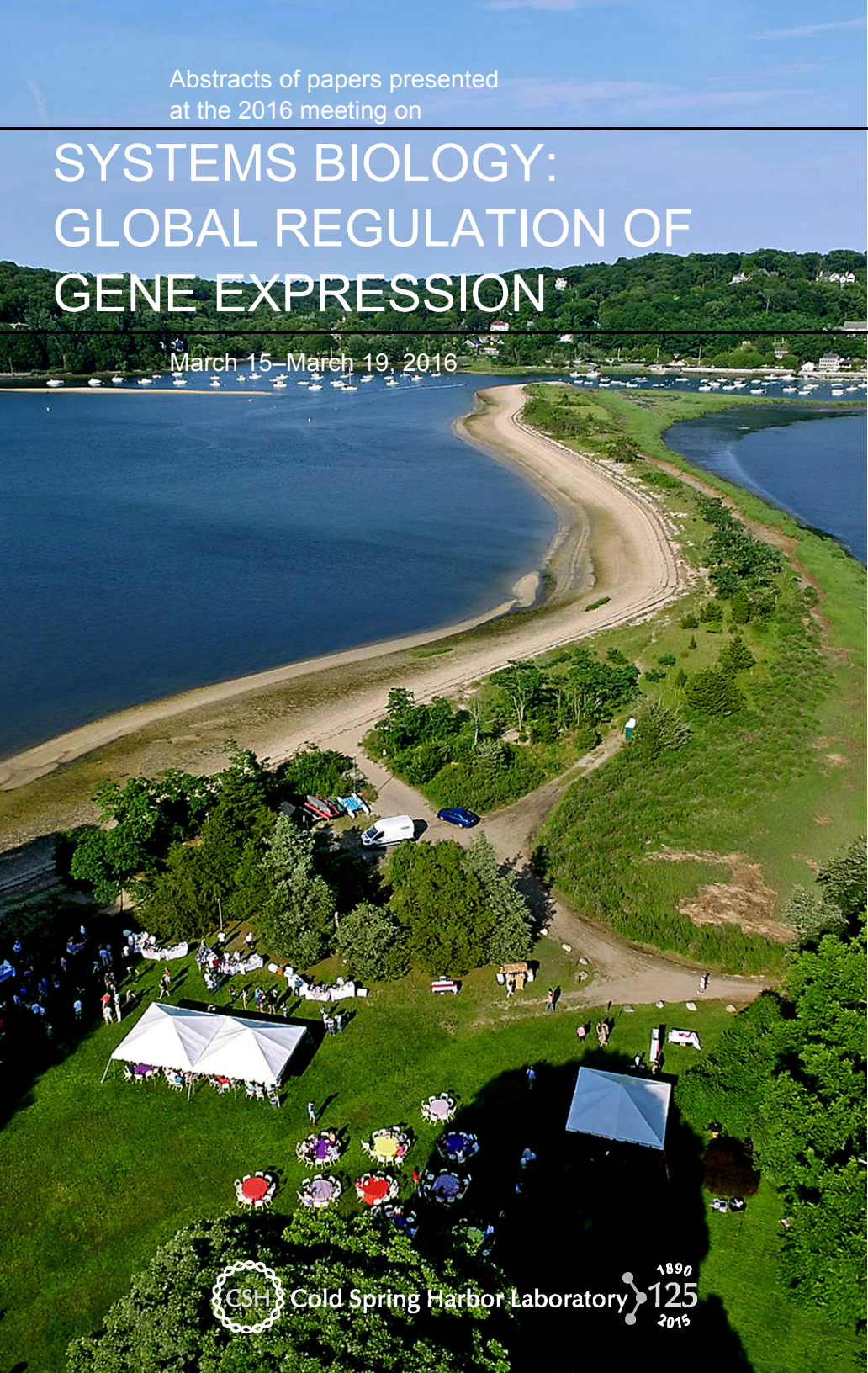
Abstracts of papers presented
at the 2016 meeting on

SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION

March 15–March 19, 2016



Cold Spring Harbor Laboratory
1890 125 2015

Abstracts of papers presented
at the 2016 meeting on

SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION

March 15–March 19, 2016

Arranged by

Barak Cohen, *Washington University, St. Louis*
Christina Leslie, *Memorial Sloan-Kettering Cancer Center*
John Stamatoyannopoulos, *University of Washington*
Sarah Teichmann, *WT Sanger Institute & EMBL-EBI, UK*

This meeting was funded in part by the **National Human Genome Research Institute,** a branch of the **National Institutes of Health.**

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

Corporate Sponsors

Agilent Technologies
Bristol-Myers Squibb Company
Genentech
Life Technologies (part of Thermo Fisher Scientific)
New England BioLabs

Plant Corporate Associates

Monsanto Company

Front Cover: Aerial view of the CSHL 125th Anniversary Picnic. Photo by Drew Mendelsohn.

SYSTEMS BIOLOGY: GLOBAL REGULATION OF GENE EXPRESSION
Tuesday, March 15– Saturday, March 19, 2016

Tuesday	7:30 pm	**1** Cis-Regulation
Wednesday	9:00 am	**2** Transcription Factors
Wednesday	2:00 pm	**3** Chromatin and 3D Architecture
Wednesday	5:00 pm	*Wine and Cheese Party*
Wednesday	7:30 pm	**Keynote Speaker**
Wednesday	8:15 pm	**4** Poster Session I
Thursday	9:00 am	**5** Evolution and Variation
Thursday	2:00 pm	**6** RNA and Its Regulation
Thursday	7:30 pm	**Keynote Speaker**
Thursday	8:15 pm	**7** Poster Session II
Friday	9:00 am	**8** Cellular Decision Making
Friday	2:00 pm	**9** Emerging Technologies
Friday	6:00 pm 7:00 pm	Concert Banquet
Saturday	9:00 am	**10** Networks and Global Analyses

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 15—7:30 PM

SESSION 1 CIS-REGULATION

Chairpersons: **Raluca Gordan,** Duke University, Durham, North Carolina
Nadav Ahituv, University of California, San Francisco

WEDNESDAY, March 16—9:00 AM

SESSION 2 TRANSCRIPTION FACTORS

Chairpersons: **Bas van Steensel,** Netherlands Cancer Institute, Amsterdam
Dirk Schubeler, Friedrich Meischer Institute, Basel, Switzerland

SESSION 3 CHROMATIN AND 3D ARCHITECTURE

Chairpersons: **David Arnosti,** Michigan State University, East Lansing
Asifa Akhtar, Max Planck Institute of Immunology and Epigenetics, Freiburg, Germany

WEDNESDAY, March 16—5:00 PM

Wine and Cheese Party

WEDNESDAY, March 16—7:30 PM

KEYNOTE SPEAKER

Wendy Bickmore
University of Edinburgh

"Enhancers—Complexities in their evolution and mechanisms of action"

SESSION 4 POSTER SESSION I

THURSDAY, March 17—9:00 AM

SESSION 5 EVOLUTION AND VARIATION

Chairpersons: **Barbara Engelhardt,** Princeton University, New Jersey
Yoav Gilad, University of Chicago, Illinois

SESSION 6 RNA AND ITS REGULATION

Chairpersons: **Sean Eddy,** HHMI, Harvard University, HHMI, Cambridge, Massachusetts
Julie Ahringer, Wellcome Trust/Cancer Research UK Gurdon Institute, Cambridge, United Kingdom

Widespread shortening of 3' UTRs and increased exon inclusion are evolutionarily conserved features of innate immune response to infection
Athma A. Pai, Golshid Baharian, Ariane Page Sabourin, Jessica F. Brinkworth, Yohann Nedelec, Jean-Christophe Grenier, Katherine J. Siddle, Anne Dumaine, Vania Yotova, Zachary P. Johnson, Robert E. Lanford, Christopher B. Burge, Luis B. Barreiro.
Presenter affiliation: MIT, Cambridge, Massachusetts. 94

Non-coding isoforms of coding genes in B cell development and malignancies
Irtisha Singh, Shih-Han Lee, Christina S. Leslie, Christine Mayr.
Presenter affiliation: Memorial Sloan Kettering Cancer Center, New York, New York. 95

Uncovering the RNA-protein interaction network that control vertebrate embryogenesis
Daniel Cifuentes, Charles Vejnar, Romain Christiano, Stephanie Lau, Alfredo Castello, Matthias Hentze, Tobias Walther, Antonio Giraldez.
Presenter affiliation: Yale University, New Haven, Connecticut. 96

THURSDAY, March 17—7:30 PM

KEYNOTE SPEAKER

John T. Lis
Cornell University

"Genome-wide views into the networks and mechanisms of transcription regulated by the stress response"

THURSDAY, March 17—8:15 PM

SESSION 7 POSTER SESSION II

Sequence-based predictions of *cis*-regulatory DNA elements active in the human heart
Dongwon Lee, Ashish Kapoor, Aravinda Chakravarti.
Presenter affiliation: Johns Hopkins University School of Medicine, Baltimore, Maryland. 97

FRIDAY, March 18—9:00 AM

SESSION 8 CELLULAR DECISION MAKING

Chairpersons: **Rob Mitra,** Washington University, St. Louis, Missouri
Aviv Regev, Broad Institute, Cambridge, Massachusetts

FRIDAY, March 18—2:00 PM

SESSION 9 EMERGING TECHNOLOGIES

Chairpersons: **William Greenleaf,** Stanford University, California
Steve McCarroll, Harvard Medical School, Boston, Massachusetts

FRIDAY, March 18—6:00 PM

CONCERT
Grace Auditorium

Fei-Fei Dong, piano

Praised for her ”bountiful gifts and passionate immersion into the music she touches” (The Plain Dealer), Chinese pianist Fei-Fei Dong is a winner of the 2014 CAG Victor Elmaleh Competition and a top six finalist at the 14th Van Cliburn International Piano Competition. She continues to build a reputation for her poetic interpretations, charming audiences with her “passion, piquancy and tenderness” and “winning stage presence” (Dallas Morning News).

Fei Fei’s burgeoning career includes a number of prominent engagements in the 2015-16 season, including New York City debut recitals at Weill Recital Hall at Carnegie Hall (CAG Winners Series), and at Merkin Concert Hall (Tuesday matinees series). Additional recitals include the Gilmore Rising Stars Series (Kalamazoo, MI), Macon Concert Association (Georgia), and St. Vincent College College Concert series (near Pittsburgh), where she will receive the Father Joseph Bronder Memorial Piano Prize.

Featured concerto engagements in 15-16 include a special benefit performance with the Kansas City Symphony and Music Director Michael Stern, as well as the Lexington (KY), Brazos Valley (TX) and Northeastern Pennsylvania Philharmonic orchestras.

2015 Summer festival highlights include Bravo! Vail Valley, the Highlands Chamber Music and Lake George Music Festivals, and a recital for Chicago’s Dame Myra Hess Concert Series. She is also showcased prominently in the new documentary film about the 2013 Cliburn Competition, Virtuosity, which premiered on PBS in August 2015.

Fei-Fei has performed at Alice Tully Hall in recital as the winner of Juilliard’s 33rd Annual William Petschek Recital Award and as a soloist with the Juilliard Orchestra under the baton of Jeffrey Kahane. She has appeared as a soloist with the Aspen Music Festival Orchestra, Fort Worth Symphony, Hong Kong Philharmonic Orchestra, Corpus Christi Symphony, Hudson Valley Philharmonic, Austin Symphony, Youngstown Symphony, Fort Collins Symphony, Denver Philharmonic, and in China with the Shanxi and Shenzhen Symphony Orchestras, working with conductors such as Leonard Slatkin, Randall Craig Fleisher, and John Giordano. Notable recitals include those at Warsaw Philharmonic Concert Hall and the Louvre, as well as The Cliburn’s spring 2015 Chopin Festival. She has also been featured numerous times on New York’s WQXR radio.

She is a member of the Aletheia Piano Trio, which debuted at the Kennedy Center in February 2014 as part of its Conservatory Project. Deeply committed to sharing her joy for music and connecting with communities, Fei-Fei also engages students and audiences through school and outreach concerts and master classes.

Born in Shenzhen, China, Fei-Fei began piano lessons at the age of 5. She moved to New York to study at The Juilliard School, where she earned her Bachelor and Master of Music degrees under the guidance of Yoheved Kaplinsky.

FRIDAY, March 18

BANQUET

Cocktails 7:00 PM Dinner 7:45 PM

SATURDAY, March 19—9:00 AM

SESSION 10 NETWORKS AND GLOBAL ANALYSES

Chairpersons: **Shirley Liu,** Dana-Farber Cancer Institute, Boston, Massachusetts
Dana Pe'er, Columbia University, New York, New York

HOW DO CLOSELY RELATED TRANSCRIPTION FACTORS TARGET DISTINCT GENOMIC REGIONS?

Ning Shen[1,2], Josh Schipper[2,3], Tristan Bepler[2], John Horton[2,3], Raluca Gordan[2,3]

[1]Duke University, Department of Pharmacology and Cancer Biology, Durham, NC, [2]Duke University, Center for Genomic and Computational Biology, Durham, NC, [3]Duke University, Department of Biostatistics and Bioinformatics, Durham, NC

Most eukaryotic transcription factors (TFs) are part of large protein families, with several TF family members (i.e. paralogous TFs) being expressed at the same time in the cell but targeting different sets of genes and performing different regulatory functions. Closely related TFs, with amino acid similarity of 70% or more in the DNA binding domain (DBD), are generally believed to have identical DNA binding specificities. However, their in vivo genomic binding patterns are markedly different. Currently, we do not have a good understanding of the general molecular mechanisms by which TFs with highly similar DBDs select distinct in vivo targets.

We show that, in general, closely related TFs interact differently with their putative genomic targets even in vitro, in the absence of any additional factors. Our study is focused on eleven paralogous factors from 4 protein families: bHLH, E2F, ETS, and RUNX. For each pair of related TFs, we used genomic-context protein-binding microarray (gcPBM) assays to compare the binding affinities of the two factors for ~25,000 putative genomic binding sites. We found that for most pairs of paralogous TFs, the two factors interact differently with their genomic sites in vitro, despite having identical PWMs. The only two exceptions were: (1) E2F1 and E2F3, which play similar regulatory roles and can partially substitute for each other in the cell, and (2) Runx1 and Runx2, which are typically not expressed at the same time in the cell; in addition, Runx2 is known to compensate for the loss of Runx1 in leukemia cells, which is consistent with our finding that these paralogous TFs have identical specificities.

The way in which paralogous TFs differ is specific to each protein family: E2F1 and E2F4 prefer the same core GCGC/GCGG and differ in their flanking preferences for high affinity sites, ETS factors ETS1 and ELK1 different in specificity for medium and low affinity sites, while bHLH factors c-Myc and Mad1 prefer different flanks for their highest affinity site CACGTG, and differ significantly in their affinity for alternative cores CACATG/CACGCG. Overall, we believe that differences in genomic binding specificity between paralogous TFs are due both to direct recognition of DNA bases in the core binding sites (i.e. base readout), and to indirect recognition of different structural features in the flanking regions (i.e. shape readout).

Importantly, the differences in intrinsic binding specificity between paralogous TFs, as identified in vitro by gcPBM, can partly explain differential in vivo binding, measured by ChIP-seq. While we cannot expect the in vitro specificities of paralogous TFs to completely explain their in vivo binding patterns, our work shows that intrinsic preferences of TFs for genomic sites represent an important mechanism by which closely related factors achieve their regulatory specificity.

LOCAL SEQUENCE CONTEXT DETERMINES AP-1 BINDING SITE ACTIVITY.

Hemangi G Chaudhari, Barak A Cohen

Washington University in St. Louis, Genetics, St. Louis, MO

The AP-1 (Activator Protein-1) family of proteins regulates important cellular processes such as proliferation, apoptosis, and differentiation. Consensus AP-1 binding sites are found in the 5-kb upstream regions of 40% of all human genes. However, only a tiny fraction of these sites are induced in any cell type. What distinguishes functional AP-1 sites from the large number of non-functional sites? We seek to discover features in the local flanking sequences that specify functional AP-1 sites. We identified a set of twenty AP-1 dependent regulatory elements in K562 cells that drive HIGH or LOW expression in a gene reporter assay. Importantly, all the elements have an identical consensus AP-1 binding site in the center of their sequence. To discover features in the flanking sequence we made all possible single base variants of the elements and measured their activity using CRE-seq, a massively-parallel reporter gene assay. We also dissected AP-1 mutant versions of these elements in which the AP-1 site was mutated in order to determine interactions between AP-1 sites and their flanking sequence. The majority of the variants had similar effects on activity in both wild-type and mutant backgrounds, suggesting these features work independently of AP-1. A small fraction of variants drove differential activity between wild-type and mutant backgrounds and likely mark sequence features that specify functional AP-1 sites. I will discuss features in the local flanking sequence such as DNA shape parameters, extended motifs, activators and repressor motifs that explain why only a small subset of genomic AP-1 sites are functional.

GENOMIC CONTROL OF ENDODERM-SPECIFICALLY EXPRESSED POSTERIOR HOX GENE

Miao Cui, Isabelle Peter, Eric Davidson

California Institute of Technology, Biology and Biological Engineering, Pasadena, CA

Transcriptional Factors (TFs) expressed during early embryonic development can interpret maternal and signaling inputs to achieve lineage-specific expression. However, those maternal and signaling inputs are often broadly and dynamically expressed. Thus, understanding the instructions for the specific expression of embryonic TFs in the genome is particularly interesting. Here we chose to study a posterior Hox gene, *hox11/13b*, in sea urchins embryos. Sea urchin Hox11/13b is crucial for the endoderm specification. Its spatial expression is highly regulated and exhibits two phases. In the early phase, it is initially expressed in the endomesoderm and soon restricted to the anterior endoderm. In the Later phase, its transcripts level is stabilized and exclusive to the posterior endoderm. In order to decode its genomic control, we conduct a cis-regulatory analysis. We scanned a ~150kb genomic region for active enhancers across extensive developmental stages using high-through reporter system. We discovered the activity in the early phase comes from an intronic fragment call module A. Further dissecting on module A found that it is operated by an "AND" logic: its expression depends on ETS and EVE for activation and TCF/β-catenin for spatial restriction. Interestingly, although a BAC encompassing the entire scanned region is able to drive the expression in both early and late phases, none of the fragments in the screen is active in the late phase. Surprisingly, we discovered that the late phase expression requires both module A and a fragment that is ~1kb away called module B. Thus, neither module alone is sufficient to drive the late expression, but they are both necessary. We also identified functional CTCF sites upstream of the basal promoter which are required to loop both modules to the Transcriptional Start Site. In summary, we showed the endoderm-specific expression of *hox11/13b* in sea urchin embryos is regulated by "AND" logic in two tiers: between functional TF sites within the same module for the early phase and between two modules for the late phase. More importantly, the study uncovered an interesting case in which one cis-regulatory module turns active only in the context of another module. This phenomenon is contrary to what is well accepted that enhancers are highly modular.

INTEGRATION OF REGULATORY INFORMATION AT PROMOTERS THROUGH COMBINATORIAL CONTROL OF THE TRANSCRIPTIONAL CYCLE

Clarissa Scholes[1], Angela H DePace[1], Alvaro Sanchez[2]

[1]Harvard Medical School, Department of Systems Biology, Boston, MA,
[2]Harvard University, Rowland Institute at Harvard, Cambridge, MA

Combinatorial regulation of gene expression by multiple transcription factors (TFs) enables cells to carry out sophisticated computations that are key to cellular and developmental decision-making. The rate of transcription is determined by the action of TFs that bind to cis-regulatory sequences of genes. How is the information contained in multiple TF binding sites integrated to dictate the rate of transcription, and thereby the level of gene expression? The dominant model is that direct physical interactions between different TFs combinatorially recruit (or inhibit the recruitment of) each other and RNA polymerase (RNAP) to the promoter. Here we explore an alternative model, where combinatorial gene regulation results from TFs working on different kinetic steps of the transcription cycle. We present a quantitative framework that links the effect of TFs on specific kinetic steps to the overall rate of transcription. This framework clarifies the null hypotheses for independent action of TFs under the recruitment model and our kinetic model. We show how combinatorial kinetic control of the transcription cycle can be used to implement a wide range of analog and Boolean computations, analogous to those accomplished by recruitment of TFs and RNAP. Our results emphasize the importance of deciphering the function of TFs beyond activation and repression, the role of the basal promoter in processing regulatory information, and suggest qualitative explanations for the flexibility of regulatory evolution.

FUNCTIONAL CHARACTERIZATION OF GENE REGULATORY ELEMENTS

Fumitaka Inoue[1,2], Kircher Martin[3], Chenling Xiong[1,2], Beth Martin[3], Daniela M Witten[4], Gregory M Cooper[5], Jay Shendure[3], Nadav Ahituv[1,2]

[1]UCSF, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA, [2]UCSF, Institute for Human Genetics, San Francisco, CA, [3]University of Washington, Department of Genome Sciences, Seattle, WA, [4]University of Washington, Department of Biostatistics, Seattle, WA, [5]HudsonAlpha Institute for Biotechnology, HudsonAlpha Institute for Biotechnology, Huntsville, AL

Nucleotide variation in gene regulatory elements is a major determinant of human phenotypes. Despite continual progress in the cataloging of these elements, little is known about the code and grammatical rules that govern their function. Deciphering the code and their grammatical rules will enable high-resolution mapping of regulatory elements, accurate interpretation of nucleotide variation within them and the design of sequences that can deliver molecules for therapeutic purposes. To this end, we are using massively parallel reporter assays (MPRAs) to simultaneously test the activity of thousands of gene regulatory elements in parallel. By designing MPRAs to learn regulatory grammar or to carry out saturation mutagenesis of every possible nucleotide change in disease causing gene regulatory elements, we are increasing our understanding of the phenotypic consequences of gene regulatory mutations.

QUANTITATIVE CHARACTERIZATION OF THE INFLUENCE OF CONTEXT ON DEVELOPMENTAL ENHANCER ACTIVITY AND ROBUSTNESS WITH CELLULAR AND TEMPORAL RESOLUTION

Amanda L Zacharias, Teddy D Lavon, Elicia Preston, John I Murray

Perelman School of Medicine, University of Pennsylvania, Genetics, Philadelphia, PA

Signaling pathways achieve specificity in target gene activation through the dependence on additional inputs, commonly referred to as "context", such cell-specific transcription factors and appropriate chromatin states. To evaluate whether quantitative differences in activity of the Wnt signaling pathway could also contribute to context-specific regulation, we measured the nuclear localization of the Wnt effectors, TCF and β-catenin, in all cells throughout development in *C. elegans* embryos. We found that the effect of Wnt compounds over successive exposures: we observed significantly increased nuclear localization of TCF and β-catenin in Wnt-signaled cells whose parents had also received a Wnt signal. This trans-generational "memory" of Wnt signaling influences target gene regulation suggesting that the level of signaling pathway activity can act an additional form of context.

To better understand how context information is encoded in Wnt target enhancers, we are investigating the dependence of enhancer activity on the organization and affinity of binding sites for TCF and other co-regulatory, cell-specific transcription factors. We identified 20 novel targets of Wnt signaling in *C. elegans* embryonic development. This set of targets is unbiased with regards to the identity of the co-regulatory transcription factors. We used bioinformatics to identify 88 putative Wnt target enhancers and created a pipeline to test their activity in embryos.

Transgenic worms carrying a fluorescent enhancer reporter construct and a ubiquitously expressed GFP-histone are imaged using time-lapse confocal microscopy with ~1 minute resolution. Image analysis software identifies the nuclei and measures the reporter intensity, generating quantitative expression data for each individual cell at each timepoint from a single-celled zygote to a 600-cell elongating embryo. Automated lineage reconstruction generates cell identities that are mapped onto the invariant *C. elegans* lineage to create expression profiles for each enhancer that can be directly compared. We used this approach to characterize the robustness of expression patterns driven by the same enhancer in different transgenic strains and in different embryos of the same strain. We found consistent expression in cells presumed to be the "true" location of enhancer activity, but also ectopic expression that varied in frequency between strains. These results provide a foundation to study the importance of context in enhancer function and robustness quantitatively with cellular and temporal resolution.

MAPPING CAUSAL REGULATORY INTERACTIONS ON A GLOBAL SCALE THROUGH DYNAMIC PERTURBATIONS IN A EUKARYOTIC SYSTEM

Dave Hendrickson[1], Bernd Wranik[1], Yuko Arita[2], David Botstein[1], Charles Boone[2], R. Scott McIsaac[1]

[1]Calico Life Sciences, LLC, South San Francisco, CA, [2]University of Toronto, Donnelly Centre, Toronto, Canada

Organisms rely on complex networks of regulatory interactions to appropriately adapt to environmental cues and thrive. How do we identify causal interactions within these networks on a global scale? We will describe a synthetic system that enables perturbing gene expression levels in a switch-like fashion in yeast with single-gene precision. We will then describe the construction of a collection of strains in which every gene in the yeast genome is made inducible. Finally, we will show that our system allows us to distinguish direct from indirect regulatory connections in the yeast transcription factor network, identify novel binding motifs, and elucidate instances of combinatorial regulation.

MULTIPLEX ENHANCER-REPORTER ASSAYS REVEAL UNSOPHISTICATED P53 ENHANCER LOGIC

Annelien Verfaillie, Dmitry Svetlichnyy, Stein Aerts

University of Leuven, Department of Human Genetics, Leuven, Belgium

Transcription factors establish and maintain the specific transcriptome of a cell by binding to genomic regulatory regions, thereby regulating the transcription of their target genes. Like many transcription factors the DNA sequence-specific binding preferences of p53 are known. However, it remains largely unclear what distinguishes functional enhancers from other bound genomic regions that have no regulatory activity. In addition, the genome is scattered with seemingly perfect recognition sequences that remain unoccupied. To disentangle the rules of genome-wide p53 binding, we employed two complementary techniques of multiplex enhancer-reporter assays, one using barcoded reporters and the other using enhancer self-transcription. We compared the activity of more than one thousand candidate p53 enhancers under loss and gain of p53 conditions and identified several hundred high-confidence p53 responsive enhancers. Strikingly, the large majority (99%) of these target enhancers can be characterized and distinguished from negative sequences by the occurrence of a single p53 binding site. By training a machine learning classifier on these data, and integrating the resulting genome-wide predictions with fifteen publicly available human p53 ChIP-seq data sets, we identified a consensus set of 1148 functional p53 binding sites in the human genome. Unexpectedly, this direct p53 cistrome is invariably used between cell types and experimental conditions, while differences between experiments can be largely attributed to indirect non-functional binding. Our data suggest that direct p53 enhancers function in a context-independent manner and do not contain obvious combinatorial complexity of binding sites for multiple transcription factors. They represent a class of unsophisticated cell-autonomous enhancers with a single binding site, distinct from complex developmental enhancers that integrate signals from multiple transcription factors. This suggests that context-dependent regulation of p53 target genes is not encoded in the p53 enhancer, but at different upstream or downstream layers of the cell's gene regulatory network.

UNRAVELING HUMAN PROMOTER REGULATION BY AN ULTRA-HIGH THROUGHPUT FUNCTIONAL ASSAY

Joris van Arensbergen[1], Vincent D FitzPatrick[2], Marcel de Haas[1], Ludo Pagie[1], Harmen J Bussemaker[2], Bas van Steensel[1]

[1]Netherlands Cancer Institute, Division of Gene Regulation, Amsterdam, Netherlands, [2]Columbia University, Department of Biological Sciences and Department of Systems Biology, New York, NY

Genomic patterns of transcription are driven by proximal promoters, which often interact with nearby cis-regulatory modules such as enhancers, as well as the local chromatin environment. To dissect causality in this complex regulatory system, we developed a method named SuRE to survey ~100 million sequences (each ~0.2-2kb in size) along the entire human genome for their ability to drive transcription in the absence of other regulatory elements.

The resulting "autonomous promoter activity" maps explain ~35% of the genomewide variance in endogenous gene expression, indicating that a substantial proportion of gene regulation is encoded in proximal sequences and independent of genomic and chromatin context. Furthermore, we find that autonomous antisense and sense transcription are both predominantly driven by the same ~150bp region upstream of the transcription start site. Enhancers also generally exhibit autonomous transcription initiation activity, which correlates with endogenous enhancer strength in the same cell type.

Thousands of genes reside inside lamina-associated domains (LADs). Most of these genes are inactive, which has led to the proposal that they are repressed by the heterochromatic state of LADs. Our SuRE data provided an ideal opportunity to test this hypothesis in a genomewide manner. Surprisingly, most promoters still do not drive expression in the SuRE assay, where they have been taken out of their LAD context. These promoters must lack the proper cis-regulatory elements or trans-acting factors. A subset of LAD promoters do exhibit increased activity when taken out of their native LAD context. We conclude that two complementary mechanisms explain the global low activity of genes in LADs: (1) LAD genes have low autonomous promoter activity and (2) transcription is poorly supported by LAD chromatin.

Together, these results illustrate that our SuRE method offers new opportunities to conduct genomewide functional studies of regulatory elements in large genomes.

RNA Pol II MODELING PREDICTS ACTIVE TRANSCRIPTION FACTORS

Joseph G Azofeifa[1], Mary A Allen[2], Jonathan Rubin[3], Josephina Hendrix[4], Timothy Read[2,4], Robin D Dowell[1,2,4]

[1]University of Colorado, Computer Science, Boulder, CO, [2]University of Colorado, BioFrontiers Institute, Boulder, CO, [3]University of Colorado, Chemistry and Biochemistry, Boulder, CO, [4]University of Colorado, Molecular, Cellular and Developmental Biology, Boulder, CO

Enhancers are known to contain numerous transcription factor binding sites as well as short unstable transcripts (eRNAs), but the precise nature of the relationship between transcription factors and eRNAs remains uncharacterized. We have developed a detailed mathematical model of the behavior of polymerase. Using this model to analyze nascent transcription data (both global run-on (GRO-seq) and precision run-on (PRO-seq)) we have uncovered a tight relationship between the motif of active transcription factors and the origin of enhancer transcripts. Given this relationship, we can predict which transcription factors are active in a given cell type based entirely on the pattern of enhancer RNAs transcribed. Furthermore, our results shed light into the relationship between pioneering transcription factors and subsequent activity by stimulus responding transcription factors.

STRUCTURAL INSIGHT INTO ADAPTIVE DNA BINDING BY A MULTI-PROTEIN HOX COMPLEX OBTAINED THROUGH SELEX-SEQ ANALYSIS

Judith F Kribelbauer[1,2], Namiko Abe[3], Gabriella D Martini[1], Chaitanya Rastogi[1,2,4], Richard S Mann[2,3], Harmen J Bussemaker[1,2]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]Columbia University Medical Center, Department of Systems Biology, New York, NY, [3]Columbia University Medical Center, Department of Biochemistry and Molecular Biophysics, New York, NY, [4]Columbia University, Department of Applied Physics and Applied Mathematics, New York, NY

An emerging theme in regulatory genomics is the existence of multiple DNA binding modes for multi-protein transcription factor complexes. Our limited mechanistic and quantitative understanding of such allosteric behavior is an impediment to our ability to predict binding and gene regulation from sequence. To address this problem, we investigated how Hox proteins achieve target specificity through simultaneous complex formation with the homeodomain co-factors Homothorax (Hth/Meis) and Extradenticle (Exd/Pbx). We combined the high-throughput DNA binding assay SELEX-seq (Slattery et al., Cell, 2011) with novel computational analysis methods to obtain detailed structural insight into adaptive DNA binding. We found that distinct configurations for the Hth-Exd-Hox-DNA complex occur that vary with respect to distance and relative orientation between homeodomain binding sites. The DNA sequence of the spacer between binding sites impacts the thermodynamic stability of the complex, even though this stretch of DNA is not expected to be directly contacted by these proteins. We used generalized linear models based on DNA base and shape features to dissect the effect of spacer sequence on the overall binding free energy across different binding modes. Our analyses indicated that Exd and Hth both read out the shape of the DNA minor groove outside their known binding sites. We confirmed this by performing follow-up experiments with mutated proteins. Substitutions that neutralize positively charged residues in the N-terminal arm of both Exd and Hth affected the binding affinity and shape readout of selected DNA molecules, demonstrating that these residues are crucial for stability. Together these findings suggest that N-terminal arm-mediated shape recognition is not only important for the “latent specificity” of Hox proteins, as has been previously shown, but represents a more general readout mechanism for homeodomain proteins. We speculate that multi-domain transcription factor complexes use this mechanism to read complex DNA topographies that emerge from the precise arrangement of multiple binding sites.

DOUBLE STRANDING AGILENT DNA MICROARRAYS WITH 5mC AND 5hmC—EFFECTS ON TRANSCRIPTION FACTOR BINDING.

Syed Khund-Sayeed[1], Ximiao He[1], Matthew T Weirauch[2], Charles Vinson[1]

[1]National Cancer Institute, Laboratory of Metabolism, Bethesda, MD, [2]University of Cincinnati College of Medicine, Center for Autoimmune Genomics and Etiology, Division of Biomedical Informatics and Division of Developmental Biology, Cincinnati, OH

Mammalian cytosine methylome data indicate that cytosine can be methylated outside of CG dinucleotides especially in brain and stem cells. To examine how this may change the sequence-specific DNA binding of transcription factors, we have used Agilent DNA microarrays and done the double-stranding reaction with 5 methylcytosine (5mC) or 5 hydroxymethylcytosine (5hmC) creating DNA with a single strand containing cytosine modifications as occurs in vivo. We have focused on mouse B-ZIP and B-HLH families. 5mC and 5hmC can each promote DNA binding or inhibit DNA binding. Details will be presented at the meeting.

READING AND WRITING DNA METHYLATION

Dirk Schübeler

Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

How is chromatin and DNA methylation involved in gene regulation? We are using mammalian stem cell models to monitor the epigenome and its dynamics in an unbiased way and to identify its dependency on DNA sequence. Our goal is to generate regulatory, which we test in cellular models by genetic perturbation and genome editing approaches.

To create functional genomic binding maps we use a biotin tagging approach that we first utilized to map all MBD domain proteins and a set of isoforms (Baubec et al., Cell 2013). We further determined chromosomal binding and site-specific activity of the mammalian de novo DNA methyltransferases DNMT3A and DNMT3B. While binding explains sites of enzymatic activity, DNMT3B furthermore preferentially methylates active genes. This targeting requires SETD2-mediated methylation of lysine 36 on histone H3 revealing reveal how sequence and chromatin cues guide de novo methyltransferase activity to ensure methylome integrity (Baubec et al. Nature 2015).

In order to identify transcription factors that are sensitive to DNA methylation we have mapped DNaseI hypersensitive sites that change upon global loss of DNA methylation. Among others this identified NRF1 as a TF that occupies many additional sites in the unmethylated genome. Importantly, restoring de novo methyltransferases initiates remethylation at these sites and ablates NRF1 binding. This illustrates that binding of DNA methylation-sensitive transcription factors critically depends on other factors to induce local hypomethylation (Domcke, Bardet et al., Nature 2015).

These results highlight the interplay and dynamics between the machineries that set and read DNA methylation. The furthermore illustrate how transcription factors are involved in causing local sites of reduced methylation and how other factors require this local hypomethylation for binding.

DIFFERENTIAL SPECIFICITY AND ENERGY LANDSCAPES (DI-SEL) UNMASK SEQUENCE PREFERENCES OF TRANSCRIPTION FACTORS THAT BIND INDISTINGUISHABLY TO CONSENSUS DNA MOTIFS

Aseem Z Ansari

University of Wisconsin-Madison, Genome Center of Wisconsin & Department of Biochemistry, Madison, WI

Sequence specificity landscapes of DNA-Protein interactomes acquired via diverse experimental platforms reveal that high affinity binding sites are identified with remarkable congruence and fidelity. Similarly, most computational methods generate sequence motifs that maximize information content. However, biologically relevant specificity differences between transcription factors are increasingly observed at sites that display "sub-maximal" affinity. We therefore developed Differential Specificity and Energy Landscapes (DiSEL) to compare interactomes and motifs obtained by cutting-edge methods and to unmask nuanced sequence preferences displayed by different transcription factors. DiSELs readily reveal binding sites that are differentially preferred even by highly homologous transcription factors that "read" DNA using identical amino acid side chains. In one example, homolog-specific side chains packed against the protein core, rather than those facing DNA, reposition the DNA recognition helix and alter the energetic dependence on specific positions within the binding site. Our results suggest that widely overlooked allosteric effects contribute significantly to differential binding and biological function of homologous members of transcription factor families.

CISTROME PROFILING PINPOINTS COOPERATIVE REGULATION OF ANTI-INFLAMMATORY NODES BY THE GLUCOCORTICOID RECEPTOR AND NF-κB

Vineela Kadiyala*[1], Sarah K Sasse*[1], Tzu L Phang[2], Anthony N Gerber[1,2]

[1]National Jewish Health, Medicine, Denver, CO, [2]University of Colorado, Medicine, Aurora, CO

The glucocorticoid receptor (GR) is a nuclear receptor, the only major family of transcription factors that have been successfully targeted pharmacologically. Whereas DNA binding to palindromic sites by dimeric GR is energetically favorable in comparison to interactions between monomeric GR and DNA, direct inhibition of the activity of pro-inflammatory transcription factors by ligand-activated monomeric GR has long been viewed as integral to glucocorticoid (GC)–based repression of inflammation. We previously defined surprising synergistic cooperation between GR and NF-kB at a novel enhancer of A20 (TNFAIP3), a potent negative feedback regulator of inflammation. We have now applied CRISPR-based deletion to establish the importance of this cooperative regulatory element in mediating anti-inflammatory effects of GCs. Moreover, we have applied ChIP-seq to cells treated with dexamethasone (dex, a potent GC), TNF, or both to define additional sites of cooperation between GR and NF-kB (p65) that drive regulation of anti-inflammatory genes; the majority of these genes had not previously been associated with therapeutic effects of GCs. Energetically favorable sites for dimeric GR recruit p65 across the genome, frequently in association with increased RNAP2 occupancy at distal regulatory regions, suggestive of enhancer activation and looping. In contrast, reduction in p65 occupancy with dex + TNF treatment tended to be promoter-proximal and only sporadically associated with changes in GR occupancy in comparison to TNF treatment alone, with no clear enrichment for either monomeric, dimeric, or so-called negative GR binding sequences observed at these sites. Repression by GR was thus most consistent with perturbed transcriptional complex assembly in conjunction with repressive effects of targets of GR-p65 cooperation, such as A20, as opposed to a traditional DNA-coded response in which interactions between GR and defined binding sequences dictate repressive regulatory outcomes. The complex opposing effects of positive and negative feedback on NF-kB activity, coupled with diversity in the DNA sequences associated with different occupancy patterns of GR and p65, may have obfuscated the importance of GR-p65 cooperative gene induction in other genome-wide studies. Our data are consistent with a hierarchical model for anti-inflammatory responses to GCs in which a limited set of primary effector nodes are subject to intense regulatory control and associated expression changes, with many additional sites of GR occupancy serving to fine-tune expression of secondary effectors.

SYSTEMS-LEVEL ANALYSIS OF THE REGULATION AND FUNCTION OF P53 DYNAMICS

Joshua Porter, Brian Fisher, Eric Batchelor

National Cancer Institute, National Institutes of Health, Laboratory of Pathology, Bethesda, MD

Living cells use complex signaling pathways to detect environmental stimuli and generate appropriate responses. As methods for quantifying intracellular signaling have improved, several signaling pathways have been found to transmit information using signals that pulse in time. The transcription factor p53 is a key tumor suppressor and stress-response regulator that exhibits pulsatile dynamics. In response to DNA double-strand breaks, the concentration of p53 in the cell nucleus increases in pulses with a fixed amplitude, duration, and period; the mean number of pulses increases with DNA damage. p53 regulates the expression of over 100 target genes involved in a range of cellular stress responses including apoptosis, cell cycle arrest, and changes in metabolism. p53 pulsing directly impacts p53 function: altering p53 dynamics by pharmacologically inhibiting p53 degradation changes patterns of target gene expression and cell fate. While p53 pulsing serves an important signaling function, it is less clear what it accomplishes mechanistically. Here we show that p53 pulses generate different temporal patterns of target gene expression, which can be predicted by the mRNA decay rates of the genes and are not specific to gene function. Moreover, using transcriptional profiling of 1680 single cells, we show that p53 coordinates expression of a subset of its target genes in a time-varying manner and that p53 pulsing lessens this coordination compared with the response to a constant high p53 level. These results help delineate how p53 orchestrates the complex DNA damage response and give insight into the function of an increasing number of pulsatile signaling pathways.

DYNAMIC ACTION OF REPRESSORS ON NON-TARGET ELEMENTS CHALLENGES ASSUMPTIONS FOR GENOME-WIDE IDENTIFICATION OF ENHANCERS

Kurtulus Kok[1,2], Ahmet Ay[3], Li M Li[4], David N Arnosti[1,2]

[1]Michigan State University, Biochemistry and Molecular Biology, East Lansing, MI, [2]Michigan State University, Program in Genetics, East Lansing, MI, [3]Colgate University, Mathematics and Biology, Hamilton, NY, [4]Janelia Research Campus, 19700 Helix Drive, Ashburn, VA

Metazoan transcriptional repressors regulate chromatin through diverse histone modifications. Using a precisely regulated embryonic system, we studied the activities of the conserved HES family repressor Hairy, analyzing histone marks and gene expression on a genome-wide scale. Consistent with a previously proposed "spreading" model for regulation, this long-range repressor mediates histone deacetylation and demethylation in large blocks, with highly context-specific effects on target genes. Most strikingly, Hairy exhibits biochemical activity on many loci that are uncoupled to changes in gene expression, including constitutively active and inactive genes. Rather than representing inert binding sites, as suggested for many eukaryotic factors, many regions are targeted "errantly" by Hairy to modify the chromatin landscape. The involvement of conserved core corepressors in the Hairy mechanism of action suggests that these effects are likely to be widely shared by core metazoan transcriptional regulators. Our findings emphasize that identification of active cis regulatory elements must extend beyond the survey of prototypical chromatin marks. We speculate that this errant activity may provide a path for creation of new regulatory elements, facilitating the evolution of novel transcriptional circuits.

CELLULAR ANTHROPOLOGY CAN BE USED TO IDENTIFY CIS-REGULATORY DIVERGENCE IN THE HUMAN AND CHIMP NEURAL CREST

Sara L Prescott[1], Rajini Srinivasan[1], Antoine Zalc[1], Hannah Long[1], Maria Carolina Marchetto[2], Irina Grishina[3], Inigo Narvaiza[2], Licia Selleri[3], Fred H Gage[2,4], Tomek Swigut[1], Joanna Wysocka[1,5,6]

[1]Stanford University School of Medicine, Depts. of Chemical and Systems Biology and Developmental Biology, Stanford, CA, [2]Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, CA, [3]Weill Cornell Medical College, Dept. of Cell and Developmental Biology, New York, NY, [4]University of California, San Diego, Center for Academic Research and Training in Anthropogeny (CARTA), La Jolla, CA, [5]Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA, [6]Stanford University School of Medicine, Institute of Stem Cell Biology and Regenerative Medicine, Stanford, CA

Cis-regulatory changes play a central role in morphological divergence, yet the regulatory principles underlying emergence of human traits remain poorly understood. Recently, we used epigenomic profiling from human and chimpanzee cranial neural crest cells to systematically and quantitatively annotate divergence of human and chimp craniofacial cis-regulatory landscapes. We found that this epigenomic divergence is often attributable to genetic variance within TF motifs at orthologous enhancers, and found global correlations between chromatin marks associated with active enhancer states and mutations that disrupt motifs for both known activators and, interestingly, putative repressors. From this analysis, we identified a long novel motif that was both highly enriched at species-biased sites and was the most predictive of activity biases, allowing us to use extinct ancestral genomes for our outgroup comparisons to deduce how these enhancers may have functionally diverged over recent evolutionary periods. We further explored properties of this cis-regulatory change, revealing the role of particular retroelements, uncovering broad clusters of species-biased enhancers near genes associated with human facial variation, and demonstrating that cis-regulatory divergence is linked to quantitative expression differences of crucial neural crest regulators. We are now exploring whether and how these changes in the epigenomic landscapes are associated with changes in higher order chromatin organization. Our work provides a wealth of candidates for future evolutionary studies and demonstrates the value of ‘‘cellular anthropology,’’ a strategy of using in-vitro-derived embryonic cell types to elucidate both fundamental and evolving mechanisms underlying morphological variation in higher primates.

FINDING HIERARCHICAL TOPOLOGICAL DOMAINS USING A UNIFIED PROBABILISTIC MODEL OF HI-C DATA

Gil Ron, Tommy Kaplan

The Hebrew University, School of Computer Science and Engineering, Jerusalem, Israel

Technological advances in the past decade allow us to explore the structure of the DNA in vivo, using proximity-ligation methods such as 3C and Hi-C. With Hi-C, DNA-DNA physical interactions between two loci are mapped along the genome in high-throughput.

Analysis of Hi-C data in human and mouse cells identified the packaging of DNA in 200Kb to 4Mb-long topological domains (TADs), suggesting a typical organization of chromosomal DNA in localized structures of concentrated DNA. There are several computational methods for identifying TADs along the genome, e.g. by using HMM segmentation of interaction directionality (Dixon et al, 2012). These models are often skewed by the high number of short-range interactions, leading to fragmented TAD calling and insensitivity to higher-order hierarchies of topological domains.

Here, we propose a new computational approach for the identification of topological domains. Our statistical framework uses two probabilistic models to calculate the likelihood of the observed interactions within and outside of topological domains. Each model learns different power law parameters, to capture the different frequency of Hi-C interactions for each genomic distance. Then, we compute the log-posterior ratio for each genomic region, and use a dynamic programming algorithm to optimally segment the genome into topological domains, and inter-domain boundaries regions. Finally, by dynamically increasing the penalty score for each TAD, our algorithm repeatedly “merges” neighboring TADs, to form an optimal hierarchical structure of topological domains.

As we show, this unified probabilistic model of Hi-C data, identifies both basic and complex hierarchical domains along the genome, and facilitates differential studies of genome organization in various conditions. Moreover, by reconstructing Hi-C data from our hierarchical model, we obtain a robust background model, used to highlight and identify over-represented local DNA-DNA interaction within TADs, e.g. between distal regulatory enhancers and the genes they regulate.

