PRESENTATION TITLE: Transcription factor binding site prediction in vivo using DNA sequence and shape features

AUTHORS:

Anthony Mathelier, University of British Columbia, Canada
Lin Yang, University of Southern California, United States of America
Tsu-Pei Chiu, University of Southern California, United States of America
Remo Rohs, University of Southern California, United States of America
Wyeth Wasserman, University of British Columbia, Canada

ABSTRACT:

The interaction between transcription factors (TFs) and DNA results from a complex interplay between the DNA structure and nucleotide-amino acid contacts. We previously developed a sequence-based flexible hidden Markov model approach for TF binding site (TFBS) prediction, the Transcription Factor Flexible Model (TFFM). We describe a novel approach which builds upon the TFFMs by integrating four DNA shape features derived from the DNAshape prediction method: minor groove width, roll, propeller twist, and helix twist. The model is an ensemble machine learning framework combining the sequence and shape information of the DNA at TFBSs. Our results from 400 human ENCODE ChIP-seq data-sets show that adding DNA shape features to the TFFM scores in the machine learning framework consistently improves the prediction of TFBSs. Namely, combining DNA sequence and shape information perform better than current methods to discriminate ChIP-seq from background sequences. We observe significant improvements in TFBS predictions, which are not specific to TFFMs as they are reproduced when using the classical position-specific scoring matrices in the same framework. From the TFs under study, the results highlight that incorporating DNA shape information is most beneficial when applied to the E2F and MADS-domain TF families. A further validation on plant ChIP-seq data underline the importance of the propeller twist at a specific position of the TFBSs for seven plant MADS-domain TFs. Our approach corresponds to the in vivo counterpart of previous in vitro studies underscoring the importance of DNA shape for TFBS recognition.

PRESENTATION TITLE: Evaluating Genetic Variation Impact on Transcription Factor Binding Sites
**Abstract:**

Current clinical sequence analysis focuses on exomes that highlight protein coding regions, despite awareness that cis-regulatory variations can cause human genetic disorders. Genome-wide association studies have identified thousands of disease-related variations most of which fall within cis-regulatory regions. Whole genome sequencing is now widely used in clinical genetics research, but the bioinformatics methods for the identification of functional regulatory changes are inadequate. The need to interpret and prioritize regulatory variations is becoming urgent for clinical genome analysis.

In this project, we focus on prioritizing variations likely to disrupt transcription factor binding sites in cis-regulatory elements. In developing the methods for cis-regulatory sequence analysis, we focus on differential transcription factor (TF) binding between two alleles distinguished by single nucleotide alterations. In ChIP-seq data, allele-specific binding (ASB) events, which indicate a TF selectively binds to one of two alleles at heterozygous positions, directly reveal the impact of cis-regulatory variation on TFBS within the same cellular context. We extracted ASB events from ENCODE ChIP-Seq data coupled with available WGS data in the corresponding cells. This key ASB reference collection exhibits a strong relationship between the predicted strength of TF-DNA interactions (as scored with position weight matrices (PWM)) and observed TF binding in vivo. DNase I accessibility differences between two alleles are also strongly associated with TF binding difference across multiple TFs and cells. In a TF-specific manner, cofactors can be quantitatively identified based on the differential overlap of cofactor ChIP-seq peaks between ASB and non-ASB events. Combining the available feature data, a classifier model trained to distinguish between ASB and non-ASB events achieves good accuracy (e.g. 78% for CTCF).

---

**Presentation Title:** Precision drug rescue and drug repurposing using structural systems pharmacology

**Authors:**

Lei Xie, Hunter College, The City University of New York, United States

**Abstract:**

Precision medicine is an emerging method for disease treatment and prevention that takes into consideration individual genetic and environmental variability for each person. However, the advance of precision medicine is hindered by a lack of mechanistic understanding of the energetics and dynamics of drug-target and genetic interactions in the context of the whole human genome and interactome. To
address this challenge, we have developed a novel structural systems pharmacology approach to elucidate molecular basis and genetic biomarkers of drug action. Our approach combines big data analytics and mechanism-based modeling through integrating structural genomic, functional genomic, metabolomics, and interactomic data. By searching for all structurally-characterized human proteins and applying molecular modeling and machine learning, we are able to construct genome-scale high-resolution drug-target interaction models. Subsequently, we link the putative off-targets to genome-scale biological networks to identify drug modulation pathways and cryptic genetic factors. A novel algorithm k-DELADD is developed to model the impact of edge modification of biological network on network robustness, and is used to identify hidden genetic biomarkers of drug phenotypic response.

As proof-of-concept studies, we have applied our structural systems pharmacology approach to drug rescue and drug repurposing for precision medicine. We have identified cryptic genetic factors that account for the side effect of Torcetrapib, a cholesterol-lowering drug that failed in phase III clinical trial due to serious side effects. Recently, we have revealed molecular and genetic mechanisms of metformin, enabling us to repurpose metformin as a precision anti-cancer therapy. Metformin is a drug prescribed to treat type-2 diabetes, and exhibits anti-cancer effects in a portion of patients. However, the direct molecular and genetic interactions which produce this pleiotropic effect of metformin have not yet been fully described. Using our approach, we predict a set of potential molecular targets of metformin, and experimentally validate the interaction between metformin and a majority of our top-ranked off-targets. Notably, SGK1 was identified as a key target in mediating metformin's effect. Moreover, key genetic network biomarkers are identified. They are largely consistent with existing experimental evidence. Our results shed new light on repurposing metformin as safe, effective, personalized therapies, and demonstrate that structural systems pharmacology is a potential powerful tool to facilitate the development of precision medicine.

Presentation Title: Tracking the Evolution of 3D Gene Organization

Authors:
Alon Diament, Tel Aviv University, Israel
Tamir Tuller, Tel Aviv University, Israel

Abstract:
One of the most fundamental open biological questions is what determines the eukaryotic genomic organization. It has been shown that the distribution of genes in eukaryotic genomes is not random; however, formerly reported large scale relations between gene function and genomic organization were relatively weak.

Previous studies have demonstrated that codon usage bias is related to all stages of gene expression and to protein function. Here we apply a novel tool for assessing functional relatedness, codon usage frequency similarity (CUFS), which measures similarity between genes in terms of codon and amino acid usage. By analyzing Hi-C
data, describing the three dimensional conformation of the DNA, we show that the functional similarity between genes captured by our metric is directly and very strongly correlated with their three dimensional (3D) distance in five eukaryotes ($r > 0.74; p<1e-323$ in all cases; Diament et al. Nature Commun. 2014).

We utilize this result to propose a novel approach for improving the accuracy of 3D genome reconstructions by introducing additional predicted physical interactions to the model, based on orthologous interactions in an evolutionary-related organism and based on predicted functional interactions between genes (e.g. based on CUFS). We demonstrate in the eukaryote S. cerevisiae that this approach indeed leads to the reconstruction of improved models (Diament et al. PloS Comput. Biol. 2015).

We have previously shown that some level of conservation of genomic organization exists between organisms. However, almost all studies of 3D genomic organization analyzed each organism independently from others. Here we propose a novel approach for inter-organismal analysis of the organization of genes. By utilizing Hi-C data from two fungi – S. cerevisiae and S. pombe – we detect orthologous gene families that underwent changes in their 3D co-localization during evolution. We show that this approach enables identifying various biologically meaningful modules of co-evolving genes with shared function (Diament et al. under-review 2015).

Our results emphasize the importance of three-dimensional genomic organization in eukaryotes and suggest that the evolutionary mechanisms that shape the 3D organization of genes are affected by their functionality and expression pattern. In addition, we provide novel algorithms for 3D genome reconstructions and for deciphering gene function and organization.

**Presentation Title:** The germline genetic component of drug sensitivity in cancer cell lines

**Authors:**

Michael P. Menden, European Bioinformatics Institute (EMBL-EBI), United Kingdom
Francesco P. Casale, European Bioinformatics Institute (EMBL-EBI), United Kingdom
Johannes Stephan, European Bioinformatics Institute (EMBL-EBI), United Kingdom
Graham R. Bignell, Wellcome Trust Sanger Institute, United Kingdom
Francesco Iorio, European Bioinformatics Institute (EMBL-EBI), United Kingdom
Ultan McDermott, Wellcome Trust Sanger Institute, United Kingdom
Mathew J. Garnett, Wellcome Trust Sanger Institute, United Kingdom
Julio Saez-Rodriguez, RWTH Aachen University Hospital, Germany
Oliver Stegle, European Bioinformatics Institute (EMBL-EBI), United Kingdom

**Abstract:**

Understanding differential therapeutic efficiency of cancer drugs is critical to deliver personalised treatments. While the utility of somatic mutations as drug sensitivity
biomarker has been extensively studied, much less is known about the relevance of the germline genetic background. Herein, we develop a joint analysis approach that leverages both types of mutations in the context of drug susceptibility, before applying it to the systematic genetic analysis of 993 cell lines and 265 drugs from the Genomics of Drug Sensitivity in Cancer resource (GDSC). We find that the germline contribution to drug efficacy can be as large or larger than effects due to somatic mutations and demonstrate clear improvements in the prediction of drug response. In addition, we identify epistatic interactions between germline variants and somatic mutation signatures, revealing widespread interconnections between both variant types. Finally, we unpick the molecular mechanism of a mutation in NQO1, which affects drug susceptibility of the HSP90 inhibitor 17-AAG, thereby demonstrating the value of germline genetic variations as an additional dimension to inform treatment decisions.

**Presentation Title:** Bringing big genomic data into focus for studying complex diseases in specific biological contexts

**Authors:**

Arjun Krishnan, Princeton University, United States
Ran Zhang, Princeton University, United States
Victoria Yao, Princeton University, United States
Chandra Theesfeld, Princeton University, United States
Aaron Wong, Simons Foundation, United States
Alicja Tadych, Princeton University, United States
Natalia Volfovsky, Simons Foundation, United States
Alan Packer, Simons Foundation, United States
Alex Lash, Simons Foundation, United States
Olga Troyanskaya, Simons Foundation, United States

**Abstract:**

A big challenge in genomics is characterizing the genetic and functional dysregulation in complex diseases. Addressing this problem requires systematic computational approaches that can harness the explosion of data and bring ever‐finer biological contexts into focus e.g. tissue, cell-type, sex and age. Towards this goal, we recently developed a Bayesian framework that integrates thousands of gene-expression, protein-interaction and regulatory-sequence datasets to predict tissue-specific functional relationships between genes in each of 144 specific human cell-types and tissues.

Here, using autism spectrum disorder (ASD) as an example, we demonstrate how tissue-specific networks provide a valuable apparatus for generating hypotheses about the molecular basis of human diseases. ASD has a strong genetic basis that remains poorly characterized by sequencing and quantitative genetics studies. Using an evidence-weighted machine learning approach that utilizes the human brain-specific functional gene network, we generated the first genome-wide prediction of autism-associated genes. These predictions were validated using an independent
large case-control sequencing study. Leveraging these genome-wide predictions and the brain-specific network, our analyses demonstrate that the large set of ASD genes, including a host of novel candidates, converges on a smaller number of key cellular pathways and specific early developmental stages of the brain.

Manifesting in early development and being five times more common among boys than among girls, ASD is also an exemplar of diseases whose incidence or severity varies dramatically across the human lifespan and between the sexes. Therefore, our next goal lies in expanding our genomics toolkit to address age- and sex-specificity in addition to tissue/cell-type-specificity. we will conclude with preliminary results that demonstrate the promise of some of our approaches towards this goal.

