr. Yunlong Liu
Indiana University School of Medicine
yunliu@iupui.edu

Ms. Marta Luksza
Max Planck Institute for Molecular Genetics
luksza@molgen.mpg.de

Dr. Nicholas Luscombe
EMBL Outstation - Hinxton
luscombe@ebi.ac.uk

Dr. Lesley MacNeil
University of Massachusetts Medical School
les.macneil@gmail.com

Dr. Robert Majovski
Genome Research
majovski@cshl.edu

Mr. Joseph Maranville
University of Chicago
jcmaranville@uchicago.edu

Mr. Georgi Marinov
California Institute of Technology
georgi@caltech.edu

Mr. Inigo Martincorena
EMBL-EBI
martinco@ebi.ac.uk

Mr. Toshiyuki Matsui
Institute for Virus Research, Kyoto University
tmatui@virus.kyoto-u.ac.jp

Dr. Andreas Mayer
Gene Center / LMU
mayer@lmb.uni-muenchen.de

Dr. Molly Megraw
Duke University
molly.megraw@duke.edu

Dr. Eric Mendenhall
Harvard Medical School
ericm@broadinstitute.org

Dr. Marc Meneghini
University of Toronto
marc.meneghini@utoronto.ca

Mr. Wouter Meuleman
Netherlands Cancer Institute
w.meuleman@nki.nl

Dr. Leonid Mirny
MIT
leonid@mit.edu

Dr. Masato Miyake
AIST,JAPAN
masato-miyake@aist.go.jp

Dr. Yifan Mo
Cold Spring Harbor Lab
ymo@cshl.edu

Dr. Ilario Mogno
Washington University in St. Louis
mognoi@genetics.wustl.edu

Mr. Michal Mokry
Hubrecht Institute
m.mokry@hubrecht.eu

Dr. Beth Moorefield
Fred Hutchinson Cancer Research Center
bmoorefi@fhcrc.org

Dr. Kengo Morohashi
The Ohio State University
morohashi.1@osu.edu

Dr. Randy Morse
Wadsworth Center
randall.morse@wadsworth.org

Dr. Ali Mortazavi
California Institute of Technology
alim@caltech.edu

Dr. John Murray
University of Pennsylvania School of Medicine
jmurr@mail.med.upenn.edu

Mr. Anand Murugan
Princeton University
anandm@princeton.edu

Dr. Felix Naef
Ecole Polytechnique Federale de Lausanne (EPFL)
felix.naef@epfl.ch

Ms. Renuka Nayak
University of Pennsylvania
nayakr@mail.med.upenn.edu

Dr. Zuqin Nie
NIH/NCI
niezu@mail.nih.gov

Dr. Boris Noyvert
University of Warwick
b.noyvert@warwick.ac.uk

Mr. Matthew O'Connor
OSI Pharmaceuticals
matthewo@bcm.tmc.edu

Dr. Daria Onichtchouk
University of Freiburg
daria.onichtchouk@biologie.uni-freiburg.de

Dr. Yuriy Orlov
Genome Institute of Singapore
orlovy@gis.a-star.edu.sg

Dr. Sascha Ott
University of Warwick
s.ott@warwick.ac.uk

Mr. Zhengqing Ouyang
Stanford University
zouyang@stanford.edu

Dr. Gabriela Paiva e Silva
Universidade Federal do Rio e Janeiro
gosilva@bioqmed.ufrj.br

Dr. Lorenzo Pasquali
Genomic Programming of Beta Cells, IDIBAPS
pasquali@clinic.ub.es

Dr. Dana Pe'er
Columbia University
dpeer@biology.columbia.edu

Dr. Frederic Pio
Simon Fraser University
fpio@sfu.ca

Dr. Roger Pique-Regi
Univeristy of Chicago
rpique@gmail.com

Mr. Sebastian Pott
Genome Institute of Singapore
sebastianp99@gis.a-star.edu.sg

Dr. Jonathan Pritchard
University of Chicago
pritch@uchicago.edu

Dr. Jiang Qian
Johns Hopkins School of Medicine
jiang.qian@jhmi.edu

Dr. Steve Qin
Uuniversity of Michigan
qin@umich.edu

Ms. Nisha Rajagopal
UCSD
nirajago@ucsd.edu

Dr. Oliver Rando
UMass Medical School
Oliver.Rando@umassmed.edu

Dr. Aviv Regev
Broad Institute
aregev@broad.mit.edu

Mr. John Reid
University of Cambridge
john.reid@mrc-bsu.cam.ac.uk

Dr. Mark Reimers
Virginia Commonwealth University
mreimers@vcu.edu

Dr. Bing Ren
Ludwig Institute for Cancer Research
biren@ucsd.edu

Dr. Todd Riley
Columbia University
tr2261@columbia.edu

Dr. John Rinn
Broad Institute of MIT and Harvard Medical School
jrinn@bidmc.harvard.edu