EPIGENOMIC MECHANISMS DRIVING hESC CONVERSION INTO SKELETAL MUSCLE CELLS

Sole Gatto[1], Sonia Albini[1], Stefania Dell'Orso[2], Paula Coutinho Toto[1], Vittorio Sartorelli[2], Pier Lorenzo Puri[1]

[1]Sanford Burnham Prebys Medical Discovery Institute, Development, Aging, and Regeneration Program, La Jolla, CA, [2]National Institute of Arthritis, Musculoskeletal and Skin Diseases, NIH, Laboratory of Muscle Stem Cells and Gene Regulation, Bethesda, MD

Direct generation of skeletal myoblasts from human embryonic stem cells (hESC) and formation of three-dimensional contractile structures for disease modeling in vitro are the current challenges of regenerative medicine. As loss of pluripotency and lineage commitment are events coordinated by epigenetic changes that alter the chromatin organization, the understanding of the epigenetic network that regulates hESC-commitment towards the myogenic lineage cells is crucial to establish protocols of direct conversion of hESC into skeletal muscle cells. We found that direct conversion of hESC by MyoD into skeletal muscle is precluded because of the absence of BAF60C (encoded by SMARCD3) - the component of SWI/SNF chromatin remodeling complex that mediates SWI/SNF interactions with MyoD. BAF60C upregulation during hESC differentiation into embryoid bodies (hEBs) confers the ability to adopt the myogenic lineage to these cells. Consistently, we found that ectopic expression of BAF60C enables MyoD to directly activate skeletal myogenesis in hESC. The dominant effect of MyoD/BAF60C in hESC relies on the ability to bias the epigenetic landscape toward activation of muscle gene expression, by instructing MyoD positioning and allowing chromatin remodeling at target genes. Interestingly, MyoD ChIP-seq analysis reveals a genome-wide pervasive binding of MyoD in undifferentiated hESC in absence of BAF60C and a more localized binding in presence of BAF60C on the majority of muscle regulatory regions. Moreover, the gene expression data from the RNA-seq reveals a propensity of MyoD alone to activate general transcription in a non-specific way, while the co-expression of BAF60C leads to the repression of alternative pathways, such as cardiac, nervous system, smooth muscle and lymphatic system. These data support the hypothesis that BAF60C instructs MyoD to adopt the optimal chromatin binding profile to promote an epigenetic landscape conducive to the activation of skeletal myogenesis in hESC and identify BAF60C as a key epigenetic determinant of hESC commitment to the myogenic lineage.

EPIGENETIC REGULATION BY MOF CONTAINING COMPLEXES: OLD PLAYERS, NEW FACTS.

Asifa Akhtar

Max Planck Institute of Immunobiology and Epigenetics, Department of Chromatin Regulation, Freiburg, Germany

Our lab is studying the chromatin and epigenetic mechanisms regulated by histone acetylation using evolutionary conserved complexes associated with MOF, a MYST family of histone acetyl transferase. In files and mammals MOF is associated with the MSL and NSL complexes, which are important regulators of gene expression. In flies the MSL complex is well known for regulation of the X chromosome by the process of dosage compensation, while the NSL complex regulates expression of house keeping genes. In mammals, both complexes appear to be involved in regulating diverse cellular processes.
In our recent work, we have used chromosome conformation technologies to study the contribution of nuclear organization in X chromosome regulation. Hi-C and 4C analyses revealed that high affinity sites (HAS), scattered along the X chromosome, are sites for enriched long-range contacts that occur in a sex-independent manner. We also observed that roX1 and roX2 lncRNAs target HAS on the X chromosome in trans, and via spatial proximity, induce spreading of the MSL complex in cis, leading to increased expression of neighboring autosomal genes. We show that the MSL complex regulates nucleosome positioning at HAS, therefore acting locally rather than influencing the overall chromosomal architecture. We therefore propose that the sex-independent, three-dimensional conformation of the X chromosome poises it for exploitation by the MSL complex, thereby facilitating spreading in males. Our progress will be presented.

GENOMIC MECHANISMS OF LSD1 INHIBITORS

John E Lazar[1], R. Scott Hansen[2], Shane Neph[3], Jessica Hallow[3], Jemma Nelson[3], John Stamatoyannopoulos[1,2,3]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]University of Washington, Medicine, Seattle, WA, [3]Altius Institute for Biomedical Sciences, Seattle, WA

Dysregulation of chromatin modifiers is a common feature of cancer development, and therefore these enzymes are attractive targets for therapeutic interventions. However, it is not clear how perturbation of general regulatory factors produces such specific phenotypic effects, especially since the response is highly dependent on the cellular context in which it occurs. In small cell lung cancer (SCLC), inhibition of LSD1—a histone demethylase whose overexpression is associated with an invasive phenotype— causes growth arrest in a subset of cell lines. Directly measuring the changes in regulatory activity upon LSD1 inhibition using DNaseI-Seq identifies the specific cis and trans regulatory factors that mediate the growth response. Enhancers associated with neural differentiation are affected, and specific changes in the chromatin landscape discriminate cells that show growth arrest after LSD1 inhibition from those that are not susceptible. Sequence analysis of the affected regulatory elements identifies the transcription factors (TFs) driving this response, which include both factors that are known to directly recruit LSD1, as well as TFs whose effects appear secondary and driven by changes in TF expression. A regression model based on the basal abundance of the affected TFs partially predicts cell response to drug inhibition, but in light of the changes in TF expression induced by treatment, using TF-TF networks to incorporate information about the cell type specific regulatory interactions between the factors improves predictions of sample response.

RANDOM MONO-ALLELIC REGULATION IN DIPLOID GENOME BY ALLELIC SPECIFIC ATAC-SEQ

Jin Xu*[1], Ava C Carter*[1], Mikael Attia[2], Edith Heard[2], Howard Y Chang[1]

[1]University School of Medicine, Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford, CA, [2]Institut Curie, Genetics and Developmental Biology, Paris, France

* contributed equally

Mono-allelic expression from diploid genome plays an essential role in generating cellular and functional diversity, which has important implications for the onset and severity of human diseases. Genome-wide gene expression assaying has revealed that there are more genes than expected that are subject to mono-allelic expression in mammalian genome. However, known mechanisms for establishing and maintaining random mono-allelic expression apply only to small subsets of particular genes and no unifying mechanism has been discovered.
To understand the mechanism of random mono-allelic regulation, we apply ATAC-seq, a new method that captures the open chromatin regions, in 16 clonal mouse neural progenitor cell lines and 2 mouse embryonic stem cell lines, which was hybrid from high divergent parental mouse lines (129S1 x Cast). With the polymorphism information and our allele specific ATAC-seq pipeline, we are able to distinguish opening signal to each parental genome for 77% open chromatin regions on average. The mono-allelic accessible regions increase from 2% to ~10% during the differentiation from ESCs to NPCs, which is consistent to previous observation by RNA-seq. Combining the allelic accessible patterns in all 16 NPC clones, we get 607 regions, which show mono-allelic accessible from 129S1 genome in at least one clone and mono-allelic accessible from Cast genome in another clone. The distribution of random mono-allelic accessible regions across 16 NPC clones, exhibits a stochastic pattern with less representative of zero allele and exclusive allele. These random mono-allelic accessible regions are mitotically stable over long periods in culture and enriched at promoter regions with Pol2 single by ChIP-seq. The genome-wide profile of random mono-allelic accessibility gives closer epigenetic information to understand the mechanism of how random mono-allelic expression are established and maintained.

COOPERATIVE TRANSCRIPTION FACTOR BINDING DYNAMICS UNDERLIE DIRECT MOTOR NEURON PROGRAMMING

Shaun Mahony[1], Akshay Kakumanu[1], Silvia Velasco[2], Mahmoud Ibrahim[3], Uwe Ohler[3], Esteban Mazzoni[2]

[1]Penn State University, Center for Eukaryotic Gene Regulation, University Park, PA, [2]New York University, Dept of Biology, New York, NY, [3]Max Delbrück Center for Molecular Medicine, Berlin, Germany

Cellular programming approaches allow direct conversion of one cell type into another via induction of particular combinations of transcription factors (TFs). While such approaches hold high potential for enabling regenerative medical applications, we currently know little about how induced cellular programming TFs interact with each other and the established regulatory landscape to transform cellular identity. We have recently shown that the expression of Ngn2, Isl1, and Lhx3 (the NIL factors) efficiently programs spinal motor neurons from mouse embryonic stem cells within 48 hours. To understand how cells transition from rapidly dividing pluripotent stem cells to postmitotic motor neurons, we investigate the dynamics of the transcriptome, chromatin landscape, and programming TF binding during the programming process.

We find that NIL factors bind to both active and inactive genomic loci during early programming, suggesting that their activity is not entirely determined by pre-existing chromatin signatures. However, NIL factor binding also shifts during the programming process; 31% of categorized NIL binding sites are bound only during early programming, while 21% are bound only in the late stages. NIL binding dynamics are correlated with co-temporal increases in histone marks associated with enhancer activity and the timing of nearby gene activation. Therefore, to understand motor neuron programming, we must first understand why NIL factors can override existing chromatin structures at some sites while others are not bound until late in the programming process.

Using a novel DNA motif discovery platform called SeqUnwinder, we characterized sequence features associated with each NIL dynamic binding category (i.e. early-only, constant, late-only) while deconvolving confounding features associated with other categories (e.g. sites active or inactive in pluripotent cells). We characterize several motifs associated with late-only binding and thereby predict the identities of TFs that may cooperate with programming TFs in later programming stages. We demonstrate that these predicted TFs are indeed induced during programming and bind coincidently with programming factors at late-only sites. Taken together, our results suggest a model of motor neuron programming in which programming TFs activate motor neuron specific genes in multiple waves, dependent on cooperative interactions with other TFs activated during the programming process.

PROTEIN−DNA BINDING IN THE ABSENCE OF SPECIFIC BASE-PAIR RECOGNITION SIGNIFICANTLY AFFECTS TRANSCRIPTIONAL REGULATION

Ariel Afek[1,2], Masahiko Imashimizu[3], Joshua L Schipper[2], Raluca Gordan[2], David B Lukatsky[1]

[1]Ben-Gurion University of the Negev, Chemistry, Be'er Sheva, Israel, [2]Duke University, Biostatistics and Bioinformatics, Durham, NC, [3]National Cancer Institute, Center for Cancer Research, Frederick, MD

Protein-DNA interactions are usually considered to be energetically dominated by specific structural (enthalpic) interactions, while the dynamic, nonspecific nature of bound proteins is usually not taken into account.

Based on statistical mechanics modeling, we have identified a new nonspecific protein−DNA binding mechanism induced by repetitive DNA sequences in the absence of specific base-pair recognition [1]. We experimentally demonstrated that this mechanism significantly influence proteins binding preferences and RNA polymerase (RNAP) pausing during transcription:

1. We used high-throughput protein−DNA binding assays to measure the binding levels and free energies of binding for several human transcription factors (TFs) to tens of thousands of short DNA sequences containing the same specific motif within varying repetitive flanks. We showed that nonspecific repetitive DNA sequences, when present outside of specific TF binding sites, significantly affect TF−DNA binding preferences [1].

2. We have also found that nonspecific repetitive DNA sequences in the vicinity of pause-inducing elements of RNAP, significantly enhance RNAP pausing via stabilizing RNAP backtracking [Imashimizu, et al., In preparation].

Contrary to the case of specific protein−DNA binding, which stems from a single protein−DNA binding site, the nonspecific effect characterized in our studies is nonlocal, as it stems from multiple nonspecific interactions between proteins and repeated DNA sequence patterns.

Our results provide several new insights into the biophysical mechanisms of transcriptional regulation, and are likely to improve our ability to predict how genetic sequence variations in regulatory DNA regions influence protein binding and gene expression programs in living cells.

[1] Afek Ariel, et al. "Protein− DNA binding in the absence of specific base-pair recognition." Proceedings of the National Academy of Sciences 111.48 (2014): 17140-17145.

CHARACTERIZING THE DNA-BINDING DIFFERENCES BETWEEN IRF TRANSCRIPTION FACTORS.

Kellen K Andrilenas, Brandon Leung, Jesse Kurland, Trevor Siggers

Boston University, Department of Biology, Boston, MA

The Interferon Regulatory Factors (IRFs) are essential regulators of the innate immune response to viruses. IRFs 3, 5, and 7 drive type-1-interferon and cytokine production in a cell-type and stimulus-specific manner suggesting regulatory complexity. IRF-3, -5 and -7 bind DNA as phosphorylation-activated dimers and are classically described as binding a doublet of the canonical IRF binding site: 5'-AANNGAAA-3'. IRF-3/5/7 have both common and distinct gene targets and the simple canonical motif does not capture the regulatory complexity of IRF-dependent gene expression. Previous studies, both low and high throughput, have refined our understanding of IRF regulation, yet a systematic comparison of DNA binding preferences for active, dimeric IRF complexes has not been performed. To address this, we designed an IRF-specific custom Protein-binding Microarray (PBM) that includes synthetic test sequences as well as genomic binding sites in their native context. Using constitutively active, phosphomimetic IRF-3, 5 and 7 dimers, we identified key DNA binding differences between these factors. Leveraging our custom IRF PBM data, we are generating IRF-paralog specific models of DNA binding in order to understand the complexities of IRF-3/5/7 dependent gene regulation at a genome scale. With our detailed biochemical models we hope to understand the impact of IRF binding-site SNPs and their potential effects on autoimmune disease and immune function.

BAYESIAN MODELING OF CELL SUB-POPULATIONS USING SINGLE-CELL TRANSCRIPTOMIC DATA

Elham Azizi*, Sandhya Prabhakaran*, Ambrose Carr, Dana Pe'er

Columbia University, Biological Sciences & Systems Biology, New York, NY

*authors contributed equally

We propose a hierarchical Bayesian mixture model for inferring sub-populations of cells using single cell RNA-seq data, while also taking into account the experimental and technical variations. We assume that the single cell data is distributed according to a multivariate log Normal with class-dependent mean and covariance parameters. To correct for variations in library size, we build in scaling parameters to adjust the mean and covariance These values are subsequently inferred using the hierarchical Bayesian model which is a flavor/an extension of the well known Dirichlet Process Mixture Model.

Through simultaneous learning of classes and variability, we take advantage of similarity in co-expression structures to correct cell-type-dependent variability. We applied this method to single-cell RNA-seq data from cells in the mouse somatosensory cortex and hippocampal CA1 region from Zeisel et al. 2014. We show that this method improves identification of sub-populations compared to common clustering methods and approaches involving normalization as a pre-processing step. Furthermore, we show the inferred model presents interpretable structures among genes and regulatory interactions that are dependent on cell sub-populations. We also show capabilities of imputing missing transcripts and dropouts through correcting for variability.

Zeisel, Amit, et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq." Science 347.6226 (2015): 1138-1142.

AUPAIRWISE: BIOLOGICALLY FOCUSED RNA-SEQ QUALITY CONTROL USING CO-EXPRESSION

Sara Ballouz, Jesse Gillis

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Woodbury, NY

A principal claim for RNA-sequencing has been greater replicability, typically measured in sample-sample correlations of gene expression levels. Replicability of transcript abundances in this way will provide misleading estimates of the replicability of conditional variation, which is what is of interest in expression analyses. Heuristics which implicitly address this problem have emerged in quality control measures to obtain 'good' differential expression results. However, these methods involve strict filters such as discarding low expressing genes or using technical replicates to remove discordant transcripts, and are costly or simply ad hoc.

As an alternative, we've performed a series of sample and replicate-based analyses of RNA-seq data and show that gene-gene correlations of expression levels across conditions and between replicates are a more useful measure of replicability. Through the re-analysis of reference RNA-seq expression sets, we are able to recapitulate SEQC guidelines for replicability, but using substantially less data. Thus, our gene-level method of replicability allows for flexibility and specific tailoring of the analyses in even small data sets that cannot be achieved with the general recommended filters. We then show that this gene-level replicability of differential activity can be modeled in a co-expression framework, using known co-expressing gene pairs as pseudo-replicates instead of true replicates. We use this as a quality control metric: by modelling the effects of noise that perturbs a gene's expression, we can then measure the aggregate effect of this perturbation on these co-expressing gene-pairs or 'housekeeping interactions'. Perturbing expression by only 5% is readily detectable (AUROC~0.73), which makes this a straightforward method to help customize experiments. We have named this method AuPairWise, and is available as a set of R scripts (github.com/sarbal/AuPairWise).

Our ongoing assessment is to apply AuPairWise as a method of validating methodological pipelines on real data. In particular, we are testing the parameter search space of the RNA-seq alignment STAR, to find which choices gives the output most closely aligned to known biology (as determined through expected housekeeping co-expression), an approach differing substantially from ones using simulated datasets.

DYNAMICS OF ESTROGEN STIMULATED REGULATORY NETWORKS IN BREAST CANCER

Jeanette Baran-Gale, Praveen Sethupathy, Jeremy Purvis

University of North Carolina, Curriculum in Bioinformatics and Computational Biology, Department of Genetics, Chapel Hill, NC

Breast cancer remains a leading cause of cancer and mortality in women worldwide. Estrogen receptor α (ERα) plays a key role in the disease and is both a biomarker of cancer severity and a therapeutic target. Current anti-estrogen therapies inhibit all ERα signaling and thus interfere with ERα's well-studied promotion of cell proliferation and survival. However, studies have demonstrated that ERα also suppresses key transcriptional regulators of metastatic transformation. As such, while current anti-estrogen therapies reduce tumor mass, they may also eliminate safeguards against metastatic transformation. A deeper understanding of the programs activated by estrogen in breast cancer cells is necessary to develop more specific therapies.

The cellular response to estrogen is dynamic and multilayered. ERα binds to estrogen and transcriptionally regulates expression of both coding and non-coding targets. In addition, several ERα-stimulated microRNAs (miRNAs) have been identified that post-transcriptionally regulate both ERα and its targets. Previous studies have identified ERα-responsive genes or miRNAs at single time points following estrogen stimulation, but have not examined the dynamical response to estrogen stimulation. We hypothesize that miRNA networks exert temporal control of ERα signaling, thereby regulating the timing and extent of specific cellular responses to estrogen. To test this hypothesis, we have examined a time course of paired miRNA and RNA sequencing data from MCF7 cells exposed to 10mM estradiol. Integrating these data allowed us to construct a dynamic regulatory map detailing the temporal profile of genes and miRNAs that respond to estrogen stimulation. By studying this map we have: (1) identified temporally regulated genes and miRNAs; (2) predicted miRNA master regulators of ERα-stimulated pathways; and (3) identified candidate therapeutic targets that will interfere with the ERα's mitogenic pathways without eliminating its anti-metastatic functions.

Transcriptome-wide profiling provides an unprecedented systems-level understanding of the dynamics of both coding and non-coding RNAs to the estrogen response in breast cancer cells. Importantly, these data provide a more detailed map of ERα-signaling that could guide future therapeutic strategies for breast cancer.

COMPARATIVE ASSESSMENT OF TRAINING DATA AND COMPUTATIONAL METHODS FOR PREDICTING CELL-TYPE SPECIFIC ENHANCER ACTIVITY AND THE IMPACT OF REGULATORY MUTATIONS

Felix Yu[1], Dongwon Lee[2], Aliso Mo[3,4], Jeremy Nathans[3,4,5,6], Michael A Beer[1,2]

[1]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD, [3]Johns Hopkins University, Department of Molecular Biology and Genetics, Baltimore, MD, [4]Johns Hopkins University, Department of Neuroscience, Baltimore, MD, [5]Johns Hopkins University, Department of Ophthalmology, Baltimore, MD, [6]Johns Hopkins University, Howard Hughes Medical Institute, Baltimore, MD

Most SNPs associated with common human disease by GWAS are intergenic and contribute to disease susceptibility through modulation of enhancer function. To help identify functional SNPs from among those in linkage disequilibrium (LD), we have recently developed a sequence-based model to predict the impact of regulatory variation on enhancer activity (Lee et al., 2015). This framework uses epigenetic data to train a gapped k-mer SVM (gkm-SVM) whose scoring function encodes the relative regulatory importance of individual sequence features in a specific cell-type (Ghandi et al., 2014). We use the change in these sequence feature scores induced by a sequence variant to determine the predicted impact of the variant, a score we call deltaSVM, which accurately identifies validated SNPs from GWAS studies when trained on a relevant cell-type. When trained on DHS and used predict the expression change in massively parallel reporter assays (MPRA), deltaSVM is in good agreement with recent data sets in mouse liver, K562 cells, and HepG2 cells. Here, we show that a gkm-SVM trained on ATAC-seq from mouse retina can predict retinal expression with comparable accuracy to a gkm-SVM trained on retina DHS (Vierstra et al., 2014), using a recently published MRPA in mouse retina (Shen et al., 2015), while training on unrelated cell types has no predictive value, providing strong evidence that ATAC-seq and DHS are of comparable utility in predicting enhancer activity. Finally, we compare the accuracy of deltaSVM to other computational approaches, including PWMs, other kmer-based approaches, and deep neural networks (Zhou and Troyanskaya, 2015), and explore how sensitivity is affected by simultaneously training on multiple cell types.

- Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput Biol 10, e1003711.
- Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. 47, 955–961.
- Shen, S.Q., Myers, C.A., Hughes, A.E., Byrne, L.C., Flannery, J.G., and Corbo, J.C. (2015). Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res. gr.193789.115.
- Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–1012.
- Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 12, 931–934.

TRANSCRIPTOME-WIDE IDENTIFICATION OF MUSASHI-2 TARGETS REVEALS NOVEL REGULATORY MECHANISMS AND FUNCTIONS

Christopher G Bennett*[1], Kent Riemondy*[1], Douglas A Chapnick[2], Eric Bunker[2], Xuedong Liu[2], Rui Yi[1]

[1]University of Colorado, Molecular, Cellular, and Developmental Biology, Boulder, CO, [2]University of Colorado, Chemistry and Biochemistry, Boulder, CO

The Musashi-2 (Msi2) RNA binding protein is a widely expressed, post-transcriptional regulator of cell proliferation. Studies in hematopoietic and gastrointestinal systems revealed an oncogenic potential for Msi2. However, Msi2's function in stratified epithelia progenitor cells, keratinocytes, is largely unknown. Furthermore, the mechanisms that govern Msi2 target binding and regulation are poorly understood. Using Msi2 HITS-CLIP we identified Msi2-bound targets and found a strong preference for 3'UTRs consistent with its role as a post-transcriptional regulator. A search of these Msi2 bound regions found an enrichment for a trinucleotide UAG motif core element. Using this data in conjunction with ribosome profiling and RNA-seq we demonstrate that Msi2 primarily regulates its targets through RNA stability without affecting translation efficiency. A high confidence list of targets was generated from this data and was enriched for genes involved in proliferation, survival, and motility. Indeed, loss of Msi2 shows reduction in cellular proliferation and survival while increasing migration. These findings provide new insights into the molecular mechanisms of Msi2-mediated gene repression and identify a novel function of Msi2 in restricting epithelial cell migration.

* Authors contributed equally

SHADOW ENHANCERS CONTROL GENE EXPRESSION DYNAMICS IN DROSOPHILA DEVELOPMENT

Kelly M Biette[1], Clarissa Scholes[1], Zeba Wunderlich[2], Angela H DePace[1]

[1]Harvard Medical School, Systems Biology, Boston, MA, [2]University of California Irvine, Developmental and Cell Biology, Irvine, CA

Developmental genes are often regulated by multiple enhancers, each driving a subset of an expression pattern in space and time. In some cases, a gene is controlled by shadow enhancers that drive spatially overlapping patterns and influence the precision of gene expression and its robustness to environmental perturbation. Our lab has shown that shadow enhancers are not simple redundancies. Rather, they encode complex regulatory circuits by responding to different sets of transcription factors. The functional significance of this organization is unknown, but our preliminary evidence suggests it may influence the timing of mRNA production. Therefore, our central hypothesis is that gene expression timing is controlled by the integration of information from multiple enhancers.

We have been testing this hypothesis using the MS2 system in live Drosophila embryos to measure the dynamics of transcription driven by the shadow enhancers of the *Kruppel* gene. Critically, each enhancer in this pair is active at a slightly different time and the activators and repressors that bind each regulatory sequence are known. We hypothesize that (1) activators control the onset timing of *Kruppel* mRNA production from different enhancers, and (2) Kruppel itself terminates expression in this locus via autorepression. Our experiments will map the distribution of regulatory information between a shadow enhancer pair, providing insight into how a single promoter integrates information from multiple regulatory elements. Deciphering how shadow enhancers impart precision in developmental regulatory networks is an important step in predicting how changes in regulatory DNA alter phenotype in evolution and human disease.

FUNCTIONAL CHARACTERIZATION OF GENE REGULATORY ELEMENTS OF EPILEPSY-ASSOCIATED GENES

Ramon Y Birnbaum, Reut Eshel, Aviad Idan, Noga Moshe, Rachel Sharan

Ben-Gurion University of the Negev, Department of Life Sciences, Beer-Sheva, Israel

Epilepsy is one of the most common neurological disorders. It is a complex and heterogeneous disorder which makes it difficult to precisely diagnose and provide effective treatment. A major and underexplored cause of epilepsy could be mutations in gene regulatory elements, such as enhancers. Disruption of these elements, and subsequently of the gene regulatory networks that are involved in brain development, can lead to early onset of epilepsy, such as infantile spasms (IS). This type of epilepsy is linked to an imbalance between glutamatergic projection neurons (excitatory) and GABAergic interneurons (inhibitory). With many IS-associated genes being important for forebrain inhibitory GABAergic interneuron development and function, disruption of the gene regulatory networks in these inhibitory interneurons could lead to IS. Using chromatin immuneprecipitation followed by deep sequencing (ChIP-Seq) with enhancer marks (H3K27ac, RNAPoll2), we identified transcriptional enhancer candidates during differentiation of human embryonic stem cells into inhibitory interneurons. Using RNA-seq, we determined the expression of IS-associated genes and their potential regulated enhancers. In addition, several candidate sequences showed spatiotemporal neuronal activity with high similarity to their potential regulated genes using zebrafish transgenic enhancer assays. Combined, this study provides a novel dataset of the regulatory landscape that controls the spatiotemporal gene expression during inhibitory interneurons and highlights genomic regions in the human genome where mutations could lead to IS and other epilepsies.

NOVEL NGS-BASED APPROACH TO ANALYZE mtDNA TRANSCRIPTION

Sarah Dadon*[1], Amit Blumberg*[1], Jose Antonio Enriquez[2], Dan Mishmar[1]

[1]Ben-Gurion University of the Negev, Department of Life Sciences, Beer Sheva, Israel, [2]Centro Nacional de Investigaciones Cardiovasculares, Department of Cardiovascular Development and Repair, Madrid, Spain

The mitochondrial oxidative phosphorylation system (OXPHOS) is the major energy source in eukaryotic cells. OXPHOS is operated by genes encoded by both the mitochondrial genome (mtDNA) and by the nuclear genome (nDNA). Although most OXPHOS genes are in the nDNA, the mtDNA-encoded subunits are fundamentally important for OXPHOS activity. It has been long shown that all 37 mtDNA genes are co-transcribed in polycistrones, similar to their bacterial ancestor. Previous analysis of *In-Organello* transcription enabled semi-quantitative assessment of mtDNA nascent transcripts. However, the traditional *In-Organello* transcription assay is radioactive, requires much starting material and cannot provide information as to the coding strands of the resultant transcripts. These drawbacks are particularly problematic when quantification is required and while applying the assay to non-model organisms in which the coding contents of each strand may diverge.

Here we present a novel approach that takes advantage of results generated by the established *In-Organello* transcription assay coupled to next generation sequencing (IOTS). Firstly, *IOTS* enabled detection of the post cleavage mature transcripts. Secondly, our stranded sequencing analysis of *IOTS* data clearly identified the known human light-strand and heavy strand transcription units, with striking similarity between the levels of protein coding transcripts in the two strands. Taken together, our approach paves the path towards a quantitative, reference sequence free analysis of mtDNA transcription initiation in diverse cell types and organisms. We anticipate that our approach paves the path towards novel insights into quantitative assessment of the mtDNA transcription process, and will enable the analysis of mtDNA transcription in understudied organisms, thus potentially expanding our knowledge of functional mtDNA genomics to all eukaryotes.

*equal contribution

EXPLOITING SEQUENCE SIMILARITY BETWEEN DNA BINDING DOMAINS SIGNIFICANTLY IMPROVES THE ACCURACY OF TF NETWORK MAPPING

Michael R Brent, Yiming Kang

Washington University, Center for Genome Sciences and Systems Biology and Dept. of Computer Science, St Louis, MO

A central problem in regulatory systems biology is mapping transcription factor (TF) networks - that is, identifying the TFs that bind and regulate transcription of each gene in a genome. From a technology perspective, the goal is to make TF network mapping as straightforward and reliable as genome sequencing has become. This suggests that mapping algorithms should rely primarily on scalable, commodity data sources such as RNA-seq and genome sequencing, rather than methods that, like TF-ChIP, require affinity purification. NetProphet 1.0 is a state-of-the-art algorithm for identifying the direct, functional targets of TFs using only gene expression profiles from cells in which the activity of a single TF has been perturbed.

Here, we show that a second source of easily accessible data – the amino acid sequences of the DNA binding domains (DBDs) of each TF – can be used to significantly increase the accuracy of NetProphet 1.0. The fundamental idea is to exploit the fact that TFs with similar DNA binding domains typically bind similar DNA sequences. The first step of the algorithm is to build a preliminary network map based on expression data in which each potential TF-target pair receives a score reflecting NetProphet's confidence that the TF directly binds and regulates the target. In a second pass, the score for each TF-target pair is replaced by a weighted average of the scores of similar TFs paired with the same target. The weighting is determined by the degree of similarity between the DBD of the original TF and that of the supporting TF. In this way, the final score borrows strength from the initial scores of TFs that are likely to bind similar targets.

By its nature, NetProphet 1.0 assigns high scores only to functional targets of each TF. The challenge, therefore, is to score direct functional targets above indirect functional targets. We evaluate an algorithm's success at this endeavor by using two assays: (1) the percentage of high-scoring TF-target interactions that are supported ChIP data and (2) the percentage of high-scoring TF-target interactions that are supported the in vitro binding specificity of the TF. Using both assays and data from both yeast and fly, we show that this weighted averaging significantly improves the accuracy of NetProphet 1.0. We also demonstrate that further gains can be achieved by inferring the DNA-specificity of each TF from the second-round map and the promoter sequences of likely targets. Together, these ideas constitute NetProphet 2.0, a significant step toward the goal of making TF network mapping as straightforward and reliable as genome sequencing.

MULTIPARAMETER FUNCTIONAL DIVERSITY OF HUMAN C2H2 ZINC FINGER PROTEINS

Laura F Campitelli[1], Ernest Radovani[1], Frank W Schmitges[2], Hamed S Najafabadi[3], Marjan Barazandeh[2], Guoqing Zhong[2], Hongbo Guo[2], Andrew Emili[1,2], Jack Greenblatt[1,2], Timothy R Hughes[1,2]

[1]Department of Molecular Genetics, University of Toronto, Toronto, Canada, [2]Donelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada, [3]Department of Human Genetics, McGill University and Génome Québec Innovation Centre, Montréal, Canada

The human genome contains over 700 C2H2 zinc finger proteins (ZFPs), comprising by far the largest class of transcription factors (TFs). The sequence preferences of C2H2-ZFPs evolve rapidly and vary dramatically, due to their modular and flexible DNA recognition mechanism. Much less is known about how they regulate transcription, but most human C2H2-ZFPs also contain KRAB, SCAN, or BTB domains, suggesting limited diversity in effector function. Here, we employed affinity purification with mass spectrometry (AP-MS) to map protein-protein interactions (PPIs) from cell extracts for 118 DNA-binding C2H2-ZFPs. Surprisingly, each C2H2-ZFP typically interacts with a unique spectrum of nuclear factors, often encompassing both co-activators and co-repressors. KRAB, SCAN and BTB domain subclasses recruit expected interaction partners based on previous investigations, but additional and alternative PPIs are pervasive within each group. Strikingly, C2H2-ZFPs most commonly interact with other TFs (primarily other C2H2-ZFPs), followed by post-translational modifiers, scaffold proteins and proteins with RNA-related functions. Thus, PPIs are a major contributor to the functional diversity of C2H2 ZFPs. We present evidence that the evolution of PPIs is independent of evolution in DNA sequence preferences, consistent with “multiparameter” functional diversification. Future work will focus on the means by which PPI specificity is acquired across this highly diverse class of human TFs.

UNSUPERVISED LEARNING OF FEATURES IN MULTIPLE GENE EXPRESSION DATASETS REVEALS A POTENTIAL DRIVER FOR TUMOR-ASSOCIATED STROMA

Safiye Celik[1], Benjamin A Logsdon[2], Stephanie Battle[3], Charles W Drescher[4], Mara Rendi[5], David Hawkins[3,6], Su-In Lee[1,3]

[1]University of Washington, Computer Science & Engineering, Seattle, WA, [2]Sage Bionetworks, Seattle, WA, [3]University of Washington, Genome Sciences, Seattle, WA, [4]Fred Hutchinson Cancer Research Center, Public Health Sciences, Seattle, WA, [5]University of Washington, Anatomic Pathology, Seattle, WA, [6]University of Washington, Medicine, Seattle, WA

Introduction: Discovering patient subtypes and molecular drivers of a subtype is a difficult and driving problem underlying most modern disease expression studies collected across patient populations. The *expression features* conserved across multiple gene expression datasets from independent disease studies are likely to represent important molecular events underlying the disease. Extracting such robust and biologically relevant features will effectively reduce the dimensionality of gene expression data to solve these problems.

Methods: We present a novel feature extraction method, INSPIRE (**IN**ferring **S**hared modules from multi**P**le gene exp**RE**ssion datasets), to infer highly coherent and robust *modules* of co-expressed genes and the dependencies among modules from multiple expression datasets. Focusing on identifying *module-based features* conserved across multiple gene expression datasets is important for several reasons. First, inferring features from multiple datasets will lead to increased power to detect relevant and robust features because of a greater pooled sample size. Second, our module-based feature extraction model enables the use of multiple datasets that are *not synchronized* due to discrepancies between the platforms in which genes are measured (e.g., platform specific differences). Many conventional feature extraction methods cannot integrate these datasets to infer a single combined model, whereas our approach is able to naturally model the dependencies among the latent features even when a large proportion of genes are not observed on a certain platform.

Results: We evaluated INSPIRE on synthetically generated datasets with known underlying dependence structure among features, and gene expression datasets from multiple ovarian cancer studies. We show that the features inferred by INSPIRE can explain unseen data better and can reveal prior knowledge on gene function more accurately than alternative methods. We demonstrate that applying INSPIRE to nine ovarian cancer datasets leads to the identification of a new marker and potential molecular driver of tumor-associated stroma - HOPX. We also demonstrate that the HOPX module strongly overlaps with the genes defining the mesenchymal patient subtype identified in The Cancer Genome Atlas ovarian cancer s. We provide evidence for a previously unknown molecular basis of tumor resectability efficacy involving tumor-associated mesenchymal stem cells represented by HOPX.

Discussion: The INSPIRE algorithm provides a more robust feature extraction method across datasets that naturally accounts for a lower dimensional feature space. The discovery of a new tumor-associated stroma marker, HOPX, and its module suggests a previously unknown mechanism underlying tumor-associated stroma.

GENE NETWORK DYNAMICS OF MAMMALIAN ROD PHOTORECEPTOR

Vijender Chaitankar, Jung-Woon Kim, Hyun-Jin Yang, Matthew J Brooks, Linn Gieser, Tiziana Cogliati, Gokhan Karakulah, Anand Swaroop

Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD

Photoreceptors are primary sensory neurons that convert light into electrical signals, which are processed by brain in to images. Two types of photoreceptors in the mammalian retina, rods and cones, are responsible for scotopic (night, low light) and photopic (day, color) vision, respectively. The Maf-family leucine zipper transcription factor NRL is required for determining rod cell fate from photoreceptor precursors. We therefore performed RNA-seq profiling of NRL-positive (rods) and NRL-negative (S-cone like) mouse photoreceptors at six distinct stages of development from P2 to P28 to elucidate transcriptome dynamics associated with rod differentiation. NRL interacts with other transcription factors including homeodomain protein CRX, orphan nuclear receptor NR2E3, and estrogen-related receptor ESRRB to complete the morphogenesis and functional maturation of photoreceptors. To uncover underlying regulatory network architecture that elucidates development of a rod photoreceptor, we applied a systems biology approach integrating data from multiple targetome and transcriptomics studies. We observed a transcriptome wide, sharp shift in expression profiles of genes during outer segment formation and synaptogenesis that occurs between P6 to P10 during photoreceptor development. Our studies establish a framework for facilitating biological studies to further dissect gene regulatory networks underlying photoreceptor development and disease. Our analysis has yielded a set of high confidence novel genes associated with maturation and homeostasis of rod photoreceptors. The established rank combined based systems integration approach that we implemented to integrate multiple data sets efficiently handles experimental and quantitative differences between next generation sequencing and microarray data sets.

PREDICTING CELL TYPE-SPECIFIC LONG-RANGE REGULATORY INTERACTIONS IN RARE CELL TYPES

Deborah Chasman[1], Chang Wang[2], Sushmita Roy[3,1,2]

[1]University of Wisconsin--Madison, Wisconsin Institute for Discovery, Madison, WI, [2]University of Wisconsin--Madison, Department of Computer Sciences, Madison, WI, [3]University of Wisconsin--Madison, Department of Biostatistics and Medical Informatics, Madison, WI

Long-range regulatory interactions among enhancers and gene promoters play a major role in tissue-specific gene regulation. New advances in chromosome conformation capture (3C) technology are making it increasingly possible to identify such interactions in high-resolution on a genome-wide scale. However, data are only available for a limited number of cell lines due to the cost of sequencing and the required number of cells for making reliable measurements.

Recently, we developed a supervised machine learning tool, Regulatory Interaction Prediction for Promoters and Long-range Enhancers (RIPPLE), which predicts interactions between enhancers and promoters using a small number of chromatin marks, expression of the promoter's gene, and binding profiles of a few transcription factors. While RIPPLE can accurately predict interactions in novel cell lines, a major challenge is the lack of ChIP-seq datasets for transcription factor binding in rare and primary cell types. We find that DNaseI-filtered binding motifs can be used in place of the ChIP-seq features without substantial loss in accuracy, demonstrating that RIPPLE is able to reconstruct enhancer-promoter interactions in cell lines with limited availability of transcription factor ChIP-seq data. Further, we tested RIPPLE's ability to predict interactions among general regulatory regions that are computationally predicted to have high regulatory potential as well as non-coding disease SNPs in gene poor regions. We find that high confidence target genes are often not the closest expressed gene to the SNP, nor the closest promoter marked for active transcription (H3K4me3).

These improved abilities of RIPPLE to predict interactions from limited data opens up an abundance of cell lines for reconstruction of enhancer-promoter interactions, namely those that have been profiled by NIH ENCODE and Roadmap Epigenomics consortia, and enables a comparative study of cell line-specific long-range interactions and improved interpretation of regulatory sequence variants.

MULTIPLE ENHANCER-DRIVEN POLYMORPHISMS DISRUPT THE ENTIRE *RET* GENE REGULATORY NETWORK (GRN) IN HIRSCHSPRUNG DISEASE

Sumantra Chatterjee[1], Ashish Kapoor[1], Dallas Auer[1], Jennifer Akiyama[2], Len Pennacchio[2], Aravinda Chakravarti[1]

[1]Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Baltimorre, MD, [2]Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, CA

Polymorphisms in cis-regulatory elements (CREs) are the suspected etiological causes of most complex disorders but how they cause disease is unknown. To answer this question, we studied Hirschsprung disease (HSCR: congenital aganglionosis), the commonest form of functional intestinal obstruction in neonates, whose major cause is loss of function variants in the gene encoding for the receptor tyrosine kinase *RET*. We examined all common (>10% minor allele frequency) non-coding variants within the ~153kb *RET* locus, and lying within a 225 kb topologically associated domain (TAD), to identify all associations with HSCR. Functional fine mapping led to identification of three polymorphisms residing within 3 distinct enhancers that increased risk of the disease by 4-, 2- and 1.7-fold. Haplotypes for these three independent variants display wide variation and synergistic effects on risk. By studying both human and mouse fetal gut tissues, at developmental time points relevant to enteric innervation, we demonstrate that: (i) the three CREs are *Ret* enhancers with distinct temporal activities during mouse gut development; (ii) the CREs are bound by the transcription factors Rarβ, Gata2/3 and Sox10, respectively, each developmentally expressed concordant with its cognate enhancer activity; (iii) variants in these CREs lead to loss of enhancer activity and reduced *Ret* expression; (iv) Ret is a positive feedback regulator of Sox10 and Gata2/3, but not Rarβ, transcription; and, (v) additional feedback interactions affect its ligand *Gdnf*, co-receptor *Gfra1* and the E3 ligase signal terminator *Cbl*.

These results explain how individually common but small-effect non-coding polymorphisms can lead to large genetic effects in HSCR, since small transcription attenuation of *RET* from enhancer mutations is amplified through the disruption of its own entire GRN. These results implicate RET as a key rate limiting step in early enteric nervous system development and explains why >95% of HSCR cases have at least one *RET* deficiency allele. Our study also demonstrates that critical GRNs are conserved between the mouse and human reflecting an evolutionary constraint on feedback mechanisms as well as on primary function. Our study suggests a general structure for studying the phenotypic impact of non-coding variants on a complex disorder by examining the consequent genetic changes within a primary genes' GRNs.

CONSERVATION OF THE ENHANCER SEQUENCE CODE ACROSS MAMMALS

Ling Chen*[1], Alexandra E Fish*[2], John A Capra[1,2]

[1]Vanderbilt University, Department of Biological Sciences, Nashville, TN,
[2]Vanderbilt University, Vanderbilt Genetics Institute, Nashville, TN

Enhancers are distal regulatory elements that bind transcription factors (TFs) to regulate gene expression. In mammals, enhancers are essential for the proper tissue-specific expression of genes. However, enhancers are not stable over evolutionary time; there is significant turnover in enhancer activity across mammals. In spite of this, gene expression patterns are largely conserved within tissues across species, as are the DNA binding specificities of TFs. We hypothesize that this seeming contradiction—the conservation of gene expression and TF binding preferences, but the lack of conservation of enhancer activity—is explained by conservation of a core mammalian "regulatory code", but not the specific genomic regions that exhibit this code.

To explore the conservation of the regulatory code, we quantified the sequence content of liver and developing limb enhancers across six diverse mammalian species (human, macaque, mouse, dog, cow, and opossum) by computing their k-mer spectra—the frequency of all possible short DNA patterns in their sequences. We then took a machine learning approach based on spectrum kernel support vector machines (SVMs) to demonstrate that DNA sequence patterns are able to distinguish liver and limb enhancers from matched regions from the genomic background in diverse mammals. To evaluate the conservation of the regulatory code across species, we then applied the human-trained classifier to predict liver enhancers in each of the other mammals and evaluated its performance relative to classifiers trained within each species. The classifier trained on human liver enhancers predicted liver enhancers in macaque, mouse, cow, dog and opossum, nearly as well as classifiers trained within each species (ROC AUC decrease of $< 4.5\%$ for all). We then applied enhancer SVM classifiers trained in the other five selected mammals across species and found that all the classifiers generalized well. Limb enhancer classifiers also showed similarly accurate generalization. To explore the DNA sequence features predictive of enhancer activity across species, we analyzed the weights assigned to specific 5-mers in the liver enhancer classifiers. We found that though there was little agreement in the 5-mers most predictive of enhancer activity across species, they matched to a common set of TFs enriched for liver expression. These results suggest that though enhancers evolve rapidly between species, much of the enhancer sequence code recognized by the transcriptional machinery is shared across mammals.

*Authors contributed equally.

MEMORY OF INFLAMMATION IN REGULATORY T CELLS

Joris van der Veeken[1,2], Alvaro J Gonzalez[3], Hyunwoo Cho[3], Aaron Arvey[1,2], Christina S Leslie[3], Alexander Y Rudensky[1,2]

[1]Howard Hughes Medical Institute, Memorial Sloan Kettering Cancer Center, New York, NY, [2]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY, [3]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY

Regulatory T (Treg) cells are a specialized lineage of suppressive CD4 T cells that act as critical negative regulators of inflammation in various biological contexts. Treg cells exposed to an inflammatory environment undergo numerous transcriptional and epigenomic changes, acquire highly enhanced suppressive capacity, and show altered tissue homing potential. Whether these changes represent stable differentiation akin to memory T cells, or a transient adaptation to the inflammatory environment, is currently unclear.

We used an inducible lineage tracing system to analyze the long-term stability of inflammation-induced transcriptional, epigenomic, and functional changes in inflammation-experienced Treg cells. To this end, we performed an integrative computational analysis of ATAC-seq, histone modification (H3K27ac, H3K27me3, H3K4me1) ChIP-seq, and RNA-seq profiles of Treg cells before, during, and two months after exposure to an acute inflammatory environment. We found that Treg cells, in contrast to memory T cells, showed a striking ability to revert activation-induced transcriptional and epigenomic changes and maintained only a selective and specific memory of inflammation. Genes undergoing stable expression changes underwent qualitatively similar but more dramatic chromatin remodeling than genes undergoing transient changes. Stable gene expression changes were further reinforced during secondary Treg cell activation, while genes undergoing transient expression changes were similarly regulated during primary and secondary responses. Moreover, transiently expressed genes did not maintain stable chromatin modifications that would facilitate their reactivation. Importantly, while the activation-induced increase in Treg cell suppressive function was transient, inflammation-experienced Treg cells acquired a stable non-lymphoid tissue preference characterized by differential expression of tissue homing molecules. These data suggest that memory of inflammation allows Treg cells to preferentially localize to non-lymphoid organs to dampen ongoing tissue inflammation, without becoming stably hyperactive and causing an immunosuppressed state.

ASSESSING THE REPRODUCIBILITY OF SINGLE CELL RNA-SEQ CO-EXPRESSION

Megan Crow, Anirban Paul, Z. Josh Huang, Jesse Gillis

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY

Co-expression analysis using single cell RNA-seq data has the potential to reveal new pathways and gene interactions that occur at single cell resolution, but it is unclear how complex technical variation will affect results. To address this, we sequenced two well-characterized inhibitory interneuron cell-types in an experimental design which allows us to test the reproducibility of co-expression networks across a major source of technical variation, library preparation.

Co-expression networks were built from Spearman correlations of raw counts or count-per-million (CPM) normalized data from each library batch, then functional connectivity was assessed using neighbor voting in cross-validation to learn Gene Ontology functions. We found that meta-analytic aggregation of count-based scRNA-seq networks increases performance modestly for most functions (mean AUROC of individual networks=0.538±0.005, aggregate AUROC=0.556±0.007, $p<0.05$ Wilcoxon rank-sum test) however normalized networks have lower performance, which does not improve with aggregation (highest individual network AUROC=0.514, aggregate AUROC=0.511±0.003). This indicates that our count-based networks contain reproducible functional connectivity.

Interestingly, the count-based aggregate network had high performance for synaptic genes (AUROC=0.793). To explore this we plotted both count- and CPM-based networks and discovered very strong expression level dependencies in both. Where count-based networks show an extreme preference for high co-expression among highly expressed genes, CPM-normalization induces covariation among lowly expressed genes. Consistent with this, we find that performance in count-based networks is dependent on expression level: controlling for expression reduces synaptic gene learnability to baseline levels (mean AUROC=0.53±0.02).

Our results indicate that while single-cell RNA-seq co-expression networks contain reproducible functional connectivity, results are highly sensitive to normalization and strongly dependent on expression level. We suggest that researchers explicitly control for technical variation in their experimental design, and use meta-analysis in combination with expression-level matching and thresholding to assess the robustness of co-expression results.

NO MORE SILOS: THE CYTOSCAPE CI ENABLES BEST OF BREED FUNCTIONALITY AND IMPROVED PRODUCTIVITY

Barry Demchak, Dexter Pratt, Keiichiro Ono, Christian Zmasek, Trey Ideker

UC San Diego, Medical Genomics, La Jolla, CA

In Systems Biology, insights are often driven by a virtuous combination of collaboration, rich data sets, and robust computational and visualization algorithms. The application of Internet technologies in each of these areas has empowered researchers to discover and leverage prior work more quickly and effectively than ever before. At the same time, counterproductive technological and cultural silos have arisen, where resources are spent integrating incompatible data sources and reprising existing computations and visualizations. In this abstract, we describe how the Cytoscape Cyberinfrastructure (CI) can improve research productivity by enabling the integration of siloed collaboration, data, and computational systems while also providing a path to scalability and evolvability.