---

**Presentation Title:** Creating a library of genome-wide chromatin state patterns during B lymphopoiesis

**Authors:**

Mark Maienschein-Cline, University of Illinois at Chicago, United States
Pinal Kanabar, University of Illinois at Chicago, United States
Neil Bahroos, University of Illinois at Chicago, United States
Malay Mandal, University of Chicago, United States
Marcus Clark, University of Chicago, United States

**Abstract:**

B lymphopoiesis proceeds through several stages, during which the cell undergoes rearrangements in its antibody gene content that change the specificity of its antigen recognition mechanisms. This process forms a crucial underpinning of our adaptive immune system. In addition to changes in the antibody genetic content, these stages are also defined by a number of important epigenetic modulations. However, only a fraction of these are currently understood, often only in the context of specific transcription factors in certain developmental stages. Obtaining more general patterns of chromatin state on a broader, lymphopoiesis-wide context would thus provide an invaluable resource underpinning our understanding of epigenomics changes in B lymphopoiesis.

To this end, we consider nucleosome positioning and chromatin accessibility, which play an important role in determining regions of the genome that regulatory factors can interact with. Methodologies like ATAC-seq, DNase-seq, and FAIRE-seq can be employed to detect loci with open chromatin. In particular, ATAC-seq is a newly developed and particularly powerful tool, as it involves a simpler protocol that can be applied to smaller populations of cells; ATAC-seq can also be used to infer nucleosome positioning (the converse of open chromatin) if paired-end sequencing is used. Additionally, the activity of specific transcription factors can be inferred using bioinformatic techniques like motif enrichment, making open chromatin measurements a valuable basis for additional epigenetic studies.
We have profiled eight B lymphopoietic stages using ATAC-seq and describe the exciting preliminary results and analysis strategy here. We first identified regions of interest genome-wide from open chromatin enrichment, and then used patterns of chromatin state changes to separate the loci into functionally differentiated groups. In particular, we used an unsupervised clustering approach to discover clusters of loci with concordant chromatin state changes, and employed consensus clustering to determine the number of distinct patterns that can be identified reliably. In our data, we discovered 11 distinct patterns that describe changes in chromatin state across >100,000 differentiated loci. In addition to revealing important changes to the regulatory landscape across B lymphopoiesis, we believe that these patterns and the associated loci can be used as a valuable reference “library” of chromatin state for future B lymphopoiesis studies.

**Presentation Title:** ReMiCs: Regularized Regression Model for Inference of Condition-specific Gene Regulatory Networks from Multiple Data Sets

**Authors:**
Nooshin Omranian, Max-Planck Institute of Molecular Plant Physiology, Germany
Zoran Nikoloski, Max-Planck Institute of Molecular Plant Physiology, Germany

**Abstract:**
Reconstruction of gene regulatory networks (GRNs) from gene expression levels alone often yields consensus regulatory interactions over multiple data sets. Yet, insights in the molecular basis of an organism’s response to developmental and environmental changes can only be obtained by identifying condition-specific GRNs. Here we propose a two-step approach, termed ReMiCs, based on regularized regression to infer condition-specific GRNs: In the first step, we determine the consensus regulatory interactions by employing least absolute shrinkage and selection operator (LASSO) with fusion penalty. The consensus GRN is reconstructed from multiple data sets by imposing three constraints modeled as fused LASSO: (1) sparsity of GRNs, (2) similarity of GRNs inferred from individual data sets, and (3) concordance of differential behavior between regulatory gene and its targets. In the second step, we determine the residual of each condition-specific data set upon removing the consensus effects found in the first step. We then infer the condition-specific interactions by employing weighted LASSO regression on the respective residual data set. We tested the performance of ReMiCs on two transcriptomics time-series data from Mycobacterium tuberculosis under hypoxia and reaeration as well as on transcriptomics time-series data from Arabidopsis thaliana under eight combinations of light and temperature perturbations. The comparative analysis against condition-specific regulatory interactions extracted from literature demonstrated the advantages of the ReMiCs over the existing state-of-the-art approaches applied on condition-specific data sets. The approach can readily be extended to include ChIP-Seq data used for pre-selecting likely interactions. The study indicates that the combination of sparse regression techniques with other biologically meaningful constraints is a promising framework.
for reconstructing consensus as well as condition-specific GRNs solely based on transcriptomics data.

**Presentation Title**: Nucleotide Sequence Composition Adjacent to Intronic Splice Sites Improves Splicing Efficiency and Reduces Translation Costs in Fungi

**Authors**:  
Zohar Zafrir, Tel Aviv University, Israel  
Tamir Tuller, Tel Aviv University, Israel

**Abstract**:  
RNA splicing is the central process of intron removal in eukaryotes known to regulate various cellular functions. The canonical sequence elements which are essential for intron recognition are well-known. However, the role of various sequence features affecting splicing efficiency, intronic retention, and translation regulation has yet to be thoroughly studied. Focusing on four fungi as model organisms (S. cerevisiae, S. pombe, A. nidulans, and C. albicans) we performed for the first time a comprehensive high resolution and large scale systems biology study, aimed at characterizing how splicing efficiency of introns and the crosstalk between gene splicing and translation are encoded in transcripts and affect their evolution. Our analysis suggests that pre-mRNA local folding strength at intronic boundaries is under selective pressure, as it directly affects splicing efficiency and improves recognition of intronic boundaries (Yofe* and Zafrir* et al., PLoS Genetic, 2014; Zafrir and Tuller, RNA, 2015). In addition, when considering the reading frame of exons upstream and adjacent to introns we find evidence of preference for intronic STOP codons close to the intronic 5’end and that the beginning of introns are selected for ‘codons’ with higher translation efficiency, presumably to reduce translation and metabolic costs in cases of non-spliced introns. Ribosomal profiling data analysis in S. cerevisiae supports the conjecture that in this organism intron retention frequently occurs; thus, introns are partially translated, and their translation efficiency affects organismal fitness (Zafrir and Tuller, under revision, 2015). These new discoveries are contributory steps towards a broader understanding of splicing regulation, mRNA translation, intron evolution, and the effect of silent mutations on gene expression and organismal fitness.

**Presentation Title**: Understanding gene regulation: from genetic variants to higher order chromatin structure

**Authors**:  
Inkyung Jung, Ludwig Institute of Cancer Research, United States  
Bing Ren, Ludwig Institute of Cancer Research, United States
Abstract:

Higher order chromatin structure is emerging as an important regulator of gene expression. However, chromatin three-dimensional dynamics in mammalian development and cell-type specification have yet to be fully explored. To address this question, we generated the first genome-wide chromatin interaction maps across a broad set of human tissues, human embryonic stem cells (hESC) and hESC-derived multiple lineages. From these data sets we uncovered extensive chromatin reorganization between distinct cell/tissue-types through alteration of active and inactive chromosomal compartment. Furthermore we delineate allelic chromatin interactions, chromatin modifications, and transcriptomes amongst a broad set of human tissues and cell lines. Using these data sets we explored allele biased gene expression and the mechanisms underlying it, enabled by a chromosome-spanning haplotype reconstruction strategy. The haplotype-resolved transcriptomes revealed extensive allelic biases in the transcription of human genes. The extensive allele biased gene expression correlates with allele biased chromatin states of linked promoters and enhancers and allelically biased higher-order chromatin structures. The integrative analyses of chromatin interaction maps and haplotype-resolved epigenome and transcriptome data sets shed light on the regulatory effect of genetic variants and higher-order chromatin structure.

References

Presentation Title: Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet

Authors:
Coby Viner, University of Toronto and Princess Margaret Cancer Centre, Canada
James Johnson, The University of Queensland, Australia
Nicolas Walker, University of Cambridge, United Kingdom
Marcela Sjoberg, Wellcome Trust Sanger Institute, United Kingdom
David J. Adams, Wellcome Trust Sanger Institute, United Kingdom
Anne C. Ferguson-Smith, University of Cambridge, United Kingdom
Timothy L. Bailey, The University of Queensland, Australia
Michael M. Hoffman, University of Toronto and Princess Margaret Cancer Centre, Canada
Abstract:

Many transcription factors (TFs) initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a TF might encounter. Cytosine is often present in the modified forms: 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC). TFs have been shown to distinguish unmodified from modified bases. Modification-sensitive TFs provide a mechanism through which widespread changes in DNA methylation and hydroxymethylation, found in many cancers, can dramatically shift active gene expression programs.

To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites in DNA with covalent modifications. Our models expand the standard A/C/G/T alphabet, adding m (5mC) and h (5hmC). We adapted the well-established position weight matrix formulation of TF binding affinity to this expanded alphabet.

We engineered several tools to work with expanded-alphabet sequence and position weight matrices. First, we developed a program, Cytomod, to create a modified sequence, using data from bisulfite and oxidative bisulfite sequencing experiments. Cytomod decides between multiple modifications at a single locus, using a configurable evidence model. Second, new versions of MEME (Multiple EM for Motif Elicitation), DREME (Discriminative Regular Expression Motif Elicitation), and MEME-ChIP enable de novo discovery of modification-sensitive motifs. A new version of CentriMo enables central motif enrichment analysis to infer direct DNA binding in an expanded-alphabet context. These versions permit users to specify new alphabets, anticipating future alphabet expansions.

We created an expanded-alphabet genome sequence using whole-genome maps of 5mC and 5hmC in naive ex vivo mouse T cells from BLUEPRINT. Using this sequence, expanded-alphabet position weight matrices, and ChIP-seq data from Mouse ENCODE and others, we identified cis-regulatory modules active only in the presence or absence of cytosine modifications. We reproduced various known methylation binding preferences, including the preference of ZFP57 and C/EBP for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.

Presentation Title: Genome-wide mapping of histone marks at single-nucleosome resolution

Authors:

Marcelo Rivas-Astroza, University of California, San Diego, and University of Illinois, Urbana-Champaign, United States
Shu Xiao, University of California, San Diego, United States
Abstract:

Histone post-translational modifications (PTM) control much of the workings of the genome. However, it remains unclear how different histone PTM combine within individual nucleosomes to encode regulatory functions. Regular protocols of chromatin immunoprecipitation followed by sequencing (ChIP-seq) are unsuited to shed light on this matter as they are unable to trace histone PTM to individual nucleosomes. Alternative ChIP-seq protocols employing enzymatic digestion of DNA can overcome this limitation and produce histone PTM footprints at single-nucleosome resolution, but analytical obstacles still precludes determining whether a nucleosome is marked or not by a histone PTM. In particular, care has to be taken to avoid co-founding effects stemming from the relation between nucleosomes co-localization levels and nucleosomal enrichment of histone marks. To overcome these problems, here we devised computational and statistical methods to exploit the power of ChIP-seq protocols based on enzymatic digestion of DNA to trace histone PTM to individual nucleosomes. Using a ChIP-seq protocol where DNA was MNase-digested and our analytical methods we generated genome-wide maps where H3K4me3, H3K27Ac, H3K9me3, and H3K27me3 were traced to individual nucleosomes in mouse embryonic stem cells. A significant number of nucleosomes were marked by two or more of histone marks. Nucleosomes marked simultaneously by H3K4me3 and H3K27me3 were prevalent among bivalent domains compared to the genomic background. Nucleosomes having the repressive marks H3K27me3 and H3K9me3 were enriched at the transcription starting site of highly active genes only if they were also co-localized with the activating mark H3K27Ac. Inclusion of alternatively spliced exons on the final mRNA was correlated with nucleosomes marked by H3K4me3, H3K27Ac, or H3K27me3, but was largely unaffected by nucleosomes marked by H3K9me3. Together, these findings indicate that combinatorial patterns of histone PTM within individual nucleosomes are fundamental units of regulatory information.