Dr. Ana Rodrigues
Salk Institute
arodrigues@salk.edu

Dr. Yijun Ruan
Genome Institute of Singapore
ruanyj@gis.a-star.edu.sg

Dr. Alla Rynditch
Institute of Molecular Biology and Genetics
rynditch@imbg.org.ua

Mr. Felix Sanchez Garcia
Columbia University
fs2282@columbia.edu

Dr. Rickard Sandberg
Karolinska Institutet
Rickard.Sandberg@ki.se

Dr. Amartya Sanyal
University of Massachusetts Medical School
amartya.sanyal@umassmed.edu

Dr. Simone Schimpf-Linzenbold
Centre for Ophthalmology
Simone.Schimpf-Linzenbold@med.uni-tuebingen.de

Dr. Amy Schmid
Duke University
amy.schmid@duke.edu

Dr. Gary Schroth
Illumina, Inc.
gschroth@illumina.com

Ms. Julia Schulze
UBC/CMMT
juliamschulze@gmail.com

Dr. Bjoern Schumacher
University of Cologne
bjoern.schumacher@uni-koeln.de

Ms. Rachel Sealfon
MIT
rsealfon@mit.edu

Dr. Roza Selimyan
NIH / NIA
selimyanr@grc.nia.nih.gov

Dr. Patricia Severino
Albert Einstein Research and Education Institute
patisev@yahoo.com

Dr. Noa Sher
Whitehead Institute
sher@wi.mit.edu

Dr. Idit Shiff
The Hebrew University of Jerusalem
idit.shiff@mail.huji.ac.il

Dr. Jay Shin
RIKEN Yokohama
jay.shin@gsc.riken.jp

Dr. Yuichi Shiraishi
RIKEN
yshira@riken.jp

Dr. Arend Sidow
Stanford University School of Medicine
arend@stanford.edu

Dr. Trevor Siggers
Brigham & Women's Hospital/Harvard Medical School
tsiggers@rics.bwh.harvard.edu

Dr. Eric Siggia
Rockefeller University
siggiae@rockefeller.edu

Dr. Mikael Sigvardsson
Linkoping University
miksi@ibk.liu.se

Dr. Michael Snyder
Stanford University School of Medicine
michael.snyder@yale.edu;
mpsnyder@stanford.edu

Ms. Sarah Spencer
Washington University
sjspence@wustl.edu

Dr. John Stamatoyannopoulos
University of Washington
jstam@uw.edu

Dr. Alexander Stark
IMP
stark@imp.ac.at

Dr. Angelike Stathopoulos
California Insitute of Technology
angelike@caltech.edu

Dr. Christopher Streck
Illumina
cstreck@illumina.com

Dr. Kevin Struhl
Harvard Medical School
kevin@hms.harvard.edu

Dr. Priya Sudarsanam
Washington University School of Medicine
psudarsanam@genetics.wustl.edu

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Ms. Ewa Szczurek
University of Warsaw
szczurek@molgen.mpg.de

Dr. Lara Szewczak
Cell Press - Cell
lszewczak@cell.com

Dr. Sylvia Takacova
Faculty of Medicine, Palacky University
sylviatak@gmail.com

Dr. Taku Tanaka
National Institute of Advanced Industrial Science
tacu-tanaka@aist.go.jp

Mr. Bogdan Tanasa
HHMI / TSRI
bogdan.tanasa@gmail.com

Dr. Stephen Tapscott
Fred Hutchinson Cancer Research Center
stapscot@fhcrc.org

Dr. Ronald Taylor
Pacific Northwest National Laboratory
ronald.taylor@pnl.gov

Dr. Scott Tenenbaum
University at Albany-SUNY
stenenbaum@albany.edu

Mr. Ronald Tepper
Columbia University
rgt29@columbia.edu

Ms. Desiree Tillo
University of Toronto
desiree.tillo@utoronto.ca

Dr. Jennifer To
GrassRoots Biotechnology
jenn.to@grassrootsbio.com

Ms. Tatjana Trcek
Albert Einstein College of Medicine
ttrcek@aecom.yu.edu

Mr. Alexander Tsankov
Massachusetts Institute of Technology/Harvard Med.
atsankov@mit.edu