Systems Biology researchers increasingly create network analysis and visualization workflows using modern programming systems such as R, MATLAB, and iPython. These systems provide reusable components, common data formats, collaboration features, and user communities. However, by nature they also discourage collaboration *between* communities and component reuse *across* systems, thus creating silos. Web-based data sets create their own silos by delivering data in their own formats, different from all others.

The CI is a framework organized as a Service Oriented Architecture (SOA) where workflows and algorithms are each written in the language that best suits their function. Algorithms are packaged as Microservices that exchange network data in the common and extensible CX format, and which can execute on servers distributed across the Internet. For example, the NDEx service allows network data to be stored, retrieved, and shared between users and groups of users. Other services include an ID mapper (e.g., Gene Symbol to Entrez ID), heat dissipation, network layout, and Network Based Stratification, with others on the way.

We present a prototype of a CI-enabled web application that demonstrates how services can be organized into a workflow that fetches a network from an NDEx database, merges it with experiment data, visualizes it, and then writes it back to NDEx. The application is organized as a collection of user interface elements (called widgets) that call the CI services and are themselves reusable for building new Systems Biology applications.

By enabling the use of bioinformatic services regardless of the language in which they are written, CI applications encourage the creation and reuse of best-of-breed functionality while enabling the integration of siloed communities into a larger, more productive community. It incentivizes the constant sharing and iteration of information, thereby enabling more fluid, agile, reproducible, and opportunistic bioinformatic research.

GENOME-WIDE MAPS OF PUTATIVE REGULATORY REGIONS AND GENOMIC SIGNAL INTERACTIONS

Klev Diamanti[1], Husen M Umer[1], Marcin Kruczyk[1], Michal J. Dabrowski J Dabrowski[2], Marco Cavalli[3], Claes Wadelius[3], Jan Komorowski[1,2]

[1]Uppsala University, Cell and Molecular Biology, Uppsala, Sweden, [2]Polish Academy of Sciences, Institute of Computer Science, Warszawa, Poland, [3]Uppsala University, Immunology, Genetics and Pathology, Uppsala, Sweden

Transcription factors (TF) govern transcription through binding to DNA regions. Binding maps of TFs may be used to identify regulatory regions on a genome-wide scale. Here we used TF ChIP-seq datasets for five human cell lines in order to prove that clusters of TF binding sites (TFBS) define regulatory regions. We also provided an insight of the TFs' synergistic mechanism by creating three TF-interaction models based on the physical genomic locations of TFs: neighboring, overlapping and co-occurring. As a result, we can outline tethering and antagonistic TF-complexes, and protein-protein interactions.

We identified ~144k putative regulatory regions, with the majority of them being ~300bp long. We found ~20k putative regulatory elements lying in the ENCODE heterochromatic domains, which indicates a large regulatory potential in the presumed transcriptionally silent regions. Among the most significant TF interactions that we identified in the heterochromatic networks were CTCF and the components of the cohesin complex, RAD21 and SMC3, and the strong cooperation between proteins of the same family such as USF1-USF2 and MAFF-MAFK. Finally, we investigated the implication of the obtained regions on genome looping formation. We discovered that the putative regulatory regions were present in more than 90% of the 3-dimensional (3D) interacting domains and that the heterochromatic regions participate in this mechanism. We found a significant enrichment of single nucleotide polymorphisms in the proposed putative regulatory regions and their presence in 3D interacting domains may signify a perturbing role in the genomic structural formation.

ANALYZING THE INTERPLAY BETWEEN ENHANCERS AND GENES IN THE TRANSCRIPTIONAL RESPONSE TO CELASTROL

Noah Dukler[1,3], Greg Booth[2], Ed Rice[2], Nate Tippens[2], Charles Danko[2], John Lis[2], Adam Siepel[1]

[1]CSHL, Simons Center, Laurel Hollow, NY, [2]Cornell U., BSCB, Ithaca, NY, [3]Weill Cornell Medical College, Physiology, NY, NY

Non-coding regions of the genome have recently been found to host a wide variety of regulatory elements that affect gene expression but their mechanism(s) and dynamics are still poorly understood. To study the dynamic relationship between enhancers and promoters, we generated time course PRO-seq data after treatment of human K562 cells by the small molecule celastrol. Celastrol is a steroid derivative with therapeutic potential in obesity, cancer and inflammatory disorders. It has been shown to induce a highly complex cellular response, including elements of the heat shock response (HSR) and the unfolded protein response (UPR). PRO-seq measures the 3' end of nascent transcripts, allowing us to simultaneously quantify gene expression and detect the unstable enhancer RNAs (eRNAs) that mark divergent transcription start sites (dTSS) and identify active enhancers. Since PRO-seq measures nascent transcripts, we can immediately detect concordant changes in PolII activity between disparate elements without the delay required by RNA-seq. By fine mapping the architecture of dTSS and computationally associating enhancers and promoters, we look for shared and disjoint sets of regulatory sequence elements. Using PRO-seq we observe transcriptional activity changing within a wide variety of genes as soon as 10 minutes after induction, allowing us to separate first-order and higher order responses to celastrol for the first time. We observe several co-regulated groups of genes immediately respond to celastrol treatment, including those regulating mitochondrial metabolism, the HSF1 response, and translational regulation. Roughly 60 minutes after treatment we observe the upregulation of a number of genes associated with DNA repair, consistent with the induction of a broad cellular stress response. These results allow us to explore the early regulatory underpinning of later responses and phenotypes observed in prior studies. Our results begin to illustrate the complex and multilayered signalling response to celastrol in the regulation of both coding and non-coding elements.

DISCOVERING THE AUTONOMOUS RULES BY WHICH SEQUENCE DRIVES TRANSCRIPTION USING A NOVEL MASSIVELY PARALLEL BARCODED REPORTER ASSAY

Vincent D FitzPatrick[1,2,3], Joris van Arensbergen[4], Ludo Pagie[4], Marcel M de Haas[4], Bas van Steensel[4], Harmen J Bussemaker[1,2]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]Columbia University Medical Center, Department of Systems Biology, New York, NY, [3]Columbia University, Graduate Program in Molecular Biophysics, New York, NY, [4]Netherlands Cancer Institute, Division of Gene Regulation, Amsterdam, Netherlands

DNA-binding proteins regulate expression through sequence-specific interactions with gene promoters. These interactions are modulated by local chromatin context features extrinsic to the promoter sequence, making it difficult to separate sequence-dependent direct regulatory mechanisms from indirect contextual factors. To address this problem in a comprehensive and unbiased manner, we developed SuRE (Survey of Regulatory Elements). SuRE is an ultra-high-throughput barcoded reporter assay that quantifies the autonomous ability to drive expression for millions of genomic elements in parallel. The unique nature of these data allows us to dissect transcriptional regulation in novel ways. Having SuRE readouts for fragments that partially and randomly overlap with the region around annotated transcriptional start sites (TSS) allowed us to uncover detailed spatial rules. We find that sequence signals that autonomously drive expression in the sense direction are largely contained within the first 150bp upstream of the TSS. Reporter expression driven by antisense promoter fragments is about half as intense on average, and driven by signals in the same 150bp sequence window. We also asked whether the known dependence of expression on CpG content still holds in the SuRE context, when proximal promoter sequence is isolated from its genomic context. When we categorize SuRE fragments in terms of (i) the observed CpG density and (ii) the expected CpG density given the G and C density in the same fragment, we observe a striking trend of decreasing expression with increasing CpG depletion. Our results extend and refine existing ideas about the role of CpG 'islands' in transcriptional regulation. To perform the above analyses, we developed a flexible framework based on generalized linear models (GLM) and poisson counting statistics, which we use to quantify the contribution of various (spatial or sequence) features associated with each SuRE fragment to the expression level. The same approach can be used to perform motif-based analyses of how transcription binding sites contribute to expression in a context-specific manner.

ABNORMAL DNA METHYLATION AT REGULATORY REGIONS SUGGESTS CANCER CELL IDENTITY CRISIS

Jennifer Flynn, Bo Zhang, Jing Li, Xiaoyun Xing, Joseph Costello, Paul Goodfellow, Ting Wang.

Washington University, Dept. of Genetics, St. Louis, MO

DNA methylation has long been known to alter gene expression in normal cells, where methylation at regulatory regions of genes has been associated with silencing their expression. More recently, trends in altered DNA methylation have been observed across human cancers, where global hypomethylation and local hypermethylation events are present. Although many cancers share these alterations, little is known about the specific differences in DNA methylation changes between cancer types. Here, we present a comprehensive view of DNA methylation changes present in two cancer types, endometrioid adenocarcinoma and glioblastoma multiforme, in order to better define the rules of DNA methylation disregulation common to distant cancers and those changes unique to cancers of specific cell types.

Unexpectedly, we observe a much higher percentage of shared hypermethylated regions within the cancers than hypomethylated regions, suggesting these two cancer types may epigenetically silence many more of the same genes than they activate. Further investigation of these differentially methylated regions between the cancers and normal samples reveals a striking pattern where hypermethylated regions containing motifs for overrepresented transcription factors, such as GMEB1 and FOXO3 in GBM and TEAD1 and TCF12 in EAC, enrich for biological terms associated with the cell type of origin; whereas hypomethylated regions containing motifs for overrepresented transcription factors, such as TCF3 and MEIS1 in GBM and TFAP2A and HIF1A in EAC, enrich for terms relating to a suite of cell types. Interestingly, we also find that many of the differentially methylated regions in both cancers overlap to transposable elements, where the transposable elements, such as the MIR52A subfamily, within hypomethylated regions may serve as active enhancers and promoters in the cancers. These results suggest a possible mechanism for tumerogenesis whereby a cancer cell loses its identity as a fully differentiated cell and activates regions of the genome specific for other fully differentiated cell types, in part possibly through the activation of regulatory regions harbored within transposable elements.

A TRANSCRIPTION FACTOR-WIDE PROTEIN-DNA INTERACTION NETWORK FOR ~15% OF *C. ELEGANS* GENES

Juan I Fuxman Bass[1], Carles Pons[2], Lucie Kozlowski[1], John S Reece-Hoyes[1], Shaleen Shrestha[1], Akihiro Mori[1], Chad L Myers[2], Albertha J Walhout[1]

[1]University of Massachusetts Medical School, Program in Systems Biology and Program in Molecular Medicine, Worcester, MA, [2]University of Minnesota-Twin Cities, Department of Computer Science and Engineering, Minneapolis, MN

Gene regulatory networks (GRNs) play critical roles in development, physiology and responses to environmental cues. Comprehensive GRN mapping requires identifying all possible protein-DNA interactions (PDIs). In recent years, chromatin immunoprecipitation (ChIP) assays greatly expanded our view of the genomic regions occupied by individual transcription factors (TFs). However, only ~10% of the ~900 *C. elegans* TFs have been assayed by ChIP. Here, we use gene-centered enhanced yeast one-hybrid (eY1H) assays to delineate a network comprising 21,714 PDIs between 366 TF and 2,576 target genes in *C. elegans*. Analysis of this network identified links between TF essentiality and target redundancy between TFs. By integrating the PDI network with co-expression data, we characterize TFs as transcriptional activators or repressors. Finally, we illustrate how connectivity in the PDI network can help study the function of uncharacterized target genes and TFs.

INTEGRATING SIGNALING NETWORKS INFERRED FROM STRESS-INDUCED TRANSCRIPTOMIC AND PHOSPHO-PROTEOMIC RESPONSES IN YEAST REVEALS AN ANCIENT SIGNALING NETWORK LINKED TO DISEASE BIOLOGY

Matthew E MacGilvray [1], Evgenia Shishkova[2], Deborah Chasman[3], Yi-Hsuan Ho[1], Mark Craven[3], Joshua J Coon[2], Audrey P Gasch[1]

[1]University of Wisconsin-Madison, Laboratory of Genetics, Madison, WI, [2]University of Wisconsin-Madison, Department of Chemistry, Madison, WI, [3]University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics, Madison, WI

Cells react to stressful situations by orchestrating complex responses customized for each situation. Such multi-faceted responses include changes in transcript abundance, protein levels and modification, metabolic fluxes, and cellular structures. These changes are generally coupled with arrest of growth and cell-cycle progression, consistent with the idea that stress defense and proliferation represent competing interests in the cell. How disparate physiological processes are coordinated is poorly understood but likely critical for surviving and acclimating to stressful conditions. We previously developed an interger linear programming approach to integrate gene-fitness contributions, normal and mutant transcriptome changes, phospho-proteomic changes triggered by stress, and protein-protein interactions to infer the stress-activated regulatory network that controls transcriptome changes in *Saccharomyces cerevisiae*. The resulting network implicated new regulators in the yeast stress response, revealed extensive cross connections between regulatory pathways, and pointed to decision points in the defense-versus-growth decision. The final network included only a third of yeast proteins with stress-dependent phosphorylation changes, likely because many of the phospho-proteins represent regulators or effectors involved in other physiological processes. Here we have developed a computational pipeline to infer the signaling networks that control phospho-proteomic changes to stress in yeast. The results reveal hierarchical arrangements of kinases and phosphatases, implicate new kinase-substrate relationships, and point to extensive feedback and auto-phosphorylation. Integration of the signaling networks controlling phospho-proteomic changes and transcriptome programs is providing new insights into how cells coordinate disparate physiological responses with transcriptome changes. Remarkably, the resulting networks are enriched for yeast genes whose human orthologs cause cancer when mutated in somatic tissues or encode proteins that interact with the tumor suppressor p53. We propose that the stress-activated signaling network in yeast represents an ancient signaling system that is orthologous to cancer-relevant networks in humans.

SPECULATIONS ON GENE REGULATION IN EARLY EMBRYOGENESIS BY EXPLORATORY BIOINFORMATICS

Steven X Ge

South Dakota State University, Mathematics and Statistics Department, Brookings, SD

Biomedical research community values narrowly-focused, hypothesis-driven projects. But little attention has been paid to how such hypotheses are formulated. To go beyond intuition, we can take advantage of the massive amount of public data following John Tukey's philosophy of exploratory data analysis. Through open-ended analysis of genomics data aimed at a systematic understanding of a biological process, we may gain insights that lead to specific hypotheses. As an example, I conducted such a study on gene regulation in mouse pre-implantation development (PD) from fertilized egg to blastocyst. Combining biological reasoning with exploratory analysis of a large amount of genomics data, especially single-cell RNA-seq data, I observed: (1)long terminal repeats (LTRs) of the ERVL family are significantly enriched in the promoters of the genes upregulated at early 2-cell stage; (2) these LTRs are probably bound by Obox homeobox transcription factors (TFs) such as Obox1/2/3/5, which are highly and transiently expressed; (3) presence of B1 and B2 SINEs (Short Interspersed Nuclear Elements) repeats in promoter regions are highly correlated with the upregulation of thousands of genes at late 2-cell stage) in a dosage- and distance-dependent manner. Similar associations are found for human Alu and bovine tRNA SINE repeats, as well as DNA transposons in zebrafish. In addition, B1 and B2 elements in promoters are stronger predictor of broad gene expression in mouse embryonic stem cells (ESCs) than in differentiated progenitor cells. (4) Systematic study on the distribution of transposable elements (TEs) in the mouse genome reveals enrichment of SINEs in promoters and introns, DNA transposons in introns, and some LTRs in promoters, some with strong strand-preference. Most intronic retrotransposons are more likely to be found on opposite strands as the host genes. Genes with multiple B1 and B2 elements in promoters are more likely to be evolutionarily conserved and code for proteins that constitute intracellular parts. Thus there is strong evidence linking TEs with gene regulation during PD, suggesting potential function of repetitive "junk DNA". (5) Organisms with longer generation time seems to have bigger genomes, mostly due to accumulation of TEs. The expansion of TEs may be necessary for promoting genotypic diversity and adaptation in slow-reproducing organisms. (6) Analysis of enriched TF binding motifs in promoters of genes co-regulated at various PD stages identifies several known TFs such as Sox2, KLF4, and Oct4, as well as many homeobox TFs that might be involved in gene regulation in PD. Overall, this study shows that exploratory bioinformatics investigation (EBI) can produce many evidence-based, novel hypotheses that can be further studied to bridge knowledge gaps.

COMBINATORIAL ANALYSIS OF TRANSCRIPTION FACTOR BINDING SITES IDENTIFIES HIGHER ORGANIZATION PATTERNS AT REGULATORY ELEMENTS

Ilias Georgakopoulos-Soares[1], Naman Jain[2], Guillermo Parada[1], Martin Hemberg[1]

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [2]Imperial College London, Department of Life Sciences, London, United Kingdom

Transcription Factors (TFs) are proteins that can recognize their cognate DNA motifs, known as TF binding sites (TFBSs), and influence the expression of associated genomic regions. TFBSs are typically clustered at cis-regulatory elements such as promoters and enhancers. To understand the mechanisms within regulatory elements that are involved in the control of gene expression, the specific interactions between TFs, co-factors and the associated DNA need to be elucidated. However, computationally predicting TF interactions or identifying potential complexes at these sites remains very difficult. We present Motif Combinatorics Analysis (MCA), a new method for systematic identification of enrichment and depletion of single motifs or motif combinations at cis-regulatory elements. MCA examines potential hetero- and homo- dimers based on the frequency and the distance between motif occurrences.

Using MCA, a differential enrichment analysis was performed to identify both single motifs and motif combinations depleted or enriched at promoters, enhancers or CTCF sites. Interestingly, we report a set of motifs and motif combinations present across different cis-regulatory elements, at the flanking regions of these elements, but not in the center and which have narrowly defined absolute distances in homotypic and heterotypic clusters. Some of these pair-wise patterns are found outside the center of the regulatory elements, suggesting that the functional region is wider than what was implicated by studying motifs individually.

To further illustrate the strength of MCA, we analyzed co-factor ChIP-seq data. Although co-factors do not directly bind DNA, MCA was able to infer the binding motifs of associated TFs. Analysis of Rad21 identified CTCF as the most prominent partner, a previously known interaction. Studying another co-factor Sin3a, we found RFX1, MECP2 and TGIF1 all of which are known partners.

MCA is easy to use as we provide both a web-interface (freely available at www.genomegeek.com/MotifAnalysis) and a Python package. In the web-interface, the user inputs the motifs of his interest or selects among several popular motif databases to investigate differentially enriched motifs and motif combinations, as well as potential dimers. The package allows for high throughput, parallelized combinatorial motif analysis.

TIME-COURSE ANALYSIS OF GENE REGULATORY NETWORK DYNAMICS DURING MESENCHYMAL LINEAGE COMMITMENT

Deborah S Gérard, Thomas Sauter, Lasse Sinkkonen

University of Luxembourg, Life Sciences Research Unit (LSRU), Belvaux, Luxembourg

Gene regulatory networks (GRNs) play an important role in cell survival and cell differentiation. Mesenchymal stem cells are able to differentiate into different lineages such as adipocytes and osteoblasts. Both differentiation processes are mutually exclusive and, therefore, adipocytes and osteoblasts exhibit a clear relationship at both physiological and pathological level. To identify the GRNs underlying these differentiation processes, and to better understand their dynamics over time, we have generated time-series transcriptomic and epigenomic data during differentiation of both lineages from the same precursor cells using RNA-seq and ChIP-seq, respectively. Multipotent ST2 stromal cells were differentiated for 15 days either into adipocytes or osteoblasts and RNA and chromatin were collected at 6 different time points. RNA-seq analysis identifies over 30,000 expressed transcripts and ChIP-Seq for histone H3 lysine 27 acetylation (H3K27ac) identifies over 46,000 active enhancer regions controlling their expression. Expression levels of most transcripts correlate with their H3K4me3 and H3K36me3 signals, suggesting that their expression is largely controlled at the transcriptional level. Over 10% of the transcripts and enhancers exhibit significant changes over time with osteoblasts showing somewhat slower dynamics than adipocytes. Using a systems biology approach we are integrating the time-series transcriptomic and time-series epigenomic data to get further insights into the GRNs controlling the reciprocal differentiation of ST2 cells towards adipocytes or osteoblasts, with particular focus on key nodes linked to super-enhancers and regulation of metabolism.

HIGH-THROUGHPUT FUNCTIONAL COMPARISON OF PROMOTER AND ENHANCER ACTIVITIES

Thomas A Nguyen[1], Richard D Jones[1], Andrew Snavely[1], Andreas Pfenning[2], Martin Hemberg[3], Jesse M Gray[1]

[1]Harvard Medical School, Genetics, Boston, MA, [2]Carnegie Mellon University, Computational Biology, Pittsburgh, PA, [3]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Promoters initiate RNA synthesis, and enhancers stimulate promoter activity. Whether promoter and enhancer activities are encoded distinctly in DNA sequences is unknown. We compared the enhancer and promoter activities of 2,671 short (139 bp) genomic DNA fragments transduced into mouse neurons. We find that the same sequences typically encode both enhancer and promoter activities. However, gene promoters generate more promoter activity than distal enhancers, despite generating similar enhancer activity. Surprisingly, the greater promoter activity of gene promoters is not due to conventional core promoter elements or splicing signals. Instead, we find that particular transcription factor binding motifs are intrinsically biased toward the generation of promoter activity, while others are not. Our results suggest that specific complements of transcriptional activators endow gene promoters with stronger promoter activity than distal enhancers.

A COMPREHENSIVE ANALYSIS OF 3' END SEQUENCING DATA SETS REVEALS NOVEL POLYADENYLATION SIGNALS AND THE REPRESSIVE ROLE OF HNRNPC ON CLEAVAGE AND POLYADENYLATION

Andreas J Gruber[1], Ralf Schmidt[1], Andreas R Gruber[1], Georges Martin[1], Souvik Ghosh[1], Manuel Belmadani[2], Walter Keller[1], Mihaela Zavolan[1]

[1]University of Basel, Biozentrum, Basel, Switzerland, [2]University of British Columbia, Michael Smith Laboratories, Vancouver, Canada

Alternative cleavage and polyadenylation is a pervasive mechanism of transcript diversification in mammals, which has been recently linked to proliferative states and cancer [1,2]. Transcript isoforms that differ in their 3' untranslated regions (3' UTRs) interact with distinct sets of RNA-binding proteins (RBPs), which are able to modify the translation, stability and subcellular localization of the corresponding transcripts, and even the localization of the encoded proteins. Although several core factors regulating cleavage and polyadenylation were reported to impact the usage of poly(A) sites, to date, the precise mechanisms that underlie the observed transcriptome-wide, systematic changes in 3' UTR lengths are poorly understood. Applying a novel computational methodology we were able to shed light on the regulatory role of specific factors in 3' end processing. Through a uniform analysis of a large number of 3' end sequencing data sets we have established a comprehensive, high-confidence catalog of 3' end processing sites and uncovered 18 conserved signals, 6 of which novel, whose positioning with respect to pre-mRNA cleavage sites indicates a role in pre-mRNA 3' end processing. Moreover, we have found that binding of the heterogeneous ribonucleoprotein C (HNRNPC) to the poly(U) motif, whose frequency also peaks in the vicinity of poly(A) sites, causes transcriptome-wide changes in poly(A) site usage. The HNRNPC-regulated 3' UTRs are rich in binding sites of the ELAVL1 RBP, and they include the alternatively regulated region of the CD47 gene, which is a target of the recently discovered 3' UTR-dependent protein localization (UDPL) mechanism [3]. In summary, our work establishes a high-confidence catalog of 3' end processing sites and novel variants of the poly(A) signal. It further uncovers an important role of HNRNPC in the regulation of 3' end processing and approves that U-rich elements interact with multiple RBPs that regulate different stages in a transcript’s life cycle.

[1] Masamha, C. P. et al., 2014, Nature 510, 412-416
[2] Mayr, C. & Bartel, D. P., 2009, Cell 138, 673-684
[3] Berkovits, B. D. & Mayr, C., 2015, Nature 522, 363-367

PROGRAMMING CHROMATIN ACCESSIBILITY FOR TRANSCRIPTION FACTOR BINDING AND GENE EXPRESSION IN THE EARLY EMBRYO

Katharine N Schulz[1], Eliana R Bondra[1], Markus Nevil[1], Jason E Lieb[2], Tommy Kaplan[3], Daniel J McKay[4], Melissa M Harrison[1]

[1]University of Wisconsin Madison, Biomolecular Chemistry, Madison, WI, [2]University of Chicago, Human Genetics, Chicago, IL, [3]The Hebrew University Jerusalem, Computer Science and Engineering, Jerusalem, Israel, [4]University of North Carolina Chapel Hill, Biology and Genetics, Chapel Hill, NC

Transcription factors drive cell-fate specification by binding to cis-regulatory regions and controlling gene expression. Despite their broad impact, our understanding of how they access the correct sites in the genome remains incomplete. While transcription factors bind specific DNA motifs, only a small fraction of their potential binding sites are occupied. Thus, in addition to DNA sequence, other features influence where transcription factors bind. By studying the processes that shape the initial activation of the zygotic genome, we are working to understand how chromatin structure is reconfigured during early Drosophila development to allow for transcription factor binding and gene expression. During these first stages of development, the genome is transcriptionally quiescent, and transcription does not initiate until hours later. We are focused on the activities of two essential transcription factors, Grainy head and Zelda. We have identified genome-wide binding sites for both factors throughout multiple stages of development. While binding by these factors remained largely stable, there were dramatic changes in the network of genes each regulated. Furthermore, unlike many other transcription factors, Zelda binding is driven predominantly by the underlying DNA sequence, and early binding is predictive of where additional transcription factors will later bind. This suggests that Zelda establishes or maintains a chromatin environment permissive for subsequent transcription factor binding. We directly tested this hypothesis by measuring chromatin accessibility in the presence or absence of Zelda. We found that Zelda is specifically required for regions of open chromatin. This Zelda-mediated chromatin accessibility facilitated transcription factor recruitment and was a driver of early gene expression. Unexpectedly, chromatin at a large subset of Zelda-bound regions remained open even in the absence of Zelda. We demonstrated that both the GAGA factor-binding motif and sites bound by GAGA-factor in the embryo were specifically enriched in these regions. Based on these data, we propose that the activities of both Zelda and GAGA factor are necessary to specify sites of open chromatin and determine transcription-factor binding. These data show that there is a class of transcription factors instrumental in programming the chromatin environment of the early embryo to allow for the tightly regulated cascade of gene expression necessary for proper development.

SC3 - CONSENSUS CLUSTERING OF SINGLE-CELL RNA-SEQ DATA

Vladimir Y Kiselev[1], Kristina Kirschner[2], Michael Schaub[3,4], Tallulah Andrews[1], Tamir Chandra[5], Kedar N Natarajan[1,6], Wolf Reik[1,5,7], Mauricio Barahona[8], Anthony R Green[2], Martin Hemberg[1]

[1]Wellcome Trust Sanger Institute, Computational Genomics, Hinxton, United Kingdom, [2]Universtity of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Stem Cell Institute and Department of Haematology, Cambridge, United Kingdom, [3]University of Namur, Department of Mathematics and naXys, Namur, Belgium, [4]Université catholique de Louvain, ICTEAM, Louvain-la-neuve, Belgium, [5]The Babraham Institute, Epigenetics Programme, Babraham, United Kingdom, [6]EMBL, European Bioinformatics Institute, Hinxton, United Kingdom, [7]Univesity of Cambridge, Centre for Throphoblast Research, Cambridge, United Kingdom, [8]Imperial College London, Department of Mathematics, London, United Kingdom

Cell type diversity is a fundamental characteristic of multicellular organisms. Traditionally, cell types are defined based on morphological features and surface markers. With the advancement of modern experimental tools, such as single-cell RNA-sequencing (scRNA-Seq), it has become possible to acquire the full transcriptome of individual cells, thereby making it possible to define cell types based on the similarity of expression profiles in a data-driven manner. However, due to the large degree of variability in gene expression and the high dimensionality of the transcriptome, the clustering of cells based on their expression profiles remains a challenging problem. We present a novel method, Single-Cell Consensus Clustering (SC3), for the unsupervised clustering of cells from scRNA-Seq experiments. SC3 explores a large number of different solutions in parallel and integrates them through a consensus approach. When tested on nine published datasets, SC3 outperforms existing methods. In addition, SC3 can handle datasets with tens of thousands of cells (e.g. from Drop-Seq experiments) by using Support Vector Machines (SVM). A key feature of SC3 is a user-friendly interface which allows one to validate the clustering in real-time, not only by objective criteria but also through direct visual examination. SC3 also provides biological insights by identifying differentially expressed genes, marker genes and cell outliers (e.g. rare cell types) across the obtained cell clusters. Finally, we show how SC3 can identify clusters corresponding to different subclones of neoplastic cells from patients, thereby providing novel biological insight. In summary, SC3 is fast, accurate, interactive and scalable, and we believe it could provide a platform of wide applicability in experimental laboratories.

SURVEYING CHROMATIN ACCESSIBILITY AND GENE EXPRESSION DYNAMICS IN RESPONSE TO DNA-BINDING PROTEIN ACTIVATION

Dave Hendrickson, Ilya Soifer, Bernd Wranik, R. Scott McIsaac

Calico Life Sciences, LLC, South San Francisco, CA

Chromatin accessibility is a hallmark of transcriptionally active genomic regions. However, the extent to which changes in accessibility correspond to changes in gene expression is unknown. Here we employ ATAC-seq and gene expression analysis across high-resolution transcription factor and chromatin modifier induction time courses to address this question. For many genes (across different induced factors) we observe a simple concordant relationship in changes over time for both measures. However, we also observe a subset of genes and induced factors where the relationship is discordant. From these data we can begin to unravel the salient features both general and specific of how various DNA binding proteins alter chromatin accessibility and gene expression.

FOXP2 MODIFIES THE CHROMATIN LANDSCAPE OF DEVELOPING NEURONS

Stephanie L Hickey, Stefano Berto, Genevieve Konopka

University of Texas Southwestern Medical Center, Neuroscience, Dallas, TX

FOXP2 is the only gene to be implicated in a heritable form of verbal dyspraxia. Because the encoded transcription factor is expressed in the brain during development, its role in speech and language has been of particular interest. While genes regulated by FOXP2 have been identified in many studies, data describing FOXP2 binding has been limited to promoter regions. By analyzing FOXP2 whole-genome binding and gene regulation in human neural progenitors (hNPs), we have found evidence that FOXP2 may actively modify the chromatin landscape. Weighted gene coexpression network analyses of RNA sequencing data from hNPs over-expressing FOXP2 identified a coexpression module negatively correlated with FOXP2. This module is enriched for gene ontology terms such as “chromatin”, “repressor”, and “regulation of transcription”, suggesting that FOXP2 regulates a network of genes involved in epigenetic control in hNPs. Moreover, chromatin immunoprecipitation sequencing (ChIP-seq) of FOXP2 binding sites in these cells revealed that repressed direct targets of FOXP2 are specifically enriched for genes encoding transcriptional repressors or co-repressors. Activated targets, however, are enriched for neuronal differentiation genes. We hypothesize that by modifying the chromatin landscape of neural progenitors FOXP2 turns off cellular programs that maintain an undifferentiated state while turning on programs that drive a cell towards a neuronal fate. To further understand the role of FOXP2 in chromatin modification during neural differentiation, we used ChIP-seq to profile genome wide occupancy of FOXP2 and multiple histone marks in proliferating and differentiated hNPs with and without FOXP2. Additionally, we identified areas of nucleosomal depletion using an assay for transposase-accessible chromatin using sequencing (ATAC-seq) in the same conditions, and correlated epigenetic changes to changes in gene expression over differentiation. Together, these data examine epigenetic regulation by FOXP2 in neural progenitors over development, a novel role for FOXP2 outside of its previously studied function as a canonical transcription factor.

HIGH-THROUGHPUT, GENOMEWIDE DETECTION OF PHARMACODYNAMIC CIRCUITS IN HUMAN BRAIN USING EPIGENOME PATHWAY ANALYSIS

Gerald A Higgins[1,2], Ari Allyn-Feuer[1], Edward Barbour[1], Brian D Athey[1]

[1]University of Michigan Medical School, Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]Assurex Health, Inc., Pharmacogenomic Science, Mason, OH

Noncoding variants from GWAS associated with the pharmacodynamic properties of lithium, ketamine, citalopram, quetiapine and their side effects were investigated to identify the most probable causal polymorphisms using a multivariate, multi-level analysis pipeline. Methods were applied for prediction of causal variants for drug response and adverse events, including data from machine learning for detection of regulatory variants, transcription factor annotation, QTL analysis, Hi-C mapping, transcript localization in postmortem human brain, and knowledge-based discovery from published and unpublished clinical data. These methods were used to define regulatory pathways in human brain that mediate the psychotropic properties of neuropsychiatric medications. Two types of regulatory pathways have been detected in human brain. The first, as exemplified by the lithium response pathway, which includes *CACNA1C, DLG4, GRIA2, GSK3B, HTR1A, SESTD1* and *SLC1A2*, exhibits a preponderance of intersecting intra-pathway polymorphisms associated with lithium response, risk of bipolar disorder and adverse events associated with lithium therapy. The second, as characterized by the ketamine and quetiapine response pathways, exhibit numerous *cis* and *trans* interactions which cause extensive chromatin remodeling involved in neuroplasticity correlated with positive response symptoms, and direct enhancer-promoter interactions with genes found in other tissues that appear to mediate adverse events such as quetiapine-induced QT prolongation. Subsequent studies in human populations have validated these predictions using our computational approach.

GENE REGULATORY CHANGES IN CARCINOGENESIS

Di Huang, Ivan Ovcharenko

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

To understand the changes of gene regulation in carcinogenesis, we explored the signals of DNA methylation – a stable epigenetic mark of gene regulatory elements — and designed a computational model to identify loss and gain of regulatory elements (RE, including enhancers and promoters) in cancer cells. After applying this model to genome-wide DNA methylation profiles of chronic lymphocytic leukemia (CLL), we detected the REs differently-activated during CLL carcinogenesis (deREs), including 6802 gained and 4606 lost deREs in CLL. The identified perturbations of the gene regulatory landscape coincide with concordant changes in the expression of target genes. In particular, tumor suppressors and DNA methyltransferases, such as PAX3, BCOR and MGMT, harbor multiple lost deREs, which indicates that their damaged regulatory architecture might be one of the key causes of tumor formation. deREs display significantly elevated density of GWAS SNPs associated with CLL and CLL-related traits. We examined in depth deRE SNPs associated with CLL or CLL-related traits and observed that the cancer-associated alleles of most of these SNPs (71%) display significant haplotype association with the risk alleles at the corresponding SNPs reported in GWAS. deREs are enriched for the binding sites of NFkB, AP2 and P53 – the well-established B-cell and CLL transcription factors (TFs). Also different deRE groups display distinct TF binding features – gained deREs contain abundant binding sites of TCF3 and PPARs (TFs known to promote proliferation of CLL cells), while lost deREs are enriched with the binding sites TFs involved into B-cell differentiation and maintainance (such as E2F1, PAX5 and SP1). Also, the binding sites of these TFs are more likely damaged in the lost deREs or generated in the gain deREs than expected. Moreover in normal B-cells, while the lost deREs have later replication timing than the RE activated in both CLL and normal B-cells (shared REs), the gained deREs have replication timing similar with the shared REs, highlighting a potential interplay between these genomic regions when the gained deREs become active during CLL initiation.

DYNAMIC CONTROL OF ENHANCER REPERTOIRES DRIVES LINEAGE AND STAGE-SPECIFIC TRANSCRIPTION DURING HEMATOPOIESIS

Jialiang Huang*[2,3], Xin Liu*[1], Dan Li*[4], Zhen Shao*[5], Hui Cao[1], Yuannyu Zhang[1,5], Eirini Trompouki[2], Teresa V Bowman[2], Leonard I Zon[2], Guo-Cheng Yuan[3], Stuart H Orkin[2,6], Jian Xu[1]

[1]University of Texas Southwestern Medical Center, Department of Pediatrics, Dallas, TX, [2]Boston Children's Hospital, Division of Hematology/Oncology, Boston, MA, [3]Dana-Farber Cancer Institute, Department of Biostatistics and Computational Biology, Boston, MA, [4]Harvard College, Cambridge, MA, [5]Chinese Academy of Sciences, CAS-MPG Partner Institute for Computational Biology, Shanghai, China, [6]Howard Hughes Medical Institute, Boston, MA

Enhancers are the primary determinants of cell identity, but the regulatory components controlling enhancer turnover during lineage commitment remain largely unknown. Here we compare the enhancer landscape, transcriptional factor occupancy and transcriptomic changes in human fetal and adult hematopoietic stem/progenitor cells and committed erythroid progenitors. We find that enhancers are modulated pervasively and direct lineage and stage-specific transcription. GATA2-to-GATA1 switch is prevalent at dynamic enhancers and drives erythroid enhancer commissioning. Examination of lineage-specific enhancers identifies TFs and their combinatorial patterns in enhancer turnover. Importantly, by CRISPR/Cas9-mediated genomic editing, we uncover unexpected functional hierarchy of constituent enhancers within the *SLC25A37* super-enhancer. Despite indistinguishable chromatin features, we reveal through genomic editing the functional diversity of several GATA switch enhancers in which enhancers with opposing functions cooperate to coordinate transcription. Thus, genome-wide enhancer profiling coupled with *in situ* enhancer editing provide critical insights into the functional complexity of enhancers during development.

A SCALABLE FRAMEWORK FOR INFERRING FITNESS CONSEQUENCES OF NONCODING MUTATIONS IN THE HUMAN GENOME

Yifei Huang[1], Brad Gulko[1,2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cornell University, Graduate Field of Computer Science, Ithaca, NY

Genome-wide association studies and evolutionary studies suggest that a large proportion of variants important to human diseases, phenotypes, and evolutionary adaptation are located in noncoding regions. However, methods for identifying causal noncoding variants important to human evolution and disease have only achieved limited success. Here we report a scalable framework, L-INSIGHT, for inferring fitness consequences of noncoding mutations by integrating a variety of comparative and functional genomic data. This novel framework is based on a previously developed model, INSIGHT, which was used in the fitCons framework to infer fitness consequences of noncoding mutations in human genome. As a generalized linear regression model with a special likelihood function borrowed from the INSIGHT model, L-INSIGHT significantly improved the scalability of the fitCons method and allows the integration of vastly larger numbers of genomic features. Unlike most existing methods, including GWAVA, DeepSEA, CADD, and FunSeq2, our framework is explicitly defined in evolutionary terms and can distinguish between weak and strong deleterious effects. Using noncoding disease variants curated by the ClinVar and HGMD databases, we show that our new method performs favorably compared to the state-of-the-art.

GENE FUSIONS IN CANCER: DETERMINING THE ONCOGENIC POTENTIAL OF A FUSION EVENT AND ITS REGULATORY EFFECT ON PATHWAY NEIGHBORS

Katelyn J Hughes[1], Oleksandr Narykov[2], Nathan T Johnson[1], Dmitry Korkin[1,2]

[1]Worcester Polytechnic Institute, Bioinformatics and Computational Biology Program, Worcester, MA, [2]Worcester Polytechnic Institute, Department of Computer Science, Worcester, MA

Cancer is a leading cause of death worldwide, estimating in 8.2 million mortalities in 2012 (National Cancer Institute). Gene fusions, hybrid genes formed from two originally separated genes, have been known as cancer drivers (Mertens, et al. 2015), but are also present in healthy cells due to routine errors in replication (Janz, et al., 2003). Can an automated approach determine whether a fusion event comes from cancer or healthy samples? Recently two studies have attempted to address this question for various cancer types (Abate, et al., 2014; Shugay, et al. 2013). However, they did not consider the fact that there may exist overlapping fusion events in both cancer and healthy cells.

In this project, we first develop an accurate computational approach, called Fusion Enriched Learning of Cancer Mutations (FuEL-Can). FuEL-Can is a novel semi-supervised model that uses iterative self-learning random forest classifier. To train this model, only those fusion events that are exclusive either to healthy samples (negative dataset) or to cancer samples (positive dataset) are labeled, the rest of the data is defined as unlabeled. Our learning model is a feature-based and relies on the features generated by the gene fusion detection software, FusionMap (Ge, et al., 2011) and in-house fusion functional annotation features. In our preliminary study, we designed two accurate FuEL-CAN classifiers for the Acute Lymphoblastic Leukemia (ALL) and breast cancer (BC) samples. The first dataset includes 20 healthy and 10 ALL human samples, with 533 gene fusions (Almamun, et al., 2015). The second dataset includes 56 healthy and 112 BC samples. Our semi-supervised learning, has achieved the peak accuracy rate of 87.6% and an F-score of 84.6% (ALL data), while the top performing supervised learning classifier was able to reach only 81.1% of accuracy and 76.6% of f-measure.

Finally, we hypothesize that the gene fusion events associated with disease affect their functional neighbors (i.e., genes that occur upstream and downstream of the same pathway the fused genes are a part of) as well as structural neighbors (i.e., genes that are co-localized in the same chromosomal region) by altering their gene expression levels. As a case study, we explored the fusion event TEL-AML1, which is well characterized in ALL. We found a significant downregulation of the genes co-localized with AML1, including TCR, GZMB, GM-CSF, IL-3 MPO and ELANE. Knowledge of the fusion-affected gene expression patterns will be used as novel features for FuEL-Can as a future step.

DYNAMIC REGULATION OF RNA EDITING IN HUMAN BRAIN DEVELOPMENT AND DISEASE

Taeyoung Hwang[1,2], Joo Heon Shin[2], Daniel Weinberger[2]

[1]Johns Hopkins University School of Medicine, Department of Biomedical Engineering, Baltimore, MD, [2]Lieber Institute for Brain Development, Baltimore, MD

RNA editing is increasingly recognized as a molecular mechanism regulating RNA activity and recoding proteins. Here, we surveyed the global landscape of RNA editing in human brain tissues and identify three unique patterns of A-to-I RNA editing during cortical development: stable highly-edited, stable lowly-edited and increasingly-edited. RNA secondary structure and the temporal expression of adenosine deaminase acting on RNA (ADAR) contribute to cis- and trans- regulatory mechanisms of these RNA editing patterns, respectively. Interestingly, the increasing pattern is brain-specific and conserved in mouse brain development. The increasing pattern associates with the growth of cortical layers and neuronal maturation. Gene ontology analyses implicate the increasing pattern in vesicle or organelle membrane-related genes and glutamate signaling pathways. We also show that the increasing pattern is significantly perturbed in spinal cord injury and glioblastoma. Our findings reveal dynamic and functional aspects of RNA editing in brain, providing new insight into epigenetic regulation of sequence diversity.

INTEGRATION AND MODELING OF GENOME-WIDE TIME-COURSE DATA REVEALS PRINCIPLES OF DIRECTED MOTOR NEURON PROGRAMMING

Mahmoud M Ibrahim[1,2], Silvia Velasco[3], Akshay Kakumanu[4], Shaun Mahony[4], Esteban O Mazzoni[3], Uwe Ohler[1,2]

[1]Max-Deblrueck-Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, Berlin, Germany, [2]Humboldt University Berlin, Department of Biology, Berlin, Germany, [3]New York University, Department of Biology, New York, NY, [4]The Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA

Programmed differentiation of cell types via the activity of transcription factors holds great promise for various medical applications. However, the gene regulatory mechanisms of such drastic cell fate transitions are not known, impeding rational design of transcription factor programming cassettes. Ectopic expression of Ngn2, Isl1 and Lhx3 (NIL) in murine Embryonic Stem Cells (ESCs) induces the differentiation of spinal motor neurons with >90% efficiency within 48 hours (Mazzoni et al. 2013). In this study, we profile the NIL-induced neuronal differentiation process at multiple time points after NIL expression using RNA-seq, histone modifications and transcription factor ChIP-seq data as well as ATAC-seq data. We found that NIL transcription factors binding is highly dynamic, occurring mostly in distal regions and in cascades of early only, stable and late only binding. Furthermore, each binding cascade is associated with a different set of cobound transcription factors which are not necessarily directly induced.

To enable the analysis of our highly dimensional data without ignoring time dependencies, we developed a Dynamic Bayesian Network model that can learn and classify combinatorial trajectories of multiple histone modification datasets across time or across lineage trees, such that a given cluster represents the dynamic trajectories of all analyzed histone modifications. The model makes minimal assumptions about the data and can be easily applied to multiple data sets even if the time-points assayed do not match. Clustering histone modification trajectories of all distal Isl1 and Lhx3 binding sites using this approach, we discovered a highly dynamic enhancer landscape with ATAC-seq, H3K27ac and H3K4me2/1 dynamics that match Isl1/Lhx3 binding dynamics, especially in sites that were inactive before NIL expression. Therefore, Isl1/Lhx3 binding likely directly controls the chromatin dynamics of its binding sites. We then applied our model to co-cluster H3K4me2/3 and H3K27ac/me3 time trajectories in promoter regions. We found that despite the short time frame of the differentiation process, NIL expression leads to a wide range of chromatin changes activating a specific set of neuron genes and inactivating pluripotency genes via a H3K27ac/H3K27me3 antagonistic switch, but leaves other developmental genes poised in the well-characterized +H3K4me3+H3K27me3 bivalent state. Assigning enhancer classes to promoter classes by distance proximity reveals that neuron genes activated during differentiation associate with Isl1/Lhx3-activated enhancers, indicating that the most direct effect of NIL factors expression is to selectively activate the motor neuron genes mainly via activating their enhancers.

GLOBAL PREDICTION OF CHROMATIN ACCESSIBILITY USING RNA-SEQ FROM SMALL NUMBER OF CELLS

Weiqiang Zhou, Zhicheng Ji, Hongkai Ji

Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD

Conventional high-throughput technologies for mapping regulatory element activities such as ChIP-seq, DNase-seq and FAIRE-seq cannot analyze samples with small number of cells. The recently developed ATAC-seq allows regulome mapping in small-cell-number samples, but its signal in single cell or samples with ≤500 cells remains discrete or noisy. Compared to these technologies, measuring transcriptome by RNA-seq in single-cell and small-cell-number samples is more mature. Here we show that one can globally predict chromatin accessibility and infer regulome using RNA-seq. Genome-wide chromatin accessibility predicted by RNA-seq from 30 cells is comparable with ATAC-seq from 500 cells. Predictions based on single-cell RNA-seq can more accurately reconstruct bulk chromatin accessibility than using single-cell ATAC-seq by pooling the same number of cells. Integrating ATAC-seq with predictions from RNA-seq increases power of both methods. Thus, transcriptome-based prediction can provide a new tool for decoding gene regulatory programs in small-cell-number samples.

TRANSCRIPTION AT INTRAGENIC ENHANCERS ATTENUATE GENE EXPRESSION

Senthilkumar Cinghu*, Pengyi Yang*, Justin Kosak, Amanda E Conway, Karen Adelman, Raja Jothi

Epigenetics & Stem Cell Biology Laboratory, NIEHS, NIH, Research Triangle Park, NC

Eukaryotic transcription is regulated at multiple rate-limiting steps, including RNA Polymerase II (RNAPII) recruitment, transcription initiation, elongation, and termination. Emerging evidence points to much of the transcriptional regulation occurring well after RNAPII recruitment, through controlled pause and release of promoter-proximal RNAPII during early elongation. After its regulated escape into productive elongation, RNAPII is generally assumed to processively progress through the gene, terminate and eventually reinitiate transcription. Although RNAPII elongation rates correlate with gene expression, mechanisms regulating transcription during productive elongation remain enigmatic. Here we report a novel class of intragenic enhancers, marked by transcription initiation, with a role in attenuating gene transcription during productive elongation. The levels of transcription at these intragenic enhancers in the anti-sense direction correlate positively with their capacity to attenuate sense strand transcription of genes encoding them. Notably, nascent RNA expression is higher upstream of these intragenic enhancers than downstream, supporting transcriptional interference/collision model. Using CRISPR/Cas9-mediated deletions, we demonstrate a physiological role for intragenic enhancer-mediated transcription attenuation in the maintenance of pluripotency in embryonic stem cells. Our findings suggest intragenic enhancers may have evolved to not only enhance transcription of one or more genes from distance but also fine-tune transcription of their host gene through transcription attenuation, much like a transcriptional rheostat, facilitating differential utilization of the same regulatory element to perform distinct functions.