Presentation Title: Estimating the Number and Diversity of Cancer Mutations In the Overall Population from 5,319 Complete Cancer Genomes

Authors:

Prathik Naidu, Thomas Jefferson High School for Science and Technology, United States
Joseph Kaplinsky, Beth Israel Deaconess Medical Center, United States
Ramy Arnaout, Beth Israel Deaconess Medical Center and Harvard Medical School, United States
Abstract:

Cancer is a genetic disease. To understand the link between cancer and genetics, large-scale efforts such as The Cancer Genome Atlas (TCGA) have begun to catalog cancer-related mutations, using hundreds of samples across many cancer types and subtypes. However, because even this sample size is small compared to the tens of millions of people with cancer worldwide, there is a risk for substantial sampling bias: the number and diversity of cancer mutations in the sample may not reflect their number and diversity in the overall population. To address this issue, we used the Recon (Reconstruction of Estimated Communities from Observed Numbers) algorithm to estimate the overall number and diversity of coding and non-coding cancer mutations for 14 common cancer types, including breast cancer, prostate cancer, and glioblastoma, and four clinically important subtypes of breast cancer (e.g. luminal A, luminal B Her2, and Basal). Our results suggest that while most common mutations have been discovered, the majority (78,186 – 1,887,539) remain unknown. Interestingly, the number of undiscovered mutations is not obvious from observations for a given cancer. For example, although glioblastoma and prostate-cancer samples exhibit relatively few mutations, our results suggest that overall, glioblastoma—but not prostate cancer—is likely to have as many mutations as all breast cancers. Thus, our algorithm may reveal aspects of cancer that are not obvious from direct observations (e.g., the potential presence of genetically different subtypes).

Presentation Title: tRNA-derived fragments in Drosophila and their potential targets

Authors:

Spyros Karaiskos, Rutgers University, United States
Ammar Naqvi, Rutgers University, United States
Karl Swanson, Rutgers University, United States
Andrey Grigoriev, Rutgers University, United States

Abstract:

In our paper, "Age-driven modulation of tRNA-derived fragments in Drosophila and their potential targets" (1), we presented rather unusual suspects in the world of gene regulation – transfer RNA fragments (tRFs). Such fragments have previously escaped detection or have been generally ignored as noise due to low count numbers in of small RNA libraries. While the focus in the analysis of such libraries has been primarily on microRNAs (miRNAs), recent studies have reported findings of fragments of transfer RNAs (tRFs) across a range of organisms.

We had originally investigated the age-related dynamics of miRNA loading into different RNA-induced silencing complexes (RISC) of Drosophila melanogaster using small RNA-Seq (2). Here we describe tRFs found in the same transcriptome libraries, and focus on their structural and functional features that make these
fragments similar to miRNAs. Similar to miRNAs tRFs have distinct isoforms with precise ends preferentially originating from 5’ or 3’ end of a precursor molecule (tRNA). Analogously to the seed sequences in miRNAs, we observe that tRF ends possess short 7-mer sequences matching conserved regions across 12 Drosophila genomes, preferentially in 3’ UTRs but also in introns and exons. Like miRNAs, tRFs display specific isoform loading into Ago1 and Ago2 and thus likely function in RISC complexes. And finally, as is the case with miRNAs, we observe the levels of tRF loading into Ago1 and Ago2 to differ considerably and both tRF expression and loading appear to be age-dependent, indicating potential regulatory changes from young to adult organisms.

We found that Drosophila tRF reads mapped to tRNA genes for all 20 amino acids, while previous studies have usually reported fragments from only a few tRNAs. Moreover, we detected fragments of both nuclear and mitochondrial tRNAs, while only the former have been described. Following the similarities with miRNAs and based on complementarity with conserved Drosophila genome regions we described seed sequence found in the most abundant tRFs. Further, we identified their possible targets with matches in the Drosophila melanogaster 3'UTR regions. Strikingly, these potential target genes of the most abundant tRFs show significant Gene Ontology enrichment in development and neuronal function. This observation suggests that involvement of tRFs in the RNA interfering pathway may play a role in brain activity or brain changes with age.


Presentation Title: Quantitative modeling of gene expression from sequence, using DNA shape-based model of binding sites

Authors:
Pei-Chen Peng, University of Illinois at Urbana Champaign, United States
Saurabh Sinha, University of Illinois at Urbana Champaign, United States

Abstract:
Motivation: Prediction of gene expression levels driven by regulatory sequences is one of the major challenges of genomic biology. A major current focus in transcriptional regulation is sequence-to-expression modeling, which interprets the enhancer sequence in light of transcription factor concentrations and DNA binding specificities and predicts precise gene expression levels in varying cellular contexts. Such models have so far exclusively relied on the position weight matrix (PWM) model for transcription factor (TF)-DNA binding. Several reports have pointed out deficiencies in the PWM model and presented alternative models, including DNA shape-based models, that are claimed to be in greater agreement with TF-DNA binding data. However, it is not known if alternative models of DNA binding, such as DNA shape models, can also improve prediction of gene expression.

Results: Here, we adapted a statistical thermodynamics model to develop a
quantitative model of gene expression interprets enhancer sequences using DNA shape features of binding sites, as opposed to a PWM-based scoring of sites. We used rigorous methods to evaluate the fits of expression readouts of more than 35 enhancers regulating spatial gene expression patterns in the blastoderm-stage Drosophila embryo, and we show that DNA shape-based models perform at least as well as, and arguably better than PWM-based models. We objectively characterized the relationship between DNA shape-based models and PWM models of binding site affinity, and observed that DNA shape features carry information that is complementary to the PWM and useful for sequence-to-expression modeling. In addition, we combined DNA shape and PWM into a single model and tested if it would achieve better predictions than using either binding model independently. The integrative model did not perform consistently better than either DNA shape or PWM based model alone.

Conclusion: Our work shows that quantification of TF binding site affinity using DNA shape is not only justified by binding affinity data, it is also effective in interpreting enhancer sequence to accurately predict gene expression. With the growing availability of data sets describing TF-DNA binding affinities comprehensively, we expect that it will be possible to train such models more accurately and to utilize them to better predict gene expression and the functional effects of single nucleotide polymorphisms in the non-coding genome.

---

**Presentation Title:** A linear time-invariant model of phenotype dynamics in breast cancer cell populations

**Authors:**
Margaret Chapman, UC Berkeley, United States
Tyler Risom, Oregon Health and Science University, United States
Roel Dobbe, UC Berkeley, United States
Rosalie Sears, Oregon Health and Science University, United States
Claire Tomlin, UC Berkeley, United States

**Abstract:**

Background: Phenotypic heterogeneity, or cellular diversity on the phenotypic level, poses a fundamental challenge to effective treatment of certain cancers, including triple negative breast cancer (TNBC). Thus, the discovery of strategies to control heterogeneity is a priority in cancer biology. Understanding the phenomena that govern drug-induced changes in cell population phenotypes is needed to design therapeutic approaches that systematically control heterogeneous tumors. Here we present a mathematical framework to suggest why observed trends in phenotype dynamics occur and why they may change under therapeutic perturbation.

Methods: Phenotype and dead cell quantities were recorded from 4 replicate wells containing the TNBC cell line, HCC1143, at 6 time points every 12 hours under three distinct conditions. The conditions were application of a PI3K/mTOR inhibitor, a
MEK inhibitor, or a control containing the drug vehicle DMSO but no medicinal agent. A linear time-invariant model of phenotype dynamics was derived to investigate the therapeutic effects on phenotypic heterogeneity. Cell division and death of each phenotype in addition to switching between phenotype pairs were represented explicitly. A convex optimization program was formulated and solved to estimate values of model parameters using measurements from two wells. Data from the two remaining wells were used for model validation. Parameter values were analyzed in the context of existing biological hypotheses. This methodology will be compared to an alternative approach, involving a stochastic process, from the literature.

Results: The model predicts monotonic trends well; predicting oscillatory behavior is reserved for future work. Analysis of model parameters agrees with existing knowledge on therapeutic effects and provides rationale for observed trends. For example, the model indicates that the PI3K/mTOR inhibitor reduces the prominence of basal phenotypes due to specific switching behavior and low cell division, rather than selective elimination. Further, the model suggests that the MEK inhibitor promotes the prominence of basal phenotypes via switching phenomena, instead of increased cell division or reduced death. Remarkably, the model results specify bidirectional switching between two phenotypes that precisely reflects the hypothesized effects of the MEK inhibitor.

Conclusion: Our framework motivates further experimentation to study model predictions and the design of precise metrics for therapeutic effectiveness. Short-term future work includes: analyzing statistical significance, investigating measurement noise, and identifying gaps in the model’s descriptive power. The long-term goals are to explore richer models to capture oscillatory trends (e.g., stochastic linear time-varying or non-linear) and incorporate additional sources of biological data into the methodology.

Presentation Title: DREISS: dynamics of gene expression driven by external and internal regulatory networks based on state space model

Authors:
Daifeng Wang, Yale University, United States
Fei He, Brookhaven National Laboratory, United States
Sergei Maslov, Brookhaven National Laboratory, United States
Mark Gerstein, Yale University, United States

Abstract:
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 (dreiss.gersteinlab.org) [1]. Given a subsystem, the "state" and "control" in the model refer to its own (internal) and another subsystem's (external) gene expression levels. The state at a time is determined by the state and control at previous time. Because typical time-series data do not have enough samples to estimate the model's parameters, DREISS uses dimensionality reduction, and identifies canonical temporal expression trajectories (e.g., degradation, growth, oscillation) representing the regulatory effects coming from various subsystems.

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.


Presentation Title: An Experimentally Supported Model of the Bacillus subtilis Global Transcriptional Regulatory Network

Authors:

Mario L. Arrieta, New York University, United States
Christoph Hafemeister, New York University, United States
Ashley Rose Bate, New York University, United States
Timothy Chu, New York University, United States
Alex Greenfield, New York University, United States
Bentley Shuster, New York University, United States
Samantha N. Barry, New York University, United States
Matthew Gallitto, New York University, United States
Brian Liu, New York University, United States
Thadeous Kacmarczyk, New York University, United States
Francis Santoriello, New York University, United States
Jie Chen, New York University, United States
Abstract:

Organisms from all domains of life use gene regulation networks to control cell growth, identity, function, and responses to environmental challenges. Although accurate global regulatory models would provide critical evolutionary and functional insights, they remain incomplete, even for the best studied organisms. Efforts to build comprehensive networks are confounded by challenges including network scale, degree of connectivity, complexity of organism-environment interactions, and difficulty of estimating the activity of regulatory factors.

Taking advantage of the large number of known regulatory interactions in Bacillus subtilis and two transcriptomics datasets (including one with 38 separate experiments collected specifically for this study), we use a new combination of network component analysis and model selection to simultaneously estimate transcription factor activities and learn a substantially expanded transcriptional regulatory network for this bacterium. In total, we predict 2258 novel regulatory interactions and recall 74% of the previously known interactions. We obtained experimental support for 391 (out of 635 evaluated) novel regulatory edges (62% accuracy), thus significantly increasing our understanding of various cell processes, such as spore formation.