Dr. Nikoleta Tsvetanova
Stanford University
nina1984@stanford.edu

Ms. Soile Tuomela
University of Turku
soile.tuomela@btk.fi

Dr. Jernej Ule
MRC Laboratory of Molecular Biology
jule@mrc-lmb.cam.ac.uk

Mr. Joris van Arensbergen
Genomic Programming of Beta Cells Laboratory
joris.van.arensbergen@gmail.com

Dr. Harm van Bakel
University of Toronto
hvbakel@gmail.com

Dr. Nynke Van Berkum
University of Massachusetts Medical School
nynke.vanberkum@umassmed.edu

Dr. Alexander van Oudenaarden
Massachusetts Institute of Technology
avano@mit.edu

Dr. Bas Van Steensel
Netherlands Cancer Institute
b.v.steensel@nki.nl

Dr. Nadine Vastenhouw
Harvard University
vastenhouw@mcb.harvard.edu

Dr. Marian Walhout
UMass Medical School
marian.walhout@umassmed.edu

Ms. Yue Wan
Stanford
ywan@stanford.edu

Dr. Wei Wang
UCSD
wei-wang@ucsd.edu

Dr. Junwen Wang
the University of Hong Kong
junwen@hku.hk

Dr. Ting Wang
Washington University
twang@genetics.wustl.edu

Mr. Lucas Ward
Columbia University
ldw2105@columbia.edu

Dr. Doreen Ware
CSHL USDA-ARS
ware@cshl.edu

Dr. Matt Weirauch
University of Toronto
weirauch@soe.ucsc.edu

Dr. Brian White
Cornell University
bsw27@cornell.edu

Dr. Bartosz Wilczynski
EMBL
wilczyns@embl.de

Dr. Barbara Wold
California Institute of Technology
woldb@caltech.edu

Prof. Scot Wolfe
UMass Medical School
scot.wolfe@umassmed.edu

Mr. Weisheng Wu
the Pennsylvania State University
wuw111@psu.edu

Mr. Adam Yadon
Fred Hutchinson Cancer Research Center
yadona@uw.edu

Dr. Richard Young
Whitehead Institute/M.I.T.
young@wi.mit.edu

Mr. Scott Younger
UT Southwestern Medical Center
scott.younger@utsouthwestern.edu

Dr. Jindan Yu
Northwestern University
jindan-yu@northwestern.edu

Dr. Guo-Cheng Yuan
Dana-Farber Cancer Institute
gcyuan@jimmy.harvard.edu

Dr. Feng Yue
Ludwig Institute for Cancer Research, UCSD
fyue@ucsd.edu

Ms. Yue Yun
Washington University in St. Louis
yyun@wustl.edu

Dr. Laura Zahn
AAAS
lzahn@aaas.org

Ms. Einat Zalckvar
Weizmann Institute of Science
einatz@weizmann.ac.il

Mr. Chongzhi Zang
The George Washington University
zang@gwmail.gwu.edu

Dr. Daniel Zenklusen
Université de Montréal
dzenklus@aecom.yu.edu

Mr. Yang Zhang
Michigan State University
zhangy16@msu.edu

Dr. Chaolin Zhang
Rockefeller University
czhang@rockefeller.edu

Mr. Yue Zhao
Washington University in St. Louis
yue.zhao@wustl.edu

Mr. Xu Zhou
Harvard University
xzhou@fas.harvard.edu

Dr. Xiaobo Zhou
Brigham & Women's Hospital, HMS
xzhou@rics.bwh.harvard.edu

Dr. James Zhu
US Department of Agriculture
James.Zhu@ars.usda.gov

Dr. Heng Zhu
Johns Hopkins University
hzhu4@jhmi.edu

Ms. Yaara Zwang
Weizmann Institute of Science
yaara.zwang@weizmann.ac.il

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers
BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers
Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
Helpful tips - Obtain PIN from Meetings & Courses Office to enter Library after hours. See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail only
Lower level: Word processing and printing.
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar
Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, west wing, lower level
PIN#: **Press 64310 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($8.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)
Executive Limo	631-696-8000 (**1047**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322