THE EPIGENETIC LANDSCAPE OF T CELL EXHAUSTION

Debattama Sen[1,2], James Kaminski[3], R. Anthony Barnitz[1], Makoto Kurachi[4,5], E. John Wherry[4,5], Nir Yosef[3,6], W. Nicholas Haining[1,7,8]

[1]Dana-Farber Cancer Institute, Department of Pediatric Oncology, Boston, MA, [2]Harvard Medical School, Division of Medical Sciences, Boston, MA, [3]UC Berkeley, Center for Computational Biology, Berkeley, CA, [4]University of Pennsylvania, Institute of Immunology, Philadelphia, PA, [5]University of Pennsylvania, Department of Microbiology, Philadelphia, PA, [6]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Boston, MA, [7]Children's Hospital, Pediatric Hematology and Oncology, Boston, MA, [8]Broad Institute, Cambridge, MA

Exhausted T cells in cancer and chronic viral infection have distinctive patterns of gene expression, including sustained expression of the inhibitory receptor PD-1, but the regulation of gene expression in exhausted T cells is poorly understood. Here we define the epigenetic landscape in exhausted CD8+ T cells and show that it is profoundly different from functional memory CD8+ T cells. CD8+ T cell differentiation during infection occurs with a massive reshaping of accessible chromatin regions organized into functional modules of enhancers. Exhausted CD8+ T cells acquire unique modules of enhancers not found in functional CD8+ T cells, and an extensive network of state-specific transcription factor binding. One enhancer, -23kb from the Pdcd1 locus, is found only in exhausted T cells and other lymphocytes with sustained PD-1 expression. Genome editing shows it to be required for high PD-1 expression. Thus T cell exhaustion occurs with state-specific regulation of genes critical to their dysfunction, offering targets for genome editing that could alter gene expression preferentially in exhausted CD8+ T cells.

TRANSCRIPTION FACTOR BINDING VARIABILITY IN HUMAN CELL TYPES AND ASSOCIATION WITH COOPERATIVE MOTIFS

Mehran Karimzadeh[1], Michael M Hoffman[1,2,3]

[1]University Health Network, Princess Margaret Cancer Centre, Toronto, Canada, [2]University of Toronto, Department of Medical Biophysics, Toronto, Canada, [3]University of Toronto, Department of Computer Science, Toronto, Canada

Introduction. Transcription factor binding site occupancy is inherently limited by the concentration of the appropriate transcription factor. We investigated how variable transcription factor expression affected binding preferences. Using ENCODE ChIP-seq and RNA-seq data from multiple cell types, we divided transcription factor binding sites into one of four *binding variability* categories: dynamic, static, expression-sensitive and low-variability. We annotated binding variability using mRNA expression and binding characteristics of the transcription factors in different cell types. We then investigated differential enrichment of JASPAR database vertebrate motifs in binding sites that have different variability.
Results. Our results indicate that BRCA1 expression-sensitive binding sites were enriched for a CTCF motif in HeLa, HepG2, and H1-hESC ChIP-seq peaks but not in ChIP-seq peaks from A549 or GM12878. Additionally, several transcription factors showed enrichment patterns specific to particular cell types. For example in ChIP-seq peaks of ESRRA, while MCF-7 cell line peaks were enriched for several TFAP motifs, K562 peaks showed enrichment in several motifs including those of ESRR (ESRRA, ESRRB and ESRRG) and RUNX (RUNX1, RUNX2 and RUNX3). We also found previously known co-binding patterns such as those between RAD21 and CTCF, JUN and FOS, MYC and MAX. Our results suggested co-binding between different transcription factors in a manner specific to binding variability categories.
Discussion. As expected, binding variability categories of each transcription factor were often enriched in motifs of the same transcription factor. While ENCODE has provided ChIP-seq profiles of a handful of different transcription factors, there are still many important transcription factors with little ChIP-seq data in the literature. Using JASPAR database of vertebrate transcription factor motifs, our approach allows inference of the regulatory network of such transcription factors where they are cobound with other DNA binding proteins including transcription factors.

RNA DEPENDENT PROTEIN BINDING REVEALED BY CHIP-SEQ

Yuki Katou[1], Katsuhiko Shirahige[1]

[1]The University of Tokyo, Institute of Molecular and Cellular Biosciences, Tokyo, Japan, [2]The University of Tokyo, Institute of Molecular and Cellular Biosciences, Tokyo, Japan

ChIP-seq technique is very usuful techniqe and this technique gives us many information. However some of proteins are difficult to analyze by ChIP-seq. One of the reason why it is difficult to detect localization is bad signal- noise ratio.Recently highly expressed genes are source of noise in ChIP-seq (L.Teytelman et al.,PNAS,2013).
Scc2 is one of the difficlut protein to analyze its localization. Scc2 is known as a cohesin loading factor. However data of Scc2 doesn' t colocalize Scc1 any cell cycle.To determin whether Scc2 localization by ChIP-seq is true or not, we treated RNaseA after DNA fragmentation.When we treated RNaseA after DNA fragmentation, Scc2 localization by ChIP-seq was disappeared. Interestingly not only Scc2 localization but also replisome were disappeared. These data shows the possibility of the mechanisim of RNA dependent protein binding.

USING A NETWORK AND PATHWAY CENTRIC MODEL FOR THE COMPREHENSIVE AND UNBIASED HIT SELECTION FROM GENOME-WIDE GENETIC SCREENS IN HUMAN AND MOUSE MACROPHAGES TO INFER COMPREHENSIVE REGULATORY LANDSCAPES.

Samuel Katz[1,2], Jing Sun[1], Ning Li[1], Bhaskar Dutta[1], Clare Bryant[2], Iain Fraser[1]

[1]National Institutes of Health, National Institute of Allergy and Infectious Diseases, Laboratory of Systems Biology, Signaling Systems Unit, Bethesda, MD, [2]University of Cambridge, Department of Veterinary Medicine, Cambridge CB3 0ES, United Kingdom

The accelerated advancement of genome-wide screening technologies such as RNAi and CRISPR has made it possible to generate extensive lists of genes implicated in the regulation of various signal transduction pathways. However, the further validation of the gene candidates generated by these studies is significantly hindered by the analytical challenge of correcting for the noise in the data that emerges from, both, the system being studied as well as the method used to achieve the gene perturbation. Additionally, given the varying sensitivities across different studies, the prevailing approach of using the highest scoring candidates as "hits" has led to widely divergent conclusions and limited general overlap across different screens studying similar pathways. These combined challenges have constrained the full potential of these technologies to identify broad and comprehensive networks of the regulatory factors associated with specific signaling pathways. To address this challenge we developed an analytical model that uses our recently developed CARD software platform to integrate genome-wide screening data with predicted protein-protein interaction data and canonical pathway databases while incorporating statistical corrections for possible noise in the experimental readout and varying readout sensitivity across different studies. For our study we used an siRNA system for genome-wide perturbation in human and mouse macrophages and the cellular response to bacterial LPS through the Toll-like Receptor (TLR) signaling system as our pathway. We demonstrate that our model selects candidate regulators that show a markedly increased validation rate in secondary screens than hits picked from simple score thresholds. Our analytical model provides a novel approach towards the interpretation of data from genome-wide studies and permits broader and more comprehensive insight into the network of genes and cellular process that regulate specific signal transduction pathways.

This research was supported in part by the Intramural Research Program of the NIH, National Institute of Allergy and Infectious Disease (NIAID).

RET-SEQ: A RETINA CENTRIC ONLINE PLATFORM EMPOWERING DISCOVERY BY INTEGRATION OF NEXT GENERATION SEQUENCING RESOURCES

Koray Dogan Kaya, Vijender Chaitankar, Jung-Woong Kim, Hyun-Jin Yang, Matthew J Brooks, Linn Gieser, Tiziana Cogliati, Gokhan Karakulah, Anand Swaroop

National Eye Institute, National Institutes of Health, Neurobiology-Neurodegeneration and Repair Laboratory, Bethesda, MD

Next generation sequencing (NGS) has led to major transformation in the field of genomics and genetics through different techniques and torrent of data. One of today's major challenges is integration of the data in a coherent manner thus enabling discovery and the generation of new hypotheses. Although a few tools exist, they are limited in scope and integration. We have produced multiple distinct NGS data sets for investigating retinal development and disease. To accomplish our goals, we have constructed a relational database and developed a Java based web tool to make the data accessible in an interactive and user friendly manner. The Java layer combines Ensembl Core, Ensembl variation, Ensembl Functional Genomics, and NCBI Gene databases with the retina centric NGS data. NGS data encompass many genomic features including transcription factor and Histone modification ChIP-seq, DNA accessibility ATAC-seq, and transcriptome profiling by RNA-seq. We have combined many popular tools with our Java based application, such as a genome browser (Bio-dalliance) and a JQuery javascript framework to make it compatible with all web browsers. The database is being tested internally and will be available in the near future at http://retseq.nei.nih.gov.

A HIGH-THROUGHPUT APPROACH FOR DESIGNING NOVEL CELL-BASED SIGNAL INTEGRATORS

Jessica L Keenan[1], Trevor Siggers[2]

[1]Boston University, Bioinformatics Program, Boston, MA, [2]Boston University, Department of Biology, Boston, MA

Understanding the transcriptional mechanisms used by cells to integrate and respond to environmental signals is essential not only for understanding healthy and disease states, but also for engineering cells to detect and respond to pathological conditions. We are taking a synthetic biology approach to understand transcriptional integration, and are developing a high-throughput biophysically-motivated technique to design novel signal integrators. To do this, we use protein-binding microarrays (PBMs) to screen tens of thousands of candidate DNA response elements (DREs) for sequences that are cooperatively bound by two signal-dependent transcription factors (TFs) in the presence of an engineered scaffold protein, called miniCBP. We anticipate this approach will both enable a deeper understanding of the design principles underlying cellular signal integration and provide a generalizable platform for regulating transgene expression in response to environmental stimuli.

As a proof-of-concept for this technique, we address the problem of sensing the tumor microenvironment. Delivery of protein-based therapeutics to solid tumors remains a fundamental obstacle to the treatment of many cancers; however cell-based delivery systems that capitalize upon the homing ability of *ex vivo* engineered immune cells to deliver protein-based therapeutics in a targeted manner represent a potentially transformative approach. We attempt to achieve tumor-specific gene expression in a macrophage-based delivery vehicle by requiring that transgene expression is dependent on the macrophage coincidently encountering two signals indicative of the tumor environment: (1) hypoxia, which is sensed via the hypoxia-inducible factor HIF-2 and (2) activation of the heterodimeric nuclear receptor PPARγ:RXR, which occurs in the immunosuppressive tumor environment. We are currently using PBMs to identify miniCBP/DRE combinations that bind cooperatively in the presence of both HIF-2 and PPARγ:RXR. We will test candidate combinations in primary macrophages and evaluate the specificity of the system and robustness to perturbing signals. In addition to providing insight into signal integration at the TF/DNA level, we anticipate that this work will be highly extensible and could be utilized in many applications for the targeted delivery of therapeutic proteins.

PHYLOGENETIC AND EPIGENETIC FOOTPRINTING OF THE PUTATIVE ENHANCERS OF THE PEG3 DOMAIN

Joomyeong Kim, An Ye

Louisiana State University, Biological Sciences, Baton Rouge, LA

The Peg3 (Paternally Expressed Gene 3) imprinted domain is predicted to be regulated through a large number of evolutionarily conserved regions (ECRs) that are localized within its middle 200-kb region. In the current study, we characterized these potential cis-regulatory regions using phylogenetic and epigenetic approaches. According to the results, the majority of these ECRs are potential enhancers for the transcription of the Peg3 domain. Also, these potential enhancers can be divided into two groups based on their histone modification and DNA methylation patterns: ubiquitous and tissue-specific enhancers. Phylogenetic and bioinformatic analyses further revealed that several cis-regulatory motifs are frequently associated with the ECRs, such as the E box, PITX2, NF-κB and RFX1 motifs. A series of subsequent ChIP experiments demonstrated that the trans factor MYOD indeed binds to the E box of several ECRs, further suggesting that MYOD may play significant roles in the transcriptional control of the Peg3 domain. Overall, the current study identifies, for the first time, a set of cis-regulatory motifs and corresponding trans factors that may be critical for the transcriptional regulation of the Peg3 domain.

EVOLUTION OF MODULAR ORGANIZATION OF CHROMATIN STATES

Sara A Knaack[1], Sushmita Roy[1,2]

[1]University of Wisconsin at Madison, Wisconsin Institute for Discovery, Madison, WI, [2]University of Wisconsin at Madison, Department of Biostatistics and Medical Informatics, Madison, WI

Changes in gene regulation have been repeatedly hypothesized to contribute to phenotypic complexity of organisms. The epigenome includes post-translational modifications to the chromatin environment of DNA in eukaryotic cells, and provides an important layer of control in specifying context-specific gene expression. With advances in next generation sequencing techniques epigenomic profiles are becoming increasingly available for multiple species. However, computational approaches to systematically compare such functional genomics signals across multiple species are not well developed. In this work we present a novel multi-clustering computational approach to examine chromatin state, defined by the combination of chromatin marks, across multiple species. We apply our multi-clustering analysis to study the modular organization of the chromatin state in five species: human, mouse, pig worm and fly, spanning several 100 million years of evolutionary history. Our approach successfully identifies modules of co-modified genes with conserved epigenomic states, based on five histone modification marks. We find that the shapes of the profiles are conserved between species and furthermore modules associated with different shapes can explain a significant amount of variation in gene expression. Gene ontology analysis further associates the observed clustering patterns to biological processes such as immune system development and RNA processing. Through these functional associations we can begin to unravel the relationship between evolution of epigenetic states and different biological processes.

MUTATED MOTIFS FOR CTCF AND OTHER TRANSCRIPTION FACTORS AFFECT NEARBY CANCER GENES

Jan Komorowski[1,2], Husen M Umer[1], Marco Cavalli[3], Michal J Dabrowski[2], Klev Diamanti[1], Marcin Kruczyk[1], Gang Pan[3], Claes Wadelius[3]

[1]Uppsala University, Cell and Molecular Biology, Uppsala, Sweden, [2]Polish Academy of Sciences, Institute of Computer Science, Warszawa, Poland, [3]Uppsala University, Immunology, Genetics and Pathology, Uppsala, Sweden

Somatic mutations are known to drive cancer. Previous studies have mainly focused on coding sequences. However, there is growing evidence that some regulatory mutations are over-represented in cancer. We have developed a novel strategy to find mutations in experimentally established regulatory motifs for transcription factors (TFs). In total we found 1552 mutations predicted to significantly reduce binding affinity of many TFs. In hepatocellular carcinoma (HCC) 905 mutations were found, most of them in motifs for CTCF. A specific pattern of mutations was found at a high affinity position of CTCF motifs in HCC and in esophagus, gastric and pancreatic cancer. Near mutated motifs there is a significant enrichment of 1) genes mutated in cancer, 2) tumor suppressor genes, and a highly significant enrichment of 3) genes in KEGG cancer pathways and 4) sets of genes previously associated to cancer. Experimental and functional validations support the findings. Our strategy can be applied to any cell type with established TF motifs and will aid identifications of genes contributing to cancer.

SYNTHESIZING A GENE REGULATORY NETWORK FOR SEA URCHIN DEVELOPMENT

Nicola Paoletti[1], Boyan Yordanov[2], Youssef Hamadi[3], Christoph M Wintersteiger[2], Hillel Kugler[4]

[1]University of Oxford, Computer Science Department, Oxford, United Kingdom, [2]Microsoft Research, MSRC, Cambridge, United Kingdom, [3]LIX, Ecole Polytechnique, Palaiseau, France, [4]Bar-Ilan University, Faculty of Engineering, Ramat Gan, Israel

Seminal work from the Davidson lab [1, 2] has shown how computational models of gene regulatory networks provide system-level causal understanding of the developmental processes of the sea urchin, and enable powerful predictive capabilities. A crucial aspect of the work is empirically deriving plausible models that explain all the known experimental data, a task that becomes extremely challenging due to the inherent complexity of the system. We present results that utilize a computational approach to analyze and synthesize data constrained models using formal reasoning methods, and apply it to the sea urchin gene regulatory network, demonstrating scalability of the methods, and the ability to systematically derive predictive models. The underling methods have also been applied effectively to study other developmental and stem cell systems.

[1] I. S. Peter and E. H. Davidson. A gene regulatory network controlling the embryonic specification of endoderm. Nature, 474(7353), 2011.

[2] I. S. Peter, E. Faure, and E. H. Davidson. Predictive computation of genomic logic processing functions in embryonic development. Proc. of the National Academy of Sciences, 109(41), 2012.

EXPANSION OF GA DINUCLEOTIDE REPEATS ON THE X-CHROMOSOME PROMOTES THE EVOLUTION OF *DROSOPHILA* DOSAGE COMPENSATION

Guray Kuzu[1,2], Emily Kaye[2], Jessica Chery[3], Trevor Siggers[4], Lin Yang[5], Jason Dobson[6], Sonia Boor[2], Jacob Bliss[2], Wei Liu[7], Gerwald Jogl[2], Remo Rohs[5], Nadia D Singh[8], Martha Bulyk[9,10], Michael Y Tolstorukov[1], Erica Larschan[2]

[1]Massachusetts General Hospital, Molecular Biology, Boston, MA, [2]Brown University, Molecular Biology, Providence, RI, [3]Massachusetts General Hospital Cancer Center, Cell Biology, Boston, MA, [4]Boston University, Biology, Boston, MA, [5]University of Southern California, Biological Sciences, Los Angeles, CA, [6]Novartis Institutes for Biomedical Research, Cambridge, MA, [7]Zhejiang Academy of Agricultural Sciences, Plant Protection and Microbiology, Hangzhou, China, [8]North Carolina State University, Biological Sciences, Raleigh, NC, [9]Brigham and Women's Hospital and Harvard Medical School, Medicine, Boston, MA, [10]Brigham and Women's Hospital and Harvard Medical School, Pathology, Boston, MA

Dosage compensation is an essential process that equalizes transcript levels of X-linked genes between sexes by forming a domain of coordinated gene expression. Throughout the evolution of Diptera, many different X-chromosomes acquired the ability to be dosage compensated. Once each newly evolved X-chromosome is targeted for dosage compensation in XY males, its active genes are upregulated two-fold to equalize gene expression with XX females. Furthermore, simple dinucleotide repeats are enriched on the X-chromosome when compared with autosomes. However, the mechanism by which dinucleotide repeats promote X-chromosome identification has remained unknown. In Drosophila melanogaster, the CLAMP zinc finger protein links the dosage compensation complex to the X-chromosome. Here, we combine biochemical, genomic and evolutionary approaches to reveal that expansion of GA-dinucleotide repeats likely accumulated on the X-chromosome over evolutionary time to increase the occupancy of CLAMP, thereby driving the evolution of dosage compensation. Overall, we present new insight into how subtle changes in genomic architecture, such as expansions of a simple sequence repeat, promote the evolution of coordinated gene expression.

TISSUE-SPECIFIC EXPRESSION QTLs

Barbara E Engelhardt

Princeton University, Department of Computer Science, Center for Statistics and Machine Learning, Princeton, NJ

Gene expression profiles that differ across various covariates, such as age, sex, or disease status, have proven useful to study condition-specific and condition-differential processes. Differential gene co-expression networks, which capture pairs of genes that are co-expressed differentially across conditions, enable the study of condition-specific processes at a higher fidelity by capturing differential gene interactions rather than expression levels alone. Here, we study condition-specific gene co-expression networks, which capture genes that are co-expressed uniquely in one condition, and enable the study of condition-specific processes that have no parallels in other conditions. We describe our method for identifying complex, connected condition-specific gene co-expression networks globally from gene expression data in a supervised setting. Our method involves sparse supervised estimation of precision matrices from gene expression data, controlling for all other gene co-expression signals in the data. We apply this method to more than 8,500 gene expression profiles from the Genotype Tissue Expression (GTEx) project v6 data, specifically focusing on co-expression networks specific to one or a subset of tissues. We found a number of known and novel co-expressed genes specific to difficult-to-assay tissues including brain regions and liver and kidney. We validated our networks by finding trans-eQTLs that were identified using the condition-specific co-expression networks. In particular, for genes with known cis-eQTLs, we tested for association of the cis-eQTL SNP with each of the gene's neighbors in the tissue-specific network, limiting the association tests to samples with that condition (e.g., liver samples only). We show that trans-eQTLs in the network neighbors of the cis-eQTL target gene are enriched in samples with the specific condition but not across conditions, validating our condition-specific networks.

WIDESPREAD STABILIZING AND POSITIVE SELECTION GOVERN PRIMATE ENHANCER EVOLUTION

Charles G Danko[1], Zhong Wang[1], Edward J Rice[1], Tinyi Chu[1], Andre L Martins[1], Elia Tait-Wojno[1], John T Lis[1], Lee W Kraus[2], Adam Siepel[3]

[1]Cornell University, Ithaca, NY, [2]UT Southwestern, Dallas, TX, [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

The dramatic diversity of phenotypes observed in nature derives in large part from changes in gene expression across species. Evolutionary differences in the activity of transcriptional regulatory elements (TREs), including promoters and enhancers, drive many evolutionary differences in gene transcription. Here we report results from a genome-wide examination of the general mechanisms of TRE evolution in primates and their influences on gene transcription.

We used PRO-seq to precisely map the locations of RNA polymerases in resting and activated CD4+ T-cells across five vertebrate species, including human, chimpanzee, rhesus macaque, mouse, and rat. Using the dREG tool, we identified 70,586 putative TREs, including promoters and enhancers, that are active in at least one species. We found that the transcriptional activity of distal enhancers evolves surprisingly rapidly (30% conserved expression), consistent with recent reports of rapid evolutionary changes in histone modifications. To investigate the relationship between changes at enhancers and target genes we integrated cell-type-matched ChIA-PET data in humans, profiling 6,507 high-confidence TRE-TRE loop interactions. Enhancers that form at least one long-range chromatin loop are conserved in expression at more than twice the rate of matched enhancers that do not loop, demonstrating a stronger influence from stabilizing selection on a subset of TREs. In contrast to distal enhancers, gene promoters are highly conserved in transcriptional activity (75% in primates). Promoters having ChIA-PET-supported interactions are more conserved in expression than those without such interactions. Each additional loop interaction increases the chance a promoter is conserved between primate species by more than 5%, and the number of loops is strongly correlated with conservation of expression ($R2 = 0.84$). This implies that stabilizing selection acts to preserve gene transcription levels in spite of frequent evolutionary changes at the level of individual TREs. However, when changes in gene transcription do arise, the activity of distal TREs plays a central role. Changes in TRE activities explain the majority of the variance in gene-body transcription level ($R2 = 0.72$), accounting for activities of distal looped TREs, nearby TREs, and the proximal promoter. Analysis of polymorphism and divergence using INSIGHT reveals that positive selection has played an important role in the evolution of human-specific TREs, accounting for at least 35% of fixed nucleotide substitutions. Indeed, we find several examples of recent transcriptional changes in regions with exceptionally high sequence differentiation since the divergence of human and neanderthal, some of which may explain phenotypic differences between primate CD4+ T-cells.

Our analysis demonstrates that (i) active enhancers which loop to target genes are highly constrained in transcriptional activity, (ii) redundancy within these enhancers confers robustness to gene transcription, but (iii) positive selection nevertheless drives changes in transcription levels between closely related species.

AN ARCHAEAL HISTONE FUNCTIONS AS A TRANSCRIPTION FACTOR REGULATING CELL SHAPE

Keely A Dulmage[1,2], Amy K Schmid[1,2]

[1]Duke University, Biology, Durham, NC, [2]Duke University, Program in Genetics and Genomics, Durham, NC

The evolutionary roots of histone can be traced to the Archaea, the third domain of life. Currently, a dominant hypothesis is that archaeal histones resemble those of eukaryotes in compacting DNA and forming nucleosomes. We test this hypothesis using quantitative phenotyping, transcriptomic, proteomic, and DNA binding location analysis to demonstrate that the sole histone protein of a model archaeal species is required to regulate cell shape and is not necessary for survival or genome compaction. Specifically, genome-wide expression changes in histone deletion strains primarily included genes involved in cell envelope biogenesis. Identification of proteins from chromatin enrichments yielded levels of histone and putative nucleoid associated proteins similar to those of transcription factors, consistent with an open and transcriptionally active genome. These results suggest that the function of histone in this model archaeal species more closely resembles a transcription factor than a canonical, eukaryotic chromatin organizing protein despite strong conservation of hallmark residues with eukaryotic histones. In the context of studies from other archaeal species, our results are consistent with varying functions for histones across archaeal lineages. The implications of this study for the transcription regulatory network and histone evolutionary trajectory are discussed.

INTEGRATING GENOMIC AND ENVIRONMENTAL FEATURES TO MAP CRITICAL FACTORS IN INDIVIDUAL DEVELOPMENT OF CHRONIC DISEASES

Marie-Julie Favé[1,2], Alan J Hodgkinson[1,3], Elias Gbeha[1,2], Jean-Christophe Grenier[1,2], Héloise Gauvin[1], Audrey Smargiassi[4], Markey Johnson[5], Veronica Y Sabelnykova[2], Vanessa Bruat[1], Paul C Boutros[2], Youssef Idaghdour[6], Philip Awadalla[1,2,7]

[1]Sainte-Justine Hospital, University of Montréal, Peadiatrics, Montréal, Canada, [2]Ontario Institute for Cancer Research, Bioinformatics and Biocomputing, Toronto, Canada, [3]King's College, Bioformatics and population genomics, London, United Kingdom, [4]University of Montréal, School of Public Health, Montréal, Canada, [5]Health Canada, Air Health Science Division, Ottawa, Canada, [6]NYU Abu Dhabi, Biology, Abu Dhabi, United Arab Emirates, [7]CARTaGENE Biobank, Montréal, Canada

Phenotypic variation is the result of the combined effect of genetic variation with environmental influences. Gene-by-environment interactions are thought to be pervasive and may be responsible for a large fraction of the unexplained variance in heritability and disease risk. Yet, a general understanding of how regulatory variation is modulated by environmental factors is lacking. To systematically survey genetic, environmental, and interaction effects on whole blood transcriptome, we combined whole transcriptome RNASeq profiling with whole genome genotyping on 1,000 deeply endophenotyped individuals selected from over 40,000 participants in the CARTaGENE resource. We document substantial geographic variation in whole blood gene expression in this founder population that follows a south-north cline in the province of Quebec. Using haplotype-based methods on genome-wide genotyping, we detected fine-scale genetic structure, and were able to identify individuals that have migrated within the province. Expression profiles of migrants are more similar to those of individuals currently residing in the same region than to those of individuals with the same ancestry but residing in a different region, indicating a substantial impact of regional environment on gene expression profiles overpowering that of ancestry. Furthermore, we combined our gene expression profiles with high resolution environmental measures and found that the environmental effect is likely mediated by regional air composition. We also report several instances of significant transcriptional gene-environment interactions for genes implicated in inflammation and cardiovascular traits. This environmental effect modulating the transcriptome may contribute to the different levels of respiratory and cardiovascular conditions between regions in Quebec with distinct air composition. Our findings illustrate how environmental variation can significantly alter disease risk and call for placing regulatory variants in the context of their geographic distribution and associated environmental exposures.

HIGH-THROUGHPUT ALLELE-SPECIFIC EXPRESSION ACROSS 250 ENVIRONMENTAL CONDITIONS

G Moyerbrailean[1], O Davis[1], C Harvey[1], C Kalita[1], D Kurtz[1], A Richards[1], A Alazizi[1], D Watza[1], Y Sorokin[1], N Hauff[1], X Zhou[2], X Wen[2], R Pique-Regi[1], F Luca[1]

[1]Wayne State, School of Medicine, Detroit, MI, [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI

Recent studies have shown that GxE interactions can be detected when studying molecular phenotypes that are relevant for complex traits (e.g. infection response eQTLs in immune cells). Despite these relevant examples, the extent to which the environment can modulate genetic effects on quantitative phenotypes is still to be defined. Here we took a high-throughput approach to achieve a comprehensive characterization of GxE interactions in humans through allele-specific expression (ASE) analysis. To this end we have investigated the transcriptional response to 50 treatments in 5 different cell types (for a total of 250 cellular environments). To deeply characterize the transcriptional response and identify ASE in the cellular environments with largest gene expression changes, we deep sequenced 423 libraries (130M reads/sample on average). Across these 89 environments we identified 11,772 instances of ASE (10% FDR), corresponding to 1,530 unique ASE genes. We found that in an individual sample, on average, 0.96% of genes with heterozygous SNPs are ASE genes. The majority of ASE is consistent across conditions ("shared" ASE), confirming previous conditional eQTL analyses. Overall, we find 170 genes with evidence for GxE interaction (conditional ASE, cASE). Because our study design includes multiple vehicle-control treatments, we could empirically estimate the FDR. We estimated FDR<10% for the 64 cASE signals with largest differences in the allelic imbalance between treatment and control. We found a significant negative correlation between gene expression levels and cASE and a positive correlation between differential gene expression in response to treatments and cASE. These results suggest that cASE tends to occur in genes that respond to treatments rather than in highly expressed house-keeping genes. We then used the GxE interactions we identified in vitro to learn about organismal level GxE interactions and identify putative molecular mechanisms for observed risk/protective environmental factors for complex traits. We used GEMMA to jointly analyze summary statistics from 18 GWAS meta-analysis studies with annotations of regulatory variation. For 12 traits, SNPs in genes with cASE show the strongest heritability enrichment, with the highest values observed for lumbar spine bone mineral density and blood total cholesterol (23 and 12-folds, respectively). These results show that genetic effect on trait is larger for cASE, compared to ASE and other genic regions. Specific examples of cASE associated with GWAS traits include the gene ERAP1, which is in a locus associated with Behcet disease and shows cASE in response to selenium. Our cASE result suggests a putative mechanism connecting the reported low selenium levels in Behcet patients with the elevated disease risk observed in individuals homozygous for the selenium-specific low expression allele.

MAPPING VARIATION IN GENE REGULATION: FROM DNA TO PROTEIN

Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, Jonathan K Pritchard

University of Chicago, Chicago, IL

Noncoding variants play a central role in the genetics of complex traits, but we still lack a full description of the main molecular pathways through which they act. Here we used molecular data to quantify the contribution of cis-acting genetic effects at each major stage of gene regulation from chromatin to proteins, within a population sample of Yoruba lymphoblastoid cell lines (LCLs). As expected, we found an important contribution of genetic variation via chromatin, contributing 65% of eQTLs (expression Quantitative Trait Loci). The remaining eQTLs, which are not associated with chromatin-level variation, are highly enriched in transcribed regions, and hence may affect expression through co- or post-transcriptional processes. Further, using a novel method to detect variation in mRNA splicing, we identified an unprecedented number of splice QTLs (2,942 at 10% false discovery rate), most of which have little or no effect on gene-level expression. Thus mRNA splicing is a third primary target of common genetic variation. Remarkably, we found that splice QTLs are major contributors to complex traits, roughly on a par with variants that affect gene expression levels. RNA splicing is therefore a critical regulatory mechanism by which common genetic variation modulates complex disease risk.

SEQUENCE VARIATION AFFECTING TRANSCRIPTION FACTOR OCCUPANCY IN HIGHLY DIVERGED MOUSE STRAINS

Matthew T Maurano[1,2], Jessica M Halow[2,3], Rachel Byron[4], Mark Groudine[4], M A Bender[4], John A Stamatoyannopoulos[2,3]

[1]Institute for Systems Genetics, New York University Langone Medical Center, New York, NY, [2]Department of Genome Sciences, University of Washington, Seattle, WA, [3]Altius Institute for Biomedical Sciences, Seattle, WA, [4]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA

Assessment of the functional consequences of sequence variation at non-coding regulatory elements is complicated by their high degree of context sensitivity to both the local chromatin and nuclear environments. Allelic profiling of chromatin accessibility across individuals has shown that a minority of sequence variation affects transcription factor (TF) occupancy, yet the low sequence diversity in human populations means that no experimental data are available for the majority of disease-associated variants. We generated high-resolution maps of variants affecting TF activity at their native locus and chromatin environment using deep allelic profiling of chromatin accessibility in liver, kidney, lung and B cells from 5 increasingly diverged strains of F1 hybrid mice. The high density of heterozygous sites in these strains enables a precise quantification of the effect size and cell-type specificity of variants associated with altered chromatin accessibility throughout the genome. We show that functional variation delineates characteristic sensitivity profiles for hundreds of TF motifs, representing nearly all important TF families. We develop a compendium of TF-specific sensitivity profiles to account for genomic context effects and model cell-type specific perturbations of the protein-DNA interface. Taking advantage of the high conservation of TF coding sequence among mammals, we apply these models to the classification of disease-associated single nucleotide polymorphisms in human. These models enable quantitative prediction of the effect of non-coding variation on TF activity across a variety of cellular contexts, facilitating both fine-mapping and systems-level analyses of common disease-associated variation in human genomes.

IDENTIFICATION OF FUNCTIONAL ENHANCERS AT *SCN5A* THAT MODULATE QT INTERVAL VARIATION

Ashish Kapoor, Dongwon Lee, Aravinda Chakravarti

Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

Genome-wide association studies (GWAS) have indicated that sequence variation in cis-regulatory elements (CREs) play important roles in common disease risk and trait variation. However, identification of the majority of these causal variants has remained a significant challenge. We performed reporter assays for all common variants at the QT interval associated *SCN5A* GWAS locus, with the goal of identifying the causal variants. A target region of ~500kb at *SCN5A* was defined based on recombination hotspots (rate>10cM/Mb; estimated from HapMap) flanking the 5 independent QT interval GWAS hits. Within the target region, all common variants (minor allele frequency >5%), from the 1000 Genomes European ancestry populations, in moderate linkage disequilibrium (r^2>0.3) with any of the 5 independent GWAS hits, were selected for reporter assays. Of a total 121 variants selected, 112 variants in 104 amplicons passed primer design (amplicon size range 256-617bp; median 397bp), and we successfully cloned both alleles of 106 variants along with flanking sequences in 98 amplicons upstream of a minimal promoter-driven firefly luciferase gene in pGL4.23. Human cardiomyocyte cells, AC16, were transfected with test constructs and Renilla luciferase vector (for transfection normalization) in triplicate; luciferase assays were performed 24h later. Reporter assays on a subset of variants were repeated for assessing replication of allelic difference in regulatory activity. All cloning and reporter assays were performed in 96- and 24-well plates. In reporter assays, compared to empty vector, 24 and 40 amplicons showed enhancer (>2-fold) and suppressor (<0.5-fold) activities in AC16 cells, respectively. Of these 64 CREs, 17 were observed or predicted to be open chromatin regions in human heart tissue by DNase-seq and ATAC-seq. Overall, 12 variants showed nominally significant allelic difference (P<0.05) in reporter activity and were repeated with 18 replicates; 7 variants were identified to have repeated significant allelic difference in regulatory activity. Thus, independent of the publicly available epigenomic data, which are of limited cell-type relevance, an unbiased *in vitro* reporter screen for CREs overlapping all common variants associated with QT interval at the *SCN5A* GWAS locus identified 7 common cis-regulatory variants. Our immediate next goals are to a) evaluate the effect of these 7 CREs on *SCN5A* cardiac expression, b) evaluate the effect of deleting these 7 CREs on *SCN5A* expression in AC16 cells and c) identify the trans-acting factors regulating their functions. We believe that these types of exemplar systems are necessary to test the efficiency of high-throughput genomic assays.

SURVEY OF VARIATION IN HUMAN TRANSCRIPTION FACTORS REVEALS PREVALENT DNA BINDING CHANGES

Luis A Barrera[1,2,3,4], Anastasia Vedenko*[1], Jesse V Kurland*[1], Julia M Rogers[1,2], Stephen S Gisselbrecht[1], Elizabeth J Rossin[3,5,6], Jaie Woodard[1,2], Luca Mariani[1], Kian Hong Kock[1,7], Sachi Inukai[1], Trevor Siggers[1,13], Leila Shokri[1], Raluca Gordân[1,14], Nidhi Sahni[8,9], Chris Cotsapas[5,6,12], Tong Hao[8,9], Song Yi[8,9], Manolis Kellis[4,6], Mark J. Daly[5,6, 10], Marc Vidal[8,9], David E. Hill[8,9], Martha L. Bulyk[1,2,3,6,7,8,11#]

[1]Brigham and Women's Hospital and Harvard Medical School, Div. of Genetics, Dept. of Medicine, Boston, MA, [2]Harvard University, Committee on Higher Degrees in Biophysics, Cambridge, MA, [3]Harvard Medical School, Harvard-MIT Division of Health Sciences and Technology, Boston, MA, [4]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [5]Massachusetts General Hospital and Harvard Medical School, Analytic and Translational Genetics Unit, Dept. of Medicine, Boston, MA, [6]Broad Institute of Harvard and MIT, Cambridge, MA, [7]Harvard University, Program in Biological and Biomedical Sciences, Cambridge, MA, [8]Dana-Farber Cancer Institute, Center for Cancer Systems Biology, Boston, MA, [9]Dana-Farber Cancer Institute, Dept. of Cancer Biology and Harvard Medical School, Dept. of Genetics, Boston, MA, [10]Massachusetts General Hospital, Center for Human Genetics Research and Center for Computational and Integrative Biology, Boston, MA, [11]Brigham and Women's Hospital and Harvard Medical School, Dept. of Pathology, Boston, MA, [12]Current address: Yale School of Medicine, Depts. of Neurology and Genetics, New Haven, CT [13]Current address: Boston University, Dept. of Biology, Boston, MA, [14]Current address: Duke University, Depts. of Biostatistics and Bioinformatics, Computer Science, and Molecular Genetics and Microbiology, Institute for Genome Sciences and Policy, Durham, NC
** These authors contributed equally to this work.*
 # Corresponding author. E-mail: mlbulyk@receptor.med.harvard.edu

Sequencing of exomes and genomes has revealed abundant genetic variation affecting the coding sequences of human transcription factors (TFs), but the consequences of such variation remain largely unexplored. We developed a computational, structure-based approach to evaluate TF variants for their impact on DNA-binding activity and used universal protein binding microarrays to assay sequence-specific DNA-binding activity across 41 reference and 117 variant alleles found in individuals of diverse ancestries and families with Mendelian disease. We found 77 variants in 28 genes that affect DNA-binding affinity or specificity and identified thousands of rare alleles likely to alter the DNA-binding activity of human sequence-specific TFs. Our results suggest that most individuals have unique repertoires of TF DNA-binding activities, which may contribute to phenotypic variation.

A STATISTICAL TEST OF RNA BASE PAIR COVARIATION APPLIED TO PROPOSED LNCRNA STRUCTURES

Elena Rivas[1], Sean R Eddy[2]

[1]Harvard University, Molecular & Cellular Biology, Cambridge, MA, [2]HHMI/Harvard University, Molecular & Cellular Biology, Cambridge, MA

Long noncoding RNAs (lncRNAs) are controversial, and many may be transcriptional noise or unrecognized short protein-coding genes. Even for well-studied functional lncRNAs like Xist, how they function as an RNA remains largely unknown. One line of positive evidence that helps show that an RNA functions as an RNA, and helps guide specific mutational analyses, comes when the RNA has an evolutionarily conserved secondary structure. Evolutionary conservation of RNA base pairs induces pairwise covariation in homologous RNA sequence alignments. Comparative RNA sequence analysis has long been used to infer consensus RNA secondary structures from pairwise covariation data. Based on sequence analysis, the existence of evolutionarily conserved secondary structures has been proposed now for some of the most-studied lncRNAs, including HOTAIR, steroid receptor activator (SRA) RNA, and the Xist RepA domain.

Asking whether a lncRNA has an evolutionary conserved structure or not is different from consensus structure prediction. In structure prediction, one tries to find the structure most compatible with the data, assuming that a structure is present. To test whether the existence of a conserved secondary structure is supported by the data, it is necessary to test the statistical significance of observed pairwise covariations against a null hypothesis of no conserved structure. There is history of mathematical work on statistical significance of RNA pairwise covariation data, but those methods are not readily available in software tools, and they were apparently not used for the published HOTAIR, SRA, or Xist analyses.

We have developed a method for calculating the statistical significance of base pair covariations in RNA sequence alignments, under a null hypothesis that considers confounding covariation that arises from phylogenetic correlations. We have implemented the method both as software and as a web server, R-scape. R-scape analysis supports the presence of evolutionarily conserved secondary structures for many recently identified RNAs, including autoregulatory ribosomal protein mRNA leaders in gamma-proteobacteria, and a proposed control region for alternative splicing of the *Drosophila* Dscam1 gene. R-scape analysis of HOTAIR, SRA, and Xist Rep-A finds no statistically significant support for the evolutionary conservation of the published proposed secondary structures of these RNAs, nor for any alternative structures.

MANY PROMOTERS OF lncRNAs AND mRNAs ACT AS ENHANCERS IN LOCAL GENE REGULATORY NETWORKS

Jesse M Engreitz[1,2], Jenna E Haines[1], Glen Munson[1], Patrick McDonel[1], Mitchell Guttman[3], Eric S Lander[1,4,5]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]MIT, Division of Health Sciences and Technology, Cambridge, MA, [3]Caltech, Division of Biology and Biological Engineering, Pasadena, CA, [4]MIT, Division of Biology, Cambridge, MA, [5]Harvard Medical School, Department of Systems Biology, Boston, MA

Thousands of noncoding loci in mammalian genomes produce multi-exonic, polyadenylated lncRNAs that emanate from defined promoters. Mechanistic studies have demonstrated that several of these loci, such as *Xist*, encode functional lncRNAs that recruit transcriptional regulatory complexes to influence the expression of spatially proximal genes. While it has been proposed that many other lncRNAs similarly regulate gene expression, an alternative possibility is that these noncoding loci act through mechanisms that do not require the lncRNA itself - for example, as enhancers. Notably, most lncRNAs do not have syntenic orthologs across mammalian species, which has led to controversy regarding the functional importance of these RNAs.

Here we dissect the *cis* regulatory functions of 14 genomic loci that produce nuclear lncRNAs. We find that promoter knockouts for 6 of these 14 loci affect the expression of neighboring genes. However, none of these regulatory functions involve the lncRNA transcripts themselves. Instead, the promoters, transcription, or splice structures of lncRNAs directly act to regulate the expression of neighboring genes through mechanisms that do not involve the specific lncRNA sequence. Interestingly, knocking out the promoters of protein-coding genes also impacts local gene regulation. Thus, promoter-promoter regulation is a general property of the *cis* architecture of transcription. These results reveal local networks of regulatory connections between spatially proximal genes, both protein-coding and noncoding, and highlight diverse mechanisms that may contribute to the evolution of noncoding transcription in mammalian genomes.

RNACOMPETE-S: COMPLEX RNA SEQUENCE/STRUCTURE MODELS DERIVED FROM A SINGLE-STEP IN VITRO SELECTION

Kate B Cook, Shankar Vembu, Debashish Ray, Hong Zheng, Quaid D Morris, Timothy R Hughes

University of Toronto, Donnelly Centre, Toronto, Canada

RNA-binding proteins recognize RNA sequences and structures, but there is currently no systematic and accurate method to derive both types of preferences that are large, complex, and reflect direct binding. To address this absence, we introduce RNAcompete-S, which couples a single-step competitive binding reaction with an excess of random RNA 40-mers to a custom computational pipeline for interrogation of the bound RNA sequences and derivation of SSMs (Sequence and Structure Models). RNAcompete-S confirms that HuR, QKI, and SRSF1 prefer binding sites that are single stranded, and recapitulates known 8-10 bp sequence and structure preferences for Vts1, RBMY, and Drosophila SLBP. The SSM derived for SLBP is 18 bases long, and accurately discriminates replication-dependent histone mRNAs from other cellular transcripts. Thus, RNAcompete-S enables accurate identification of large, complex, and intrinsic specificities with a uniform assay, and these motifs can be sufficient for specific recognition of target RNAs regardless of context or cofactors.

CAPTURING THE KINETICS OF PRE-mRNA SPLICING *IN VIVO* BY SINGLE MOLECULE RNA-SEQ

Lydia Herzel*[1,2], Fernando Carrillo Oesterreich*[1], Korinna Straube[1], Katja Hujer[2], Karla M Neugebauer[1]

[1]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT, [2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

The expression of protein-coding genes requires transcription by RNA polymerase II (Pol II) and the removal of introns from pre-mRNA by the spliceosome. Research in diverse organisms from yeast to man has firmly established that most pre-mRNA splicing occurs during transcription, but the dynamic relationship between transcription and pre-mRNA splicing *in vivo* is still poorly understood. It is currently believed that spliceosome assembly occurs slowly and that catalysis occurs when Pol II has reached downstream gene regions, implying a considerable delay between intron synthesis and splicing.
Using long read sequencing and a targeted paired-end sequencing strategy (Single Molecule Intron Tracking, SMIT) of nascent RNA, we directly determine the progress of splicing with regard to Pol II position. With SMIT we derive kinetics that track the progression of transcription and splicing for almost 100 endogenous yeast genes. These resemble kinetic curves seen in every biochemistry textbook, showing the conversion of pre-mRNA (precursor) to mRNA (product) on the timescale of transcription. We find that splicing is 50% complete when Pol II is only 45 nt downstream of introns, with the first spliced products observed as introns emerge from Pol II. Perturbations that slow the rate of spliceosome assembly or speed up the rate of transcription caused splicing delays, showing that regulation of both processes determines *in vivo* splicing profiles. We propose that matched rates streamline the gene expression pathway, while allowing regulation through kinetic competition.

PROPERTIES AND ACTIVITIES OF ENHANCERS

Julie Ahringer

University of Cambridge, The Gurdon Institute, Cambridge, United Kingdom

Gene expression is controlled through regulatory elements such as promoters and enhancers. Enhancers are classically bound by sequence-specific transcription factors that regulate transcription from promoters. The interactions, activities, and regulatory events that occur at and between enhancers and promoters are poorly understood.

We mapped the genome-wide landscape of RNA Polymerase II transcription initiation and elongation in *C. elegans*. We find that transcription initiation at protein coding promoters is generally bidirectional, but productive elongation is oriented. In addition to initiation sites at promoters, we also observe widespread bidirectional transcription initiation at enhancers, and find that enhancers generate long nuclear specific RNA. This pattern, which is also seen at mammalian enhancer regions, indicates a surprising similarity between enhancers and promoters. We are conducting in vivo assays to investigate and compare activities of enhancers and promoters.

Through studying genome domain organization, we find that enhancers are enriched at boundaries between domains of active and inactive chromatin, suggesting that transcriptional activity may have a role in separating chromatin domains. To further investigate domain structure and interactions between regulatory regions, we have developed a variant Hi-C method for high resolution mapping of regulatory site interactions.