Presentation Title: Transcriptional regulatory network inference for rare immune cell populations from gene expression and chromatin accessibility measurements

Authors:

Emily R. Miraldi, NYU School of Medicine, United States
Maria Pokrovskii, NYU School of Medicine, United States
Jason A. Hall, NYU School of Medicine, United States
Nicholas Carriero, Simons Center for Data Analysis, United States
Dan R. Littman, NYU School of Medicine, United States
Richard Bonneau, Simons Center for Data Analysis, United States

Abstract:

Innate lymphoid cells (ILCs) compose a newly discovered and relatively rare population of immune cell lineages (ILC1, ILC2, ILC3) that play many important roles, including early defense against infections, tissue homeostasis and repair, and autoimmune disease. These diverse ILC physiological functions require coordination of complex gene expression patterns, involving thousands of genes. Our goal here is the inference of transcriptional regulatory networks (TRNs) to
model these gene expression responses as multivariate functions of transcription factor (TF) activities. Although TF mRNA can serve as a proxy for TF activity, this assumption is imperfect, given the importance of downstream processing (e.g., translation, chemical modification) to TF. Other groups have proposed using prior knowledge of TF target gene expression to estimate TF activity, and the Bonneau lab has recently integrated this TFA estimation procedure into our parsimonious TRN inference method, the Inferelator, yielding great performance advantages in the model organism B. subtilis, for which a database of known TF-target interactions exists (Ortiz, Hafemeister, et al., submitted). Unfortunately, very few transcriptional regulatory interactions are known for ILCs, and, thus, we developed a method to derive a “prior” of putative TF-gene interactions from ATAC-seq chromatin accessibility measurements and gene expression in ILCs. The ATAC-seq data, when integrated with TF motif information, can provide a list of TF candidate regulators that bind cis to target genes. However, given that motifs among TF families can be degenerate and often more than one TF in a given family is expressed per condition, a key challenge with ATAC-seq data is determining which transcription factors (TFs) contribute to accessible regions (e.g., peaks) for a given experimental condition. Here, I develop methods to harness the quantitative signal in ATAC-seq datasets across conditions to best ascribe ATAC-seq signal to the activity of a single TF. These ATAC-seq-derived priors then serve as input to the Inferelator, which uses context likelihood of relatedness (CLR) and Bayesian best subset regression (BBSR) to infer TRNs from gene expression and transcription factor activity estimates.

Importantly, we validate our integrative method for TRN inference from ATAC-seq and gene expression data in the more well-studied context of in vitro Th17 cells, for which we also generated ATAC-seq data and for which a gold-standard network based on TF knockout and ChIP-seq data exists. We then apply our methods to ILCs and use the resulting TRNs to understand ILC regulation in host health and pathophysiology.

Presentation Title: An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data

Authors:
Mark Carty, Memorial Sloan Kettering Cancer Center, United States
Lee Zamparo, Memorial Sloan Kettering Cancer Center, United States
Alvaro Gonzalez, Memorial Sloan Kettering Cancer Center, United States
Raphael Pelossof, Memorial Sloan Kettering Cancer Center, United States
Olivier Elemento, Weill Cornell Medical College, United States
Christina Leslie, Memorial Sloan Kettering Cancer Center, United States

Abstract:
We present an integrated model for estimating the significance of Hi-C interactions that accounts for systematic sources of variation in interaction read counts, including the dependence of random polymer ligation on genomic distance and GC content and mappability bias. Additionally, we explicitly model the zero-inflation
and over-dispersion of counts in the contact matrix by using a generalized linear model approach based on hurdle regression. By learning a null model that incorporates all these statistical properties of Hi-C contact matrix counts, we show that our estimates of significance (P values) are less inflated than previous methods, yielding fewer false discoveries. We performed an analysis on a recent very high resolution in situ Hi-C data set for the lymphoblastoid cell line GM12878 and found that our method can identify significant interactions at the sub-topologically associating domain level, including DNA loops mediated by CTCF and/or cohesin and enhancer-promoter interactions. We also found that CTCF-associated interactions were more strongly enriched in the middle range of genomic distance (~700Kb-1.5Mb), while interactions involving actively marked DNase accessible elements were enriched both at short range (<500Kb) and longer range (>1.5Mb). Strikingly, we found a dramatic enrichment of longer range gene-gene interactions connecting the replication-dependent histone genes on chromosome 6, potentially representing the chromatin architecture at the histone locus body.

**Presentation Title:** Large-scale models of signal propagation derived from phosphoproteomic data to study kinase inhibitors

**Authors:**

Julio Saez-Rodriguez, EMBL-EBI & RWTH Aachen University Hospital, United Kingdom  
Camille Terfve, EMBL-EBI, United Kingdom  
Edmund Wilkes, Queen Mary University of London, United Kingdom  
Pedro Casado, Queen Mary University of London, United Kingdom  
Pedro Cutillas, Queen Mary University of London, United Kingdom

**Abstract:**

Mass spectrometry is widely used to probe the proteome and its modifications in an untargeted manner, with unrivalled coverage. Applied to phosphoproteomics, it has tremendous potential to interrogate phospho-signalling and its therapeutic implications. However, this task is complicated by issues of undersampling of the phosphoproteome and challenges stemming from its high-content but low-sample-throughput nature. Hence, methods using such data to reconstruct signalling networks have been limited to restricted data sets and insights.

We propose PHONEMeS, a new method to handle high-content discovery phosphoproteomics data on perturbation by putting it in the context of kinase/phosphatase-substrate knowledge, from which we derive and train logic models [1]. We show, on a data set obtained through perturbations of cancer cells with small-molecule inhibitors [2], that this method can study the targets and effects of kinase inhibitors, and reconcile insights obtained from multiple data sets, a common issue with these data [1].

**References:**
Presentation Title: AEF: A methodology based on Assortativity of Epigenetic Features in promoter centered chromatin interaction networks identifies Polycomb and RNA Polymerase as main players

Authors:
Vera Pancaldi, Spanish National Cancer Research Centre (CNIO), Spain
Enrique Carrilo de Santa Pau, Spanish National Cancer Research Centre (CNIO), Spain
Peter Fraser, The Babraham Institute, United Kingdom
Mikhail Spivakov, The Babraham Institute, United Kingdom
Daniel Rico, Spanish National Cancer Research Centre (CNIO), Spain
Alfonso Valencia, Spanish National Cancer Research Centre (CNIO), Spain

Abstract:
With the advent of genome-wide chromatin interaction mapping techniques, numerous methods have been proposed to analyze the resulting datasets. However, only few studies have focused on the comparison of different datasets and even fewer exploit the numerous tools of network theory in the analysis of these networks.

Assortativity is a widely used metric in the field of social networks, which characterizes the extent to which similar nodes are more likely to interact. We propose to analyze assortativity of a number of epigenetic features in chromatin interaction networks with the aim of identifying key features that associate with chromatin contacts.

We consider two very recent promoter-centered interaction maps for mouse Embryonic Stem Cells obtained with two variations of promoter-capture HiC (PCHiC and HiCap), a novel technique that detects chromatin contacts involving at least one promoter. To characterize the interacting chromatin fragments, the nodes of the network, we exploit a collection of 79 epigenetic features (cytosine and histone modifications and binding peaks of chromatin related proteins). We project the abundance of these features on the network as node attributes and calculate their assortativity.

Plotting assortativity versus abundance of each epigenetic feature we identify the Polycomb Group Proteins and associated histone marks as particularly important in the chromatin contacts. The results are reproduced in the two datasets, despite the technical differences in the experimental techniques and in the topology of the mapped chromatin networks. Moreover, we distinguish between two subnetworks: the network of contacts amongst promoters (PP) and the network of contacts between promoters and other genomics fragments (PO), presumably distal elements.
We observe substantial differences in the assortativity of certain features in these subnetworks, for example the active promoter mark H3K4me3 has negative assortativity in the PP contacts and positive assortativity in the PO contacts. Interestingly, we find a difference in assortativity of different forms of RNA Polymerase II in PO contacts, with the actively elongating form showing much higher assortativity compared to inactive or poised forms. These findings, reproduced in two independent datasets and validated on additional ChIA-PET networks, are inline with the model in which enhancers contact the elongating form of polymerase ,which would be key in the regulation of transcription. To conclude, we propose assortativity of features in general or in subnetworks as a topology-aware metric that can be used to compare multiple contact maps and to assess their biological relevance.

Presentation Title: Global chromatin defects and spurious transcription defines a novel subset of human cancers

Authors:

Navneet Singh, Cincinnati Children’s Hospital Medical Center, United States
Kwangmin Choi, Cincinnati Children’s Hospital Medical Center, United States
Nathan Salomonis, Cincinnati Children’s Hospital Medical Center, United States
Matthew Weirauch, Cincinnati Children’s Hospital Medical Center, United States
Kakajan Komurov, Cincinnati Children’s Hospital Medical Center, United States

Abstract:

Through a comprehensive computational pan-cancer analysis of The Cancer Genome Atlas mRNA-seq datasets, we found that a subset (ranging in 2% to 30%) of almost every human cancer is characterized by severely defective gene body chromatin remodeling and RNA Polymerase II (Pol II)-mediated transcription, resulting in extensive genome-wide aberrant shortening of mRNA transcripts (GTS: Global Transcript Shortening). GTS was characterized by extensive cryptic spurious transcription and intron splicing defects, such that, more than half of all expressed genes were producing truncated transcripts. Strikingly, GTS mostly affected large genes with specific chromatin structures, while small genes involved in housekeeping processes were mainly overexpressed in these tumors. The mRNA-level observations, including pathway enrichments and gene length bias in expression, were also reproduced at the protein level in these tumors. Importantly, pro-inflammatory cell death pathways, which are primarily regulated by large genes, were significantly repressed in all GTS+ tumors due to excessive truncation of their mRNAs. Accordingly, GTS+ melanoma and clear cell renal cell carcinoma (ccRCC) patients had significantly poor response to immunotherapy with interferon (IFN), interleukin (IL) and the newer drugs such as ipilimumab in the clinic, but not to other forms of therapy. We found that some cancer cell lines in vitro exhibit the GTS phenotype. GTS+ cell lines had severe defects in genic histone methylation and acetylation, as well as in Pol II transcription initiation and elongation functions, and had a transcriptional and pathway activation profile highly consistent with GTS+ clinical samples. Consistent with defective immune cell death response phenotype, GTS+ cancer cells had impaired response to pro-inflammatory stimuli and FasL-
mediated killing. However, we show that GTS imposes specific metabolic and proteotoxic vulnerabilities in cancer cells with potential implications for therapeutic targeting in the clinic. Overall, our integrated computational and experimental analyses reveal a previously unknown clinically significant major phenotype of human cancers with clear implications in patient stratification for therapies, especially with immune modulatory agents, and therapeutic strategies to target this subset of cancers in the clinic.