WIDESPREAD SHORTENING OF 3' UTRS AND INCREASED EXON INCLUSION ARE EVOLUTIONARILY CONSERVED FEATURES OF INNATE IMMUNE RESPONSE TO INFECTION

Athma A Pai*[1], Golshid Baharian*[2,3], Ariane Page Sabourin[3], Jessica F Brinkworth[2,3,4], Yohann Nedelec[2,3], Jean-Christophe Grenier[3], Katherine J Siddle[5], Anne Dumaine[3], Vania Yotova[3], Zachary P Johnson[6], Robert E Lanford[7], Christopher B Burge[1,8], Luis B Barreiro[3,9]

[1]MIT, Dept of Biology, Cambridge, MA, [2]University of Montreal, Dept of Biochemistry, Montreal, Canada, [3]CHU Sainte-Justine, Montreal, Canada, [4]University of Illinois, Dept of Anthropology, Urbana-Champaign, IL, [5]Harvard University, Dept of OEB, Cambridge, MA, [6]Emory University, Yerkes & Dept of Human Genetics, Atlanta, GA, [7]Texas Biomedical Research Institute, Dept of Virology and Immunology, San Antonio, TX, [8]MIT, Dept of Biological Engineering, Cambridge, MA, [9]University of Montreal, Dept of PediatricsMontreal, Canada

Changes in gene regulation have long been known to play important roles in both innate and adaptive immune responses. However, post-transcriptional mechanisms involved in mRNA processing have been poorly studied despite emerging examples of their role as regulators of immune defenses. We sought to investigate the role of mRNA processing in the cellular responses of human macrophages to live bacterial infections. We used mRNA sequencing to quantify gene expression and isoform abundances in primary macrophages from 60 individuals, before and after infection with *Listeria monocytogenes* and *Salmonella typhimurium*. We show that immune responses to infection are accompanied by pervasive changes in isoform usage that lead to an overall increase in isoform diversity after infection. In response to both bacteria, we see global shifts towards *(i)* the inclusion of cassette exons and *(ii)* shorter 3' UTRs, with near-universal shifts towards usage of more upstream polyadenylation sites. Using complementary data collected in non-human primates, we show that these features are evolutionarily conserved among primates. Following infection, we identify candidate RNA processing factors whose expression levels are associated with individual-specific variation in isoform abundance. Substantial inter-individual variation allowed us to map genetic loci underlying quantitative changes in RNA processing in response to infection and further identify regulatory elements and factors regulating differential isoform usage. Finally, by profiling microRNA levels, we further show that 3' UTR regions with reduced abundance after infection are significantly enriched for target sites for specific miRNAs - including miR-146b, miR-3661, miR-151b, and miR-125a - that are up-regulated following infection. These results suggest that the prominent shift towards shorter 3' UTRs might be a cellular control mechanism to escape repression by immune-activated miRNAs. Overall, our results support concerted regulation of transcription and RNA processing in the control of gene regulatory programs engaged in the response to immune stressors.

NON-CODING ISOFORMS OF CODING GENES IN B CELL DEVELOPMENT AND MALIGNANCIES

Irtisha Singh[1], Shih-Han Lee[2], Christina S Leslie[1], Christine Mayr[2]

[1]Memorial Sloan Kettering Cancer Center, Computational Biology Program, New York, NY, [2]Memorial Sloan Kettering Cancer Center, Cancer Biology and Genetics Program, New York, NY

Alternative cleavage and polyadenylation (ApA) is most often viewed as the selection of alternative pA signals in the 3'UTR, generating 3'UTR isoforms that code for the same protein. However, ApA events can also occur in introns, generating either non-coding transcripts or truncated protein-coding isoforms due to the loss of C-terminal protein domains, leading to diversification of the proteome. Since previous studies have demonstrated the cell type and condition specific expression of 3'UTR isoforms, we decided to investigate the cell type specificity and potential functional consequences of isoforms generated by intronic ApA. We therefore carried out an analysis of 3'-seq and RNA-seq profiles from chronic lymphocytic leukemia (CLL) and multiple myeloma (MM) samples as compared to mature human B cells (naïve and CD5+) and plasma cells, respectively, together with our previous 3'-seq atlas generated from a wide variety of tissues and cell lines. This analysis showed that intronic ApA is a normal and regulated process, most widely used in immune cells, with intronic ApA events enriched near the start of the transcription unit, yielding non-coding transcripts or messages with minimal coding sequence (CDS). These early intronic ApA events preferentially occur in transcription factors, chromatin regulators, and ubiquitin pathway genes. De novo assembly of RNA-seq data supports ~60% of the intronic ApA events from plasma cells and MM samples, leading to >2000 candidate alternative transcripts arising from intronic ApA, with ~900 transcripts ending near the start of the transcription unit, retaining less than 25% of the coding sequence. Our analysis showed that two thirds of these intronic ApA isoforms have minimal coding potential, likely generating non-coding isoforms from protein coding genes. CLL cells increase the expression of early intronic ApA events relative to mature B cells, while MM cells decrease the expression of these events relative to plasma cells. For a fraction of genes, increased expression of isoforms generated by intronic ApA coincides with reduced expression of the full length mRNA in CLLs compared to mature B cells; conversely, lower expression of intronic ApA events coincides with higher full length mRNA expression for some genes in MM samples compared to plasma cells. In these genes, expression of the intronic event may function as a switch to alter full-length mRNA expression. The other fraction of these non-coding isoforms may potentially act as scaffold for recruiting regulatory factors to the locus.

UNCOVERING THE RNA-PROTEIN INTERACTION NETWORK THAT CONTROL VERTEBRATE EMBRYOGENESIS.

Daniel Cifuentes[1], Charles Vejnar[1], Romain Christiano[2], Stephanie Lau[1], Alfredo Castello[3], Matthias Hentze[3], Tobias Walther[2], Antonio Giraldez[1]

[1]Yale University, New Haven, CT, [2]Harvard University, Cambridge, MA, [3]EMBL, Heidelberg, Germany

Regulation of mRNAs deposited in the oocyte is a fundamental step during embryogenesis. The clearance of these maternal mRNAs allows the embryo to take control of its own fate. To understand this post-transcriptional regulatory mechanism we have investigated I) What are the RNA-binding proteins (RBPs) that mediate mRNA clearance during embryogenesis? II) What are their mRNA targets? III) What are the sequence elements that mediate maternal mRNA regulation?

Using Interactome Capture in zebrafish embryos, we have detected 165 proteins that directly bind to RNA during early embryogenesis. Interestingly, most of these proteins are maternally provided, rising the question of how exquisitely timed post-transcriptional regulation is achieved if the RNA-binding proteins are present since oocyte maturation and before maternal mRNA clearance. To address this point we found that the binding to RNA of 45 of these proteins is developmentally regulated.

To decipher the RNA regulatory code during early embryogenesis, we have captured the RNAs targeted by 22 RNA-binding proteins using iCLIP analysis. We identified some RBPs as preferential 3'-UTR binders and others as coding-sequence binders. To uncover their sequence recognition elements, we performed motif analysis. Interestingly, *in vivo* RBP binding is not predicted by a search of in vitro-defined motif revealing higher level of binding regulation. To understand the role of these RBPs, we are investigating which proteins bind to each functional class of mRNAs (destabilized targets, microRNA targets, highly translated mRNAs). Furthermore, we are capitalizing on our large number of RBPs assayed to understand how mRNA fate is cooperatively regulated.

In summary, the concerted identification of RNA-protein interaction dynamics provides an invaluable tool to decipher the post-transcriptional regulatory network that shapes gene expression during early vertebrate embryogenesis and to establish the framework to understand post-transcriptional regulation in other biological transitions.

SEQUENCE-BASED PREDICTIONS OF *CIS*-REGULATORY DNA ELEMENTS ACTIVE IN THE HUMAN HEART

Dongwon Lee, Ashish Kapoor, Aravinda Chakravarti

Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Baltimore, MD

Finding *cis*-regulatory elements (CREs) is an essential step for understanding gene regulatory mechanisms. Although high-throughput assays, such as ChIP-seq and DNaseI-seq, can identify thousands of CREs in the genome with a single experiment, finding *all* potential elements remains a challenge. There are several factors that limit experimental identification of CREs: 1) the stochastic nature of CRE activity, 2) genetic and environmental factors affecting CRE activity, and 3) technical and experimental variation (i.e. batch effects) in detecting CREs. To tackle these problems, we propose to use sequence-based computational methods to predict such missing elements based on experimental data. Using human heart CREs as a framework, we systematically assessed our ability to predict known CREs from DNA sequence. With gkm-SVM[1] trained on adult Heart DHSs (DNaseI hypersensitive sites) from the Roadmap Epigenomics Project, we predicted ~6,000 putative CREs in chromosome 9 (not included in the training set), among which ~2,000 elements overlapped at least one DHS peak (~33% of precision rate). More importantly, ~500 additional predicted CREs overlapped heart DHS peaks detected by four other heart DNaseI-seq data sets (two from ENCODE, two generated here), which is equivalent to a ~42% precision rate. This result strongly suggests that sequence-based prediction can find true CREs missed by a limited number of experiments. Integration of multiple models further increases the precision rate (to ~50%), while maintaining a reasonable recall rate (~40%). Consequently, genome-wide prediction of CREs doubled the number of potential CREs and significantly expanded the repertoire of putative elements. As an independent validation, we also tested 98 distinct regions (avg. length ~400bp) from the *SCN5A* locus using *in vitro* luciferase assays in the AC16 human cardiomyocyte cell line. Among the 24 amplicons that showed significant enhancer activity (>2-fold), four overlapped DHS peaks, and five were additionally predicted by gkm-SVM (~38% recall rate). Consistent with the previous analysis, a 50% precision rate was achieved by the gkm-SVM prediction (five positive amplicons out of ten predicted elements). This result further demonstrates that our computational approach can identify many novel CREs that may have been missed based on experimental data alone. We believe that sequence-based prediction is a valuable strategy to complement such experimental data.

1) Ghandi et al, PLoS Comput Biol 10, e1003711, 2014

A MACHINE LEARNING TECHNIQUE TO LEARN THE BIOMARKER POTENTIAL IDENTIFIES NOVEL MARKERS FOR SENSITIVITY TO CHEMOTHERAPY DRUGS

Su-In Lee[1,2,3], Safiye Celik[1], Benjamin A Logsdon[4], Timothy J Martins[5], Vivian G Oehler[6,7], Elihu H Estey[6,7], Chris P Miller[7], Sylvia Chien[7], Akanksha Saxena[7], Anthony Blau[3,7], Pamela S Becker[3,6,7]

[1]University of Washington, Computer Science and Engineering, Seattle, WA, [2]University of Washington, Genome Sciences, Seattle, WA, [3]Center for Cancer Innovation, -, Seattle, WA, [4]Sage Bionetworks, Public Health Sciences, Seattle, WA, [5]Quellos High Throughput Screening Core, -, Seattle, WA, [6]Fred Hutchinson Cancer Research Center, Clinical Research, Seattle, WA, [7]Department of Medicine and Institute for Stem Cell and Regenerative Medicine, Hematology, Seattle, WA

Introduction: Identifying the genes whose expression levels are predictive of sensitivity or resistance to a chemotherapy drug is of paramount importance to find a way to better match patients to drugs. A conventional analysis is to perform association tests between each gene's expression and each drug's sensitivity measure genome-wide across all genes and against all drugs. However, due to the high-dimensionality (i.e., the number of gene-drug pairs is much greater than the number of samples) along with potential biological or experimental confounders, it is difficult to identify robust biomarkers whose expression levels drive the sensitivity or resistance to specific drugs.

Methods: Here we characterized the genome-wide gene expression and the dosage-response curves for *in vitro* cultured blasts across a panel of 160 chemotherapy treatments from 30 patients with acute myeloid leukemia (AML) and 14 commonly used AML cell lines. We present a novel computational method to prioritize gene-drug associations to identify robust gene expression biomarkers for specific drugs. We then use the cell lines as a source for validation of the identified biomarkers. Our method is a machine learning approach that estimates the *biomarker potential* of G based on six properties of G that are relevant to G's AML driver potential, such as mutation, copy number variation, and methylation profiles and whether G is likely a hub in a gene regulatory network. We conjecture that the information that a gene contains about the disease state and correspondingly the outcome of treatments is highly dependent on the gene's driver potential.

Results: We show that our method effectively reduces the search space of pairwise combinations of gene expression levels and drug sensitivity profiles by integrating information on genes to prioritize the gene-drug pairs. We present that the gene-drug pairs with a high potential are more likely to have a significant association replicated in the cell lines than when prioritizing gene-drug pairs based on the association p-values. We identify SMARCA4 expression as a biomarker and potential driver of response to topoisomerase II inhibitors. We show that cell lines transduced to have high SMARCA4 expression show dramatically increased sensitivity to mitoxantrone and etoposide.

Discussion: This methodological framework and novel patient-derived expression and drug sensitivity data provide a promising way to identify robust biomarkers for personalized therapy in leukemia. Our method provides a general novel computational framework for identifying candidate biomarkers for patient response to therapy based on high-dimensional molecular data.

HOW PROTEIN AND PROMOTER DIFFERENCES CONTRIBUTE TO THE DIVERGENCE OF THE YEAST TRANSCRIPTION FACTOR FAMILY MEMBERS MET31 AND MET32

Sylvia K Burns[1], Chelsea A Stellmach[1], Trevor W Siggers[2], Martha L Bulyk[3], Traci A Lee[1]

[1]University of Wisconsin-Parkside, Biological Sciences, Kenosha, WI, [2]Boston University, Biology, Boston, MA, [3]Harvard University, Medicine, Boston, MA

Transcription factor (TF) family members allow a broad range of transcriptional responses among their targets despite recognizing identical sequences. We study Met31 and Met32, a simple two-member TF family that plays an essential role in yeast sulfur metabolism. Expression of yeast sulfur metabolism genes is turned on by a sole activator, Met4. Met4 lacks DNA-binding ability and relies on interactions with Met31 and Met32 to bind target promoters. Even though Met31 and Met32 are redundant for some functions, microarray studies show *met32Δ* cells exhibit much less Met4-activated expression than *met31Δ* cells. ChIP studies show Met32 is found at target promoters at much higher levels than Met31. Also, the genome-wide ChIP profile of Met32 mirrors that of Met4. These findings all indicate Met32 plays a much bigger role in mediating Met4-activated expression than Met31. To determine if innate protein differences (such as DNA-binding abilities or affinity for Met4) account for Met32 being a better mediator of Met4-activated transcription, we conducted PBM studies on Met31 or Met32, alone and in combination with other components of the Met4 complex. Alternatively, the functional difference between Met31 and Met32 may be due to their different expression profiles. Met31 is constitutively expressed while Met32 is not present under non-inducing conditions but rapidly increases upon induction. These profiles are seen at both the RNA and protein levels, suggesting Met31 and Met32 levels are transcriptionally regulated. To investigate the importance of their expression profiles in dictating their roles in transcription, we swapped the endogenous promoters of *MET31* and *MET32*. Findings from both promoter swap and PBM studies will be presented.

FRAGILE ENHANCERS AND STABLE ENHANCERS TOGETHER DEFINE LINEAGE-SPECIFCITY

Shan Li, Ivan Ovcharenko

National Center for Biotechnology Information, Computational Biology Branch, Bethesda, MD

To study the relationship between genetic variant in nocoding interval and the common diseases, and to systematically prioritze the risk variants within implicated genomic regions, we identified nucleotides capable of disrupting binding of transcription factors and deactivating enhancers if mutated (dubbed candidate killer mutations or KMs) in human heart. We categorized the heart enhancers based on their density of KMs and compared the genomic features and functionality of different groups. The enhancers with dense KM positions (KMPs) are less conserved and are more ubiquitously active across tissues as compared to the ones with sparse KMPs. In contrast, the enhancers with sparse KMPs are more tissue-specific. Dense-KMP enhancers tend to have shadow enhancers in an enhancer cluster, whereas sparse-KMP enhancers tend to function on their own. Dense-KMP and sparse-KMP enhancers are bound by different cohorts of transcription factors (TFs), and homotypical TFs are more likely to bind dense-KMP enhancers. KMPs at sparse-KMP enhancers have more deleterious effect on TF binding and gene expression compared to the ones at dense-KMP enhancers.

A WAVE-LIKE MODEL FOR GENE REGULATORY NETWORKS IN DROSOPHILA EMBRYOS

G.R. Liu[1,2], Y.X. Lin[1], J. Ma[3]

[1]Cincinnati of University, Aerospace Engineering & Engineering Mechanics, Cincinnati, OH, [2]Taiyuan University of Technology, Mathematics, Taiyuan, China, [3]Cincinnati Children's Research Foundation, Biomedical Informatics and Developmental Biology, Cincinnati, OH

Extensive genetic studies have revealed that formation of the segmented pattern in the Drosophila embryo requires gap gene activities and regulatory interactions. Mutual interactions between gap genes and their products, along with the maternal inputs, result in a space-time evolution of the concentration of the various types of proteins along the anterior-posterior axis, forming the segmented patterns in the developing embryo.

The evolution of these proteins can be modeled mathematically in a form of governing partial differential equations (PDEs) with the unknown field being the concentrations of each protein that may produce, diffuse, decay and in mutual interaction. Due to the complexity of interaction term with regulatory mechanisms, the governing PDE is too complicated to be analytically solved. In this work, we attempt to develop a so-called "wave-like" model to solve the governing PDEs in a combined analytical-numerical manner. We first discretize the PDE in time domain. An analytical solution is then obtained in the form of wave modes over the embryo domain for any small time interval, because the complexity of the governing PDE is significantly reduced by the time-discretization. The space-time evolution of the protein concentrations are finally obtained via convolution using the wave modes and an external term representing the interactions. Such a convolution evaluates the outcome of the "battling-waves". Numerical examples will be presented to demonstrate the novel wave-like model.

References

[1] Liu GR and Xi ZC, Elastic Waves in Anisotropic Laminates, CRC Press LLC, 2001.
[2] Sergei Vakulenko, Manu and John Reinitz. Size Regulation in the Segmentation of Drosophila: Interacting Interfaces between localized Domains of Gene Expression Ensure Robust Spatial patterning[J]. Physical Review Letters. 2009,103,168102.
[3] Johannes Jaeger, Maxim Blagov, David Kosman, etc. Dynamical Analysis of Regulatory Interactions in the Gap Gene System of Drosophila melanogaster[J]. Genetics. 2004, 167(4):1721-37.
[4] Jonathan Bieler, Christian Pozzorini, Felix Naef. Whole-embryo Modeling of Early Segmentation in Drosophila Identifies Robust and Fragile Expression Domains[J]. Biophys J. 2011, 101(2):287-296.

SYSTEMATIC IDENTIFICATION OF PREDICTIVE CAUSAL GWAS SNPs WITHIN TRANSCRIPTION FACTOR BINDING SITES THAT CONTROL CHROMATIN LANDSCAPE

Yunxian Liu, Michael J Guertin

University of Virginia, Biochemistry and Molecular Genetics, Charlottesville, VA

Transcription factors (TFs) control cell fate and identity by regulating the timing and magnitude of gene expression. Between 80% and 98% of GWAS-identified loci are non-coding single-nucleotide polymorphisms (SNPs); we hypothesize that many of these non-coding SNPs alter transcription factor binding sites and mediate their affect on disease risk by modulating TF binding and gene regulation. We employed an novel integrative genomics approach to identify candidate TF binding motifs that confer breast cancer (BrCa)-specific phenotypes identified by GWAS. We developed an unbiased approach of iterative *de novo* motif analysis, measuring evolutionary constraint of identified motifs, and composite TF footprinting to identify sequence elements and cognate TFs that are controlling chromatin landscape in BrCa-relevant tissue and cell lines. We used The Cancer Genome Atlas (TCGA) RNA-seq data to show that expression of several identified TFs, including FOXA1, JUND, ELF1, and RFX3, are predictive of BrCa patient survival in a BrCa subtype-specific manner. Next, we imputed the genotypes of 937 TCGA BrCa patients using the 1000 Genomes reference panel and identified all SNPs in strong linkage disequilibrium with the top GWAS SNP. We identified top candidate SNPs that are within DNase-defined regulatory regions of breast-derived cell lines or tissue; our top candidate SNPs are within residues that are critically important for the predicted binding of TFs that dictate the chromatin landscape of BrCa-relevant cells. Indeed, the cognate TFs, including CTCF, bind to these sites with allele-specific preferences. We further identified the candidate causal genes of TF binding site SNPs that reside in expression quantitative trait loci (eQTL) of breast tissue and other tissues. This framework can be used to identify candidate causal variants with regulatory regions and TF binding sites that confer phenotypic variation and disease risk.

ROLES OF GENE LENGTH AND H3K79ME2 MODIFICATIONS IN ESTABLISHING TEMPORAL GENE EXPRESSION

Killeen Kirkconnell[1,2,4], Brian Magnuson[2,4], Michelle T Paulsen[2,4], Karan Bedi[2,4], Mats Ljungman[2,3,4]

[1]University of Michigan, Human Genetics, Ann Arbor, MI, [2]University of Michigan, Radiation Oncology, Ann Arbor, MI, [3]University of Michigan, Environmental Health Science, Ann Arbor, MI, [4]University of Michigan, Translational Oncology Program, Ann Arbor, MI

Transcriptional timing is inherently influenced by gene length, thus providing a mechanism for temporal regulation of gene expression. While size has been shown to be important for the expression timing of specific genes during early development, it has not been extensively explored whether it plays a role in the timing of other global gene expression programs. Using the nascent RNA sequencing technique Bru-seq, we have identified immediate genome-wide transcriptional changes following cellular stimulation with serum or following induction of DNA damage. We have also tracked the dynamics of enhancer activation using BruUV-seq and found that enhancer activation (generation of eRNA) occurs very rapidly after stimulation. Immediately induced genes, including genes encoding transcription factors and miRNAs, display a wide range of sizes resulting in staggered production of their gene products which in turn leads to a cascade of dynamic regulation of downstream transcription. Induced genes show a broader range of sizes than repressed genes and housekeeping genes tend to be rather compact. The exact timing of the completion of RNA synthesis of genes may be fine-tuned by regulation of specific chromatin marks and we, and others, have found that a high density of H3K79me2 modifications associates with high transcription elongation rate suggesting a role for epigenetics in the regulation of expression timing. Using BruDRB-seq, we have found that inhibition of the methyltransferase DOT1L, resulting in genome-wide reduction of H3K79 methylation, dramatically reduces the elongation rate of transcription in cells. We also show a link between this histone modification and the efficiency of splicing. Despite the energy waste of having transcription complexes translocate trough large genes mostly consisting of intronic sequences, the evolutionary conservation of gene lengths speaks to the importance of size in achieving a proper temporal expression of our genes. Furthermore, while gene lengths are fixed in cells, epigenetic modulation through H3K79 methylation allows cells to optimize the timing of gene expression by regulating elongation rates.

HIGH-THROUGHPUT ASSAYS REVEAL miR-155'S CONTEXT-SPECIFIC REGULATION OF IMMUNE SYSTEM

Gabriel Loeb[1], Yuheng Lu[2], Jing-Ping Hsin[1], Christina Leslie[2], Alexander Rudensky[1]

[1]Memorial Sloan Kettering Cancer Center and HHMI, Immunology Program, New York, NY, [2]Memorial Sloan Kettering Cancer Center, Computational Biology Program, New York, NY

MicroRNAs (miRNAs) are critical post-transcriptional regulators of gene expression that repress their target mRNAs by mediating the interaction between RISC and cognate sites in 3'UTRs. Recent studies have begun to investigate whether miRNAs, like transcription factors, regulate their targets in a cell-type and context dependent manner. For example, alternative polyadenylation (ApA) produces cell-type specific changes in 3'UTR isoform expression, and relative "shortening" or "lengthening" of 3'UTRs can lead to loss or gain of miRNA binding sites. ApA-independent mechanisms, such as the relative abundances of RNA-binding proteins, can also lead to miRNA context specificity. Previous studies in various cell lines, however, have reached conflicting conclusions, which may suggest the importance of studying miRNA regulation in physiologically relevant cellular contexts.

miR-155 is an important regulator in the immune system and is up-regulated upon activation in multiple immune cell types. To precisely map the target sites of miR-155, we performed Argonaute iCLIP (individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation) in macrophages, dendritic cells, and T and B lymphocytes from WT and miR-155 KO mice. Quantitative analysis showed that a large fraction of potential targets are differentially bound by miR-155 between the four cell types, which also result in significant differences in expression at both the gene and 3'UTR isoform levels as revealed by RNA-Seq and PolyA-Seq in the corresponding cells. Overall, miR-155 targets are strongly enriched for genes with multiple 3'UTR isoforms. Among the multi-isoform target genes, there is a significant overlap between genes differentially regulated by miR-155 and genes exhibiting ApA between cell types. In depth analysis suggested that ApA-dependent microRNA specificity may be a combination of two mechanisms: (1) lengthening or shortening of 3'UTRs can result in gain or loss of microRNA target sites; and (2) the regulatory activity of the same microRNA target site may depend on its relative position within 3'UTR isoforms. Meanwhile, we have also found that in many cases the same target 3'UTR isoform can be differentially regulated between cell types, suggesting ApA-independent mechanisms of microRNA specificity.

DEEP LEARNING METHODS PREDICT CHROMATIN ACCESSIBILITY FROM GENOME ANNOTATION OF PUTATIVE PIONEER FACTOR BINDING SITES

Bryan J Lunt, Saurabh Sinha

University of Illinois at Urbana-Champaign, Computer Science, Urbana, IL

We created and trained sequence-based deep learning architectures for the prediction of chromatin accessibility at stage 5 of D. melanogaster embryogenesis. This important epoch of development brackets the maternal zygotic transition (MZT) when large numbers of embryonic genes are expressed for the first time. It is commonly thought that “pioneer” transcription factors bind to regulatory sequences in preparation for the MZT, opening chromatin for later binding by “settler” transcription factors. Accordingly, we represented the sequence through features based on binding site annotation for six selected transcription factors (TFs), including Zelda and Trithorax-like, that are commonly held to be “pioneer” transcription factors. These feature vectors were presented to deep learning methods including a Long Short-Term Memory (LSTM) network, an LSTM atop a convolutional neural network, and recurrent neural network models. The models were trained and tested using a cross-validation scheme on ~250 bona fidé enhancers from the FlyEnhancer database that are active during stages 4-6 and have discernible chromatin accessibility peaks. In qualitative evaluation, predictions generally captured important features, such as dips and peaks in accessibility within the enhancer. Systematic assessments of accuracy were also performed and will be reported upon.

To our knowledge, this is the first attempt to predict chromatin accessibility entirely from sequence, building upon recent literature that has revealed several sequence motifs associated with accessibility. We expect that further refinements of this system will allow it to be used with other drosophilae. We also hope that further work will assist in elucidating the mechanisms of pioneer transcription factors.

A RARE GENETIC "TRANSCRIPTOMOPATHY" SYNDROME LEADING TO INSIGHTS INTO MORE COMMON NEUROLOGIC DISORDERS

Gholson J Lyon, Jason O'Rawe, Yiyang Wu, Han Fang

Cold Spring Harbor Laboratory, Genetics, Cold Spring Harbor, NY

We have recently described a new X-linked genetic syndrome associated with mutations in TAF1, presenting with global developmental delay, intellectual disability, characteristic facial dysmorphology, and variable neurologic features, all in male cases. There are thus far at least twelve families with overlapping clinical presentations and with de novo or maternally inherited mutations. Two additional families harboring large duplications involving TAF1 were also found to share phenotypic overlap with the probands harboring missense mutations, but they also demonstrated a severe neurodegeneration phenotype. RNA-seq for one of the families suggests that the phenotype is associated with down regulation of a set of genes notably enriched for genes regulated by E-box proteins, and knockdown and mutant studies for this gene in zebrafish have a quantifiable, albeit small, effect on a neuronal phenotype. TAF1 is the largest subunit of the TFIID complex, and has been ranked 53rd among the top 1,003 constrained human genes in a recent population-scale study, suggesting a critical role for this protein in normal cellular function. Previous work in Drosophila cells has shown that TAF1 depletion increases the magnitude of the initial transcription burst and causes delay in the shutoff of transcription upon removal of the stimulus. It was shown that the magnitude of the transcriptional response to the same signaling event, even at the same promoter, can vary greatly depending on the composition of the TFIID complex in the cell. Consistent with the notion that TAF1 is important in controlling the binding patterns of TFIID to specific promoter regions, the set of genes conferring increased expression were enriched for TATA-containing promoters, suggesting an association between the depletion of TAF1 and increased expression of genes with the TATA-motif. The genomic region containing TAF1 has also been suggested to play an important role in a somewhat common condition in the Philippines called X-linked dystonia-parkinsonism (XDP), although the exact mechanism remains undetermined. XDP is an X-linked recessive movement disorder characterized by adult onset dystonia and parkinsonism, which leads to eventual death due to oropharyngeal dystonia or secondary infections. Studies investigating the molecular basis of XDP demonstrated aberrant neuron-specific TAF1 isoform expression levels in neuronal tissue containing TAF1 variants. A reduction in TAF1 expression is associated with large-scale expression differences across hundreds of genes, and studies in rat and mice brain also corroborate the importance and relevance of TAF1 expression patterns specific to neuronal tissues.

DISTINCT SPECIFICITIES OF THE ANDROGEN AND GLUCOCORTICOID RECEPTORS REVEALED USING FEATURE-BASED RECOGNITION MODEL ANALYSIS OF SELEX DATA

Liyang Zhang[1], Gabriella D Martini[2], H. Tomas Rube[2], Vincent D FitzPatrick[2,3], Harmen J Bussemaker[2,4], Miles A Pufall[1]

[1]University of Iowa, Biochemistry, Iowa City, IA, [2]Columbia University, Biological Sciences, New York, NY, [3]Columbia University Medical Center, Graduate Program in Molecular Biophysics, New York, NY, [4]Columbia University Medical Center, Systems Biology, New York, NY

The androgen (AR) and glucocorticoid (GR) nuclear hormone receptors are closely related transcription factors. They are believed to bind to DNA as homodimers with indistinguishable specificity through identical DNA binding surfaces, and yet each occupies distinct genomic loci to drive distinct gene expression programs. How this functional difference ensues is not well understood. Here, by combining SELEX-seq assays on the DNA binding domain of AR and GR with statistical modeling, we show that the intrinsic DNA binding preference of the two factors differ substantially. We present an iterative algorithm that can accurately quantify the free energy parameters of a biophysically motivated recognition model over DNA footprints of unprecedented length (~30bp) by fitting a feature-based generalized linear model. Use of this algorithm allows us to analyze contributions to the binding specificity well outside the 15bp core region. In these outer flanks AR, but not GR, shows a preference for poly-A sequences. Isothermal titration calorimetry measurements confirm the difference in intrinsic specificity, and point to an AR-specific enthalpy-driven binding mechanism that derives additional binding energy from a narrowed minor groove. Our analysis shows that this mode of recognition restricts AR from binding GR sites, although the converse is not true. This contrast provides a basis for the differential genomic occupancy exhibited by AR and GR in LnCaP cells, helping to explain the finding that GR can functionally substitute for AR in androgen independent prostate cancers. Taken together, our results demonstrate that differences in the intrinsic DNA binding specificity between closely related steroid hormone receptors exist and are functionally relevant. Our computational approach is general and widely applicable.

TRANSCRIPTIONAL TIMERS REGULATING MITOSIS IN EARLY *DROSOPHILA* EMBRYOS

Amir Momen-Roknabadi[1], Stefano Di Talia[2], David Robinson[3], Eric Wieschaus[1]

[1]HHMI/ Princeton University, Department of Molecular Biology, Princeton, NJ, [2]Duke University Medical Center, Department of Cell Biology, Durham, NC, [3]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ

Embryonic development is characterized by precise regulation of cellular behaviors both in space and time. Several experiments have elucidated the mechanisms underlying the precise spatial regulation of development. However, the mechanisms of temporal patterning throughout development remain poorly understood. In this study, we aim to understand how *Drosophila* embryos achieve the precise temporal sequence of mitoses observed during gastrulation. This pattern is regulated by the transcription of the cdc25 homolog, *string* (*stg*), which drives mitosis in 25 highly reproducible mitotic domains (MD). We hypothesize that accumulation dynamics of a small number of rate-limiting regulators control the time of activation of *stg* transcription. To test this model, we examine whether changing the dosage of different chromosomal regions using heterozygous deficiency embryos alters the temporal pattern of mitosis in MD1 and 2. Using a whole-genome screen we have identified several effectors that encode the underlying clock. Heterozygote embryos develop with normal spatial pattern. In theory, a rate-limiting factor might act in all mitotic domains to set up timing globally throughout the embryo. However, we have observed very few such global regulators. Instead most of the effectors we have identified act in domain-specific manners, shifting the relative timing of mitosis in MD1 and 2. These regulators are made up of both activators and repressors. For example, the timing of MD2 is delayed in *buttonhead* (an activator) heterozygotes, but advanced in *hairy* (a repressor) heterozygotes. We conclude that in temporal regulation, the rate of accumulation of the regulators set up the developmental clock. Our results challenge prevailing models of spatial regulation of patterning, leading to the idea that the role of transcriptional repressors is mainly to set the boundary of expression of their target genes. We propose that repressors act with activators within the same region to establish the precise temporal pattern of mitosis observed during *Drosophila* gastrulation.

DECODING ANIMAL DEVELOPMENT AT SINGLE CELL RESOLUTION

Amanda L Zacharias, Travis Walton, Elicia Preston, Shaili Patel, John I Murray

Perelman School of Medicine at the University of Pennsylvania, Department of Genetics, Philadelphia, PA

In animal development the single-celled zygote divides, and its progeny differentiate to produce an incredible diversity of terminally differentiated cell types. This process is regulated in large part by transcription factors and signaling pathways whose function is conserved from humans to invertebrates. These regulators act combinatorially such that a given factor can regulate distinct targets and fates in different cellular contexts. We employ the nematode worm *Caenorhabditis elegans* embryo as a model to study developmental gene regulation because of its invariant, fully enumerated lineage, ease of experimentation, and conserved developmental regulators.

We developed confocal microscopy-based automated lineage tracing methods, allowing the definition of expression patterns and dynamics of over 220 transcription factors (TFs) *in vivo* at single cell resolution throughout embryonic development. Integrating these data has identified regulators of early lineage identity across the organism. Computational and experimental analysis of the resulting regulatory networks has identified novel mechanisms important for developmental gene regulation.

For example, the Wnt signaling pathway is a major regulator of many of these regulators, despite their being expressed in diverse stages and cell types (contexts). This raises the question of how this context-specificity is encoded in regulatory DNA. We have found that quantitative differences in Wnt pathway activity integrate information across multiple cell cycles to help ensure context-specific expression.

A striking feature of the embryonic regulatory network contains is the high degree of redundancy – most lineages express multiple paralogous TFs and mutations in multiple TFs are often required to see penetrant developmental phenotypes. Analysis of single cell defects in single and double mutants suggests confirms this redundancy and suggests it exists to ensure robust, rapid development. Transcripts of these regulators are short and accumulate rapidly and to high levels. Together this suggests that redundancy in developmental regulatory networks may buffer against noise due from simultaneous differentiation and rapid cell cycle length.

HETERO-DGF: A NOVEL ALGORITHM TO DECOMPOSE HETEROGENEOUS BINDING FOOTPRINTS OF TRANSCRIPTION FACTORS

Ryo Nakaki, Shuichi Tsutsumi, Hiroyuki Aburatani

Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, Komaba 4-6-1, Meguro-ku, Tokyo, Japan

Transcription factors (TFs) recognize different genomic sites, thereby regulating gene expressions in a state-specific manner. The recognition of the state-specific binding sites by TFs is closely controlled by various patterns of binding motifs, where various transcription complexes of the TFs are recruited. In addition to recognizing of its particular binding motif, interactions of different co-factors may select the TF binding motifs distinct at the level of nucleotides. Recently, genome-wide DNase I hypersensitivity profiles obtained by DNase-seq experiments enable us to predict the TF binding footprints at the nucleotide level. Although there are some computational algorithms to identify TF footprints, they are limited to assigning a particular DNase I cleavage pattern to each binding motif.

We propose a novel algorithm to predict the heterogeneous TF footprints, through decomposing the binding motifs with different patterns of DNase I cleavages. Compared with the existing methods to predict TF footprints, our algorithm was about 1.5-times more accurate in identification of the footprints of TF recognizing short binding motifs (65~75% at true positive rates), such as ETS and GATA factors. Subsequently, this algorithm was applied to identify the cell-specific binding footprints of GATA2 in HUVEC (human umbilical vein endothelial cells) and K562. We found that GATA2 cell-specifically recognized diverse binding motifs including 'GATAA'-sequence, each of which corresponds to different DNase I cleavage patterns. Moreover, we identified that GATA2 heterogeneously recognize cell-specific and common binding motifs in each cell-type. These results have indicated that hetero-DGF is useful to predict state/site-specific DNA-recognition mechanisms of TFs.

INTEGRATIVE EPIGENOME PROFILING REVEALS TISSUE-SPECIFIC GENE REGULATIONS OF HUMAN VASCULAR ENDOTHELIAL CELLS

Ryuichiro Nakato[1], Youichiro Wada[2], Genta Nagae[3], Yuki Katou[1], Ryo Nakaki[3], Shuichi Tsutsumi[3], Toutai Mituyama[4], Hiroshi Kimura[5], Hiroyuki Aburatani[3], Katsuhiko Shirahige[1]

[1]Inst Mol Cell Biosci, Tokyo Univ, Tokyo, Japan, [2]Isotope Sci Cen, Tokyo Univ, Tokyo, Japan, [3]Res Cen Adv Sci, Tokyo Univ, Tokyo, Japan, [4]Biotech Res Inst, AIST, Tokyo, Japan, [5]Grad Sch Biosci Biotech, Tokyo Tech, Tokyo, Japan

Endothelial cells configure the most inner layer of vasculature and serve as lining in the whole body. Their phenotypes, physiological function and gene expression pattern are regulated by developmental signals and outer stimuli. We profiled epigenome states of 7 types (20 tissues) of human vascular epithelial cells, and integrated various genome-wide analyses for histone modifications, DNA methylation and gene expression. To overcome individual variations, we developed a method to normalize signal-to-noise ratio of ChIP-seq data, and identified 11,563 vascular cell-specific enhancers, among which 30% were distally located. Motif analysis of these enhancers identified significant enrichment of key transcription factors crucial for vascular cells. Combining with RNA Polymerase II-mediated ChIA-PET data of HUVEC cells, hundreds of annotated enhancers were linked to potential target genes. Epigenome-based clustering of cell types was well consistent with anatomical proximity, while transcriptome-based clustering appeared to be more environmentally affected. Our data provide an epigenomic basis for understanding vascular disease.

DISCOVERY AND CHARACTERIZATION OF AN EPIGENOMIC SIGNATURE OF LATE-STAGE DIFFERENTIATION

Khoi T Nguyen[1,2,3], Manolis Kellis[1,2]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Broad Institute, Cambridge, MA, [3]Massachusetts Institute of Technology, Biological Engineering, Cambridge, MA

Epigenetic modifications, including histone modifications and DNA methylation, help determine cell phenotype and have been shown to contribute to the cellular differentiation process in a few individual cell lineages. In this work, we aim to discover epigenomic changes characteristic of differentiation across various lineages and stages of differentiation. Given a tree structure based on prior knowledge of the differentiation relationships between samples, we systematically search for epigenetic features that can order the samples consistent with the tree. We applied this to chromatin state segmentations for 62 epigenomes generated by the Roadmap Epigenomics Consortium, including samples from pluripotent stem cell lines, the germ layers, neural progenitors, fetal and adult brain, mesenchymal and hematopoietic adult stem cells, and mesenchymal and hematopoietic tissues. We have found that while differentiation of pluripotent stem cells is associated with loss of bivalent chromatin states, as previously reported, the signature of later stages of differentiation is distinct and primarily marked by the loss of the "ZNF/Rpts" state. This chromatin state preferentially regulates recent additions to the human genome, such as the zinc finger gene family and repetitive DNA regions. We show that the loss of "ZNF/Rpts" influences lineage-specific gene expression, especially expression of repeat-derived transcripts, and is linked to enhancer activation. This work describes a novel epigenomic signature of late-stage differentiation and may lead to the discovery of novel biological mechanisms active in cellular development.

FUNCTIONAL ANALYSIS OF *DROSOPHILA* CORE PROMOTER ARCHITECTURE USING SYNTHETIC PROMOTERS

Miroslav Nikolov[1], Christophe Jung[1], Zhan Qi[1], Peter Bandilla[1], Claudia Ludwig[1], Ulrich Unnerstall[1], Mark Heron[1], Anja Kiesel[1], Johannes Soeding[1,2], Ulrike Gaul[1]

[1]Gene Center of the LMU, Biochemistry, Munich, Germany, [2]MPI for Biophysical Chemistry, Quantitative Biology and Bioinformatics, Goettingen, Germany

The core promoter is the primary determinant of gene expression level and plasticity and comprises approximately 150 bps surrounding the transcription start site. It guides the differential recruitment of the basal transcription machinery and directs transcription initiation by RNA polymerase. Genomic studies have shown that core promoter motifs and nucleosome occupancy play important roles in regulating the recruitment of the transcription machinery. To evaluate the effects of these features, we developed a synthetic biology approach to perform high-throughput structure-activity analysis of the *Drosophila* core promoter. In order to construct a combinatorial core promoter library, we defined seven sequence blocks representing distinct functional promoter elements: blocks 1 and 7 are representative -1 and +1 nucleosome positioning sequences; block 2 is a transcriptional activator binding region for ecdysone to investigate different states of transcriptional activity of the same underlying sequence; blocks 3-6 represent the core promoter region, comprising 19 divergent native promoters and a manually curated library of 10,000 mutated versions to dissect the role of promoter motifs, their combinations, location and spacing. The promoter elements are assembled in a defined order selectively or in a semi-random fashion by means of high-throughput automated cloning and delineated using next-generation sequencing. The promoter activity measurements are carried out in *Drosophila* S2 cells using a dual luciferase assay in a fully automated robotic pipeline. The resulting data are integrated into an in-house developed quantitative computational model of transcriptional regulation that realistically captures the underlying molecular mechanisms.

Proof-of-principle experiments using a set of native promoters showed differential expression spectrum covering more than 3 orders of magnitude, measured with very high reproducibility. A first round of measurements covering 1000 individual promoter constructs revealed a broad range of effects from single bp mutations or individual or pair-wise motif knock-outs. In addition, ecdysone treatment led to various levels of activation depending on the promoter class and type of mutation present. Taken together, these results validate the synthetic screening approach presented here and are being used to direct the combinatorial assembly of a library of 10,000 synthetic promoters for a fine-grained dissection of individual core promoter features.

BASAL GANGLIA DEVELOPMENT THROUGH THE LENS OF TRANSCRIPTION FACTOR PROFILING, ENHANCER ACTIVITY, AND SINGLE-CELL TRANSCRIPTOMICS

Magnus Sandberg[1], Linda Su-Feher[2,3], Shanni Silberberg[1], Anna Rubin[1], Iva Zdilar[2,3], Susan Lindtner[1], John L Rubenstein[1], Alex S Nord[2,3]

[1]UC San Francisco, Psychiatry, San Francisco, CA, [2]UC Davis, Psychiatry and Behavioral Sciences, Davis, CA, [3]UC Davis, Neurobiology, Physiology, and Behavior, Davis, CA

The embryonic basal ganglia contains distinct progenitor zones that give rise to interneuron subtypes during brain development. Understanding interneuron specification is of major interest given the critical role interneurons play in the brain and in neurological and neurodevelopmental disorders. Genomic signaling pathways governed by the interaction between transcription factors (TFs), regulatory elements, and chromatin regulate gene expression to give rise to neuronal subtypes in the basal ganglia during early brain development. Using complementary epigenomic and transcriptomic profiling, we are dissecting the basis of interneuron specification in the mouse brain. To characterize signaling cascades at a systems biology level, we are combining ChIP-seq with bulk tissue and single cell transcriptomics to profile TF binding, epigenomic state, and gene expression. As an example, we have characterized the activity of Nkx2-1 and Lhx6, TFs that are critical in interneuron specification. Via epigenomic comparison in *Nkx2-1* conditional mice, we show that Nkx2-1 binding at distal regulatory elements causes changes in histone modifications and correlated changes in gene expression. Nkx2-1 binding can drive both transcriptional activation and repression via regulating changes in chromatin state at distal regulatory elements. Combinatorial binding by Nkx2-1 and Lhx6 establishes transcriptionally permissive chromatin and directs activation of gene expression in a cell- and region-specific manner in cortical migrating interneurons. We are now extending this integrative approach to other TF pathways. In parallel, we are applying single cell transcriptomics to further understand cell identity and transcriptional regulation. In initial experiments, we successfully assigned cell classes from heterogeneous cell populations within the basal ganglia in the context of transcriptional identity and TF signaling networks. By integrating across genomic approaches, we hope to decipher the complex transcriptional networks guiding interneuron specification and basal ganglia development at the molecular, genomic, and single cell level.

REVEALING FUNCTIONAL REGULATORY ELEMENTS CONTROLLING THE MATERNAL TO ZYGOTIC TRANSITION DURING ZEBRAFISH DEVELOPMENT

Panos Oikonomou*[1], Carter Takacs*[2], Valeria Yartseva*[2], Saeed Tavazoie[1], Antonio J Giraldez[2]

[1]Columbia University, Department of Systems Biology, New York, NY, [2]Yale University, Department of Genetics, New Haven, CT

*These authors contributed equally

The maternal to zygotic transition is a fundamental example of cellular reprogramming whereby the early embryo clears out the inherited cellular state and establishes zygotic control. During this early stage the zygotic genome is considered transcriptionally silent and the embryonic cells utilize mRNAs and proteins inherited from the egg's cytoplasm to carry out cellular functions. Due to the lack of transcriptional activity, the post-transcriptional fate of messenger RNAs contributes predominantly to the transition's requisite gene expression patterns. We seek to understand the regulatory networks that determine the fate of maternal mRNAs. The primary known contributor to this process is a microRNA, miR-430, which is expressed at the onset of zygotic transcription and has been shown to target hundreds of maternal mRNAs for deadenylation and decay. Here, we show that there are several distinct programs of regulation of the maternal mRNAs that function independently of miR-430 and lead to either decay or stability. We have developed a novel experimental and computational approach to systematically characterize the functional role of thousands of RNA *cis*-regulatory sequences embedded in the maternal zebrafish transcriptome. Through *de novo* motif discovery, we uncover a catalog of novel functional regulatory motifs associated with the distinct regulatory patterns of the early embryonic program. These post-transcriptional *cis*-regulatory elements and their regulators participate in combinatorial functions that modulate gene expression during the maternal to zygotic transition.