**Presentation Title:** Functional, chemical genomic, and super-enhancer screening identify sensitivity to cyclin D1/CDK4 pathway inhibition in Ewing sarcoma

**Authors:**

Alyssa Kennedy, Harvard Medical School, United States  
Mounica Vallrupalli, Harvard Medical School, United States  
Liying Chen, Harvard Medical School, United States  
Brian Crompton, Harvard Medical School, United States  
Glenn Cowley, Broad Institute of MIT and Harvard, United States  
Paquita Vazquez, Broad Institute of MIT and Harvard, United States  
Barbara Weir, Broad Institute of MIT and Harvard, United States  
Aviad Tsherniak, Broad Institute of MIT and Harvard, United States  
S Parasuraman, Novartis Institute for Biomedical Research, United States  
Sun Kim, Novartis Institute for Biomedical Research, United States  
Gabriela Alexe, Harvard Medical School, United States  
Kimberly Stegmaier, Harvard Medical School, United States

**Abstract:**

Ewing sarcoma is an aggressive bone and soft tissue tumor in children and adolescents, with treatment remaining a clinical challenge. This disease is mediated by somatic chromosomal translocations of the EWS gene and a gene encoding an ETS transcription factor, most commonly, FLI1. While direct targeting of aberrant transcription factors remains a pharmacological challenge, identification of dependencies incurred by EWS/FLI1 expression would offer a new therapeutic avenue. We used a combination of super-enhancer profiling, near-whole genome shRNA-based and small-molecule screening to identify cyclin D1 and CDK4 as Ewing sarcoma-selective dependencies. We revealed that super-enhancers mark Ewing sarcoma specific expression signatures and EWS/FLI1 target genes in human Ewing sarcoma cell lines. Particularly, a super-enhancer regulates cyclin D1 and promotes its expression in Ewing sarcoma. We demonstrated that Ewing sarcoma cells require CDK4 and cyclin D1 for survival and anchorage-independent growth. Additionally, pharmacologic inhibition of CDK4 with selective CDK4/6 inhibitors led to cytostasis and cell death of Ewing sarcoma cell lines in vitro and growth delay in an in vivo Ewing sarcoma xenograft model. These results demonstrated a dependency in Ewing sarcoma on CDK4 and cyclin D1 and support exploration of CDK4/6 inhibitors as a therapeutic approach for patients with this disease.
**Presentation Title:** A PBM-based glossary for motif discovery in regulatory regions

**Authors:**
Luca Mariani, Harvard Medical School, United States
Martha Bulyk, Harvard Medical School, United States

**Abstract:**
Gene regulation depends on the binding of Transcription Factors (TFs) to specific DNA sequences. Recently, several genome-wide methods (e.g., ChIP-seq and DNase-seq) have identified many putative gene regulatory regions, but the exact underlying TF:DNA interactions remain elusive. Protein binding microarrays (PBM) can profile these interactions in vitro and furnish position weight matrices (PWMs), which represent the standard model for TF binding. PWMs intuitively describe TF:DNA specificity, but cannot capture some of its features (e.g., alternative binding modes or interdependency between binding residues), unlike more complex, kmer-based models. Previous performance comparisons in motif discovery favored PWM models, but were assessed on few ChIP-seq datasets (6 in DREAM5).

Here, we develop a motif discovery approach that bypasses the limitations of PWMs through the joint use of 8mers and a non-redundant description of the TF specificity landscape, and test it extensively on ENCODE datasets. We first analyze 640 non-redundant metazoan TFs from the two main PBM repositories (UniPROBE and CisBP) and from unpublished experiments (all together 25 TF families). By developing an iterative dynamical clustering method over the entire dataset, we were able to associate the main TF (sub)families to small (<200) subsets of unique 8mers, which collectively form a “glossary”.

In contrast to previous kmer and PWM models for single TFs, our glossary exploits larger, currently available PBM repositories to globally evaluate the TF specificity repertoire. Moreover, it recapitulates and expands our knowledge of the heterogeneous specificity between, and within, TF families. The glossary outperforms PWMs in identifying enriched motifs for cofactors mediating indirect TF:DNA binding in a large collection of putative regulatory regions (240 ChIP-seq datasets for site-specific TFs from ENCODE). Gene expression profiles (10 RNA-seq datasets from ENCODE) support this enrichment, showing significant up-regulation of the cofactors identified by the glossary. In accessible chromatin regions (15 DNase-seq datasets from ENCODE) the glossary reveals cell-type specific patterns of motif enrichment, corroborated by the up-regulation of the associated TFs. Besides recapitulating well-known master TFs (e.g., SOX2/OCT4 in stem cells, and GATA1 in erythrocytes), the glossary highlights new candidates (e.g., TEAD2/FOXH1 in stem cells or FOS/ELF3 in colon cells).

In conclusion, for motif discovery our glossary improves both interpretability for the field’s “neophytes”, since it explicitly describes a manageable number (114) of modules of specifically recognized 8mers and their associated TFs, and
performance, since it minimizes PBM experimental variability and identifies distinguishing 8mers that are unique to the associated TFs.

---

**Presentation Title:** The regulation of distal enhancers and silencers from the Cebpa locus during hematopoiesis, inferred using a transcriptional model

**Authors:**
Manu, University of North Dakota, United States  
John Reinitz, University of Chicago, United States  
Eric Bertolino, University of Chicago, United States

**Abstract:**
CCAAT/enhancer binding protein, alpha (C/EBPα) plays an instructive role in the macrophage-neutrophil cell-fate decision and its expression is necessary for neutrophil development. How Cebpa itself is regulated in the myeloid lineage is not known. We decoded the cis-regulatory logic of Cebpa, and two other myeloid transcription factors, Egr1 and Egr2, using a combined experimental-computational approach. We designed reporter genes capable of detecting both distal enhancers and silencers. We measured the activity of 46 putative cis-regulatory modules (CRMs) lying in 40-80 kb regions surrounding the three genes in cells representing myeloid progenitors, and derived early macrophages or neutrophils. This analysis revealed both enhancers and silencers in the vicinity of Cebpa, although the latter outnumber the former. The abundance of silencers is consistent with the downregulation of endogenous Cebpa in the differentiated cell types and in non-myeloid hematopoietic lineages. In a novel application of sequence-based transcriptional models, we determined the—as yet unknown—regulatory roles of 15 potential transcriptional regulators of Cebpa CRMs. This was achieved by testing 32,768 alternative transcriptional models, representing all possible role combinations, against CRM activity data. Analysis of the models allowed us to infer the cis-regulatory logic—how the expression patterns of transcription factors (TFs) in the cell types control activity patterns—for most of the CRMs. In particular, we found that silencer-mediated repression of Cebpa is effected mainly by TFs expressed in non-myeloid lineages. These results imply that cross-lineage antagonism is likely to be mediated by long-distance repression and, since silencers are yet to be characterized for most genes, is more widespread than previously believed. More generally, our results demonstrate that de novo cis-regulatory dissection is feasible on a large scale with the aid of transcriptional modeling.

---

**Presentation Title:** Distinct specificities of the androgen and glucocorticoid receptors revealed using feature-based recognition model analysis of SELEX data

**Authors:**
Liyang Zhang, University of Iowa, United States  
Gabriella Martini, Columbia University, United States
Abstract:

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.

Presentation Title: MiSL: a method for mining synthetic lethal partners of recurrent cancer mutations uncovers novel mutation-specific therapeutic targets

Authors:

Subarna Sinha, Stanford University, United States
Daniel Thomas, Stanford University, United States
Yang Gao, University of California at Berkeley, United States
Steven Chan, Stanford University, United States
Diede Brunen, Netherlands Cancer Institute, Netherlands
Rene Bernards, Netherlands Cancer Institute, Netherlands
Ravindra Majeti, Stanford University, United States
David L. Dill, Stanford University, United States

Abstract:
Synthetic lethality, in which a single gene defect leads to dependency on a second gene that is otherwise not essential, is an attractive paradigm to identify targeted therapies for cancer-specific mutations. Current methods to detect synthetic lethal (SL) partners for somatic mutations rely on large-scale shRNA screens in cell-lines or use human orthologs of yeast SL interactions, both of which are not necessarily representative of primary tumors and have incomplete coverage.

We have developed MiSL, a novel Boolean implication-based algorithm that utilizes large pan-cancer patient datasets (mutation, copy number and gene expression) to identify SL partners for cancer mutations. The underlying assumption of our approach is that, across multiple cancers, SL partners of a mutation will be amplified more frequently or deleted less frequently, with concordant changes in expression, in primary tumor samples harboring the mutation. Pan-cancer analysis discovers robust biological relationships that are likely to be independent of cancer subtype and increases statistical power.

First, we sought to validate MiSL using existing knowledge and large-scale shRNA data. Consistent with prior knowledge, MiSL candidates for BRCA1 mutation (mut) in breast cancer were enriched for DNA repair genes (p=.0.006). We also found: (1) significant overlap (p=0.002) between leukemia IDH1mut MiSL candidates and essential genes in IDH1mut cells determined by a DECIPHER shRNA screen we performed in doxycycline-inducible IDH1(R132) THP-1 cells, and (2) for multiple mutations in colorectal cancer, MiSL candidates were enriched (p<0.05) with genes that were selectively essential in mutated colorectal cell-lines in Achilles data.

Secondly, we experimentally confirmed novel SL partners that are druggable in acute myeloid leukemia (AML) and breast cancer. MiSL predicted a novel SL interaction in AML between IDH1mut and ACACA, the rate-limiting enzyme that controls lipid biosynthesis. Consistent with our prediction, inhibition of ACACA with shRNA or a small molecule inhibitor TOFA prevented cell proliferation in IDH1mut (but not wildtype) AML cell-lines and primary blasts. MiSL also predicted that AKT1 is a SL partner of PIK3CAmut in breast cancer which we experimentally confirmed using 8 breast cancer lines. All four PIK3CAmut (but not wildtype) breast cancers were sensitive to AKT1 inhibition in viability and colony assays.

In conclusion, MiSL is a scalable computational solution that finds novel SL interactions. Using primary patient data allows it to capture in vivo tumor evolution, revealing SL interactions missed by existing methods. It can be widely applicable and can greatly accelerate novel target discovery for precision medicine in cancer.

**Presentation Title:** Network Maximal Correlation to Infer Nonlinear Gene Modules in Cancer

**Authors:**
Soheil Feizi, MIT, United States
Ali Makhdoumi, MIT, United States
Ken Duffy, Hamilton Institute, Maynooth University, Ireland
Identifying nonlinear relationships in large datasets is a daunting task particularly when the form of the nonlinearity is unknown. Here, we introduce Network Maximal Correlation (NMC) as a fundamental measure to capture nonlinear associations in networks without the knowledge of underlying nonlinearity shapes. NMC infers, possibly nonlinear, transformations of variables with zero means and unit variances by maximizing total nonlinear correlation over the underlying network. For the case of having two variables, NMC is equivalent to the standard Maximal Correlation metric. We characterize a solution of the NMC optimization using geometric properties of Hilbert spaces for both discrete and jointly Gaussian variables. For discrete random variables, we show that the NMC optimization is an instance of the Maximum Correlation Problem and provide necessary conditions for its global optimal solution. Moreover, we propose an efficient algorithm based on Alternating Conditional Expectation (ACE) which converges to a local NMC optimum. For this algorithm, we provide guidelines for choosing appropriate starting points to jump out of local maximizers. We also propose a distributed algorithm to compute a 1-epsilon approximation of the NMC value for large and dense graphs using graph partitioning. For jointly Gaussian variables, under some conditions, we show that the NMC optimization can be simplified to a Max-Cut problem, where we provide conditions under which an NMC solution can be computed exactly. Under some general conditions, we show that NMC can infer the underlying graphical model for functions of latent jointly Gaussian variables. These functions are unknown, bijective, and can be nonlinear. This result broadens the family of continuous distributions whose graphical models can be characterized efficiently. We illustrate the robustness of NMC in real world applications by showing its continuity with respect to small perturbations of joint distributions. We also show that sample NMC (NMC computed using empirical distributions) converges exponentially fast to the true NMC value. Finally, we apply NMC to different cancer datasets including breast, kidney and liver cancers, and show that NMC infers gene modules that are significantly associated with survival times of individuals while they are not detected using linear association measures.
**Abstract:**

Promoters are DNA sequences that play an essential role in controlling gene expression. While recent whole cancer genome analyses have identified numerous hotspots of somatic mutations within promoters, most do not appear to be functional as they do not perturb gene expression. As such, positive selection does not adequately explain the frequency of promoter mutations in cancer genomes. Our analysis indicates that increased mutation density at gene promoters is in fact linked to transcriptional activity and differential DNA repair.

By analyzing 1,163 cancer genomes, we found evidence for increased local density of somatic point mutations within the DNase I hypersensitive centre of gene promoters across 14 cancer types. Mutated promoters were strongly associated with transcriptional activity, with mutation density highest within transcription factor binding sites. By analysing genome-wide maps of nucleotide excision repair (NER), we found that NER is impaired within the DNase I hypersensitive centre of active gene promoters, inversely mirroring the increase in somatic mutation density.

Taken together, our analysis has uncovered the presence of a previously unknown mechanism linking transcription initiation and DNA repair, thereby implicating localised differential DNA repair as the underlying cause for the somatic mutation hotspots observed at gene promoters of cancer genomes.

---

**Presentation Title:** Lobular Scale Spatio-temporal Modeling of Calcium Signal Propagation in the Liver

**Authors:**

Aalap Verma, University of Delaware/Thomas Jefferson University, United States
Anil Antony, Thomas Jefferson University, United States
Joannes Hoek, Thomas Jefferson University, United States
Babatunde Ogunnaike, University of Delaware, United States
Rajanikanth Vadigepalli, Thomas Jefferson University, United States

**Abstract:**

Regulation of Ca2+ signals in hepatocytes is crucial to normal liver function. Free Ca2+ in the intracellular domain regulates a wide range of hepatocyte functions such as metabolism, bile secretion, proliferation and apoptosis. Experimental observations suggest that, in response to extracellular stimuli, the oscillatory intracellular Ca2+ release manifests as a Ca2+ wave propagating across the liver lobule. Extracellular stimulants such as hormones elicit a cascade involving phospholipase C (PLC) cleavage, IP3 generation, and subsequent Ca2+ release from intracellular stores. Gap-junction mediated transfer of IP3 molecules between adjacent hepatocytes induces Ca2+ oscillations in neighboring hepatocytes which spreads through the tissue. This cell-cell interaction results in communication of extracellular stimulus levels between hepatocytes, leading to a coordinated...
response at the tissue level for efficient liver function. Disruption of such intercellular interaction may lead to pathological conditions. We stimulated intracellular Ca2+ release by treating isolated perfused mouse livers with vasopressin and captured Ca2+ wave transduction in liver lobules using confocal intact tissue imaging techniques. We segmented the images to identify Ca2+ time series profiles in individual hepatocytes, and analyzed the resulting single cell Ca2+ intensity patterns. Our analysis revealed that the direction of signal propagation between adjacent hepatocytes lying along a sinusoid changed frequently during the course of measurement suggesting that a clear wave-like propagation of Ca2+ signals does not occur at the microscopic scale. We built a computational model based on our experimental data to further understand this non-canonical propagation of Ca2+ signals. Based on model simulations and analysis we predicted that differential expression of the hormone receptor in different regions of liver lobules explains the wave-like propagation of Ca2+ signals. Reversal of Ca2+ signal between adjacent hepatocytes at the microscopic scale could be explained by interference of waves starting at different loci and passing through the same hepatocyte. We further compared Ca2+ oscillations between control and chronic alcohol adapted mouse hepatocytes. Based on model simulations, we predict increased PLC activity and reduced intercellular communication between alcohol adapted hepatocytes. Our model thus provides a framework for generating hypotheses on lobular scale changes in intracellular Ca2+ oscillations parameters and intercellular Ca2+ signal propagation characteristics that potentially underlie normal and diseased states.

Presentation Title: Insights in binding inhibition mechanism of benzofuran salicylic acid inhibitors against Protein Tyrosine phosphatases (Ptp1B) using molecular modeling and simulation approaches

Authors:

Saleem Iqbal, University of Madras, India
Manish Kesherwani, UNIVERSITY OF MADRAS, India
Gunasekaran K, University, India

Abstract:

The protein tyrosine phosphatase B (mPTPB) is a virulent phosphatase secreted by Mycobacterium tuberculosis, which is essential for the survival and persistence of the bacterium in the host and proposed as a promising drug target against tuberculosis infection. Here insilico docking studies of series of benzofuran salicylic acid derivatives as inhibitors for PTPB were performed. Based on interactions with proximal domain, three compounds (4f,4i and 4g) were selected as best. Molecular Dynamics Simulation study was carried out for these PTP1B docked complexes for better understanding of binding mode, binding free energy and stability. MGBSA calculation showed that binding free energy in terms of enthalpy of top three PTP1B complexes (compound 4f, 4i and 4g) were well correlated with with experimental IC50 and molecular determinants in terms of binding were also identified. Based on binding free energy (G), compound 4g (6-hydroxy-2-phenyl-3-{2-[3-
(trifluoromethyl) phenyl[ethynyl]-1-benzofuran-5-carboxylic acid) which has more favorable binding free energy compared to other compounds can be chosen as a potent inhibitor against tuberculosis.

**Presentation Title:** High-throughput allele-specific expression across 250 environmental conditions

**Authors:**

Gregory Moyerbrailean, Wayne State University, United States  
Chris Harvey, Wayne State University, United States  
Omar Davis, Wayne State University, United States  
Adnan Alazizi, Wayne State University, United States  
Donovan Watza, Wayne State University, United States  
Yoram Sorokin, Wayne State University, United States  
Karoline Pruder, Wayne State University, United States  
Nancy Hauff, Wayne State University, United States  
Xiaoquan Wen, University of Michigan, United States  
Roger Pique-Regi, Wayne State University, United States  
Francesca Luca, Wayne State University, United States

**Abstract:**

Adaptations to local environments have played major roles in shaping allele frequency distributions in human populations. Yet, a mismatch between genotype and environment may be responsible for higher disease risk. 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 have taken 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 and 3 individuals per cell type). Across 56 cellular environments (cell type/treatment with large changes in gene expression) we discovered 6073 instances of ASE (FDR<10%), corresponding to 4310 unique genes. We found that in an individual sample, on average, 0.5% of genes with heterozygous SNPs are ASE genes. We observe that the majority of ASE is consistent across conditions ("shared" ASE), confirming previous conditional eQTL analyses. Overall, we find 248 loci with evidence for GxE interaction (conditional ASE), 120 with control-only ASE and 128 with treatment-only ASE genes. We used a multinomial generalized linear model with elastic net regularization (glmnet) to assess which factors influence the likelihood of conditional ASE. This model allows us to control for factors that may influence ASE and potential confounders (e.g., gene expression, cell type, treatment). Cell type seems to be an important factor for shared ASE: Melanocytes show a 30% increase in the probably of ASE, while LCLs show an 18% reduction. When we focus on treatment-only ASE, there are significant differences across treatments but these are largely explained by the changes in gene
expression. For genes that are differentially expressed, each 2-fold increase in gene expression response corresponds to a 2.22-fold increase in the probability of treatment-only ASE. Finally, integrating our results with data from 18 traits from GWAS meta-analysis revealed enrichments for genes differentially expressed in specific treatments. For example, variants associated with Crohn’s disease are enriched in genes that respond to aspirin in PBMCs and HUVECs, thus identifying candidate genes for aspirin aggravating effects on Crohn's symptoms.

**Presentation Title:** Integrating single cell transcriptional signatures and cell-cell interaction network models to study multiscale control of liver regeneration dynamics

**Authors:**
Daniel Cook, University of Delaware, United States  
Sirisha Achanta, Thomas Jefferson University, United States  
Aalap Verma, University of Delaware, United States  
Jan Hoek, Thomas Jefferson University, United States  
Babatunde Ogunnaike, University of Delaware, United States  
Rajanikanth Vadigepalli, Thomas Jefferson University, United States

**Abstract:**
Recent technological advances have allowed for transcriptional profiling of single cells from tissues. Challenges remain, however, in integrating single-cell scale transcriptional regulation with tissue-scale physiological response. We address these challenges by integrating experimentally identified transcriptional cell states into a cell-cell interaction model with physiological transitions between states. We applied this technique to the framework of liver resection and regeneration, a treatment for hepatocellular carcinoma and as a technique to enable live liver transplant. Our approach gives unique insights into the multiscale control of liver regeneration dynamics. Furthermore, the availability of clinical data and pathology samples from regenerating and non-regenerating cases present an opportunity to employ modeling approaches that integrate data from different functional scales to better understand the mechanisms underlying clinical resection responses.

We used laser capture microdissection followed by high-throughput qPCR to acquire a new experimental data set on gene regulation in hundreds of single and pooled hepatocytes and hepatic stellate cells during liver regeneration in healthy and chronically diseased livers. We analyzed the high-dimensional, single-cell scale gene regulation data to identify distinct cell populations within each cell type. We then extrapolated from the single-cell scale to cell phenotype distributions at the whole-tissue scale. Our approach characterized the transcriptional state of individual hepatocytes and hepatic stellate cells in the in vivo regenerating liver at a level of detail not previously achievable. Characterization of transcriptional profiles revealed cell-type specific, functionally distinct states for hepatocytes as well as stellate cells. Surprisingly, we found the two liver cell types distributed among all
identified functional states irrespective of the experimental condition (baseline or regenerating, healthy or chronic disease). However, relative proportions of the cells distributed among the identified functional states changed based on regeneration stage as well as healthy vs. chronic disease, suggesting a cell population balance signature of disease.

We developed a computational model that accounts for the multi-scale nature of liver regeneration by integrating physiological-scale interactions, transcriptional states of hepatocytes and liver non-parenchymal cells, and molecular signaling networks. We explored a range of model dynamics to identify parameter sets that account for experimentally observed regeneration profiles in health and disease. Model simulations suggest that aberrant activation dynamics of hepatocytes and hepatic stellate cells are key factors suppressing liver regeneration in multiple disease contexts.

Our integrated experimental and computational analysis points towards new avenues for therapeutic intervention based on renormalizing the cellular functional state balances to improve surgical outcomes.

**Presentation Title:** How do closely related transcription factors recognize distinct genomic targets?

**Authors:**

Ning Shen, Duke University, United States
Joshua Schipper, Duke University, United States
Tristan Bepler, Duke University, United States
John Horton, Duke University, United States
Raluca Gordan, Duke University, United States

**Abstract:**

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.

Here, we show that 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 nine paralogous factors from three protein families: bHLH, E2F, and ETS. 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 exception were TFs E2F1 and E2F3, which showed identical specificities. Interestingly, these factors play similar regulatory roles and can partially substitute for each other in the cell.

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, 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 TF preferences for genomic sites represent an important mechanism by which closely related factors achieve their regulatory specificity.