ChIP-ATLAS: COMPREHENSIVE AND INTEGRATIVE DATABASE FOR VISUALIZING AND MINING ALL PUBLISHED CHIP-SEQ DATA

Shinya Oki[1], Tazro Ohta[2], Go Shioi[3], Ryo Nakaki[4], Osamu Ogasawara[5], Yoshihiro Okuda[5], Hideki Hatanaka[6], Chikara Meno[1]

[1]Grad. Sch. of Med. Sci, Kyushu university, Fukuoka, Japan, [2]DBCLS, Research Organization of Information and Systems, Chiba, Japan, [3]CLST, RIKEN, Kobe, Japan, [4]RCAST, University of Tokyo, Tokyo, Japan, [5]DDBJ, National Institute of Genetics, Mishima, Japan, [6]NBDC, Japan Science and Technology Agency, Tokyo, Japan

A growing number of researchers have published studies using ChIP-seq, with the raw sequence data being deposited into public repositories as SRAs (Sequence Read Archives). Although SRAs are freely accessible to all researchers, exploiting them requires sophisticated skills with command-line interface. In addition, several tens of thousand experiments are recorded as ChIP-seq SRAs, making it difficult to extract informative findings and speculation from across the amount of ChIP-seq data.
To address these challenges, we developed ChIP-Atlas, an integrative and comprehensive database available on web-based platform. In this database, almost all of the published ChIP-seq data (> 35,000) have been collected, processed to graphical format (BigWig), and peak-called. These data are graphically and integratively displayed on genome browser, IGV, enabling users to find the proteins bound to given genomic loci. Furthermore, ChIP-Atlas provides integrative analysis data calculated from across all ChIP-seq datasets. Thus, upon selecting a transcription factor of interest, users can know its potential target genes and colocalizing proteins. In addition, users' data are also acceptable to perform enrichment analysis using all published ChIP-seq data. If a list of genes or BED-formatted genomic intervals are submitted, ChIP-Atlas returns proteins whose bindings are enriched to the given data. Therefore, based on comprehensive and integrative analysis of published ChIP-seq data, ChIP-Atlas is able to provide researchers with findings and predictions much informative to understand gene regulatory networks. In this meeting, we will also present various predictions obtained with ChIP-Atlas.

URL: http://chip-atlas.org

DECODING GENOMIC REGULATION BY THE TYPE II NUCLEAR RECEPTORS

Ashley R Penvose*, Jessica L Keenan*, Katherine Tooley, Vijendra Ramlall, Ryan Webster, Nicole Akramoff, Trevor Siggers

Boston University, Biology, Boston, MA

*Authors contributed equally

Nuclear receptors (NRs) are transcription factors (TFs) that are critical mediators of endocrine and metabolic signaling. The ability of NRs to directly couple ligand binding to transcriptional changes has made them ideal drug targets for treatment of inflammatory and metabolic disorders. A fundamental problem in NR-mediated genomic regulation is understanding how NRs regulate distinct transcriptional responses despite sharing DNA-binding site preferences.. The type II NRs bind to DNA as heterodimers with a common partner, the retinoid x receptor (RXR). The binding sites for type II NRs are composed of direct repeats of two hexad half-sites (5'-AGGTCA-3') linked by a spacer of 1-5 nucleotides (DR1-5). Multiple type II NRs are reported to recognize DRs with the same spacer length, however these NRs have distinct yet overlapping genomic targets *in vivo*, suggesting that existing models of DNA-binding specificity are incomplete. Many studies have shown that the sequence of the flanks and spacer of the DR imparts specificity for NR binding, but this has yet to be thoroughly characterized in a high-throughput manner. To build a more comprehensive model of genomic regulation, we are using protein-binding microarrays (PBMs) to characterize the DNA-binding landscape of the type II NRs. We use custom-designed PBMs to interrogate the DNA-binding specificity of the type II NRs to thousands of different DNA sequences. Comparative analyses of PBM-based NR binding profiles are used to identify both NR-shared and NR-specific DNA binding sites. We integrate these results with ChIP-seq and gene expression datasets to create a genome-scale model of NR signaling that incorporates binding data for all type II NRs.

CRISPRESSO: SEQUENCING ANALYSIS TOOLBOX FOR CRISPR-CAS9 GENOME EDITING

Luca Pinello[1], Matthew C Canver[2], Megan D Hoban[3], Stuart H Orkin[2], Donald B Kohn[3], Daniel E Bauer[2], Guo-Cheng Yuan[1]

[1]Dana-Farber Cancer Institute/ Harvard TH Chan School of Public Health, Biostatistics and Computational Biology, Boston, MA, [2]Boston Children's Hospital, Dana-Farber Cancer Institute, Pediatric Oncology, Boston, MA, [3]University of California, M.I.M.G. and Pediatrics, Los Angeles, CA

CRISPResso is a software pipeline for the analysis of targeted CRISPR-Cas9 sequencing data. This algorithm allows for the quantification of both non-homologous end joining (NHEJ) and homologous directed repair (HDR) occurrences. Compared to existing tools, CRISPResso offers several novel features, including: batch sample analysis via command line interface, integration with other pipelines, tunable parameters of sequence quality and alignment fidelity, discrete measurement of insertions, deletions, and nucleotide substitutions (ignored by other methods), tunable windows around the cleavage site to minimize false positive classification, quantification of frameshift versus in-frame coding mutations, and distinction between NHEJ, HDR, and mixed mutation events. CRISPResso automates the following steps: 1. filtering low quality reads, 2. trimming adapters, 3. aligning the reads to a reference amplicon, 4. quantifying the proportion of HDR and NHEJ outcomes, and 5. determining the proportion of frameshift and in-frame mutations as well as detecting potential splice site mutations. A graphical report is generated to visualize mutagenesis profiles and plain text output files are also produced for further integrative analyses.This pipeline can be used for assessment of on-target editing efficacy as well as of off-target editing at selected loci. In addition, CRISPResso accommodates single or pooled amplicon deep sequencing and WGS datasets. The CRISPResso suite offers flexible tools to evaluate and quantitate genome editing outcomes from sequencing experiments, and for standardizing and streamlining analyses that currently require development of custom in-house algorithms. CRISPResso can be run either as a stand-alone command line utility http://github.com/lucapinello/CRISPResso or web application: www.crispresso.rocks.

CHARACTERIZING CONDITION SPECIFIC TRANSCRIPTION FACTOR BINDING WITH ATAC-SEQ

R Pique-Regi[1], D Watza[1], M Estill[1], S Chaudhry[1], H Shim[2], F Luca[1]

[1]Wayne State University, Genetics, Detroit, MI, [2]Purdue University, Statistics, West Lafayette, IN

Specific regulatory sequences control gene transcription response when a cell is exposed to changes in the cellular environment (e.g. drug treatment). Recent technical advances in functional genomics have facilitated the profiling of regulatory sequences across many cell-types and tissues, yet we are still very far from mapping the sequences that control cell transcriptional response to many external stimuli. Profiling across different environmental conditions the binding activity of these TFs can be quickly accomplished at a genome-wide scale with the recently developed technique ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. When coupled with a computational method such as CENTIPEDE, footprint models for TFs with known motifs can be generated across the genome to detect binding. To date, there are no methods that efficiently incorporate the information provided by paired-end sequencing which allows both the identification of the library fragment length as well as the two cleavage locations that generated the fragment. We have extended CENTIPEDE to utilize fragment length information to exploit the joint statistics of cleavage pairs. Our results indicate that paired-end sequencing provides a more informative footprint model for ATAC-seq libraries which leads to greater accuracy in predicting TF binding. These results were validated with ChIP-seq data (ENCODE Project) for multiple factors including CTCF, NRSF, NRF-1, and NFkB. We then assayed TF activity in lymphoblastoid cell-lines (LCLs) across multiple treatments (selenium, copper, retinoic acid and glucocorticoids) for which we previously determined significant differences in gene expression levels. From our initial sequencing results we were able to resolve 383 actively bound motifs across all conditions. We were also able to characterize 9,263 regions that have significantly changed chromatin accessibility (FDR < 10%) in response to GRs, copper and selenium using DEseq2. These results are also validated using a more powerful multiscale Poisson based approach which can detect changes at a smaller scale (<20bp). We have extended the CENTIPEDE model hierarchical prior to detect motifs that have differences in footprint activity in treatment vs. control experiments. For both metal ions we have detected a significant increase of binding for ETS and CRE motifs. Our results demonstrate that ATAC-seq together with an improved footprint model are excellent tools for rapid profiling of transcription factor binding activity to study cellular regulatory response to the environment.

GENE EXPRESSION NETWORKS OF TRANSCRIPTIONAL RESPONSE TO 78 CELLULAR ENVIRONMENT PERTURBATIONS

D Kurtz, G Moyebrailean, O Davis, C Harvey, A Alazizi, D Watza, Y Sorokin, N Hauff, F Luca, R Pique-Regi

Wayne State University, School of Medicine, Detroit, MI

Cells respond and adapt to changes in their cellular environment by altering the expression of multiple transcripts through a programed response across a network of interconnected genes. Using RNA-seq we can measure the correlation in gene transcriptional response for the entire genome and identify sets of interacting genes that constitute the circuits regulating the molecular response to specific environmental conditions (e.g. drug treatment). Here, we have implemented a weighted correlation network approach to construct a coexpression network from RNA-seq data collected from five different cell types (human umbilical vein endothelial cells, lymphoblastoid cells, peripheral blood mononuclear cells, smooth muscle cells, and melanocytes) exposed to 33 treatments and using 3 different individuals per cell type for a total of 78 distinct cellular environments. We used the R software package Weighted Gene Coexpression Analysis (WGCNA) to construct a large coexpression network for 14,527 genes. Of the genes analyzed, 7936 clustered into 87 highly connected network modules. For each module an eigengene value was calculated to summarize the gene expression within that module, and a we fitted a linear regression model to determine the extent to which gene expression within modules could be explained by genes within that module responding to a specific treatment. 21 modules showed a significant response to treatments (|T statistic| > 10). For example, we recovered a 72 gene module that showed significant response to retinoic acid. One of the highly connected hub genes within this module is DHRS3, which is known to attenuate retinoic acid synthesis to maintain the correct balance of intracellular retinoic acid levels during body axis-formation. We also identified several modules that share a similar response profile across different treatments. For example, the same 4 modules that responded strongly to the glucocorticoid treatment also respond to the aldosterone treatment, with the known glucocorticoid induced leucine zipper TSC22D3 gene (a transcription factor also known as GILZ) being a central hub in the module with the strongest response to glucocorticoids. GILZ is also known to be upregulated in kidney cells exposed to aldosterone, suggesting that this transcription factor regulates the genes physiologically relevant to a cell's exposure to both mineralocorticoids and glucocorticoids. Finally, we have identified many novel modules within our network that contain genes with expression levels strongly correlated to specific treatments.

COMPUTATIONAL ANALYSIS OF MRNA-SEQ IN MULTIPLE MUTANTS AND GENOME-WIDE AGO2 BINDING PROFILES PROVIDES DETAILED VIEW OF miR-17~92 ACTIVITY IN MOUSE EMBRYONIC DEVELOPMENT

Yuri Pritykin[1], Xiaoyi Li[1,2], Carla Concepcion[1,2], Yoon-Chi Han[1], Joana Vidigal[1], Ping Mu[1,2], Doron Betel[2], Andrea Ventura[1], Christina Leslie[1]

[1]Memorial Sloan Kettering Cancer Center, New York, NY, [2]Weill Cornell Medical College, New York, NY

miRNAs are associated with a variety of phenotypes. By forming a complex with a protein from Ago family and binding to 3'UTR of mRNAs, each of many miRNAs expressed in a mammalian cell regulates expression of hundreds of genes post-transcriptionally, with a limited effect on each gene. This complicates discovering true miRNA-mRNA interactions and better understanding miRNA functioning.

The gene miR-17~92 encodes six different miRNAs and is associated with a number of developmental disorders and cancers. We reanalyzed an RNA-seq data set from our previous study (Han et al., 2015) from multiple tissues of developing mouse embryo for multiple miR-17~92 alleles and combined it with genome-wide Ago2 binding profiles to get new insights on miRNA functionality in vivo.

Using generalized linear modeling allows to precisely associate gene expression changes with different miRNAs. We find that most of the effect on expression upon deletion of a single miRNA can be attributed to direct targeting, with little secondary effect. The exception is a large number of genes consistently under-expressed upon miR-19 deletion. A significant fraction of them are direct targets of miR-92, suggesting a cross-talk between members of miR-17~92.

What constitutes an effective miRNA target is a subject of debate. Combining gene expression and Ago2 binding data, we are able to select a small subset of true targets from a large pool of computational predictions from popular tools. We show efficient binding of Ago2 to 6-mer sites resulting in significant down-regulation of gene expression. We also show evidence of efficient non-canonical targeting.

We find that cooperation between different members of a polycistronic cluster miR-17~92 is mostly limited to additive effect of direct targeting of the same genes by different miRNAs of the cluster.

Besides miRNA regulation of mRNA levels, we find the effect on lncRNAs. In particular, miR-92 directly targets and down-regulates a highly conserved lincRNA cyrano previously found critical in zebrafish embryonic development.

AptaTRACE: ELUCIDATING SEQUENCE-STRUCTURE BINDING MOTIFS BY UNCOVERING SELECTION TRENDS IN HT-SELEX EXPERIMENTS

Phoung Dao[1], Jan Hoinka[1], Yijie Wang[1], Mayumi Takahashi[2], Jiehua Zhou[2], Fabrizio Coscta[2], John Rossi[1], John Burnett[2], Rolf Backofen[3], Teresa M Przytycka[1]

[1]NIH, NCBI, Bethesda, MD, [2]Beckman Research Institute of City of Hope, Department of Molecular and Cellular Biology, City of Hope, CA, [3]University of Freiburg, Department of Computer Science, Freiburg, Germany

Discovering of RNA binding motifs is important for understanding of functioning of regulatory RNA binding molecules as well as for developing RNA aptamers binding to specific targets. Aptamers short synthetic RNA/DNA molecules binding specific targets with high affinity and specificity. They are utilized in an increasing spectrum of bio-medical applications. Aptamers are identified in vitro via the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) protocol. SELEX selects binders through an iterative process that, starting from a pool of random ssDNA/RNA sequences, amplifies target-affine species through a series of selection cycles. HT-SELEX, which combines SELEX with high throughput sequencing, has recently transformed aptamer development and has opened the field to even more applications. HT-SELEX is capable of generating over half a billion data points, challenging computational scientists with the task of identifying aptamer properties such as sequence structure motifs that determine binding. However currently available motif finding approaches do not have the generality or scalability required for HT-SELEX data, and they do not take advantage of important properties of the experimental procedure.

We developed AptaTRACE, a novel approach for the identification of sequence-structure binding motifs in HT-SELEX derived aptamers. Our approach leverages the experimental design of the SELEX protocol and identifies sequence-structure motifs that show a signature of selection. Because of its unique approach, AptaTRACE can uncover motifs even when these are present in only a minuscule fraction of the pool. Due to these features, our method can help to reduce the number of selection cycles required to produce aptamers with the desired properties, thus reducing cost and time of this rather expensive procedure. The performance of the method on simulated and real data indicates that AptaTRACE can detect sequence-structure motifs even in highly challenging cell SELEX data.

NON-CODING GENETIC VARIANTS RESULT IN ALLELE-DEPENDENT BINDING OF TRANSCRIPTION FACTORS IN AUTOIMMUNE DISEASES

Mario A Pujato, Xiaoting Chen, Avery Maddox, Leah Kottyan, John Harley, Matthew T Weirauch

Cincinnati Children's Hospital Medical Center, Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati, OH

Disease association studies have revealed that ~90% of the disease-associated genetic variants fall within non-coding regions of the genome. Determining the functional impact of non-coding disease-associated variants is of utmost importance for understanding disease onset and progression. One likely mechanism of these variants is alteration of gene expression through modification of transcription factor (TF) DNA binding sites in an allele specific manner. A wealth of genomic information regarding TF binding is available in the form of chromatin immunoprecipitation experiments, followed by sequencing (ChIP-seq). The high-resolution nature of these experiments should in principle allow us to determine allele-dependent binding of TFs. Further, experiments performed in cells that are heterozygous at a given variant are ideal for identifying allele-depending binding, since the presence of both alleles in the same cells offers a natural control.

We have analyzed ~100 ChIP-seq datasets that significantly overlap with multiple loci containing genetic variants associated with ~20 different autoimmune diseases. Using a combination of publicly available and in-house generated genotyping data, we have identified heterozygous variants in the same cell lines in which the ChIP-seq experiments were performed. Through a novel computational pipeline, we find allele-dependent TF binding to autoimmune-associated variants involving dozens of TFs at hundreds of loci. A significant fraction of these events are likely functional, as manifested by multiple lines of evidence. First, examination of the alleles often reveals a loss or gain of a binding site for the assayed TF that is concordant with the allelic ChIP-seq read depths. Second, the variants we describe are often expression quantitative trait loci (eQTLs) in related cell types. Third, specific sets of TFs tend to display allele-sensitive binding to the same allele at multiple loci, suggesting allele-dependent formation of entire protein complexes. The aggregation of observations from multiple affected TFs may prove to be a powerful method to delineate the possible molecular mechanisms driving autoimmune diseases.

IDENTIFICATION OF SPLICING eQTLS IN LEAD-TREATED *DROSOPHILA* MODEL

Wen Qu[1], Roger Pique-Regi[2], Douglas M Ruden[1]

[1]Wayne State University, Pharmacology, Detroit, MI, [2]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI

Lead poisoning has been a major public health issue globally. In order to better understand how lead plays a role as a neurotoxin, we utilized the *Drosophila melanogaster* model to study the genetic effects of lead exposure during development and identify the lead-sensitive genes.
In our previous studies, we have successfully identified hundreds of control-treated or lead-treated expression QTLs by using RNA-seq analysis on eight-way recombinant inbred *Drosophila* lines. This time, we focused our attention on alternative splicing events upon lead exposure.
There have been a variety ways of targeting splicing QTLs. But not many people have compared all the available methods. By using the same RNA-seq dataset as the one in search for eQTLs, we utilized several different statistical methods, such as, exon expression level, transcription ratio, exon/gene fraction of reads and multivariate approach (sQTLseekeR) to find lead-specific splicing QTLs. We also created an ANOVA-like method to capture the sQTLs. As a result, we were able to target hundreds of lead-sensitive splicing QTLs on *Drosophila*. These discoveries might help to unravel the mystery of lead neurotoxicity.

A VERSATILE FRAMEWORK FOR LEARNING FEATURE-BASED PROTEIN-DNA RECOGNITION MODELS DIRECTLY FROM SELEX DATA

Chaitanya Rastogi[1,2,3], Gabriella Martini[1], H. Tomas Rube[1], Harmen J Bussemaker[1,2]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]Columbia University Medical Center, Department of Systems Biology, New York, NY, [3]Columbia University, Graduate Program in Applied Physics and Applied Mathematics, New York, NY

SELEX-seq [1] and HT-SELEX [2,3] are sequencing-based methods for elucidating the intrinsic DNA binding specificity of transcription factor (TF) complexes at high resolution. While the amount of raw information that modern SELEX provides is unprecedented, the computational methods for building DNA recognition models ("motifs") from these data are still far from mature. The standard is to tabulate of the relative enrichment of each oligomer of a given length [4], for which we have developed efficient software [5]. Unfortunately, having to use oligomer tables as an intermediate step for feature-based analysis [6] has two key disadvantages: (i) limited range over which readout can be analyzed, as counts decrease exponentially with footprint size; and (ii) requirement for prior *ad hoc* sequence-based alignment of different oligomers. We present a new and highly versatile framework for motif discovery from SELEX data that overcomes these limitations. It uses a hierarchical maximum likelihood approach to fit a feature-based biophysically motivated protein-DNA recognition model directly to the raw SELEX data. First, this allows us to consider base and shape readout in more detail and over a larger footprint than was possible before, as we illustrate by reanalyzing Hox heterodimer data. Second, we can now for the first time analyze shape readout for TFs with low binding specificity, which we demonstrate using Hox monomer data. We find that shape readout by the Hox N-terminal arm is already seen for the monomer, but is altered by the presence of the Exd cofactor. Our method produces rich, biophysically interpretable models from only a single round of SELEX-seq data. Additionally, our flexible modeling framework should be easily extendable to other sequencing-based assays.

[1] M. Slattery, T.R. Riley, P. Liu, N. Abe, P. Gomez-Alcala, R. Rohs*, B. Honig*, H.J. Bussemaker*, R.S. Mann*. (2011) Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox proteins. Cell 147:1270-82.
[2] Y. Zhao, D. Granas, G.D. Stormo*. (2009) Inferring Binding Energies from Selected Binding Sites. PLoS Comput. Biol. 5(12): e1000590.
[3] A. Jolma, et. al. (2013) DNA-Binding Specificites of Human Transcription Factors. Cell 152: 327-339.
[4] T.R. Riley, M. Slattery, N. Abe, C. Rastogi, D. Liu, R.S. Mann*, and H.J. Bussemaker*. (2014) SELEX-seq, a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol. Biol. 1196:255-78.
[5] http://bioconductor.org/packages/release/bioc/html/SELEX.html
[6] N. Abe, I. Dror, L. Yang, M. Slattery, T. Zhou, H.J. Bussemaker, R. Rohs*, R.S. Mann*. (2015) Deconvolving the Recognition of DNA Shape from Sequence. Cell 161:307-18.

NETWORK-DRIVEN DISCOVERY AND INTERPRETATION OF CANCER DRIVER MUTATIONS

Juri Reimand[1,2]

[1]Ontario Institute for Cancer Research, Informatics and Biocomputing, Toronto, Canada, [2]University of Toronto, Medical Biophysics, Toronto, Canada

Identifying driver mutations from cancer exome and genome sequencing data is essential to deciphering tumour biology and designing precision treatments. Information on pathways and molecular interaction networks can improve interpretation of cancer mutations and associating mechanism and clinical information. We hypothesize that many cancer driver mutations precisely modify interfaces encoded in small sites in proteins and DNA, leading to interaction losses and gains in networks. We developed computational strategies to find network-associated driver mutations and infer their impact on network topology. The mutation enrichment model ActiveDriver detects proteins with site-specific positive selection, and the machine learning method MIMP infers mutations that rewire kinase signalling networks by erasing existing phosphorylation sites and creating new phosphorylation sites in substrate proteins. We conducted pan-cancer analyses of post-translational modification (PTM) networks and showed their enrichment in known driver mutations, increased functional impact, frequent rewiring of network topology, and associations to clinical characteristics. We also studied PTM networks in the human population and found that inter-individual genome variation is significantly reduced in PTM sites while inherited disease mutations are significantly enriched. This emphasizes the importance of network-related variation in human physiology and cancer. We extended our approaches to full cancer genomes and investigated site-specific mutations in gene promoters, enhancers, and transcription regulatory networks, using data from the International Cancer Genome Consortium and the Roadmap Epigenomics project. Our network-centric approaches provide novel interpretation to known cancer mutations, help find new cancer drivers and risk modifier alleles, and characterise their biological mechanisms.

THE AFFINITY DATA BANK: AN ONLINE SUITE OF TOOLS FOR INVESTIGATION OF PROTEIN-NUCLEIC ACID AFFINITY MODELS AND BIOPHYSICAL ANALYSIS OF REGULATORY SEQUENCES

Todd R Riley[1], Cory Colaneri[1], Brandon Phan[1], Aadish Shah[2], Pritesh Patel[2], Zazil Villanueva-Esperon[1]

[1]University of Massachusetts Boston, Department of Biology, Boston, MA, [2]University of Massachusetts Boston, Department of Computer Science, Boston, MA

We present The Affinity Data Bank (ADB), a suite of tools that provides biologists with novel aids to deeply investigate the sequence-specific binding properties of a transcription factor (TF) or an RNA-binding protein (RBP), and to study subtle differences in specificity between homologous nucleic acid-binding proteins. Also, integrated with Pfam, the PDB, and the UCSC database, The ADB allows for simultaneous interrogation of protein-DNA and protein-RNA specificity and structure in order to find the biochemical basis for differences in specificity across protein families. The ADB also includes a biophysical genome browser for quantitative annotation of levels of binding – using free protein concentrations to model the non-linear saturation effect that relates binding occupancy with binding affinity. The biophysical browser also integrates dbSNP and other polymorphism data in order to depict changes in affinity due to genetic polymorphisms – which can aid in finding both functional SNPs and functional binding sites. Lastly, the biophysical browser also supports biophysical positional priors to allow for quantitative designation of the level of locus-specific accessibility that a protein has to the DNA. Importantly, the use of this toolset does not require bioinformatics programming knowledge – which makes the ADB tool suite highly useful for a wide range of researchers.

Along with the protein's sequence specificity, protein concentration is an important ingredient that also greatly affects the levels of protein-nucleic acid binding. In addition, as protein concentrations increase, the saturation of the highest-affinity binding sites additionally increases the levels of occupancy for functional medium and low-affinity sites. This biophysical, nonlinear relationship between free protein concentration, binding site affinity, and resultant binding is an important part of accurately determining the level of protein-DNA and protein-RNA binding under in vivo conditions.

Lastly, accurate protein-DNA affinity models are necessary but not sufficient enough to properly model and predict the level of *in vivo* protein-DNA binding and subsequent gene regulation. For example, in the human genome most possible binding sites are not accessible for binding by a TF-protein. Tissue-specific chromatin state and accessibility is a complex, major factor that heavily influences protein-DNA binding. Because many possible binding sites are actually inaccessible for binding, methods that do not include *in vivo* accessibility when searching for putative binding sites in or near a gene have a high false positive rate. The ADB can properly model *in vivo* protein-DNA binding by integrating the effects of chromatin accessibility and epigenetic marks via the inclusion of biophysical occupancy-based and affinity-based positional priors.

COMPUTATIONAL APPROACHES TO STUDY DYNAMICS OF LONG-RANGE GENE REGULATION

Deborah Chasman[1], Alireza F Siahpirani[2], Sara Knaack[1], Chang Wang[2], Rupa Sridharan[3,1], Sushmita Roy[1,2,4]

[1]Wisconsin Institute for Discovery, Systems Biology, Madison, WI, [2]University of Wisconsin, Madison, Computer Sciences, Madison, WI, [3]University of Wisconsin, Madison, Cell and Regenerative Biology, Madison, WI, [4]University of Wisconsin, Madison, Biostatistics and Medical Informatics, Madison, WI

Long-range gene regulation is emerging as a critical player in context-specific gene expression programs and in the interpretation of regulatory variation. Such interactions are established through a complex interplay of one-dimensional signals such as transcription factor proteins, architectural proteins such as CTCF, histone modifications and three-dimensional spatial proximity of genomic loci. Recent technological advances in genomics are expanding our repertoire of regulatory genomics datasets that can measure different aspects of the gene regulation machinery. However, identification of functional long-range regulatory interactions, that is, those that are associated with downstream phenotypic and expression changes; the important determinants of these interactions; and how long-range regulation impacts the global gene regulatory network wiring, is a major challenge. To address this challenge, we are developing novel computational methods to integrate and analyze chromatin state, transcription factor occupancies and three-dimensional proximity across multiple cell types and species, as well as, novel approaches to integrate these data to better understand cell-type specific gene regulatory networks. In particular, we have developed a computational approach, Module Inference on Trees (MInT), to identify regulatory modules, defined by sets of genes with similar chromatin states on a cell lineage. Application of MinT to mouse cellular reprogramming data identified novel multivalent chromatin states that are likely important transient states during cell-fate specification. We have also developed RIPPLE, to predict enhancer-promoter interactions between pairs of genomic loci by training a classifier on available chromosome conformation capture (3C) datasets. We have also developed an ensemble-based extension of RIPPLE to predict enhancer-promoter interactions in new cell types with no available 3C datasets. To identify principles of long-range gene regulation we are developing computational techniques based on multi-task learning to compare both, chromatin state, and, three-dimensional genome organization, across multiple species. As datasets measuring chromatin state and three-dimensional proximity expand to more cell types and species, our computational tools will be increasingly important to provide insight into the principles of dynamics of long-range gene regulation.

UBIQUITIN-MEDIATED DEGRADATION OF HISTONE H3 PROMOTES NUCLEOSOME TURNOVER AT GENE PROMOTERS TO REGULATE GENE EXPRESSION

Ronen Sadeh[1], Fuqiang Geng[2], William P Tansey[3], David C Allis[4]

[1]Chemical Biology, Jerusalem, Israel, [2]Genetic Medicine, NYC, NY, [3]Cell and Developmental Biology, Nashville, TN, [4]Laboratory of Chromatin Biology and Epigenetics, NYC, NY

Nucleosomes are highly dynamic structures that undergo replication-independent turnover at specific sites on chromatin including promoters, active genes, and telomeres. The rapid exchange of nucleosomes on chromatin is necessary for epigenome maintenance. Yet, the mechanisms which drive this exchange remain obscure. Here, we provide biochemical and genomic evidence that histone exchange at promoter-bound nucleosomes is driven by ubiquitylation and subsequent proteasomal degradation of histone H3 in situ. We show that promoter-bound histones are preferentially ubiquitylated and degraded by the proteasome, and that inhibition of proteasome function results in reduced nucleosome turnover—and reduced expression—of genes displaying high promoter H3 ubiquitylation. Our results puts localized clearance of nucleosomal histones by the ubiquitin-proteasome system as the key mechanism for nucleosome replacement at promoters and maintenance of nucleosome dynamics.

SHOGoiN DATABASE: A DATABASE TO INTEGRATE STANDARDIZED DATA AND DIVERSE KNOWLEDGE OF CELLS

Kunie Sakurai[1], Midori Yuji[1], Junko Yamane[1], Kenta Kobayashi[1], Nobuko Taniyama[1], Takeaki Taniguchi[2], Koji Yamanegi[3], Wataru Fujibuchi[1]

[1]Center for iPS Cell Research and Application, Kyoto University, Departmentof Cell Growth and Differentiation, Kyoto, Japan, [2]Mitsubishi Research Institute, Inc., Tokyo, Japan, [3]Hyogo College of Medicine, Department of Clinical Pathology, Hyogo, Japan

Along with growing fields of cell-based biomedical research, stem cells have been recognized as important tools in biomedical science, as they are useful systems for modeling how tissues and organs are formed to acquire a specific function in vitro. Using advanced high-resolution cell-based assay techniques, many data have been produced to observe characteristics of cells. Here we introduce our cell-centric knowledge database called 'SHOGoiN' which is designed to integrate cellular information comprehensively for defining cell types with diverse knowledge and scientific data from biomedical research. The database contains several modules such as transcriptome and methylome data, and diverse cell type-specific information obtained from the literature, based on a backbone of sophisticated cell taxonomy. Most of cell data and information in SHOGoiN are derived from human cells, however, we are now developing another taxonomy system for defining 'orthologous cells' among human and model animals such as mouse, nematode, and sea urchin, expecting more cellular information to be effectively collected beyond species, which have been accumulated well in the past decades. Using accumulated transcriptome data, SHOGoiN also supports a cell quality check tool named CellMapper. Importantly, we have recently proposed an international guideline for describing cellular assay data, which can be used as the data standardization format in SHOGoiN and other cell-related databases worldwide. We believe that SHOGoiN will play an important role in the "Big Data" era to effectively analyze cellular information as a reference dictionary.

TISSUE AND DEVELOPMENTAL TIME STAGE SPECIFIC ENHANCER PREDICTION USING PATTERN RECOGNITION

Anurag Sethi[1,2], Jing Zhang[1,2], Sushant Kumar[1,2], Joel Rozowsky[1,2], Mark Gerstein[1,2,3]

[1]Yale University, Computational Biology and Bioinformatics, New Haven, CT, [2]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT, [3]Yale University, Department of Computer Science, New Haven, CT

Enhancers are distal regulatory sequences containing a dense cluster of transcription factor binding sites that have the potential to increase the transcription of a target gene or genes. Enhancers play a critical role and are selectively activated at different developmental stages and in different tissues. Enhancer prediction is particularly challenging because they are sparsely distributed in the vast intergenic regions within the genome and no sequence-based signatures are known to be associated with enhancers. Due to the development of a number of massively parallel assays for identifying regulatory regions in the genomes, we have been able to identify the epigenetic signatures underpinning active enhancers. In this work, we utilized comprehensive experimental datasets released from the ENCODE consortia to identify potential active enhancers in various mouse and human tissues using a signal processing algorithm. The pattern within the signal of different epigenetic datasets will be computed from regulatory regions identified using different massively parallel assays and this pattern is conserved across a diverse set of species. This method was used to predict tissue and cell line-specific active enhancers based on H3K27ac ChIP-Seq datasets generated by the Epigenome Roadmap, and ENCODE projects. We show that our models lead to highly accurate and precise prediction of enhancers. We compare enhancer usage between tissue and developmental stage specificity of various enhancers in mouse and human and the conservation of enhancer usage between the two species.

WISHBONE IDENTIFIES BIFURCATING DEVELOPMENTAL TRAJECTORIES FROM SINGLE CELL DATA

Manu Setty[1,2], Michelle D Tadmor[1,2], Shlomit Reich-Zeliger[3], Sean Bendall[4], Nir Friedman[3], Dana Pe'er[1,2]

[1]Columbia University, Dept of Biological Sciences, New York, NY,
[2]Columbia University, Dept of Systems Biology, New York, NY,
[3]Weizmann Institute of Science, Dept of Immunology, Rehovot, Israel,
[4]Stanford University, Department of Pathology, Stanford, CA

We present Wishbone, an algorithm for aligning cells along bifurcating developmental trajectories with high resolution. Wishbone uses multi-dimensional single cell data as input and simultaneously orders cells according to their developmental trajectory and labels each cell as belonging to the trunk or to one of two cell fates. Thus Wishbone can pinpoint where the bifurcation happens and characterize the differences leading up to and following bifurcation. We applied Wishbone to characterize T cell development in the mouse thymus using 30 channel mass cytometry data. Wishbone accurately recovered the various known stages in T cell development including the bifurcation point. Wishbone also generalizes to branches in human myeloid differentiation and significantly outperforms existing methods in identification of both ordering and branching of cells.

THE REGULATORY LANDSCAPE OF ENDOCRINE DISRUPTOR EXPOSURE

Kyle T Siebenthall[1,2], Richard Sandstrom[1,2], Jessica Halow[1,2], Kristen Lee[1,2], R. Scott Hansen[2,3], Peter Sabo[1,2], Rajinder Kaul[1,2], John A Stamatoyannopoulos[1,2]

[1]University of Washington, Department of Genome Sciences, Seattle, WA, [2]Altius Institute for Biomedical Sciences, Seattle, WA, [3]University of Washington, Division of Medical Genetics, Seattle, WA

Endocrine disrupting chemicals (EDCs) are a class of chemicals that serve useful functions, such as in agriculture or plastic manufacture, but have unintended environmental and health-related consequences. EDCs mimic naturally occurring hormones, which are recognized by the nuclear hormone receptor family of transcription factors (TFs). Upon ligand binding, these TFs are activated and rapidly interact with target sites to modulate chromatin structure and gene expression. While there is extensive epidemiological evidence supporting negative impacts of EDC exposure in utero on later health outcomes, the molecular mechanisms by which these effects are mediated remain vague. To shed light on gene regulatory responses mediated by EDC exposure, we have focused on bisphenol A (BPA), diethylstilbestrol, genistein, and methoxychlor, all of which bind to the estrogen receptor (ER) by mimicking its natural ligand, beta-estradiol (E2). We treated the estrogen-responsive cell lines MCF-7 and T-47D with these EDCs at three different doses (as well as with a single dose of E2) continuously for either two hours, five days or three weeks, at which point we profiled chromatin accessibility as a genome-wide measure of TF occupancy using DNase-seq. In MCF-7 alone, we uncovered nearly 90,000 DNase I hypersensitive sites (DHSs) with significantly altered accessibility (FDR 5%, vs. control) indicative of changes in TF occupancy and chromatin remodeling in response to E2/EDC exposure. The response of the regulatory landscape differed by dose as well as time for each EDC, with changes largely distinct for each EDC. For example, treatment with E2 generally resulted in increased accessibility, indicative of TF recruitment, and this response was maximal after five days of treatment. And while BPA treatment impacted many of the same DHSs as E2 treatment, most alterations were unique. The enrichment of the ER recognition sites within DHSs impacted by both E2 and BPA is consistent with the induction of ER binding, however the enrichment of recognition sequences for other TFs within DHSs impacted only by BPA points to additional modulators of cellular responses to EDCs. Our ongoing analyses of these extensive datasets will elucidate the circuitry and network of regulatory factors that are modulated upon EDC exposure.

SINGLECELLQUEST! - EASY TO USE SINGLE-CELL-RNA-SEQ SOFTWARE FOR EXPLORING AND VISUALIZING DATA.

Joshua Starmer, Terry Magnuson

University of North Carolina at Chapel Hill, Genetics, Chapel Hill, NC

RNA-seq and single-cell-RNA-seq are exciting technologies, but the associated analysis software is not. Most, if not all it, is aimed at bioinformaticians and requires considerable experience with using a command line and programming. This excludes most bench scientists from the fun parts of high-throughput experiments - generating cool graphs and exploring the data. To rectify this dire situation, we developed singleCellQuest!. Using a point-and-click interface, users easily generate a wide range of killer graphs, including PCA plots to identify cell populations and awesome jitter plots to look at individual genes. Once the graphs and charts are created, the fun really begins when users highlight regions to identify samples and genes. singleCellQuest! ensures that everyone, from seasoned bioinformaticians to bench scientists, gets to have fun exploring their data.

THE IMPACT OF PRE-GRAVID OBESITY ON NEONATAL MONOCYTE GENE REGULATION AND EXPRESSION

Suhas Sureshchandra[1], Randall M Wilson[1], Jonathan Q Purnell[2], Kent Thornberg[2], Nicole E Marshall[2], Ilhem Messaoudi[1]

[1]University of California-Riverside, Division of Biomedical Sciences, Riverside, CA, [2]Oregon Health & Science University, Department of Medicine, Portland, OR

Obesity during pregnancy is associated with adverse health outcomes for the offspring that can persist into adulthood, notably a higher lifetime risk of developing obesity, asthma, diabetes, and cardiovascular disease, which lead to a higher all-cause mortality rate compared to offsprings of lean mothers. These diseases have a significant inflammatory component, suggesting that maternal obesity impacts development and maturation of the neonatal immune system. More importantly, we have recently shown that monocytes in the human umbilical cord blood mononuclear cells (UCBMC) collected from babies born to obese mothers generate dampened responses following TLR4 and TLR1&2 agonists compared to monocytes collected from babies born to lean mothers. To uncover the molecular mechanisms underlying functional changes in monocytes, we compared the transcriptomes and DNA methylomes using RNA-Seq and targeted bisulfite sequencing respectively. Our analysis identified 4803 hyper- and 8811 hypo-methylated cytosines in monocytes purified from UCBMC collected from babies born to obese mothers compared to those born to lean mothers. These aggregated into 787 (273 hyper- and 514 hypo-methylated) differentially methylated regions (DMRs), with 537 DMRs overlapping annotated gene bodies. Broadly, the hyper-methylated DMRs mapped to genes involved in transcriptional regulation (*HDAC4*, *RERE*, *SATB1*) and signal transduction (*CD44*, *ITPR3*, *PRKCH*). On the other hand, the hypo-methylated DMRs overlapped with genes encoding kinases and components of defense response (*ITGB5*, *ADORA3*, *JAK1*, *PDE7B*). Additionally, DMRs in intergenic regions mapped to cis-regulatory regions that have been shown to regulate genes involved in toll-like receptor signaling and immune system development. At the transcriptomic level, we detected 109 differentially expressed genes (DEGs) that play a role in immune function (*IL6*, *CXCL3*, *CEBPA*, *HLADQB1*) and cell-cycle progression (*ZBTB16*, *S100A4*). Interestingly, although the methylation changes correlate inversely with gene expression on a global level, we saw a poor correlation between DNA methylation and expression at the gene level. In summary, these preliminary findings show that obesity during pregnancy exerts a significant impact on gene regulation and expression within the offspring's innate immune cells, which is likely to result in functional differences. Findings from this study will help identify potential therapeutic targets to treat pre-gravid obesity-induced inflammation in the neonate.

TRANSCRIPTIONAL NETWORKS OF M2 MACROPHAGES DERIVED FROM RECRUITED MONOCYTES OR TISSUE RESIDENT MACROPHAGE LINEAGES

Mei San Tang[1], Emily Miraldi[2,3,4], Natasha Girgis[1], Richard Bonneau[2,3,4], P'ng Loke[1]

[1]New York University School of Medicine, Microbiology, New York City, NY, [2]Center for Genomics and Systems Biology, New York University, Biology, New York City, NY, [3]Simons Foundation, Simons Center for Data Analysis, New York City, NY, [4]Courant Institute of Mathematical Sciences, New York University, Computer Science, New York City, NY

Bacterial and viral infections trigger type 1 immune responses, exemplified by cytokines such as interferons (IFNs), which activate macrophages to adopt the classical M1 phenotype. In contrast, M2 macrophages are alternatively activated by type 2 cytokines (interleukin-4 (IL-4) and IL-13) that are characteristic of the immune response to parasitic helminth infections. M2 macrophages are also involved in a wide range of biological processes, such as atherosclerosis plaque regression, obesity, asthma and cancer. Macrophages can differentiate either from circulating blood monocytes or through proliferative self-renewal of tissue-seeded embryonic precursors. We have previously found that M2 macrophages of monocytic and embryonic origins are phenotypically and functionally distinct. Here, we further investigate the mechanisms regulating the activation of M2 macrophages from different cellular origins. Using a combination of computational and experimental methods, we aim to identify transcription factors (TFs) regulating lineage-specific M2 activation by generating transcriptional regulatory networks (TRNs). We use the “Inferelator” to generate TRNs, which solves an ordinary differential equation (ODE) by regression. We combine gene expression data with TF binding motifs enriched in accessible chromatin regions (generated from ATAC-Seq analysis of M2 macrophages from different lineases) to estimate TF activity and incorporate this as prior knowledge to influence model selection during regression. Top-ranking candidate TFs regulating lineage-specific macrophage functions will be experimentally validated downstream using *in vitro* and *in vivo* models of M2 macrophages deficient in specific TFs. These studies should identify novel pathways and targets regulating M2 macrophages in various helminth infections and other chronic inflammatory responses.

DIFFERENTIALLY CO-EXPRESSED GENES ARE CRITICAL REGULATORS IN CO-EXPRESSION NETWORKS

Lina D Thomas[1], Dariia Vyshenska[2], Anatoly Yambartsev[1], Andriy Morgun[2]

[1]Universidade de Sao Paulo, Instituto de Matematica e Estatistica, Sao Paulo, Brazil, [2]Oregon State University, College of Pharmacy, Corvallis, OR

Co-expression networks are a tool vastly used to analyze Big Data which can range from microarrays to next-generation sequencing (for genotyping or transcriptome profiling) and mass spectrometry-based proteomics and metabolomics data. Such networks show how genes are interconnected and their effect on each other. Differential co-expression analysis is a recent approach that measures how the interactions between genes change when a biological system transitions from one condition to another. A network can either gain or lose an edge or switch correlation directions in a state transition. It is already know the importance of differentially co-expressed (DC) genes to identify dys-regulated pathways, but their location and role in a co-expression network is still an unanswered question. Here we investigate how close DC genes can be to causal genes in Differentially Expressed Genes (DEGs) co-expression networks and therefore influence the pathway flow changes that lead to phenotype alterations. We studied two systems which we already know the causal genes. The first one has only one causal factor: Bcell knockout. We show that the immunoglobulin genes (causal genes) are highly enriched with DC genes. The second system, which is cervical cancer, has strong heterogeneity for causal factors of gene expression alterations and their key drivers are DEGs located in chromosomal aberrations. We disclosed a very newsworthy behavior: DC genes act as bottlenecks. Most flows coming from the key drivers to the peripheral genes (connected by only one edge) must pass through them. Our in vitro experimental settings involving two DC genes, FGFR2 and CACYBP, showed that they play critical regulatory roles in cancer pathogenesis. Therefore locating DC genes in DEGs co-expression networks is an important tool to assist detecting the major genes responsible for changes in phenotype.

A BIAS-CORRECTING CAPTURE HI-C ANALYSIS REVEALED THE INTERACTIONS BETWEEN P53 BINDING SITES AND THE TARGET GENES

Shuichi Tsutsumi[1], Atsushi Okabe[2], Hiroyuki Aburatani[1]

[1]University of Tokyo, RCAST, Genome Science Div., Meguro, Japan,
[2]Chiba University, Department of Molecular Oncology, Chiba, Japan

Gene activities are controlled by a combination of proximal and distal regulatory elements that interact with each other despite their genomic distances. The keys among these are enhancers, which associate with promoters to activate cell type- or disease-specific gene expression. A large number of enhancer regions have been annotated in the human genome by their hyper- sensitivity to DNaseI digestion and binding of chromatin modifier, but the principles defining the relationships between these regulatory elements and distal target genes remain unknown. In addition, disease-associated SNPs are often located at enhancer regions; therefore, a chromatin interaction assay to identify direct targets of enhancer regions is important to elucidate disease susceptibility mechanism. We aim to elucidate the relationship between chromatin dynamics between cis-regulatory elements and transcriptional regulation.
To investigate globally how active enhancers dynamically interact with their target genes, we used Hi-C and Capture Hi-C assay in p53-positive HCT116 cells treated with a thymidylate synthase inhibitor 5-Fluorouracil (5-FU). However, capturing process makes the observed interactions skewed with the count of overlapping to the bait regions. We applied Bayes theorem (reported as 'Mango') for calculating the probabilities. Using even 'SureSelect All Exon + Regulatory' (containing 220,000 bait regions), more than 181,000 significant interactions were found ($p < 10^{-4}$). Over 96% of those were inside of a topologically associated domain (TAD). Active enhancers have a tendency to interact with active promoters. For Capture Hi-C, our pipeline and viewer are able to find feasible genomic interactions using complicated Capture baits.

TRANSIENT ELEVATIONS IN NEURONAL ACTIVITY INDUCE A SUBSET OF THE NEURONAL ACTIVITY-REGULATED GENE PROGRAM DEFINED BY MAPK/ERK-DEPENDENT ENHANCERS.

Kelsey M Tyssowski, Jesse M Gray

Harvard Medical School, Department of Genetics, Boston, MA

Cell signaling pathways translate neuronal activity into gene transcription that is required for the cell biological changes underlying long-term memory formation. The specific activity-dependent changes that take place in a neuron depend on both the pattern of activity and the corresponding set of genes that are induced. However, it is unclear how specific activity patterns deferentially induce subsets of the activity-regulated gene program. We compared two distinct patterns of activity, transient and sustained, and found that transient activity induces expression of only a subset of the activity-regulated genes that are induced by sustained activity. The genes induced by transient activity have in common a requirement for the MAPK/ERK pathway that is not shared by the remainder of the activity-regulated gene program. Transient activity-induced genes are also located near MAPK/ERK-dependent enhancers. Upon MAPK/ERK inhibition, these enhancers have reduced activity-induced eRNA expression but no change in H3K27ac. These data suggest a mechanism by which subsets of the neuronal activity-regulated gene program are expressed in response to particular patterns of stimuli.

SYSTEMS ANALYSIS OF THE AUTOIMMUNE RESPONSE IN PRIMARY SCLEROSING CHOLANGITIS

D.B.R.K. Gupta Udatha[1], Brian K Chung[1,2], Adi Mehta[3], Fridtjof Lund-Johansen[3], Evaggelia Liaskou[2], Gideon M Hirschfield[2], Tom Hemming Karlsen[1]

[1]Oslo University Hospital Rikshospitalet, Norwegian PSC Research Center, Oslo, Norway, [2]University of Birmingham, NIHR Birmingham Liver Biomedical Research Unit, Birmingham, United Kingdom, [3]Oslo University Hospital Rikshospitalet, Department of Immunology, Oslo, Norway

Primary sclerosing cholangitis (PSC) is a cholestatic autoimmune liver disorder of unknown etiology characterized by prominent hepatic immune infiltrate and extensive fibrotic strictures of the intra- and extra-hepatic biliary ducts. Genome-wide association studies (GWAS) have shown that the genetic architecture of PSC closely resembles other prototypical autoimmune disorders such as celiac disease and type 1 diabetes where the strongest genetic risk factors reside within the HLA alleles of the major histocompatibility complex. In PSC, the contribution of the overwhelming HLA association is unclear. In this work, we present the systems level analysis of the autoimmune response from PSC patients for the identification of novel PSC autoantigens and their protein targets / antigenic triggers.

DREISS: USING STATE-SPACE MODELS TO INFER THE DYNAMICS OF GENE EXPRESSION DRIVEN BY EXTERNAL AND INTERNAL REGULATORY NETWORKS

Daifeng Wang[1], Fei He[2], Sergei Maslov[2], Mark Gerstein[1]

[1]Yale University, Molecular Biophysics and Biochemistry, New Haven, CT, [2]Brookhaven National Laboratory, Biological, Environmental and Climate Sciences, Upton, NY

Gene expression is controlled by combinatorial effects of regulatory factors from different biological subsystems driving specific regulatory functions such as general transcription factors, cellular growth factors and microRNAs. A subsystem's gene expression may be controlled by its internal regulatory factors, exclusively, or by other external subsystems, or by both. It is thus useful to distinguish the degree to which a subsystem is regulated internally or externally – e.g., how species-specific regulatory factors affect the expression of conserved genes during evolution.

We developed a computational method (DREISS) for dynamics of gene expression driven by external and internal regulatory modules based on state space model to help dissect the effects of different regulatory subsystems on gene expression. To demonstrate capabilities of DREISS, we study the regulatory effects of evolutionary conserved vs. divergent transcription factors across distant species. In particular, we applied it to the time-series gene expression datasets of C. elegans and D. melanogaster during their embryonic development. We analyzed the expression dynamics of the conserved, orthologous genes (orthologs), seeing the degree to which these can be accounted for by orthologous (internal) versus species-specific (external) transcription factors (TFs). We found that between two species, the orthologs have matched internally driven expression dynamic patterns but very different externally driven patterns. This is particularly true for genes with evolutionarily ancient functions (e.g. the ribosomal proteins), in contrast to those with more recently evolved functions (e.g., cell-cell communication). This suggests that despite striking morphological differences, some fundamental embryonic-developmental processes are still controlled by ancient regulatory systems. Furthermore, after clustering orthologous genes into gene co-expression modules, we found the developmental hourglass patterns at the network level; i.e., these modules have high modularity and are tightly connected with each other around the phylotypic stage, which suggests that the conserved functions have to coordinate at the middle stages rather than the early or late stages in embryonic development.

DE-NOVO INFERENCE OF ENHANCER-GENE NETWORKS IN DIVERSE CELLULAR CONTEXTS REVEALS THE LONG-RANGE REGULATORY IMPACT OF DISEASE-ASSOCIATED VARIANTS

Jianrong Wang[1,2,4], Peyton Greenside[3], Anshul Kundaje[1,2,4], Manolis Kellis[1,2]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Lab, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Stanford University, School of Medicine, Stanford, CA, [4]Current address: Stanford University, Department of Genetics, Stanford, CA

Non-coding variants implicated in genome-wide association studies (GWAS) are enriched in enhancer elements active in disease-relevant cellular contexts. Identifying context-specific target genes and downstream pathways affected by enhancers harboring regulatory variants remains a challenge. We developed a novel mixed-membership probabilistic model and a new non-linear association statistic that leverage the modular dynamics of gene expression and multiple enhancer-associated chromatin marks (H3K27ac, H3K4me1, H3K27me3 and DNase-seq) across hundreds of diverse human cell types and tissues from the Roadmap Epigenomics and ENCODE projects to infer highly-connected, tissue-specific enhancer-gene networks. Chromatin conformation maps and expression QTLs validate the superior accuracy and tissue-specificity of our predicted networks compared to existing approaches. While half of the predicted links involve elements less than 50 kilobases apart, only a third associate enhancers with their nearest genes. Linked enhancers significantly improve tissue-specific regression models of gene expression. Distal co-association of regulatory sequence motifs suggests synergistic regulation of genes by multiple enhancers with a key role for protein-protein interactions between lineage-specific transcription factors in mediating enhancer-promoter interactions. We further predicted specific transcription factor pairs as mediators of enhancer-promoter associations in different cellular contexts, providing hypotheses for in-depth experimental studies on enhancer regulatory mechanisms. Networks of cooperating enhancers with shared motif composition and target genes are depleted of disease-associated variants, suggesting regulatory buffering mechanisms. We demonstrate the utility of our context-specific enhancer-gene links as a rich resource to predict putative target genes, biological processes and pathways of non-coding variants associated with diverse traits and diseases ranging from autoimmune disorders to colorectal cancer.

THE COVERAGE-BASED PARADIGM FOR CISTROME DISCOVERY

Lonnie R Welch, Rami Al-Ouran

Ohio University, Bioinformatics Laboratory, Athens, OH

Computational discovery of transcription factor binding sites presents several difficulties. First, motif discovery algorithms often produce an overwhelming number of putative binding sites, which cannot be tested in the lab due to financial constraints and time constraints. Second, motif discovery algorithms sometimes have poor sensitivity, specificity, and accuracy. Motif discovery ensembles can help to address the second difficulty, but may do so at the cost of exacerbating the first problem.

The authors have developed a multi-stage approach that considers accuracy, sensitivity, and specificity, while producing a minimal set of motifs. Stage 1 employs a powerful motif discovery ensemble that produces a rich set of features (addressing the problems of poor performance of individual motif discovery tools). Since motifs produced by individual motif discovery algorithms may not be optimal (with respect to coverage of TF binding sites), stage 2 combines similar motifs. The goal of this stage is to produce a set of features called super motifs, which are an aggregation of individual motifs; super motifs are constructed by combining the views of diverse motif discovery tools, and thus they are expected to be more representative of the actual TFBSs than are the individual motifs. This is a three-step process: (a) generate a motif similarity matrix, (b) cluster similar motifs, and (c) optimize each cluster (refine precision by removing noise and keeping a set of motifs that work together to enhance one or more of the following: accuracy, sensitivity and specificity). Stage 3 uses classification techniques (specifically, Random Forests and Decision Trees) to extract sets of motifs that provide high accuracy. Stage 4 offers two alternatives to further refine the set of motifs identified in stage 3. A greedy heuristic finds a minimal set of features that maximizes sensitivity. A genetic algorithm finds a minimal set of features that maximizes sensitivity and specificity.

The methods have been refined via analysis of Chip-seq data and microarray data in a variety of biological contexts, including human disease biology, nematode genome annotation, and tissue-specific gene expression in plants.

A SIMPLE REGULATORY GRAMMAR DISTINGUISHES ACTIVATING FROM REPRESSING CIS-REGULATORY ELEMENTS IN PHOTORECEPTORS

Michael A White*[1], Jamie C Kwasnieski*[1], Connie A Myers[2], Joseph C Corbo[2], Barak A Cohen[1]

[1]Washington University in St. Louis School of Medicine, Center for Genome Sciences & Systems Biology, Department of Genetics, St. Louis, MO, [2]Washington University in St. Louis School of Medicine, Department of Pathology, St. Louis, MO

*Authors contributed equally

Different cis-regulatory elements (CREs) targeted by the same transcription factor (TF) often produce strikingly different transcriptional responses in the same cell, ranging from strong activation to repression. How CREs respond to a TF depends on the collection of TF binding sites they contain, yet the cis-regulatory grammar that relates the number and arrangement of binding sites with CRE activity is usually unknown. Additionally, the activity of distal CREs is modified by their interaction with proximal promoters, and those interactions depend on the cis-regulatory grammar in ways that are currently not predictable.

We investigated these phenomena by using a massively parallel reporter assay to measure the activity of thousands of genomic and synthetic DNA sequences with binding sites for the photoreceptor TF Crx. We assayed these sequences in wild-type and $Crx^{-/-}$ murine photoreceptors, and in the presence of different basal promoters. We found that whether CREs activate or repress transcription depends critically on the total number and strength of binding sites for Crx. Using binding site number and strength as a proxy for Crx occupancy, we find that CREs with high predicted Crx occupancy repress transcription, whereas CREs with lower Crx occupancy drive activation. This “rule” is modified in predictable ways by a cooperative interaction between Crx and Nrl, which promotes strong transcriptional activation by CREs with both low and high predicted Crx occupancy. Together these two rules also explain how CRE activity is modified when paired with different promoters. These rules account for the concordance between the activity of genomic CREs and simple synthetic CREs, which demonstrates that synthetic CREs capture essential features of the regulatory grammar that governs Crx targets in the genome.

We conclude that two simple rules, the total occupancy of Crx and the interaction of Crx with Nrl, govern the behavior of both activating and repressing photoreceptor CREs, as well as the interaction of these CREs with different basal promoters. Our results show how simple rearrangements of TF binding sites direct qualitatively different but predictable cis-regulatory responses to the same TF, and suggest that a single regulatory grammar accounts for both the differences among CREs and the interaction of those CREs with proximal promoters.

A COMPUTATIONAL STRATEGY TO ADJUST FOR COPY NUMBER IN TUMOR HI-C DATA

Hua-Jun Wu[1,2], Franziska Michor[1,2]

[1]Dana-Farber Cancer Institute, Department of Computational Biology and Biostatistics, Boston, MA, [2]Harvard School of Public Health, Department of Biostatistics, Boston, MA

Our knowledge of the higher-order structure of the genome has rapidly expanded over the last decade with the development of several methods able to elucidate the non-linear spatial conformation of the genome. One important contribution was the development of a chromatin conformation capture (3C)-based method called Hi-C, which enables high-throughput analysis of spatial structures of chromatin.
Raw Hi-C data has been observed to suffer from both technical and biological biases, with three predominant sources of bias identified so far: fragment length, GC bias and mappability. To correct for these biases, many software packages have been developed in order to generate an unbiased interaction map. The predominant approach performs an iterative correction (IC), which does not require a priori knowledge of the biases.
A novel source of bias that can arise in Hi-C data is related to the copy number of genomic material. This type of bias has so far been unaccounted for since most Hi-C applications investigate normal tissue and healthy cell line samples, which have a uniform copy number of chromosomes. However, once tumor samples are analyzed, biases related to copy number alterations become important and need to be corrected in order to obtain an accurate view of the interaction map between genomic locations.
We first identified the bias caused by DNA copy number by analyzing the ENCODE K562 Hi-C data. Surprisingly, we found that the copy number bias still exists after within-chromosome iterative correction (also called ICB correction here). Further analyses demonstrated a between-chromosome copy number bias in the ICB-corrected Hi-C map, which cannot be corrected simply by using total or average contact counts of chromosomes. We thus designed a linear regression-based chromosome-level adjustment method called caICB to correct for this bias. We performed the analyses on multiple resolution contact maps (1Mb, 250Kb, 100Kb and 10Kb) and found that the performance of our caICB correction is significantly better than the original ICB correction in terms of correcting for copy number bias. Our analyses show that the three previously identified bias factors are also accurately corrected for by using caICB. Furthermore, the caICB correction is robust when using a small subset of genomic ranges instead of using the whole genome contact map, and is easy and fast to apply even for extremely high-resolution maps. Our method does not require copy number data for the samples for which Hi-C data is available, and has the potential to adjust for other biases in Hi-C data without a priori knowledge.

COUPLING PHENOTYPIC PERSISTENCE TO DNA DAMAGE INCREASES GENETIC DIVERSITY UNDER SEVERE STRESS CONDITIONS

Gilad Yaakov*@[1], David Lerner*[1], Kajetan Bentele*[1,2], Joseph Steinberger[1,3], Naama Barkai@[1]

[1]Weizmann Institute of Science, Molecular Genetics, Rehovot, Israel,
[2]Berlin Institute of Health, Core Unit Bioinformatics, Berlin, Germany,
[3]Weill Cornell Medical College, Physiology and Biophysics, NY, NY

Mutation rate balances the need to protect genome integrity with the advantage of evolutionary innovations. Microorganisms increase their mutation rate when stressed, perhaps addressing the growing need for evolutionary innovation. Such a strategy, however, is only beneficial under moderate stresses that allow cells to divide and realize their mutagenic potential. By contrast, severe stresses rapidly kill the majority of the population with the exception of a small minority of cells that are in a phenotypically distinct state termed persistence. The stochastic event triggering persistence is poorly understood. We report that DNA damage is a key trigger of persistence in budding yeast, and show that the persister subpopulation carries an increased load of genetic variants, unrelated to their stress survival. Phenotypic persistence may therefore function to increase genetic diversity specifically in severe stress conditions where division potential, and therefore the ability to generate de-novo mutations, is limited.

* equal contribution, @ corresponding authors

CX-4945, A POTENT CK2 INHIBITOR, MEDIATES ALTERATIONS IN ENHANCER AND HETEROCHROMATIN LANDSCAPES IN ACUTE MYELOID LEUKEMIA

Naomi Yamada[1], Sadie Steffens[2], Kimberly Payne[3], Shaun Mahony[1], Sinisa Dovat[2]

[1]Penn State University, Dept of Biochemistry & Molecular Biology, University Park, PA, [2]Penn State University Medical College, Dept of Pediatrics, Hershey, PA, [3]Loma Linda University, Division of Anatomy, School of Medicine, Loma Linda, CA

A precise characterization of an anti-cancer drug's molecular and epigenetic mechanisms is key to designing effective therapeutic strategies. Acute myeloid leukemia (AML) is a disorder of normal maturation and clonal expansion of hematopoietic progenitor cells and is the most common malignant myeloid leukemia in adults. Over-expression of casein kinase II (CK2) is associated with metastatic potential and poor prognosis in AML. CX-4945, a CK2 inhibitor, impairs CK2 mediated phosphorylation of tumor suppressor proteins such as Ikaros and PTEN, and induces apoptosis in hematologic malignancies including AML. Though the efficacy of CX-4945 is well demonstrated *in vitro* and *in vivo*, the effect of CX-4945 on epigenetic regulation and genome-wide transcription remain elusive.

We identified that CX-4945 leads to repression and induction of enhancer activities, possibly via modulations of H3 and H4 acetylation, and alters heterochromatin landscapes in AML. The epigenetic changes upon CX-4945 treatment suppress oncogenic pathways, restore tumor suppressor activities, and cause cell apoptosis. CX-4945 modulates H3K27 acetylation levels at enhancers, which significantly affects nearby gene expression. Notably, the down-regulated genes proximal to enhancers with reduced H3K27ac include RASD2 and MYB, which participate in the PI3K/Ras oncogenic pathway. Enhancers with increased H3K27ac upon CX-4945 treatment contain DNA-binding motifs for key hematopoiesis factors such as Ikaros, PU.1, and Runx1, suggesting their roles in modulating enhancer activities. Moreover, enhancers with increased H3K27ac are highly enriched in H3K4me2 (a mark of active promoters), and H3K56ac (implicated in DNA-repair), H3K9ac and H4K5ac. In addition to its effect on enhancers, CX-4945 treatment leads to modulated heterochromatin profiles at a subset of genomic sites. CX-4945 treatment results in orchestrated changes in enrichment of H3K27me3, H3K9me3, and H4K20me3 heterochromatin marks, as well as H4K20me2 and H3K79me2 enrichment. Taken together, our study reveals that a potent CK2 inhibitor, CX-4945, dynamically modifies epigenetic landscapes and affects tumor suppressor gene expression and apoptosis signaling.

DIFFERENTIAL CONTRIBUTION OF CIS-REGULATORY ELEMENTS TO HIGHER ORDER CHROMATIN STRUCTURE AND EXPRESSION OF THE CFTR LOCUS

Rui Yang[1,2], Jenny L Kerschner*[1,2], Nehal Gosalia*[1,2], Lidijia K Gorsic[1,2], Shih-Hsing Leir[1,2], Ann Harris[1,2]

[1]Lurie Children's Research Center, Human Molecular Genetics Program, Chicago, IL, [2]Northwestern University Feinberg School of Medicine, Pediatrics, Chicago, IL

* Authors contributed equally

Recent genome-wide Hi-C studies show that higher order chromatin structure establishes domains that organize the genome and coordinate gene expression. However, the molecular mechanisms controlling transcription of individual loci within a topological domain (TAD) are not fully understood. The cystic fibrosis transmembrane conductance regulator (CFTR) gene provides a paradigm for investigating these mechanisms. CFTR encodes a cAMP-activated chloride ion channel, and when mutated, causes the genetic disease cystic fibrosis. We showed previously that intronic and extragenic enhancers, bind specific transcription factors, and are recruited to the CFTR promoter by a looping mechanism to drive tissue specific gene expression. Here we use 4C-seq analysis to characterize the higher order chromatin structure in disease-relevant cell types. The results show a conserved TAD (or sub-TAD) boundary and cell type specific interaction profiles within the TAD. We then show that the TAD boundaries are dependent on CTCF and cohesin complex, by using siRNA depletion of these factors followed by 4C-seq. To determine the relative contribution of structural elements and enhancers to the higher order structure and expression of the CFTR locus, we performed CRISPR/Cas9 mediated deletion of important cis-regulatory elements. Removal of an upstream CTCF-binding insulator alters the interaction profile, but has little effect on CFTR expression. Loss of an intronic enhancer element dramatically decreases CFTR expression, but has minor effect on overall 3D locus structure. Our study provides insights into transcriptional regulatory mechanisms involving the interplay of architectural proteins, transcription factors, and chromatin modifications, to establish and facilitate dynamic responses of the 3D chromatin structure.

PANORAMIX/CG9754 ENFORCES piRNA-DEPENDENT COTRANSCRIPTIONAL SILENCING

Yang Yu[1], Jiaqi Gu[1,3], Ying Jin[1], Yicheng Luo[1], Jonathan Preall[1], Jinbiao Ma[3], Benjamin Czech[1,2], Gregory Hannon[1,2,4]

[1]Cold Spring Harbor Laboratory, Watson School of Biological Science, Cold Spring Harbor, NY, [2]University of Cambridge, CRUK Cambridge Institute, Cambridge, United Kingdom, [3]Fudan University, Department of Biochemistry, Shanghai, China, [4]The New York Genome Center, The New York Genome Center, New York, NY

The Piwi-interacting RNA (piRNA) pathway is a small RNA-based innate immune system that defends germ cell genomes against transposons. In *Drosophila* ovaries, the nuclear Piwi protein is required for transcriptional silencing of transposons, though the precise mechanisms by which this occurs are unknown. To address these questions, we took advantage of a λN/BoxB tethering system to mimic piRNA targeting. Surprisingly, artificial tethering of λN-Piwi to a luciferase reporter containing BoxB sites failed to silence the expression of the reporter, while guiding Piwi via artificial piRNAs is sufficient to silence their targets. These results strongly suggest that confirmation changes of Piwi RISC complexes may be required to recruit additional factors to repress transcription. Through mining the data from several independent genome-wide RNAi screens for factors required for transposon silencing, we identified a novel protein-coding gene (CG9754) that can influence global transposon expression in a similar fashion as Piwi when eliminated. The effect is not due to the defects of piRNA biogenesis since levels of piRNAs remained unchanged and Piwi maintained its nuclear localization. Strikingly, enforced tethering of CG9754 to nascent mRNA transcripts causes co-transcriptional silencing of the source locus (~1000 fold repression) and the deposition of repressive chromatin marks. Interestingly, CG9754 is a component of Piwi complexes that functions downstream of Piwi and its binding partner, Asterix. Thus, we have named CG9754 Panoramix, the mentor who empowers Asterix to perform his feats of strength. Importantly, we found that both Eggless/dSetDB1 (H3K9 methyltransferase) and dLSD1 (H3K4me2 demethylase) are required for Panoramix-mediated silencing. Therefore, we propose that Panoramix forms one of the missing links between the piRNA pathway and the general silencing machinery that it recruits to enforce transcriptional repression to protect germline from deleterious transposons.

IDENTIFYING NOVEL REGULATORS OF EARLY CARDIAC DEVELOPMENT IN ZEBRAFISH USING A HETEROLOGOUS ENHANCER

Xuefei Yuan[1,2,3], Mengyi Song[4], Anastasiia Aleksandrova[2], Patrick Devine[5], Benoit G Bruneau[5,6], Ian C Scott*[2,3], Michael D Wilson*[1,3]

[1]The Hospital for Sick Children, Genetics and Genome Biology, Toronto, Canada, [2]The Hospital for Sick Children, Developmental and Stem Cell Biology, Toronto, Canada, [3]University of Toronto, Molecular Genetics, Toronto, Canada, [4]Peking University, Life Science, Beijing, China, [5]Gladstone Institute, Cardiovascular Disease, San Francisco, CA, [6]University of California, San Francisco, Developmental and Stem Cell Biology, San Francisco, CA

*co-corresponding authors

The zebrafish is an ideal model organism for studying vertebrate heart development. However there is currently a lack of early cardiac lineage markers that will enhance our understanding of the earliest steps of vertebrate heart development. Recently a mouse enhancer (*Smarcd3*-F6) was shown to label cardiac progenitor cells in vivo. Here we asked whether this mouse *Smarcd3*-F6 enhancer could serve as an early cardiac marker in zebrafish. We created stable zebrafish lines using *Smarcd3*-F6 to label zebrafish cells with GFP. We characterized the labeled population using immunostaining, RNA-seq and ATAC-seq. The *Smarcd3*-F6 enhancer was active at early gastrula stage and enriched for cardiac progenitor cells. Despite sharing no sequence conservation with zebrafish, GFP expression driven by the mouse *Smarcd3*-F6 enhancer required the cardiac master regulator Gata5. Both RNA-seq and ATAC-seq results showed cardiac related pathways were enriched in *Smarcd3*-F6 labeled cells. Several ATAC-seq peaks near known cardiac genes drove early cardiac expression when tested in embryos, suggesting they could be novel markers facilitating early cardiac development studies. In order to further dissect the *Smarcd3*-F6+ cell population, we conducted single-cell mRNA-seq and identified a cluster of cells co-expressing known cardiac markers at the end of gastrulation, which represented the potential cardiac lineage. Using *in situ* hybridization we have tested the expression of over 20 novel genes from the single-cell cardiac cluster and confirmed the vast majority of them showed early cardiac-like expression in zebrafish embryos. On-going experiments will characterize the roles the novel genes play in cardiac development via CRISPR/Cas9 system.In addition to improving our understanding of early cardiac gene regulation in zebrafish our work underscores the conserved regulatory logic of vertebrate heart development.

PROCESSING OF BULGED pri-miRNAs BY ARABIDOPSIS DICER-LIKE 3 INHIBITED BY PHOSPHATE

Satoru Machida[1], Y. Adam Yuan[1,2]

[1]National University of Singapore, Biological Sciences, Singapore, Singapore, [2]National University of Singapore (Suzhou) Research Institute, Biological Sciences, Suzhou, China

In plants, the majority of 24 nt small RNAs are processed by Dicer-like 3 (DCL3) from double-stranded RNAs, derived from single-stranded RNAs by RDR2 and endogenous RNA fold-backs, transcribed from inverted-repeats (IRs), retrotransposons and MIR loci. Although DCL3-mediated 24 nt small RNA processing are well documented, the molecular mechanisms of size and structure specificity of DCL3-mediated small RNA biogenesis are largely unknown. Here, we performed in vitro RNA processing experiments to investigate the mode of action of recombinant DCL3 in processing pri-miRNAs with different secondary structures into 24 nt small RNAs. The in vitro processing of conserved and non-conserved plant pri-miRNAs and engineered RNA fold-backs showed that DCL3 alone is a potent cleaver of single-stranded fold-backs including those known to be preferentially processed by DCL1 in vivo. Strikingly, our data showed that double-stranded RNA binding proteins (DRBs) significantly inhibit DCL3-catalyzed pri-miRNA processing, and inorganic phosphate regulates DCL3 activity depending on the stem structure of the pri-miRNA target. These factors play a discrimination role in substrate specificity rather than enhance DCL3 catalytic activity in general. Thus, DRBs and inorganic phosphate may function as discrimination factors to regulate the processing activity of DCL3 against pri-miRNAs subclasses with remarkable different stem structures. Our findings could provide insightful hints into the existence of a non-canonical pri-miRNA processing pathway in plants.

DUAL REGULATORY SWITCH THROUGH SERIAL INTERACTIONS OF TCF7L2/TCF4 WITH STAGE-SPECIFIC PARTNERS PROPELS OLIGODENDROGLIAL MATURATION

Chuntao Zhao*[1], Yaqi Deng*[1], Wei Liu[1], Richard Lu[1]

[1]Cincinnati Children's Hospital, CBDI, Cincinnati, OH

Constitutive activation of Wnt/β-catenin signaling inhibits oligodendrocyte myelination. Tcf7l2/Tcf4, a DNA-binding transcriptional partner of β-catenin, is required for oligodendrocyte differentiation. How Tcf7l2 modifies β-catenin signaling and controls myelination processes remains elusive. Here we define a stage-specific Tcf7l2-regulated transcriptional circuitry in initiating and sustaining oligodendrocyte differentiation and myelination. Integrative analysis of genome occupancy and gene expression reveals that Tcf7l2 serially cooperates with distinct co-regulators to control stage-specific transitions during oligodendrocyte lineage progression. At the onset of oligodendrocyte differentiation, we identify a transcriptional repressor Zbtb33/Kaiso as a Tcf7l2 partner to block β-catenin signaling and initiate the differentiation process. During oligodendrocyte differentiation, Tcf7l2 recruits and cooperates with another partner Sox10 to promote myelinogenic programs. In that context, Tcf7l2 activates the genes encoding cholesterol synthetic enzymes by directly targeting their enhancers, and cholesterol supplementation can partially rescue oligodendrocyte differentiation in Tcf712 mutants. Together, our studies identify stage-specific Tcf7l2 partners Kaiso and Sox10 that sequential interact with Tcf7l2 to coordinate the switch at the transitions of oligodendrocyte differentiation initiation and maturation, and point to a previously unrecognized role of Tcf7l2 in control of cholesterol biosynthesis for CNS myelinogenesis.

* these authors contributed equally

SYSTEMATIC IDENTIFICATION OF NUCLEOSOME-DISFAVORING SEQUENCES

Fangjie Zhu[1], Lucas Farnung[2], Leena Salmela[3], Fan Zhong[1], Esko Ukkonen[3], Patrick Cramer[2], Jussi Taipale[1,4]

[1]Karolinska Institutet, Department of Biosciences and Nutrition, Stockholm, Sweden, [2]Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Göttingen, Germany, [3]University of Helsinki, Department of Computer Science, Helsinki, Finland, [4]University of Helsinki, Genome-Scale Biology Program, Helsinki, Finland

Eukaryotes assemble their genomes into nucleosomes for packaging and for expression control of genes. Nucleosome positions are dictated by multiple factors acting on a basic picture set up by the biochemical affinity of DNA sequences to the nucleosome. Great efforts have been made to elucidate the rules for nucleosome-positioning sequences, but our understanding of biochemical mechanisms that determine nucleosome occupancy are still partially incomplete. Unlike transcription factors, which bind to a very small fraction of the genome, nucleosomes are bound to almost all cellular DNA, and are only absent from specific functional sites. Therefore, we have here explored nucleosome positioning rules through specifically enriching sequences that are disfavored by the nucleosome from DNA libraries containing random sequences. Sequence preferences of nucleosome were also examined when the CpG sites on DNA are methylated, by enriching bound and unbound sequences from a random library of CpG-methylated DNA. Analysis of millions nucleosome-disfavoring sequences revealed that they are depleted of the 10.2 bp periodicity of dinucleotides observed in control nucleosome-bound sequences. They are also enriched in short subsequences containing A stretches followed by T stretches. To test the data, the disfavored and favored subsequences were used to train a hidden Markov model. This model performed better than previous models when used to predict the positions of nucleosomes on eukaryotic genomes. Our results thus indicate that analyzing random sequences at high sequencing depth, and including nucleosome disfavoring sequences improves the predictive power of nucleosome positioning models.

GENE SIMILARITY NETWORK REVEALS SUB-POPULATIONS OF CELLS IN SINGLE-CELL RNA-SEQ DATA

Bo Wang[1], Jesse Zhang[2], Junjie Zhu[2], Serafim Batzoglou[1]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Electrical Engineering, Stanford, CA

Single-cell RNA-seq is rapidly becoming a powerful method for studying cell-to-cell variation at the transcriptional level. Due to experimental limitations, single-cell datasets suffer from technical noise in addition to variation caused by culture conditions and systematic biological effects. Traditional methods such as hierarchical clustering struggle with noisy high dimensional single-cell RNA-seq data. Some studies have tackled this problem by selecting genes based on prior knowledge and biological intuition either at the beginning of the experiment or when applying clustering methods. This approach, however, introduces bias into the computational analysis and therefore can fail to capture subtle relationships in the data that would only be identifiable using the removed genes. This work introduces Sargen (**s**ingle-cell **a**nalysis using **r**obust PCA on **ge**ne similarity **n**etwork), a novel method that finds subpopulations of cells in a single-cell RNA-seq data by learning subtle gene-gene relationships in the form of a gene similarity network. Using a deep autoencoder, Sargen learns the network before using it to map the data into a reduced dimensional space. The core strategy in the method revolves around formulating a non-convex optimization problem that attempts to find a low-rank, network-aligned approximation of the input data. Sargen solves a convex relaxation of the problem and uses the solution to generate cell subpopulations. We analyzed recent single-cell datasets (Buettner et al., 2015, Kolodziejczyk et al., 2015) influenced by cell cycle effects and culture conditions, and used Sargen to successfully identifies cell subpopulations, verified by both gene ontology and prior information pertaining to the experiments. We also demonstrate from these experiments that Sargen is able to automatically select relevant genes without prior knowledge of specific gene functions in identifying cell subpopulations. Finally, we empirically show that the framework introduced by Sargen significantly improves many existing clustering (e.g., K-means and hierarchical clustering) and visualization methods (e.g., t-distributed stochastic neighbor embedding and principle components analysis), on single-cell RNA-seq data.

References:

- F. Buettner, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol., vol. 33, no. 2, pp. 155–160, 2015.
- A. A. Kolodziejczyk, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell, vol. 17, no. 4, pp. 471–485, 2015.

ELUCIDATING THE ENHANCER CODE IN 9P21 LOCUS UNDERLYING AGING-RELATED DISEASES

Yizhou Zhu, Cagdas Tazearslan, Yousin Suh

Albert Einstein College of Medicine, Department of Genetics, Bronx, NY

Since the vast majority of genetic variations identified by genome-wide association (GWA) studies are located in non-coding genome, it becomes clear that regulatory elements such as gene enhancers play crucial roles in development of complex diseases and traits. However, due to highly versatile and tissue-specific role of enhancers, current knowledge about disease risk mechanism underlying GWAS signals is rudimentary and largely incomplete. We focus on solving mechanisms of gene regulation in 9p21.3, one of the most replicated GWAS hotspot associated with multiple age-related disease phenotypes. Through usage of targeted endonuclease-assisted enhancer editing and high throughput sequencing methods, we aim to identify key enhancers in the locus and their target genes, and understand how disease causal variants interfere with enhancer-gene interactions. Our current data suggest a complex regulatory network involving multiple enhancers and downstream genes around 9p21, which may explain the broad spectrum of diseases associated with the locus.

RECORDING OLIG 2 BINDING DURING THE MOTOR-NEURON/GLIA CELL FATE DECISION.

David N Mayhew[1,3], Dylan A McCreedy[2], Shelly E. Sakiyama-Elbert[2], Rob D Mitra[1,3]

[1]Washington University School of Medicine, Genetics, St Louis, MO,
[2]Washington University, Biomedical Engineering, St Louis, MO,
[3]Washington University School of Medicine, Center for Genome Sciences and Systems Biology, St Louis, MO

A detailed understanding of the molecular mechanisms that underlie cell fate choice would not only yield insights into a fundamental biological process, but would also have practical applications, as the transplantation of differentiated cell types holds substantial therapeutic promise. We have developed an in vitro model of progenitor motor neuron (pMN) differentiation as a platform for the genomic analyses of cell fate decisions. This system is ideal because it is well-characterized, recapitulates most of the molecular events that occur in vivo, and could ultimately be useful for producing cells to treat spinal cord injury or neuropathies. We have focused on the transcription factor Olig2, which has been shown to play a major role in the motor neuron/glia fate choice. Using transposon 'Calling Cards' we have recorded the binding of Olig2 and found that this protein binds to different loci in pMNs that differentiate into motor neurons vs those that differentiate into oligodendrocytes. These loci are adjacent to several genes known to be important in neural differentiation and the data elucidates the regulation of this cell fate decision. We have validated enhancer activity at these loci using cre-mediate enhancer tracing. This approach represents a general methodology for unraveling the transcriptional circuitry that underlie cell fate decisions.

EXPLORING T CELL FATE CHOICES AT SINGLE-CELL RESOLUTION

Tapio Lönnberg[1], Kylie R. R James[2], Michael J.T. Stubbington[3], Valentine Svensson[1], Ashraful Haque[2], Sarah A Teichmann[3]

[1]European Molecular Biology, Laboratory European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom, [2]QIMR Berghofer Medical Research Institute, Brisbane, Australia, [3]Wellcome Trust Sanger Institute, Cambridge, United Kingdom

Differentiation of naive CD4+ T cells into functionally distinct T helper subsets is crucial for the orchestration of immune responses. We have explored the process of cellular decision-making in this system using single-cell RNA transcriptomics along with a novel computational approach based on overlapping mixtures of Gaussian processes. We reconstructed the bifurcating developmental trajectories of Th1 and Tfh cell populations during Plasmodium infection in mice and found that these cell fates emerged from a common, highly proliferative and metabolically active precursor. We further found that precursor T cells were coached towards a Th1 but not a Tfh fate by monocytes.

In addition, we have developed a method that can reconstruct the paired sequences of rearranged T cell receptor genes in these cells and use this to infer clonal relationships between cells that derive from a single common progenitor. We have shown that sibling cells derived from the same naive CD4+ T cell can concurrently populate both Th1 and Tfh subsets in the Plasmodium model and also that sibling cells span multiple cell states during Salmonella infection.

STRESS-DEPENDENT TRANSCRIPTOME CHANGES SERVE TO REALLOCATE TRANSLATIONAL CAPACITY DURING STRESS ACCLIMATION.

Yi-Hsuan E Ho[1], Evgenia Shishkova[2], Joshua Coon[2], Audrey Gasch[1]

[1]University of Wisconsin - Madison, Laboratory of Genetics, Madison, WI, [2]University of Wisconsin - Madison, Department of Chemistry, Madison, WI

Proper response to environmental stress is pivotal for cells to survive and adapt to constantly changing environments. Activation of stress defense systems is often coupled with arrest of growth and cell cycle progression. The transcriptomic response of *Saccharomyces cerevisiae* to stress includes a common stress transcriptomic change called the environment stress response (ESR). The ESR includes induced expression of ~ 300 genes involved in stress defense and reduced expression of ~ 600 genes required for active growth, namely genes encoding ribosomal proteins and other translation machinery. There has been a debate if activation of the ESR, and in particular reduced expression of growth-promoting ESR genes, is an active response to stress or is merely an indirect byproduct of reduced growth and cell-cycle progression. Since stress responses are generally correlated with growth and division arrest, it is hard to deconvolute responses triggered directly by stress versus indirectly by growth reduction. Here, we decouple the response to external stress and growth control by following transcriptomic and proteomic changes of arrested cells upon stresses. Our results show that activation of the ESR, including repression of growth-promoting genes, is not associated with cell-cycle phase and is independent of growth arrest. Instead, arrested cells exposed to stress actively induce the ESR, even though they are no longer making significant biomass – this indicates that reduced expression of growth-promoting genes is not necessarily related to growth control. We are using mathematical modeling of transcriptome and proteome changes, along with polysome profiling to measured global translation rates, to understand the function of transcriptome changes under different conditions. Our results suggest that differences in transcript abundance during steady-state growth serve to set the levels of protein production required for a given growth rate, whereas changes in transcript levels during active stress help to reallocate translational machinery during acclimation.

A SLOW ACTIVATOR OUTCOMPETES A FAST SUPPRESSOR TO ROBUSTLY DELAY THE TRANSITION FROM QUIESCENCE TO PROLIFERATION

Ariel Jaimovich, Steven D Cappell, Mingyu Chung, Gautam Dey, Tobias Meyer

Stanford University, Department of Chemical and Systems Biology, Stanford, CA

The decision cells make to transition from a quiescent to a proliferating state is one of the most fundamental processes in biology and its misregulation can lead to cancer and neurodegenerative diseases. One hallmark of this transition in mammalian cells is the phosphorylation of the RetinoBlastoma (RB) protein by the Cyclin Dependent Kinases 4/6 (CDK4/6) that triggers E2F-dependent expression of genes required for proliferation. To elucidate the role of transcription regulation in executing this transition we develop a high content imaging platform that enables single-cell resolution measurements of mRNA and protein levels of the key components regulating this transition in mammalian cells released from a prolonged mitogen deprivation.

We observe that both mRNA and protein levels of Cyclin D1, the main activator of CDK4/6, start increasing immediately after mitogen addition. However, we do not observe RB phosphorylation or transcription of E2F targets in the first six hours after mitogen addition. One possible explanation for such a delay would be the expression of a CDK4/6 inhibitor such as p21. Indeed, we show that p21 mRNA and protein levels show a rapid synthesis rate and thus the steady increase in Cyclin D1 levels has to overcome a much more rapid synthesis of p21 to activate CDK4/6. To check whether the kinetic behavior of such a simple circuit can explain our observations we create a detailed kinetic model and experimentally measure its parameters. We show that this model can recapitulate both the observed expression profiles as well as the robust delay in RB phosphorylation. We further validate this hypothesis by showing that in p21 knockout cells there is no robust delay in RB phosphorylation.

Finally, to check whether a similar mechanism also controls G1 entry in asynchronous cell populations we combined live-cell imaging and single-cell mRNA FISH in unperturbed mammalian cells to find that the fate of these cells is determined by post-transcription regulation of Cyclin D1 and p21. We conclude that after mitogen deprivation the transition from quiescence to proliferation is controlled by the kinetic characteristics of transcription regulation of a simple cellular circuit where a slow increase in activator levels needs to outcompete a suppressor with a rapid synthesis rate resulting in a robust delay of the cellular decision to enter a proliferative state. However, in asynchronous cells, though the same circuit controls RB phosphorylation its behavior is governed by post-transcriptional regulation.

A DEVELOPMENTAL OSCILLATOR AND ITS IMPORTANCE IN ESTABLISHING SEQUENTIAL CELL FATE TRANSITIONS IN *C. ELEGANS*

Christopher M Hammell

Cold Spring Harbor Laboratory, Genomics, Cold Spring Harbor, NY

Gene expression during animal development is remarkably dynamic. For instance, approximately ten to twenty percent of coding genes in *C. elegans* are expressed in an oscillatory fashion during larval development. The molecular mechanism(s) that generate these patterns of expression and furthermore, how these patterns are incorporated into temporal cell fate specification is unknown. In a screen to identify components that increase or dampen the expression of heterochronic microRNAs, we identified two genes, *blmp-1* and *lin-42*, that form the core of a developmental oscillator. BLMP-1, a conserved Zn-finger transcription factor, and LIN-42, the C. elegans Period homolog, not only maintain their own cyclical expression patterns during larval growth but also generate pulses of transcription for many key miRNAs and mRNAs that control major temporal transitions throughout development. We will outline the molecular details of this oscillator and demonstrate how its activity uniquely impacts sequential aspects of cell fate specification. While this system is important for the well-characterized precision of the *C. elegans* lineage in nutritionally replete conditions, we will present evidence that nutritional cues, and specifically starvation, induce a systemic developmental arrest, a suspension of BLMP-1 and LIN-42 expression and a pausing of oscillatory transcription. Finally, we will demonstrate that this developmental oscillator is essential for the resumption of normal temporal development once arrested animals resume development in more favorable environments. Therefore, the *blmp-1/lin-42* regulatory system provides a cadence development and establishes continuity of cell fate specification in constantly varying conditions.

GENOME-WIDE ANALYSIS OF HUMAN TH17 DIFFERENTIATION REVEALS NOVEL HUMAN SPECIFIC FACTORS AND REGULATORY REGIONS THAT CO-LOCALISE WITH IMMUNE DISEASE ASSOCIATED SNPS

Zhi Chen[1], Soile Tuomela[1], Subhash Tripathi[1], Antti Larjo[2], Sini Rautio[2], Verna Salo[1], Helena Ahlfors[3], Brigitta Stockinger[3], Harri Lähdesmäki[1,2], Riitta Lahesmaa[1]

[1]University of Turku and Åbo Akademi University, Turku Centre for Biotechnology, Turku, Finland, [2]Aalto University School of Science, Department of Information and Computer Science, Espoo, Finland, [3]The Francis Crick Institute, London, United Kingdom

Th17 cells have been associated with autoimmune and inflammatory diseases and cancer. Current knowledge of Th17 cells relies mainly on murine studies. Recently systems biology approaches have been successfully exploited to derive and experimentally validate the dynamic regulatory networks that control the differentiation of mouse Th17 cells. To investigate the degree to which these findings can be translated to human Th17 cell biology RNA-sequencing over multiple time points during the early human Th17 differentiation as well as ChIP sequencing and siRNA mediated perturbation of selected factors was carried out. We identified the cross-species conserved and species specific gene expression signatures and gene regulatory pathways during early Th17 cell differentiation. Importantly, we identified human specific long non-coding RNA molecules differentially regulated during human Th17 cell priming. We also identified for the first time early direct and indirect targets of STAT3 in human and showed that STAT3 is critical for transcriptional regulation of early human Th17 cell differentiation. Finally, we found that a great number of SNPs associated with immune mediated disorders were located at sites where STAT3 binds to induce Th17 cell specification. The study provides a basis for understanding and modulating Th17-mediated pathogenic immune responses in human.

CELL FATE DECISIONS IN RESPONSE TO A SHORT PULSE OF TNF

Robin E Lee[1,2,4]

[1]Dana Farber Cancer Institute, Department of Cancer Biology and CCSB, Boston, MA, [2]Harvard Medical School, 2Department of Genetics, Boston, MA, [3]McGill University, Biomedical Engineering, Montreal, Canada, [4]University of Pittsburgh, Computational and Systems Biology, Pittsburgh, PA, [5]New York University Abu Dhabi, Engineering, Abu Dhabi, United Arab Emirates

TNF is a pro-inflammatory cytokine that modulates cellular behaviors including proliferation, differentiation, and apoptotic cell death. To influence these cell fate decisions, binding of TNF to its receptor sequentially activates opposing intracellular signals. Bound receptors at the plasma membrane rapidly induce nuclear accumulation of the NF-κB transcription factor, driving transcription of anti-apoptotic genes that promote cell survival. This is followed by internalization of TNF-bound receptors that initiate signals for caspase-dependent apoptosis in the same cell.

To regulate diverse cell fates *in vivo*, exposure to TNF is strictly controlled and likely of short duration. However, it is unknown how transient exposure to TNF coordinates cell fate decisions. To understand the interplay between survival and death signals after a pulse of TNF, we monitored live cells expressing fluorescent reporter proteins by time lapse microscopy. Using cells cultured in a microfluidic flow device designed for precise spatiotemporal control over TNF delivery, we quantified single-cell timecourses of nuclear NF-κB. From a single-cell dataset of NF-κB localization dynamics and transcript numbers, determined by smFISH in the same cell, we established the threshold of nuclear NF-κB translocation required to induce gene transcription. For a high TNF concentration, we show that a TNF pulse of 10s elicits significant NF-κB translocation in a fraction of cells, although a 30s pulse or longer is required to approximate continuous exposure. We also find that the minimal TNF pulse is dose dependent, with lower concentrations requiring a longer pulse for similar NF-κB activation. To monitor caspase activity and apoptosis in single cells, we similarly imaged cells expressing a FRET-based initiator caspase reporter (IC-RP) and monitored IC-RP cleavage in addition to cell morphology. Contrary to expectations, we observed that a 1-minute pulse of TNF can be more effective at cell killing than a pulse of shorter or longer duration, a non-monotonic relationship that contrasts with NF-κB activation. We present a ligand-receptor model that suggests receptor oligomerizations can be ‘fine-tuned’ by pulse duration, leading to the observed relationship between pulse duration and cell death. Taken together, our results suggest that TNF concentration and duration together coordinate cell fate decisions.

SINGLE-CELL CHROMATIN ACCESSIBILITY REVEALS PRINCIPLES OF REGULATORY VARIATION

Jason D Buenrostro[1,2], Ryan Corces[2], Beijing Wu[1], Ulrike M Litzenburger[2], Dave Ruff[3], Michael L Gonzales[3], Michael P Snyder[1], Ravi Majeti[1], Howard Y Chang[2], William J Greenleaf[1,4]

[1]Stanford University School of Medicine, Department of Genetics, Stanford, CA, [2]Stanford University School of Medicine, Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford, CA, [3]Fluidigm Corporation, South San Francisco, CA, [4]Stanford University, Department of Applied Physics, Stanford, CA, [5]Stanford University, Department of Medicine, Stanford, CA

Recent proliferation of powerful methods for interrogating single cells has allowed detailed characterization of phenotypic molecular variation at the single cell level, and provided deep insight into characteristics underlying developmental plasticity, cancer heterogeneity, and drug resistance. In parallel, genome-wide mapping of regulatory elements in large ensembles of cells have unveiled tremendous variation in chromatin structure across cell-types, particularly at distal regulatory regions. However, the lack of methods to probe DNA accessibility within individual cells has prevented quantitative dissection of single cell regulatory variation – variation the likely drives phenotypic heterogeneity. To address this need, we have developed single-cell ATAC-seq (scATAC-seq), a robust method for mapping the accessible genome of individual cells via assay for transposase-accessible chromatin using sequencing (ATAC-seq). Maps aggregated from data generated from hundreds of single-cells i closely resemble accessibility profiles from tens of millions of cells and provides insights into cell-to-cell variation. We have developed novel single cell analysis methods that can associate variability in chromatin accessibility with sets of genomic loci. Accessibility variance is associated with specific *trans*-factors and *cis*-elements, and combinations of *trans*-factors are associated with either induction or suppression of cell-to-cell variability. We further identify sets of *trans*-factors associated with cell-type specific accessibility variance across 8 cell types. Targeted perturbations of cell cycle or transcription factor signaling evoke stimulus-specific changes in this observed variability. Surprisingly, genomic regions that are correlated in accessibility variation in *cis* across the genome are likely to lie in the same topological domain, providing a potential link between single-cell accessibility variation and three-dimensional genome organization. We have also investigated the single cell regulatory variation within the hematopoietic hierarchy and related this variation to the baseline changes that occur during differentiation during hematopoiesis. All together, single-cell analysis of DNA accessibility provides new insight into cellular variation of the "regulome" at ultimate sensitivity, with powerful potential applications to basic biology and clinical investigation.