**Presentation Title:** Three-dimensional analysis of regulatory features reveals functional enhancer-associated loops

**Authors:**

Yao Wang, University of Texas Health Science Center at San Antonio, United States
Junbai Wang, Oslo University Hospital – Norwegian Radium Hospital, Norway
Yufan Zhou, University of Texas Health Science Center at San Antonio, United States
Malaina Gaddis, University of Southern California, United States
Rohit Jadhav, University of Texas Health Science Center at San Antonio, United States
Xun Lan, Stanford University, United States
Tim Huang, University of Texas Health Science Center at San Antonio, United States
Shili Lin, The Ohio State University, United States
Peggy Farnham, University of Southern California, United States
Seth Frietze, University of Vermont, United States
Victor Jin, University of Texas Health Science Center at San Antonio, United States

**Abstract:**

Several critical gaps remain in our knowledge of the relationship of chromatin structure to gene regulation. These include 1) classifying different types of chromatin interactions (including promoter-enhancer contacts), 2) determining the relationships between classes of chromatin interactions and the epigenomic state, 3) deciphering the functional relevance of chromatin interactions, and 4) determining whether genes associated with different chromatin interaction classes are involved in disease.

To gain insight into the relationship between chromatin structure and gene
expression, we conducted chromatin conformation analysis using PANC1 pancreatic cancer and MCF7 breast cancer cells. For PANC1, we carried out Tethered Chromatin Capture (TCC) on two biological replicates, and compared correlation between replicates to validate the data quality. For MCF7, we used both TCC and in situ Hi-C protocols on replicates and also performed correlation analysis. To analyze the 3D conformation in these two cancer cells, we first detected topologically associated domains (TADs) in each chromosome, then applied a novel Hi-C analysis algorithm and identified hundreds of thousands of Interacting Loci Pairs (ILPs) in each of the two cell types. We classified ILPs according to location with respect to gene structure, gene expression, different histone modifications, DNase hypersensitivity, and RNA polymerase II and CTCF binding. Interestingly, we found that a majority of ILPs are within a particular TAD, and only 5% of the ILPs are involved in promoter regions, with even fewer promoter-enhancer loops. To further explore the potential mechanism behind 3D conformation and gene expression, we conducted TCC on PANC1 treated with drug ICG001 known as a CBP inhibitor, and C646 known as a CBP/p300 competitor. We examined the changes of TADs and ILPs in the drug-treated PANC1, and the impact of pharmacological inhibition of histone acetylation on genes having promoter-enhancer loops in PANC1 cells. We find that genes associated with promoter-enhancer loops have cell-type-specific functional annotations. We further demonstrated that genes with promoter-enhancer loops altered expression in response to drug treatment in PANC1, suggesting that the chromatin loops we identified are functional. Taken together, our study provides insights into the interdependence of three-dimensional chromatin looping and gene expression mediated by enhancer-promoter interactions.

Presentation Title: Microbes are STICKY – Large-scale Inference and Topological Analysis of Microbial Interaction Networks

Authors:
Christian L. Müller, Simons Foundation, United States
Zachary D. Kurtz, New York University School of Medicine, United States
Richard Bonneau, Simons Foundation, New York, United States

Abstract:
In recent years, 16S-rRNA and other environmental sequencing measurements of microbial communities have revealed phylogeny and the abundances of microbial populations across diverse ecosystems. While changes in microbial community composition are demonstrably associated with certain environmental conditions (from metabolic and immunological health in mammals to ecological stability in soils and oceans), the identification of underlying mechanisms and interaction patterns among microbes requires new statistical tools. Key challenges for 16S rRNA data analysis are that (i) the data are typically compositional, i.e., counts of Operational Taxonomic Units (OTUSs) are normalized to the total number of counts in the sample due to limits in sequencing capacity, and (ii) the data are underdetermined, i.e., the number of detected OTUs typically exceeds the number of available samples by an order of magnitude. Thus, traditional statistical measures
for the detection of OTU-OTU relationships such as Pearson correlation can lead to spurious results. Our recently introduced SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) framework (http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226) addresses both of these issues by combining compositional data transformations with algorithms for sparse neighborhood and inverse covariance selection that also work in the underdetermined regime.

Using the SPIEC-EASI framework we learned sparse interaction networks from a large collection of publicly available data sets across many ecological environments, ranging from gut and freshwater habitats to urban environments. Using concepts from graph theory, statistics, and theoretical ecology, we analyzed the topology of the learned interaction graphs. We identify a number of general network architecture principles that reproduce within and across habitats. We observe that all learned networks share the small-world property, common to many biological networks, and that the networks follow the May-Wigner criterion for ecological stability. In addition, we observe that the STICKY (or Chung-Lu) model is an excellent statistical network model that describes microbial networks accurately across multiple habitats. We believe that these findings shed a new light on the organization principles of microbial communities and present a promising step toward the unification of ecological theory and data-driven systems microbiology.

Presentation Title: Identifying condition specific transcription factor binding with ATAC-seq

Authors:
Roger Pique-Regi, Wayne State University, United States
Donovan Watza, Wayne State University, United States
Molly Estill, Wayne State University, United States
Sophia Chaudhry, Wayne State University, United States
Francesca Luca, Wayne State University, United States

Abstract:
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, NRSE, NRF-1, and NFκB. 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 5236 regions that have significantly changed chromatin accessibility (FDR < 10%) in response to both copper and selenium. 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 binding factor activity to study cellular regulatory response to the environment.

**Presentation Title:** Visualizing three-dimensional organization and long-range interactions of the mammalian genome with the 3D Genome Browser.

**Authors:**

Yanli Wang, Penn State College of Medicine, United States
Gal Yaroslavsky, Penn State University, United States
Tyler Derr, Penn State University, United States
Lijun Zhang, Penn State University College of Medicine, United States
Feng Yue, Penn State University, United States

**Abstract:**

The mammalian genome subscribes to a complex spatial organization that defines the three-dimensional interactions of potentially distant functional elements that control the regulation of transcription and replication. Recent advancements in sequencing and analysis techniques – specifically Hi-C, or high-throughput chromosome conformation capture– have revealed these interactions genome-wide at unprecedented resolutions. Unfortunately, navigating the Hi-C data remains a daunting feat for many biologists, as its $O(n^2)$ complexity for the already big data intrinsic to mammalian genomes poses a challenge to its analysis (time and memory usage), storage and transfer.

Our laboratory has developed and extended the functionality of the 3D Genome Browser (http://3dgenome.org), a web-based, intuitive and accessible browser of Hi-C data. The browser adopts a genome feature-searchable approach and contextualizes the region with established University of California Santa Cruz (UCSC) Genomic Browser while retaining its flexibility to customize genome tracks and load personalized UCSC sessions. Furthermore, the browser allows the visualization of the Hi-C contact matrix as both rotated heatmap, as well as virtual 4C (linear plot detailing the number of interactions between a single genomic site of
interest -bait or anchor locus- with other loci), which is further contextualized with DNase I Hypersensitive Site (DHS)-linkage, ChIA-PET and capture Hi-C tracks. The virtual 4C display is particularly valuable in the identification of cis-regulatory elements and their gene targets.

To ameliorate the \(O(n^2)\) complexity of Hi-C data, we pioneered a new file format to represent contact matrices, called BUTLR (Binary Upper Triangular matrix). This format not only compresses the high-resolution Hi-C data by \(\sim 20\) , it permits random access such that interaction values among any regions could be queried fast and directly. Therefore, the 3D Genome Browser supports the visualization of Hi-C datasets in BUTLR format hosted on any HTTP accessible server without requiring the transfer of entire files, similar to bigWig/bigBed files for the UCSC Genome Browser.

Currently, our browser boasts \(\sim 3000\) views every month and receives frequent improvements in its user interface. Several laboratories have successfully utilized the browser to visualize user-generated Hi-C interactions. With our gene-centric, binary-file browser approach, the 3D Genome Browser improves the accessibility in browsing Hi-C data. With the visualization of the spatial organization and long-range interactions of particular genomic regions along with their genetic and epigenetic context, our browser seeks to drive hypothesis-generation about and enrich the understanding of the intrinsic link between genomic organization and genetic regulation.

**Presentation Title:** Network model of normal gene expression predicts gene perturbation fold changes

**Authors:**

Sudhir Varma, HiThru Analytics, United States

**Abstract:**

Gene expression exhibits a network effect whereby perturbations of some genes (in the form of siRNA knockdown or drug treatment in vitro or mutations, changes in methylation or aneuploidy in disease) influence the expression of downstream genes. Thus the downstream genes are simply responding to the dysregulation of the root-cause genes and are themselves not the source of the perturbation. If the expression (or fold change) of a gene is explained by the expression (or change in expression) of other genes, it becomes less likely to be the source of the perturbation. Conversely, genes with a large positive or negative difference between the predicted and actual expression show evidence of being the main drivers in the experimental or disease condition.

Using a set of 4277 normal samples from various organs compiled from public datasets, we have built a network where the expression of each gene (target gene) is modeled as a linear combination of a small number of other genes (source genes). A single model for each gene was fitted to samples from all organs. For each organ we tested the fit of the model using the correlation between the predicted and actual
values. We used the network model to predict the fold changes using a set of 658 siRNA knockdown samples.

The network model predicts the expression of a median of 42% (20%-71%) of variable-expression genes (log expression range>0.5) in all of the organs with a correlation>0.80. On the siRNA knockdown samples, the model predicted the resulting fold changes of all genes with a median correlation of 0.31 across the samples (0.13-0.68).

We demonstrate that a single linear regression model (per gene) is sufficient to predict the expression of most genes for multiple organs. The relationships between source genes and target gene defines a network which is capable of quantitatively predicting the downstream effects of a perturbation. Conversely, the difference between the predicted and true expression in a disease sample points to possible root causes of the disease.

We have implemented a web tool for exploring the network predictions on a variety of disease samples (http://www.explainbio.com).

**Presentation Title:** Dysregulated transcription factor networks and clusters in breast cancer subtypes, identified by coexpression and cistromic data integration

**Authors:**
Qian Zhu, Princeton University, United States
Vessela Kristensen, University of Oslo, Norway
Olga Troyanskaya, Princeton University, United States

**Abstract:**
Transcriptional regulation mediated by transcription factor binding provides a key mechanism for the dysregulation of genes, in that they influence gene expression via proximal element (promoter) and distal element (enhancer). Previously, several attempts have been made to construct a transcription factor (TF) network to dissect the genetic programs in breast cancer. However, these approaches have not been disease subtype specific or have been based on promoter sequence evidence alone.

In this work, we sought to delineate subtype-specific TF networks in breast cancer by integrating transcriptomic data with experimental cistromic data (Chip-seq) from relevant MCF7, T47D cell lines, an epithelial (A549) and an embryonic stem cell cell-line (H1-HESC) deposited in ENCODE. Unlike previous approaches, our approach importantly is targeted by a coexpressed gene-set, which allows subtype signatures to be defined and subtype network to be constructed.

We leveraged over 170 expression datasets to first identify subtype-specific coexpression signatures, next prioritized 160 Chip-seq experiments from cell lines to identify TFs which may govern coexpressed genes, and finally utilized these Chip-seq data to identify protein-DNA regulatory interactions between TFs in each subtype. We witnessed extensive clustering of in vitro TF motifs, and confirmed
expected motifs enriched by coexpressed genes’ enhancer regions, which provide a sequence basis for protein-DNA interactions.

Importantly, the networks of TF illuminated the luminal, basal, and stem cell lineages involved in breast cancer subtypes. Transcriptional hierarchies were constructed and delineated the early molecular factors leading to the development of lineages. When we investigated the TFs’ deregulation potential, we identified 9 groups of TFs based on their shared profiles of deregulations across subtypes (i.e., SNPs, somatic mutations, DNA methylation and copy number aberrations), suggesting that these 9 TF clusters are consistently targeted by genetic and epigenetic alterations. Not only did the TF clusters illustrate deregulation differences among subtypes, but their patterns of deregulation were also highly reproducible in held-out datasets. The TF clusters included basal stem cell initiators SOX9, EN1, luminal progenitor GATA3, and luminal markers ESR1, BHLHE40, among others. Overall, we illustrate the informative value of a combined analysis involving coexpression and epigenetic, genetic alteration data in the study of subtype-specific transcriptional dysregulations.

Presentation Title: Multi-omics learning and optimal experimental design for microbial organisms

Authors:
Minseung Kim, University of California at Davis, United States
Navneet Rai, University of California at Davis, United States
Violeta Zorraquino, University of California at Davis, United States
Xiaokang Wang, University of California at Davis, United States
Ilias Tagkopoulos, University of California at Davis, United States

Abstract:
Accurate prediction of cellular and molecular state in novel environments is one of the grand challenges in modern biology. Despite the availability of omics profiles, it remains unclear how and at what degree their integration can train a predictive model, or how current datasets can guide which new conditions should be investigated. To address these challenges, we developed a framework of omics data integration, predictive modeling and optimal experimental design. We constructed a comprehensive Escherichia coli compendium specifically structured for efficient machine learning. The compendium integrates 4,389 profiles in multiple layers ranging from transcriptome, proteome, metabolome, fluxome, and phenome with in-depth characterization of profiling conditions by 612 features of strain genotypes, chemical composition of medium used, stresses exposed, and genetic perturbations. The compendium was undergone in multi-step procedure of preprocessing to correct for gene-level noises, batch-effects, and platform-biases. We used this resource to train a multi-scale statistical model that integrates four omics layers to predict expression levels of 4096 transcripts, 1001 proteins, 2382 metabolic fluxes and 356 metabolite concentrations as well growth dynamics. To guide future experimentation, we developed a methodology to identify experiments that optimally sample the experimental space and simultaneously decrease the...
uncertainty of the model. The proposed methodology takes into account two types of uncertainty in genome-scale prediction; prediction interval from bootstrapped RNNs and entropy estimated by Gaussian process. The genetic and environmental ontology that was reconstructed from the omics data is substantially different and complementary to the ontologies that are traditionally derived by using genetic and chemical information. Predictive performance (PCC) over novel conditions range from 0.54 to 0.87 for the various omics layers and their integration outperformed any single layer for growth rate prediction. Growth prediction of our model was particularly effective for novel wild type conditions (PCC=0.76). The efficacy of optimal experimental design was evaluated over 15 rounds of transcriptional profiling in novel conditions that resulted in a substantial improvement of the performance over alternative methods. The performance of genome-wide expression prediction for the condition space close to optimal conditions newly profiled was substantially improved after refinement (PCC from 0.41 to 0.61) and gradual decrease in uncertainty of the prediction model over the course of 15 rounds was significant than alternatives (P < 0.005). This work provides an integrative framework of omics-driven predictive modeling and experimentation that can be broadly applied to guide biological discovery.

Presentation Title: SplashRNA: accurate prediction of potent shRNAs with a sequential classification strategy

Authors:

Raphael Pelossof, MSKCC, United States
Lauren Fairchild, MSKCC, United States
Christina Widmer, Activision, United States
Chun-Hao Huang, MSKCC, United States
Vipin Sreedharan, MSKCC, United States
Qing Xiang, MSKCC, United States
Ralph Garippa, MSKCC, United States
Gunnar Ratsch, MSKCC, United States
Scott Lowe, MSKCC, United States
Christof Fellmann, UC Berkeley, United States
Christina Leslie, MSKCC, United States

Abstract:

Short hairpin RNAs (shRNAs) are a ubiquitously used experimental tool for knocking down genes, including now in CRISPR-shRNA combination models. Therefore, prediction of potent shRNA sequences is a critical computational problem. We have developed a novel machine learning approach to accurately predict potent shRNAs by training on massively parallel sensor assay data, pooled screens of RNAi reporters that identify shRNAs with potent knock-down. Our method, SplashRNA, achieves more than double the accuracy of DSIR, a standard algorithm in the field, in a blind sensor screen that tested the potency of 250,000 shRNAs, and 20% improvement in accuracy against DSIR and Sherwood, a commercial algorithm, on blind cell viability screens for 2,000 shRNAs. shRNA pooled screens are designed as sequential enrichment screens of RNAi
reporters, where each reporter consists of an shRNA (under control of doxycycline) and a sensor targeted by the shRNA (in the 3'UTR of a reporter gene) expressed together in a single cell, so that on-dox fluorescence enables sorting for potent shRNAs. The first published sensor assay screen started with an unbiased input library of shRNAs, and analysis of shRNAs that passed the screen established potency rules. Due to the high costs of such screens, later screens are designed to test biased sets of shRNA that are likely to be potent given the learned potency rules. We have generated such a screen for the training of our algorithms, testing approximately 20,000 shRNAs that were split into 64 shRNAs per 300 genes. Our final classifier sequentially aggregates classifiers in a soft cascade scheme. In this learning methodology each classifier has the option either to pass the shRNA to the next classifier or to stop and assign a weaker potency score to the shRNA. This approach is beneficial when each screen only tests a subset of the shRNAs, and therefore we would design the first classifier to screen unbiased sequences, and the second consecutive classifier to classify more potent shRNA sequences that it receives from the first. Each one of the classifiers of the cascade is trained with an SVM and uses a spectrum kernel and a weighted-degree kernel as feature representations for shRNA sequences.

Validation of SplashRNA on PTEN in 3T3 cells by western blot, shows that we are able achieve perfect classification ranking, implying that SplashRNA may have fully learned sequence specific potency rules of cellular shRNA processing. An implementation of our algorithm is publicly available at splashrna.mskcc.org.

Presentation Title: From phenotypic to molecular synergy: A transcriptional study of the dynamics of drug combinations based on single drug responses

Authors:

Mehmet Eren Ahsen, IBM Research and Icahn School of Medicine at Mount Sinai, United States
Jennifer E. L. Diaz, IBM Research and Icahn School of Medicine at Mount Sinai, United States
Xintong Chen, Icahn School of Medicine at Mount Sinai, United States
Bojan Losic, Icahn School of Medicine at Mount Sinai, United States
Gustavo Stolovitzky, IBM Research, United States

Abstract:

Drug combination therapies have proven to be good strategies in cancer treatment in that they may elicit less adverse effects than single drugs while overcoming the resistance to individual drugs that cancer cells tend to develop. Since screening all possible drug pairs is impractical, accurate methods for predicting synergistic drug combinations are needed. Attempts to predict the effect of drug combinations based on the transcriptional response of cells to single drugs have succeeded only partially because not enough data exists on how those transcriptional responses combine in the cellular environment. In this work we study the mechanisms whereby transcriptional responses combine to give rise to synergistic or additive responses
to combined therapies. Specifically, we used RNAseq to study the transcriptional response over time (0, 3, 6, 9, 12, and 24 h) and for three drugs (A, B and C) and their combinations (AB, AC and BC) in MCF-7 breast cancer cells. Cell viability measurements show that one of the combinations (AB) is strongly synergistic, whereas the other two (AC and BC) are additive. The number of differentially expressed genes for the synergistic combination AB was at least one order of magnitude larger than the number of the differentially expressed genes resulting from each of the individual drugs A or B, and increased over time. For the additive combinations the number of differential expressed genes was about the same as for the single drugs, and was dominated by one of the drugs (C). To explain the massive transcriptional response of the synergistic combination, we extended the concept of additivity from the phenotypic to the transcriptional level. We found that most of the genes differentially expressed in AB but not in A nor B are non-additive. Using this information in the MCF-7 specific gene regulatory network we looked for transcriptional cascades that could explain the transcriptional program in AB based on that in A and B. We found that the majority of transcription factors that get activated at a given time point remain active at later time points. We studied how the activation of transcriptional regulators in A and B activate synergistic genes explaining much of the transcriptional response to AB. These analyses can pave the way for the design of algorithms to predict the response of cells to drug combinations based on RNAseq data from single drugs.

Presentation Title: Super-enhancers delineate disease-associated regulatory nodes in T cells.

Authors:

Golnaz Vahedi, University of Pennsylvania, United States  
Vittorio Sartorelli, NIH, United States  
John O'Shea, NIH, United States

Abstract:

Enhancers regulate spatiotemporal gene expression and impart cell-specific transcriptional outputs that drive cell identity. Super-enhancers (SEs), also known as stretch-enhancers, are a subset of enhancers especially important for genes associated with cell identity and genetic risk of disease. CD4(+) T cells are critical for host defence and autoimmunity. Here we analysed maps of mouse T-cell SEs as a non-biased means of identifying key regulatory nodes involved in cell specification. We found that cytokines and cytokine receptors were the dominant class of genes exhibiting SE architecture in T cells. Nonetheless, the locus encoding Bach2, a key negative regulator of effector differentiation, emerged as the most prominent T-cell SE, revealing a network in which SE-associated genes critical for T-cell biology are repressed by BACH2. Disease-associated single-nucleotide polymorphisms for immune-mediated disorders, including rheumatoid arthritis, were highly enriched for T-cell SEs versus typical enhancers or SEs in other cell lineages. Intriguingly, treatment of T cells with the Janus kinase (JAK) inhibitor tofacitinib disproportionately altered the expression of rheumatoid arthritis risk genes with SE structures. Together, these results indicate that genes with SE architecture in T cells

42
encompass a variety of cytokines and cytokine receptors but are controlled by a 'guardian' transcription factor, itself endowed with an SE. Thus, enumeration of SEs allows the unbiased determination of key regulatory nodes in T cells, which are preferentially modulated by pharmacological intervention.


**Presentation Title:** Comparison of Methods to Predict Impact of Regulatory Variants

**Authors:**
Felix Yu, Johns Hopkins University, United States
Dongwon Lee, Johns Hopkins University, United States
Michael Beer, Johns Hopkins University, United States

**Abstract:**
The vast majority of sequence variants associated with common human disease are intergenic, enriched in open chromatin regions, and likely regulatory. To identify functional variants within GWAS associated LD blocks, we have developed a sequence-based model based on our gapped k-mer SVM (gkm-SVM) (Lee et al., *Nature Genetics*, 2015; Ghandi et al., *PLOS Comp Biol* 2014). This approach uses cell-type specific epigenetic data to train a gkm-SVM whose scoring function encodes the relative regulatory importance of individual sequence features in the disease relevant cell-type. The change in sequence feature scores induced by a regulatory variant determines its predicted impact, a score we call deltaSVM. We have shown that deltaSVM is roughly 10x more accurate at predicting dsQTLs than other methods (Kircher et al., *Nat Gen* 2014; Ritchie et al., *Nat Meth* 2014) and our previous kmer-SVM (Lee et al., *Gen Res* 2011). We have also used deltaSVM to predict the expression change in massively parallel reporter assays, which shows good agreement with high throughput datasets in mouse liver (Patwardhan et al., *Nat Biotech* 2012), K562 cells, and HepG2 cells (Kheradpour et al., *Gen Res* 2013). Here, we compare the accuracy of deltaSVM to other computational approaches, including PWMs, other kmer-based approaches, and deep neural networks (Zhou and Troyanskaya, *Nat Meth* 2015).