ELUCIDATING CHROMATIN ARCHITECTURE BY COMBINATORIAL CO-IMMUNEPRECIPITATION (CO-CHIP)

Ronan Sadeh[1,2], Roee Launer-Wachs[1], Hava Wandel[1], Ayelet Rahat[1], Nir Friedman[1,2]

[1]Hebrew University, School of Computer Science & Engineering, Jerusalem, Israel, [2]Hebrew University, Institute of Life Sciences, Jerusalem, Israel

Nucleosomal histones, the basic packaging units of DNA, are massively decorated by a large number of posttranslational modifications. These play critical roles in all genomic transactions and in particular transcriptional regulation. Chromatin Immunoprecipitation of histone modifications followed by next generation sequencing (ChIP-Seq) has been instrumental to our understanding of chromatin architecture, structure, and function. A major limitation of ChIP-Seq is assaying single histone modification at a time. This limitation restricts our understanding of the combinatorial nature of chromatin to correlations of the spatial distributions of modifications across the genome. Here we present MNase-co-CHIP, a novel, straightforward, highly-multiplexed and robust methodology that utilizes chromatin indexing to allow genome-wide sequential detection of two coexisting histone marks at a nucleosome resolution. We use this assay to determine the degree of coexistence of, and distance between several histone modifications. Importantly, we find cases where we can deconvolve relationships that are not found, or sometimes mistakenly found, using standard ChIP assays. Moreover, co-ChIP of certain histone modification combinations can predict genomic features better than combinations of single ChIPs. We next use co-ChIP to study how modifications of “fresh” nucleosomes incorporated by the replication-independent machinery evolve in time. This study provides spatial and kinetic insights on the order of events during nucleosome maturation and methodology for quantifying the activity of chromatin modifying enzymes in vivo.

HIGH-THROUGHPUT MAPPING OF REGULATORY DNA

Nisha Rajagopal[1], Sharanya Srinivasan[1,2], Kameron Kooshesh[2], Yuchun Guo[1], Matthew D Edwards[1], Budhaditya Banerjee[2], Tahin Syed[1], Bart Emons[2], David K Gifford[1], Richard I Sherwood[2]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2] Brigham and Women's Hospital and Harvard Medical School, Division of Genetics, Department of Medicine, Boston, MA

Quantifying the effects of cis-regulatory DNA on gene expression is a major challenge. Here, we present the multiplexed editing regulatory assay (MERA), a high-throughput CRISPR-Cas9–based approach that analyzes the functional impact of the regulatory genome in its native context. MERA tiles thousands of mutations across ~40 kb of cis-regulatory genomic space using guide RNAs(gRNAs) complementary to each region and uses knock-in green fluorescent protein (GFP) reporters to read out target gene activity. Using this approach, we obtain quantitative information on the contribution of cis-regulatory regions to gene expression. We identify proximal and distal regulatory elements necessary for expression of four embryonic stem cell–specific genes. We show a consistent contribution of neighboring gene promoters to gene expression and identify unmarked regulatory elements (UREs) that control gene expression but do not have typical enhancer epigenetic or chromatin features. We also perform deep-sequencing of genotypes generated by specific gRNAs discovered by the MERA screen to cause loss of gene activity. Comparison of functional and non-functional genotypes at the specific genomic location targeted by these gRNAs enables identification of base pair–resolution functional motifs of regulatory elements.

FUNCTIONAL DISSECTION OF ESTROGEN RECEPTOR A BOUND ENHANCERS AT THEIR ENDOGENOUS LOCI

Julia B Carleton, Kristofer Berrett, Jason Gertz

Huntsman Cancer Institute, University of Utah, Department of Oncological Sciences, Salt Lake City, UT

Cis-regulatory enhancers are key drivers of transcription regulation in metazoans, and the activity of enhancers is critical for proper development and the progression of disease. While more than 2,000,000 candidate enhancers have been predicted, it is unclear which loci represent true enhancers that are necessary for transcription regulation and how individual enhancers combine to affect expression. Few studies have directly investigated the causal and combinatorial relationship between enhancers and the expression of their target genes, largely due to technological limitations. In order to overcome these limitations and evaluate the importance of candidate enhancers, we have developed Enhancer-interference (Enhancer-i), a method for causally connecting enhancers to gene expression that adapts CRISPR interference (CRISPRi) to target distal regulatory elements as opposed to promoters. We have applied Enhancer-i to the functional dissection of estrogen receptor α (ER) bound loci in endometrial cancer cells. By targeting nuclease deficient Cas9 repressive fusion proteins to several ER bound loci simultaneously, we were able to attenuate the estradiol gene expression response of multiple genes. Enhancer-i also significantly reduced ER occupancy at most targeted loci. Comparison of enhancer interference to promoter interference uncovered that targeting enhancers is as effective as targeting promoters for blocking estrogen-induced expression. Enhancer-I allowed us to determine how multiple enhancers impact the expression of a single target gene and we found that different genes can exhibit different enhancer relationships (e.g. dominance, additivity). Our results establish Enhancer-i as a method for causally connecting distal regulatory regions to gene expression, which will aid in the discovery of functional enhancers.

SIMULTANEOUS MEASUREMENT OF CHROMATIN ACCESSIBILITY AND DNA METHYLATION IN SINGLE CELLS

Sebastian Pott, Jason D Lieb

University of Chicago, Department of Human Genetics, Chicago, IL

Sets of regulatory elements identified in tissues or in populations of cells grown in culture do not directly reflect the open chromatin state of individual cells within these populations. Identifying all of the regulatory elements accessible in a given single cell is a critical step toward understanding the rules that determine concomitant usage of different regulatory elements within single cells and how intercellular variation in chromatin organization relates to the observed transcriptional heterogeneity between cells. To address these questions, we have adapted the nucleosome occupancy and methylome sequencing (NOMe-seq) protocol for single cells. This approach is based on incubation of isolated nuclei with the viral DNA methyltransferase M.CviPI which methylates cytosines in GpC dinucleotides specifically in nucleosome-free regions. Bisulfite conversion of DNA isolated from treated cells followed by high-throughput sequencing therefore reveals sites of accessible chromatin. Crucially, sites of endogenous methylation, which preferentially occurs at CpGs, are identified simultaneously. We combined this approach with cell sorting to isolate individual nuclei before bisulfite conversion and library preparation. To test the feasibility of this strategy we used a well characterized lymphoblastoid cell line (GM12878) and prepared NOMe-seq libraries from more than 30 sorted samples ranging from 1000 cells to single cells. Sorted NOMe-seq samples, including 13 samples from single cells, show similar accessibility signals at transcription start sites and published DNase I hypersensitive sites compared to chromatin accessibility data obtained from bulk samples. These preliminary analyses demonstrate that NOMe-seq can be applied to single cells. This method provides a novel approach to simultaneously measure chromatin accessibility and DNA methylation in single cells and is uniquely suited to study the interplay between these epigenomic features.

TT-SEQ CAPTURES THE HUMAN TRANSIENT TRANSCRIPTOME

Patrick Cramer

Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Goettingen, Germany

A large portion of the human genome is transcribed but many RNAs are transient and can generally not be mapped. Here I report on our development of TT-Seq (transient transcriptome sequencing), a protocol that maps RNA-producing units in human cells in a nearly uniform and unbiased manner (Schwalb, Michel, Zacher et al., unpublished). Application of TT-Seq to human K562 cells recovers stable mRNAs and long intergenic non-coding RNAs, and additionally maps over 10,000 transient RNAs, including enhancer RNAs, antisense RNAs, and promoter-associated RNAs, both convergent and upstream antisense RNAs. TT-Seq analysis also provides RNA half-lives, and reveals that transient RNAs such as eRNAs are short and lack U1 motifs and secondary structure. TT-Seq further uncovers transcription termination sites and detects a DNA motif associated with RNA polymerase release at almost all transcribed units.

TT-Seq was developed based on our experience with functional genomics protocols to study mRNA metabolism in the yeast Saccharomyces cerevisiae. We have previously measured RNA synthesis and degradation rates globally after metabolic RNA pulse-labeling with the use of 4-thiouracil (4tU) in growing cells (Miller et al., 2011) and during the cell cycle (Eser et al., 2014). We established a normalization procedure (Sun et al., 2012) that revealed that the levels of mRNAs are buffered after perturbation of mRNA metabolism (Sun et al., 2013). Combining the 4tU-Seq method with PAR-CLIP and ChIP-Seq enabled us to characterize a global mechanism of transcriptome surveillance that relies on the factor Nrd1 that recognizes non-coding RNA (Schulz et al., 2013). For our work, collaborations with the groups of computational biologists Johannes Soeding, Achim Tresch, and Julien Gagneur were essential.

- Björn Schwalb*, Margaux Michel*, Benedikt Zacher*, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer (*shared first authors). TT-Seq captures the human transient transcriptome. Unpublished.
- Miller C, Schwalb B, Maier K, Schulz D, Dumcke S, Zacher B, Mayer A, Sydow J, - - Marcinowski L, Dolken L, Martin DE, Tresch A & Cramer P 2011. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol Syst Biol, 7, 458.
- Eser P, Demel C, Maier KC, Schwalb B, Pirkl N, Martin DE, Cramer P & Tresch A 2014. Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. Mol Syst Biol, 10, 717.
- Schulz D, Schwalb B, Kiesel A, Baejen C, Torkler P, Gagneur J, Soeding J & Cramer P 2013. Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell, 155, 1075-87.
- Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. TT-Seq captures the human transient transcriptome. Unpublished.
- Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, Tresch A & Cramer P 2013. Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. Mol Cell, 52, 52-62.
- Sun M, Schwalb B, Schulz D, Pirkl N, Etzold S, Lariviere L, Maier KC, Seizl M, Tresch A & Cramer P 2012. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. Genome Res, 22, 1350-9.

DECIPHERING THE ROLE OF VARIOUS MECHANISMS IN TRANSCRIPTION FACTOR-DNA BINDING ON A PROTEIN FAMILY-SPECIFIC BASIS

Beibei Xin*, Satyanarayan Rao*, Tsu-Pei Chiu*, Lin Yang*, Remo Rohs

University of Southern California, Molecular and Computational Biology, Los Angeles, CA

*These authors contributed equally.

Only a small fraction of putative transcription factor (TF) binding sites in the genome is functional and bound by a given TF, and it is widely unknown why some of these sites with equivalent core motifs are bound while others are not utilized. TFs recognize nucleotide sequence in a short core motif mainly through hydrogen bonds and hydrophobic contacts between amino acids and functional groups of the bases. DNA shape provides an additional layer that complements sequence recognition of the core motif, predominantly in flanking regions or spacers between co-binding proteins. Relative contributions of base and shape readout differ between diverse TF families and it is important to annotate readout mechanisms on a family-specific basis. We used high-throughput binding assays to probe the varying role of DNA shape readout between TF families, and feature selection to identify nucleotide positions where DNA shape is important for binding specificity. Since DNA shape is a "proxy" for electrostatic interactions in the minor groove, we asked the question if biophysical features can actually better capture the readout of minor groove electrostatic potential than DNA shape does. We found that electrostatic interactions in the narrow minor groove are well described through minor groove width but, in cases where a guanine amino group is present or the minor groove is not narrow, electrostatic potential is a better descriptor of readout interactions than DNA shape. Likewise, in the major groove, the addition of a bulky methyl group upon cytosine methylation introduces a functional group for hydrophobic contacts and indirectly induces a narrowing of the minor groove, which might in turn enhance electrostatic attraction of positively charged amino acids. In the more complex in vivo situation, additional mechanisms determine the outcome of a binding event. One mechanism that is well known to contribute to binding is the chromatin structure or accessibility of putative binding sites. Histone modifications can affect nucleosome stability and, as a consequence, binding site accessibility. As such, these important epigenetic marks contribute to the DNA binding specificity of TFs. We used experimental data from high-throughput binding assays and statistical machine learning approaches to quantify the role of these various determinants on TF binding. The importance of all these layers varies across protein families, and only the integration of knowledge derived from structural biology and genomics can lead towards a better mechanistic understanding of gene regulation.

COMPUTATIONAL AND EXPERIMENTAL DEVELOPMENT FOR CRISPR SCREENS

X. Shirley Liu[1,2]

[1]Harvard School of Public Health, Department of Biostatistics and Computational Biology, Boston, MA, [2]Dana-Farber Cancer Institute, Center for Functional Cancer Epigenetics, Boston, MA

CRISPR screen is a powerful technique for systematic genetic analysis to identify key genes for tumorigenesis and progression, biomarkers of drug response, and mechanisms underlying drug resistance. However, there are still many computational analysis challenges for experimental biologists to adopt the technology, including how to design a good gRNA library, how to analyze the CRISPR screen data, how to prioritize the CRISPR screen hits for functional validation, and how to model and visualize screen results in multiple conditions. I will discuss a number of computational methods we developed to overcome these challenges. I will also discuss our experimental efforts for functional screen of long non-coding RNAs (lncRNAs) using a lentiviral paired-guide RNA (pgRNA) library, as well as another study using primary and secondary screens to identify synthetic lethal genes and understand the mechanism underlying the hormone-independent growth and metastasis of breast cancer.

INTEGRATIVE, INTERPRETABLE DEEP LEARNING FRAMEWORKS FOR REGULATORY GENOMICS AND EPIGENOMICS

Avanti Shrikumar[1], Yonatan Israeli[2], Chuan Sheng Foo[1], Irene Kaplow[1], Nicholas Sinnott-Armstrong [3], Rahul Mohan[4], Nathan Boley[3], Anshul Kundaje[3,1]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Biophysics, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA, [4]Fremont High School, Sunnyvale, CA

We present generalizable and interpretable supervised deep learning frameworks to predict regulatory and epigenetic state of putative functional genomic elements by integrating raw DNA sequence with diverse chromatin assays such as ATAC-seq, DNase-seq or MNase-seq. First, we develop novel multi-channel, multi-modal CNNs that integrate haploid or diploid DNA sequence and chromatin accessibity profiles (DNase-seq or ATAC-seq) to predict in-vivo binding sites of a diverse set of transcription factors (TF) across cell types with high accuracy. Our integrative models provide significant improvements over other state-of-the-art methods including recently published deep learning TF binding models. Next, we train multi-task, multi-modal deep CNNs to simultaneously predict multiple histone modifications and combinatorial chromatin state at regulatory elements by integrating DNA sequence, RNA-seq and ATAC-seq or a combination of DNase-seq and MNase-seq. Our models achieve high prediction accuracy even across cell-types revealing a fundamental predictive relationship between chromatin architecture and histone modifications. Finally, we develop DeepLIFT (Deep Linear Importance Feature Tracker), a novel interpretation engine for extracting predictive and biological meaningful patterns from deep neural networks (DNNs) for diverse genomic data types. Since DNNs learn inherently distributed representations, we find that multiple convolutional filters often cooperatively represent distinct regulatory features such as TF binding preferences and hence caution against overinterpreting individual filters. Filter nullification and 'in-silico mutagenesis' are commonly used methods to score the relative importance of specific inputs such as individual bases in input DNA sequences. These approaches are not only computationally expensive but can often provide misleading results when the inputs contains redundant signals that potentially buffer each other. DeepLIFFT is the first method that can integrate the combined effects of multiple cooperating filters and compute importance scores accounting for redundant patterns. We apply DeepLIFFT on our models to obtain unified TF sequence affinity models, infer high resolution point binding events of TFs, dissect regulatory sequence grammars involving homodimer and heterodimeric binding with co-factors, learn predictive chromatin architectural features and unravel the sequence and architectural heterogeneity of regulatory elements.

SYNCHRONIZED TRANSLATION PROGRAMS ACROSS CELLULAR COMPARTMENTS

Mary Couvillion, Iliana Soto, Stirling Churchman

Harvard Medical School, Department of Genetics, Boston, MA

Oxidative phosphorylation (OXPHOS) is fundamental for life. OXPHOS complexes pose a unique challenge for the cell, because their subunits are encoded on two different genomes, the nuclear genome and the mitochondrial genome. Genomic approaches designed to study nuclear/cytosolic and bacterial gene expression have not been broadly applied to the mitochondrial system, thus the co-regulation of OXPHOS genes remains largely unexplored. Here we globally monitored mitochondrial and nuclear gene expression processes during mitochondrial biogenesis when OXPHOS complexes are synthesized. Nuclear- and mitochondrial- encoded OXPHOS transcript levels do not increase concordantly. Instead, we observe that mitochondrial and cytosolic translation are rapidly and dynamically regulated in a strikingly synchronous fashion. Furthermore, the coordinated translation programs are controlled unidirectionally through the intricate and dynamic control of cytosolic translation. Thus the nuclear genome carefully directs the coordination of mitochondrial and cytosolic translation to orchestrate the timely synthesis of each OXPHOS complex, representing an unappreciated regulatory layer shaping the mitochondrial proteome. Our whole-cell genomic profiling approach establishes a foundation for global gene regulatory studies of mitochondrial biology.

RNA POLYMERASE II STIMULATES TOPOISOMERASE 1 ACTIVITY TO PROMOTE EFFICIENT TRANSCRIPTION

Laura Baranello[1], Damian Wojtowicz[2], Kairong Cui[3], Ballachanda N Devaiah[4], Hye-Jung Chung[1], Kelli Wilson[5], Rajarshi Guha[5], Xiaohu Zhang[5], Jason Piotrowski[6], Craig Thomas[5], Dinah S Singer[4], Franklin Pugh[7], Yves Pommier[8], Teresa M Przytycka[2], Fedor Kouzine[1], Brian Lewis[6], Keji Zhao[3], David Levens[1]

[1]CCR/NCI, Pathology, Bethesda, MD, [2]NCBI/NIH, Computational Biology, Bethesda, MD, [3]NHLBI/NIH, Systems Biology, Bethesda, MD, [4]CCR/NCI, Experimental Immunology, Bethesda, MD, [5]NCATS/NIH, Preclinical Innovation, Rockville, MD, [6]CCR/NCI, Lymphoid Malignancies, Bethesda, MD, [7]PSU, Bioch. and Mol. Bio., University Park, PA, [8]CCR/NCI, Therapeutics and Molecular Pharmacology, Bethesda, MD

We report a mechanism through which the transcription machinery directly controls topoisomerase 1 (TOP1) activity to adjust DNA topology throughout the transcription cycle. By comparing TOP1 occupancy using ChIP-Seq, versus TOP1 activity using TOP1-Seq, a method reported here to map catalytically engaged TOP1, TOP1 bound at promoters was discovered to become fully active only after pause-release. RNA polymerase II was found to stimulate TOP1 catalytic activity. This transition coupled phosphorylation of the RNAPII carboxyl-terminal-domain (CTD)--52 tandem repeats of canonical or variant heptads YSPTSPS, with stimulation of TOP1 above its basal rate, enhancing its processivity. TOP1 stimulation is strongly dependent on the kinase activity of BRD4, a protein that preferentially phosphorylates Ser2-CTD of some heptads and regulates RNAPII pause-release. Thus the coordinated action of BRD4 and TOP1 overcame the torsional stress opposing transcription as RNAPII commenced elongation, but preserved negative supercoiling that assists promoter melting at transcription start sites. BET inhibitors such as JQ1 synergize with the TOP1 inhibitor camptothecin to kill Hct116 colorectal carcinoma cells. Using CRISPR/Cas9 to to replace exon 4 in the amino-terminus of TOP1 that is completely dispensable for catalysis, attenuated the interaction with RNAPII-CTD and abolished the synergy between JQ1 and camptothecin. This nexus between transcription and the control of DNA topoisomerase activity constitutes a versatile tool to mold patterns of gene expression and promises to elicit new strategies to intercept pathological gene expression.

Dana Pe'er

Columbia University, New York, NY

Single-cell RNA-seq is a powerful method for uncovering the relationships between genes. However, extracting this information is fundamentally limited by inefficient RNA-capture and noisy amplification. Effectively, per cell, only a small subset of the total transcriptome is measured. In order to solve this problem, we have developed a diffusion-map based method we call MAGIC (markov affinity-based graph imputation of cells) that models a diffusion process, or equivalently, a markov random walk on the cell-similarity graph in order for cells to learn values of undersampled genes from their high dimensional neighbors. We show that MAGIC successfully recovers correlations between measured markers and also absolute values in artificially dropped out CyTOF single-cell data, and bulk RNA-sequencing data.

We use MAGIC to study the epithelial-to-mesenchymal transition (EMT) in the HMLE human breast cancer cell-line. We find that MAGIC is able to learn the underlying trajectory, or one-dimensional manifold that forms the main axis of progression in the transition. We also show that after MAGIC we can compute high dimensional density estimates of genes associated with the transition, as well functionals on density such as entropy and mutual information in order to learn about gene-gene interactions during EMT.

NEW STRATEGIES TO IDENTIFY TRANSCRIPTION FACTOR REGULATORY RELATIONSHIPS IMPORTANT FOR ORGAN SYSTEM DEVELOPMENT

Siqi Wu[1,2], Karl Kumbier[1,2], Ann S Hammonds[2], Susan E Celniker[2], Erwin Frise[2], Bin Yu[1,3]

[1]UC Berkeley, Department of Statistics, Berkeley, CA, [2]Lawrence Berkeley National Laboratory, Department of Genome Dynamics, Division of Life Sciences, Berkeley, CA, [3]UC Berkeley, Department of Electrical Engineering and Computer Sciences, Berkeley, CA

Spatially defined gene interactions have been long known to play important roles during development. In the late 1970s, with the example of the *Drosophila* embryo, Lohs-Schardin demonstrated that developmental fates are already spatially pre-determined in the early embryo. This pioneering work led to the development of a fate map, linking early embryonic spatial regions to differentiated organ systems. We and others have demonstrated a molecular basis for the fate map. Using over 1640 spatial gene expression images systematically collected during *Drosophila* embryogenesis and building on ideas from statistical machine learning, we have produced a novel image representation that is able to separate the composite spatial data into basic elements, which we call principal patterns. These principal patterns allow for a compact and interpretable representation of spatial gene expression images and, more importantly, can be directly linked to the *Drosophila* fate map, thus allowing computational analysis of the biologically defined spatial regions.

Building on our principal patterns we developed strategies to associate genes to fate map regions, follow their roles and investigate spatially defined regulatory interactions. Our analysis mapped regulatory networks to cell-fate programs and assigned putative roles to uncharacterized genes. We used our principal patterns as the basis for reconstructing spatially local networks, among them the well studied gap gene network. We will present results from CRISPR knockout experiments for a previously uncharacterized gene that preliminary data indicate has a role in the gap gene network. Finally, for the first time, we investigated relationships of genes expressed early in development to genes expressed later in development. Our results demonstrate an association between the fate map and organ system specification but not differentiation. Additionally, we identify known and new genetic relationships within the defined organ systems and assign functions both to characterized and uncharacterized genes.

GENOME-WIDE MEASUREMENT OF SPATIAL EXPRESSION IN PATTERNING MUTANTS OF *DROSOPHILA MELANOGASTER*

Peter A Combs[1,2], Michael B Eisen[3], Hunter B Fraser[1]

[1]Stanford University, Department of Biology, Stanford, CA, [2]UC Berkeley, Biophysics Graduate Group, Berkeley, CA, [3]UC Berkeley, Department of Molecular and Cell Biology and HHMI, Berkeley, CA

Genome sequencing has become commonplace, but the understanding of how those genomes ultimately specify cell fate during development is still elusive. Extrapolating insights from deep investigation of a handful of developmentally important Drosophila genes to understanding the regulation of all genes is a major challenge. The developing embryo provides a unique opportunity to study the role of gene expression in pattern specification; the precise and consistent spatial positioning of key transcription factors essentially provides separate transcription readout experiments at a critical point in development.

We cryosectioned and sequenced mRNA from single Drosophila melanogaster embryos at the blastoderm stage to screen for spatially-varying regulation of transcription. Expanding on our previous screening of wild type embryos, here we present data from dosage mutants for key maternally provided regulators, including depletion of zelda and hunchback and both over-expression and depletion of bicoid. These data recapitulate all of the expected patterning changes driven by these regulators; for instance, we show spatially-confined up-regulation of expression in the bicoid over-expression condition, and down-regulation of those genes in the bicoid knock-down case, consistent with bicoid's known function as an anterior-localized activator. However, they also highlight many genes with previously undescribed responses to these transcription factors.

Our data highlight the role of combinatorial regulation of patterning gene expression. When comparing changes in multiple conditions, genes responsive to one mutation tend to respond to other mutations in a similar fashion. Furthermore, genes that respond differently to these mutations tend to have more complex patterns of TF binding. We also present ongoing work studying allele-specific expression from cryosliced Drosophila hybrids. This approach will specifically highlight regulatory changes by using the two different genomes in each cell as a control for each other.

BASE-PAIR RESOLUTION ATLASES OF THE ARABIDOPSIS CISTROME AND EPICISTROME

Shao-shan C Huang*[1,2], Ronan C O'Malley*[1,2], Liang Song[2], Mathew G Lewsey[2], Anna Barlett[1], Huaming Chen[1], Rosa Castanon[1], Cesar Barragan[1], Joseph R Nery[1], Joseph R Ecker[1,2,3]

[1]The Salk Institute for Biological Studies, Genomic Analysis Laboratory, La Jolla, CA, [2]The Salk Institute for Biological Studies, Plant Biology Laboratory, La Jolla, CA, [3]The Salk Institute for Biological Studies, Howard Hughes Medical Institute, La Jolla, CA

We developed a high-throughput sequencing assay for rapid transcription factor binding site (TFBS) discovery, DNA affinity purification sequencing (DAP-seq), that uses *in vitro* prepared transcription factors (TFs) to capture native genomic DNA. We applied DAP-seq to 1,812 *Arabidopsis thaliana* TFs to resolve motifs for 529 factors and genome-wide enrichment maps for 349 factors. Cumulatively, the ~2.7 million experimentally-determined TFBSs captured the Arabidopsis cistrome and predicted thousands of TF target genes enriched for known and novel functions. Base-resolution epicistrome maps were established by comparison of TF-binding to genomic DNA with native cytosine methylation patterns and genomic DNA that had been synthetically demethylated. This revealed methylcytosine inhibited binding of ~72% of factors and promoted binding of 4.3% of factors. Lastly, we showed DAP-seq binding sites provided a way to annotate genomic and epigenomic variations in natural populations and interpret results from genome-wide association studies. Overall, DAP-seq enables rapid development of base-resolution cistrome and epicistrome atlases for a wide-array of applications for eukaryotic genomes.

* These authors contributed equally to this work.

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Dr. Stein Aerts
University of Leuven
stein.aerts@med.kuleuven.be

Mr. Ariel Afek
Duke University
arielafek7@gmail.com

Dr. Nadav Ahituv
University of California, San Francisco
nadav.ahituv@ucsf.edu

Dr. Julie Ahringer
Wellcome Trust/Cancer Research UK Gurdon Institute
ja219@cam.ac.uk

Dr. Asifa Akhtar
Max Planck Institute of Immunology & Epigenetics
akhtar@immunbio.mpg.de

Ms. Katherine Allen
Warwick University
k.allen@warwick.ac.uk

Dr. Ernesto Andrianantoandro
Cell Press
eandrianantoandro@cell.com

Mr. Kellen Andrilenas
Boston University
kandril@bu.edu

Dr. Aseem Ansari
University of Wisconsin-Madison
azansari@wisc.edu

Dr. Alexander Arguello
Nature Publishing Group
a.arguello@us.nature.com

Dr. David Arnosti
Michigan State University
arnosti@msu.edu

Dr. Elham Azizi
Columbia University
elham.azizi@columbia.edu

Dr. Sara Ballouz
CSHL
sballouz@cshl.edu

Ms. Jeanette Baran-Gale
University of North Carolina at Chapel Hill
jtbaran@email.unc.edu

Dr. Eric Batchelor
NIH/NCI
batchelore@mail.nih.gov

Dr. Michael Beer
Johns Hopkins University
mbeer@jhu.edu

Mr. Christopher Bennett
University of Colorado - Boulder
christopher.g.bennett@colorado.edu

Dr. Wendy Bickmore
MRC Human Genetics Unit
Wendy.Bickmore@igmm.ed.ac.uk

Ms. Kelly Biette
Harvard Medical School
kbiette@fas.harvard.edu

Dr. Ramon Birnbaum
Ben-Gurion University of the Negev
ramonb@bgu.ac.il

Mr. Amit Blumberg
Ben-Gurion University of the Negev
amitblum@post.bgu.ac.il

Prof Michael Brent
Washington University
brent@wustl.edu

Dr. Harmen Bussemaker
Columbia University
hjb2004@columbia.edu

Ms. Laura Campitelli
University of Toronto
laura.campitelli@mail.utoronto.ca

Ms. Safiye Celik
University of Washington
safiye@cs.washington.edu

Dr. Vijender Chaitankar
NEI / NIH
vijender.chaitankar@nih.gov

Dr. Deborah Chasman
University of Wisconsin--Madison
chasman@cs.wisc.edu

Dr. Sumantra Chatterjee
Johns Hokins Medicine
sumantra@jhmi.edu

Ms. Hemangi Chaudhari
Washington University In St. Louis
hemangi@wustl.edu

Ms. Ling Chen
Vanderbilt University
ling.chen@vanderbilt.edu

Mr. Hyunwoo Cho
Memorial Sloan-Kettering Cancer Center
hyc2004@med.cornell.edu

Dr. Stirling Churchman
Harvard Medical School
churchman@genetics.med.harvard.edu

Dr. Barak Cohen
Washington University School of Medicine
cohen@wustl.edu

Dr. Clayton Collings
Northwestern University
clayton.collings@northwestern.edu

Dr. Peter Combs
Stanford University
pcombs@stanford.edu

Dr. Patrick Cramer
Max Planck Institute for Biophysical Chemistry
office.cramer@mpibpc.mpg.de

Dr. Megan Crow
Cold Spring Harbor Laboratory
mcrow@cshl.edu

Ms. Miao Cui
Caltech
miaocui@caltech.edu

Dr. Charles Danko
Cornell University
dankoc@gmail.com

Dr. Laura DeMare
Genome Research/Molecular Case Studies
ldemare@cshl.edu

Dr. Barry Demchak
UC San Diego
bdemchak@ucsd.edu

Dr. Peter Denolf
Bayer Cropscience NV
peter.denolf@bayer.com

Mr. Klev Diamanti
Uppsala University
klev.diamanti@icm.uu.se

Dr. Robin Dowell
University of Colorado
robin.dowell@colorado.edu

Mr. Noah Dukler
Cold Spring Harbor Labs
ndukler@cshl.edu

Dr. Sean Eddy
HHMI/Harvard University
seaneddy@fas.harvard.edu

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Mr. Jesse Engreitz
MIT
engreitz@mit.edu

Dr. Yuda Fang
Shanghai Institute of Plant Physiology and Ecology
yfang@sibs.ac.cn

Dr. Marie-Julie Favé
Université de Montréal and OICR
marie-julie.fave@umontreal.ca

Dr. Chris Fiore
Syros Pharmaceuticals
cfiore@syros.com

Mr. Vincent FitzPatrick
Columbia University
vdf2104@columbia.edu

Ms. Jennifer Flynn
Washington University in St. Louis
jaflynn@wustl.edu

Prof. Nir Friedman
Hebrew University
nir.friedman@gmail.com

Dr. Erwin Frise
Lawrence Berkeley National Labs
erwin@fruitfly.org

Dr. Audrey Fu
University of Idaho
audreyf@uidaho.edu

Mr. Daniel Fuentes
Stanford University School of Medicine/HHMI
dfuentes@stanford.edu

Prof. Wataru Fujibuchi
Kyoto University
w.fujibuchi@cira.kyoto-u.ac.jp

Dr. Juan Fuxman Bass
University of Massachusetts Medical School
Juan.FuxmanBass@umassmed.edu

Dr. Audrey Gasch
University of Wisconsin-Madison
agasch@wisc.edu

Dr. Sole Gatto
Sanford Burnham Prebys Medical Discovery Institute
sgatto@sbpdiscovery.org

Dr. Xijin Ge
South Dakota State University
xijin.ge@sdstate.edu

Mr. Ilias Georgakopoulos-Soares
Wellcomre Trust Sanger Institute
igs@sanger.ac.uk

Ms. Deborah Gerard
University of Luxembourg
deborah.gerard@uni.lu

Dr. Anthony Gerber
National Jewish Health/UC Denver
gerbera@njhealth.org

Dr. Jason Gertz
University of Utah
jay.gertz@hci.utah.edu

Dr. David Gifford
Massachusetts Institute of Technology
gifford@mit.edu

Dr. Yoav Gilad
The University of Chicago
gilad@uchicago.edu

Dr. Jesse Gillis
Cold Spring Harbor Laboratory
jgillis@cshl.edu

Dr. Raluca Gordan
Duke University
raluca.gordan@duke.edu

Dr. Ian Grant
The Francis Crick Institute
ian.grant@crick.ac.uk

Dr. Jesse Gray
Harvard Medical School
gray@genetics.med.harvard.edu

Prof. William Greenleaf
Stanford University
wjg@stanford.edu

Dr. Andreas Gruber
University of Basel
andreas.j.gruber@gmail.com

Dr. Jun-tao Guo
University of North Carolina at Charlotte
jguo4@uncc.edu

Dr. Christopher Hammell
Cold Spring Harbor Laboratory
Chammell@cshl.edu

Dr. Melissa Harrison
University of Wisconsin Madison
mharrison3@wisc.edu

Mr. Yu He
Washington University in St. Louis
yu.he@wustl.edu

Dr. Martin Hemberg
Wellcome Trust Sanger Institute
mh26@sanger.ac.uk

Dr. David Hendrickson
Calico Life Sciences
dgh@calicolabs.com

Dr. Lydia Herzel
Yale University
lydia.herzel@yale.edu

Ms. Stephanie Hickey
University of Texas Southwestern Medical Center
stephanie.lepp@utsouthwestern.edu

Dr. Gerald Higgins
University of Michigan Medical School
gehiggin@med.umich.edu

Dr. Elisha (Yi-Hsuan) Ho
University of Wisconsin Madison
elisha.ho@wisc.edu

Dr. Yifei Huang
Cold Spring Harbor Laboratory
yihuang@cshl.edu

Dr. Di Huang
NIH/NLM/NCBI
huangd2@ncbi.nlm.nih.gov

Dr. Shao-shan Huang
The Salk Institute for Biological Studies
shhuang@salk.edu

Dr. Jialiang Huang
Dana-Farber Cancer Institute
jhuang@jimmy.harvard.edu

Prof. Tim Hughes
University of Toronto
t.hughes@utoronto.ca

Ms. Katelyn Hughes
Worcester Polytechnic Institute
kjhughes@wpi.edu

Ms. Elizabeth Hutton
Watson School of Biological Sciences, CSHL
ehutton@cshl.edu

Mr. Taeyoung Hwang
Johns Hopkins University
taeyoungh@gmail.com

Mr. Mohamed Ibrahim
Max-Delbrueck-Center for Molecular Medicine
mahmoud.ibrahim@mdc-berlin.de

Dr. Ariel Jaimovich
Stanford University
ajaimovich@stanford.edu

Dr. Shruti Jain
Dr. B. R. Ambedkar Centre for Biomedical Research
shrutijain.25@gmail.com

Dr. James Jarvis
University at Buffalo
jamesjar@buffalo.edu

Ms. Tisha Jepson
Johns Hopkins University
tishajf@gmail.com

Dr. Hongkai Ji
Johns Hopkins University
hji@jhu.edu

Dr. Kaiyu Jiang
University at Buffalo
kaiyujia@buffalo.edu

Dr. Lingling Jiang
Columbia University
lj2388@columbia.edu

Dr. Raja Jothi
National Institutes of Health
jothi@mail.nih.gov

Mr. Akshay Kakumanu
Penn State University
auk262@psu.edu

Mr. Jim Kaminski
UC-Berkeley
jimkaminski@berkeley.edu

Dr. Tommy Kaplan
The Hebrew University
tommy@cs.huji.ac.il

Dr. Ashish Kapoor
Johns Hopkins University School of Medicine
ashish@jhmi.edu

Mr. Mehran Karimzadeh
Princess Margaret Cancer Centre
mehran.karimzadeh@uhnresearch.ca

Dr. Yuki Katou
Tokyo University
ykatou@iam.u-tokyo.ac.jp

Mr. Samuel Katz
National Institutes of Health
samuel.katz@nih.gov

Dr. Koray Kaya
NEI / NIH
kayak@nei.nih.gov

Ms. Jessica Keenan
Boston University
jlkeenan@bu.edu

Dr. Claudia Keller
Max-Planck-Institute f.Immunobiology & Epigenetics
keller@ie-freiburg.mpg.de

Dr. Joomyeong Kim
Louisiana State University
jkim@lsu.edu

Dr. Sara Knaack
University of Wisconsin Madison
sara.a.knaack@gmail.com

Dr. Jan Komorowski
Uppsala University
jan.komorowski@icm.uu.se

Dr. Briana Kozlowicz
Cargill
briana_kozlowicz@cargill.com

Dr. Anat Kreimer
UCB
anat.kreimer@berkeley.edu

Ms. Judith Kribelbauer
Columbia University Medical Center
jfk2132@cumc.columbia.edu

Dr. Hillel Kugler
Bar-Ilan University
hkugler@outlook.com

Dr. Anshul Kundaje
Stanford University
akundaje@stanford.edu

Mr. Jesse Kurland
Harvard Medical School and Brigham & Women's Hospi
jvkurland@partners.org

Dr. Guray Kuzu
Massachusetts General Hospital
kuzu@molbio.mgh.harvard.edu

Prof. Riitta Lahesmaa
University of Turku
rilahes@utu.fi

Mr. John Lazar
University of Washington
jlazar@uw.edu

Mr. Dongwon Lee
Johns Hopkins University
dwlee@jhu.edu

Dr. Su-In Lee
University of Washington
suinlee@cs.washington.edu

Dr. Robin E.C. Lee
University of Pittsburgh
robinlee@pitt.edu

Dr. Traci Lee
University of Wisconsin - Parkside
leet@uwp.edu

Dr. Christina Leslie
Memorial Sloan-Kettering Cancer Center
cleslie@cbio.mskcc.org

Dr. David Levens
National Cancer Institute
levens@helix.nih.gov

Dr. Shan Li
NIH/NLM/NCBI
lis11@ncbi.nlm.nih.gov

Prof Hyungsik Lim
Graudate Center and Hunter College, CUNY
hyungsik.lim@hunter.cuny.edu

Prof. John Lis
Cornell University
jtl10@cornell.edu

Dr. Yunxian Liu
University of Virginia
yl3pv@virginia.edu

Dr. Shirley Liu
Dana Farber Cancer Institute
xsliu@jimmy.harvard.edu

Prof. Gui-Rong Liu
University of Cincinnati
liugr@uc.edu

Dr. Mats Ljungman
University of Michigan
ljungman@umich.edu

Mr. Yuheng Lu
Memorial Sloan Kettering Cancer Center
yul2013@med.cornell.edu

Dr. Francesca Luca
Wayne State University
fluca@wayne.edu

Mr. Bryan Lunt
University of Illinois at Urbana-Champaign
bjlunt2@illinois.edu

Dr. Gholson Lyon
Cold Spring Harbor Laboratory
gholsonjlyon@gmail.com

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Dr. Shaun Mahony
Penn State University
mahony@psu.edu

Ms. Gabriella Martini
Columbia University
gdm2120@columbia.edu

Dr. Stephen Matheson
Cell Reports
smatheson@cell.com

Dr. Matthew Maurano
NYU Medical Center
maurano@nyu.edu

Dr. Steven McCarroll
Harvard Medical School, Broad Institute
mccarroll@genetics.med.harvard.edu

Dr. Scott McIsaac
Calico Life Sciences
rsm@calicolabs.com

Dr. Daniel McKay
UNC-Chapel Hill
dmckay1@email.unc.edu

Dr. Rob Mitra
Washington University School of Medicine
rmitra@genetics.wustl.edu

Mr. Amir Momen-Roknabadi
HHMI/ Princeton University
amirr@princeton.edu

Dr. Andreas Moor
Weizmann Institute of Science
andreas.moor@weizmann.ac.il

Dr. John Murray
University of Pennsylvania School of Medicine
jmurr@mail.med.upenn.edu

Mr. Juozas Nainys
Vilnius University
jaazas@gmail.com

Dr. Ryo Nakaki
RCAST
nakaki@genome.rcast.u-tokyo.ac.jp

Dr. Ryuichiro Nakato
Univ. of Tokyo
rnakato@iam.u-tokyo.ac.jp

Ms. Khoi Nguyen
Massachusetts Institute of Technology
ktnguyen@mit.edu

Dr. Miroslav Nikolov
Gene Center, LMU Munich
nikolov@genzentrum.lmu.de

Dr. Alex Nord
UC Davis
asnord@ucdavis.edu

Dr. Panos Oikonomou
Columbia University
po2194@columbia.edu

Dr. Shinya Oki
Kyushu university
soki@dev.med.kyushu-u.ac.jp

Dr. Sascha Ott
University of Warwick
s.ott@warwick.ac.uk

Dr. Athma Pai
Massachusetts Institute of Technology
athma@mit.edu

Dr. Evanthia Pashos
University of Pennsyvania
evpashos@mail.med.upenn.edu

Dr. Dana Pe'er
Columbia University
peerster@gmail.com

Ms. Ashley Penvose
Boston University
apenvose@bu.edu

Dr. Luca Pinello
Dana-Farber Cancer Institute/Harvard School of Public Health
lpinello@jimmy.harvard.edu

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Dr. Maria Polychronidou
Molecular Systems Biology
maria.polychronidou@embo.org

Dr. Sebastian Pott
University of Chicago
spott@uchicago.edu

Ms. Sara Prescott
Stanford University
sprescot@stanford.edu

Dr. Yuri Pritykin
MSKCC
pritykin@cbio.mskcc.org

Dr. Teresa Przytycka
NIH/NLM/NCBI
przytyck@ncbi.nlm.nih.gov

Dr. Mario Pujato
Cincinnati Children's Hospital Medical Center
mario.pujato@gmail.com

Ms. Wen Qu
Wayne State University
er7989@wayne.edu

Dr. Nisha Rajagopal
MIT
nish.iitg@gmail.com

Dr. Ritika Ramani
Cold Spring Harbor Laboratory
rramani@cshl.edu

Mr. Satyanarayan Rao
USC
satyanar@usc.edu

Mr. Chaitanya Rastogi
Columbia University
cr2166@columbia.edu

Dr. Aviv Regev
Broad Institute of MIT & Harvard
aregev@broadinstitute.org

Dr. Jüri Reimand
Ontario Institute for Cancer Research
juri.reimand@utoronto.ca

Dr. Todd Riley
University of Massachusetts Boston
todd.riley@umb.edu

Ms. Julia Rogers
Brigham and Women's Hospital, Harvard University
juliarogers@fas.harvard.edu

Dr. Remo Rohs
University of Southern California
rohs@usc.edu

Mr. Gil Ron
Hebrew University of jerusalem
gil.ron@mail.huji.ac.il

Dr. Sushmita Roy
University of Wisconsin Madison
sroy@biostat.wisc.edu

Dr. Ronen Sadeh
The Hebrew University
ronensadeh@gmail.com

Dr. Kunie Sakurai
Kyoto University
k.sakurai@cira.kyoto-u.ac.jp

Dr. Amy Schmid
Duke University
amy.schmid@duke.edu

Ms. Clarissa Scholes
Harvard Medical School
clarissascholes@fas.harvard.edu

Dr. Dirk Schubeler
Friedrich Miescher Institute for Biomedical Research
dirk.schubeler@fmi.ch

Dr. Anurag Sethi
Yale University
anurag.sethi@yale.edu

Dr. Manu Setty
Columbia University
manu.setty@columbia.edu

Dr. Maxwell Shapiro
Stony Brook University
maxwell.shapiro@stonybrook.edu

Ms. Jeewoen Shin
Stony Brook University
jeewoen.sin@stonybrook.edu

Dr. Kyle Siebenthall
Altius Institute for Biomedical Sciences
kts@stamlab.org

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Trevor Siggers
Boston University
tsiggers@bu.edu

Ms. Irtisha Singh
Memorial Sloan Kettering Cancer Center
is327@cornell.edu

Dr. Max Staller
Washington University School of Medicine
mstaller@wustl.edu

Dr. John Stamatoyannopoulos
University of Washington
jstam@uw.edu

Dr. Joshua Starmer
University of North Carolina at Chapel Hill
starmer@unc.edu

Dr. Mike Stubbington
Wellcome Trust Sanger Institute
mstubb@ebi.ac.uk

Mr. Suhas Sureshchandra
University of California, Riverside
ssure003@ucr.edu

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Dr. Lara Szewczak
Cell
lszewczak@cell.com

Dr. Mei San Tang
NYU School of Medicine
MeiSan.Tang@nyumc.org

Dr. Saeed Tavazoie
Columbia University
ST2744@COLUMBIA.EDU

Dr. Sarah Teichmann
European Bioinformatics Institute (EMBL-EBI)
sarah.teichmann@ebi.ac.uk

Dr. Lina Thomas
Universidade de Sao Paulo
linadth@ime.usp.br

Ms. Ngoc Tran
Cold Spring Harbor Laboratory
ntran@cshl.edu

Dr. Shuichi Tsutsumi
The University of Tokyo
tsutsumi@ja2.so-net.ne.jp

Ms. Kelsey Tyssowski
Harvard Medical School
tyssowski@fas.harvard.edu

Dr. D.B.R.K. Gupta Udatha
Norwegian PSC Research Center
gupta.udatha@outlook.com

Dr. Bas Van Steensel
Netherlands Cancer Institute
b.v.steensel@nki.nl

Dr. Charles Vejnar
Yale University
charles.vejnar@yale.edu

Dr. Charles Vinson
National Cancer Institute
Vinsonc@mail.NIH.GOV

Dr. Pavle Vrljicak
University of Warwick
P.J.Vrljicak@warwick.ac.uk

Dr. Daifeng Wang
Yale University
daifeng.wang@yale.edu

Dr. Jianrong Wang
Stanford University
wang.jianrong03@gmail.com

Dr. Toby Warnecke
Medical Research Council
tobias.warnecke@csc.mrc.ac.uk

Mr. Colin Waters
Harvard Medical School
colinwaters@fas.harvard.edu

Dr. Lonnie Welch
Ohio University
welch@ohio.edu

Dr. Michael White
Washington University in St. Louis
mwhite@genetics.wustl.edu

Dr. Hua-Jun Wu
Dana-Farber Cancer Institute
hjwu@jimmy.harvard.edu

Ms. BEIBEI XIN
UNIV OF SOUTHERN CA
bxin@usc.edu

Dr. Jin Xu
Stanford University School of Medicine
jinxu9@stanford.edu

Dr. Gilad Yaakov
Weizmann Institute of Sciene
Gilad.yaakov@weizmann.ac.il

Ms. Naomi Yamada
Penn State University
mahony@psu.edu

Mr. Jamie Yang
Columbia University
jsy2118@cumc.columbia.edu

Mr. Rui Yang
Northwestern University
ruiyang2016@u.northwestern.edu

Dr. Yong Yao
National Institute of Mental Health
yyao@mail.nih.gov

Dr. Yang Yu
Cold Spring Harbor Laboratory
yyu@cshl.edu

Dr. Y. Adam Yuan
National University of Singapore
dbsyya@nus.edu.sg

Ms. Xuefei Yuan
The hospital for sick children
xuefei.yuan@mail.utoronto.ca

Dr. Amanda Zacharias
University of Pennsylvania
azach@mail.med.upenn.edu

Dr. Chuntao Zhao
CCHMC
chuntao.zhao@cchmc.org

Dr. Fangjie Zhu
Karolinska Institute
fangjie.zhu@ki.se

Mr. Junjie Zhu
Stanford University
jjzhu@stanford.edu

Mr. Yizhou Zhu
Albert Einstein College of Medicine
yizhou.zhu@phd.einstein.yu.edu

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* – Use PIN#** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail 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 (Cash Only)
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes, ATM
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, east wing, lower level
Press PIN# (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 ($9.00 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**)
US Limousine Service	800-962-2827,ext:3 **(1047)**
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**)

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