Workshop during the Pacific Symposium of Biocomputing, Jan 3-7, 2019: Reading between the genes: interpreting non-coding DNA in high-throughput

Joanne Berghout† , Yves A. Lussier†, Francesca Vitali†
Center for Biomedical Informatics and Biostatistics, Dept. of Medicine, University of Arizona, 1230 Cherry Ave, Tucson, AZ 85719, USA
Emails: jberghout@email.arizona.edu, yves@email.arizona.edu, francescavitali@email.arizona.edu

Martha L. Bulyk†
Division of Genetics, Dept. of Medicine & Dept. of Pathology
Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
Email: mlbulyk@genetics.med.harvard.edu

Maricel G. Kann†
Dept. of Biological Sciences, 1000 Hilltop Circle
University of Maryland, Baltimore County, Baltimore, MD 02115, USA
Email: mkann@umbc.edu

Jason H. Moore†
Institute for Biomedical Informatics, 3400 Civic Center Blvd, Bldg 421
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: jhmoore@upenn.edu

Identifying functional elements and predicting mechanistic insight from non-coding DNA and non-coding variation remains a challenge. Advances in genome-scale, high-throughput technology, however, have brought these answers closer within reach than ever, though there is still a need for new computational approaches to analysis and integration. This workshop aims to explore these resources and new computational methods applied to regulatory elements, chromatin interactions, non-protein-coding genes, and other non-coding DNA.

Keywords: non-coding; bioinformatics; epigenetics; transcription factor; systems biology

1. Introduction

GWAS studies have frequently identified variation in non-coding regions as associated with a variety of complex traits and diseases. However, it remains difficult to assign and validate the functional consequences of these variants or to suggest a mechanism by which they actually influence an outcome. These challenges become even more difficult to address at scale, or in high-throughput, when a clear biological candidate molecule and hypothesis cannot be readily tested through experimentation. The most commonly considered mechanisms for altered function are

† all workshop co-chairs contributed equally, and are listed alphabetically

altered regulatory activity of an effector molecule (including eQTL, transcription factor binding, enhancers/insulators, epigenetic marks, etc.), alternative splicing, changes to chromosome conformation, or altered biology of non-coding RNA genes. 2017 saw the public release and extension of multiple major data resources exploring biological and biochemical functions of non-coding regions at genome scale, including across multiple tissue and cell contexts (e.g., GTEx (1), ENCODE(2)). Emerging genetic engineering and molecular editing technologies have also accelerated, with use of CRISPR expanding beyond hypothesis-driven gene knockouts into targeting of non-coding elements, unbiased tiling assays, and genome-wide screening applications (3).

Advances in computational methods are required to analyze these new data types, identify patterns, integrate across biological scales, and derive biologically and/or clinically useful insights during primary analyses, by secondary exploration of data in publicly-available resources, and/or by integrating across data sets and using new models. In addition to targeted biomedical questions, there also arises a unique opportunity for computational biologists to identify network and systems properties of non-coding DNA, linking evidence from these assays, genetics, and evolutionary biology with other datasets(4).

2. Speakers and abstracts

From genetics to therapeutics: uncovering and manipulating the circuitry of non-coding disease variants

Manolis Kellis, *Professor, MIT Computer Science and Artificial Intelligence Lab*
Institute Member, Broad Institute of MIT and Harvard

Perhaps the greatest surprise of human genome-wide association studies (GWAS) is that 90% of disease-associated regions do not affect proteins directly, but instead lie in non-coding regions with putative gene-regulatory roles. This has increased the urgency of understanding the non-coding genome, as a key component of understanding human disease. To address this challenge, we generated maps of genomic control elements across 127 primary human tissues and cell types, and tissue-specific regulatory networks linking these elements to their target genes and their regulators. We have used these maps and circuits to understand how human genetic variation contributes to disease and cancer, providing an unbiased view of disease genetics and sometimes re-shaping our understanding of common disorders. For example, we find evidence that genetic variants contributing to Alzheimer's disease act primarily through immune processes, rather than neuronal processes. We also find that the strongest genetic association with obesity acts via a master switch controlling energy storage vs. energy dissipation in our adipocytes, rather than through the control of appetite in the brain. We also combine genetic information with regulatory annotations and epigenomic variation across patients and healthy controls to discover new disease genes and regions with roles in Alzheimer's disease, heart disease, prostate cancer, and to understand their pleiotropic effects by integration with electronic health records. Lastly, we develop systematic technologies for systematically manipulating these circuits by high-throughput reporter assays, genome editing, and gene targeting in human cells and in mice, demonstrating tissue-autonomous therapeutic avenues in Alzheimer's disease, obesity, and cancer. These results provide a roadmap for translating genetic findings into mechanistic insights and ultimately therapeutic treatments for complex disease.

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

Michael M. Hoffman, *Principal investigator, Princess Margaret Cancer Centre & Assistant Professor, Departments of Medical Biophysics and Computer Science, University of Toronto*

Introduction: 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. Recent chemical probing and sequencing methods provide the opportunity to assess a variety of DNA modifications. Modification-sensitive TFs provide a mechanism by which widespread changes in DNA methylation and hydroxymethylation can dramatically shift active gene expression.

Methods: 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), h (5hmC) and other symbols—permitting computational representations of modified sequence. We also enhanced parts of the MEME Suite and RSAT to handle custom alphabets, expanding the position weight matrix (PWM) formulation of TF binding affinity and enabling clustering of modified PWMs.

Results: We created an expanded-alphabet sequence using whole-genome maps of 5mC and 5hmC in mouse naive T cells and human K562 cells. Using this sequence and ChIP-seq data from ENCODE and others, we identified modification-sensitive *cis*-regulatory modules. We reproduced known binding preferences, including the preference of ZFP57 for methylated motifs and the preference of c-Myc for unmethylated motifs. We have made several novel predictions, and are validating them using ChIP-BS-seq and CUT&RUN. (5)

Quantifying the impact of non-coding mutations on transcriptional regulation

Raluca Gordân, *Assistant Professor, Biostatistics and Bioinformatics, Duke University*

Most disease-associated genetic variants occur in non-coding regions where they can alter gene regulation, rather than gene sequence. Focusing on putative regulatory variants that can affect transcription factor (TF) binding to the genome, I will present new methods for quantifying the change in TF binding due to binding site variants, as well as the statistical significance of the predicted change. Briefly, using as input high-throughput in vitro data for hundreds of mammalian TFs, we developed regression models of TF-DNA binding that implicitly take into account the quality of the training data. Thus, in the case of low-quality data that leads to a large variance in the estimated model parameters, only large changes in TF binding will reach statistical significance; in contrast, high-quality training data sets allow us to identify even subtle changes in TF binding due to genetic variants. To assess the quality of our predictions, we leverage high-throughput enhancer assay data where all possible single base-pair mutations in specific regulatory regions have been tested directly for their effect on gene expression. We find that our TF binding models can explain about ~50% of the variation in gene expression. We are currently using the TF binding change predictions in collaborative GWAS studies to prioritize non-coding variants for further computational and experimental analyses.

CRISPR-SURF: Discovering regulatory elements by deconvolution of CRISPR tiling screen data

Luca Pinello, *Principal Investigator and Assistant Professor, Massachusetts General Hospital & Harvard Medical School*

Tiling screens using CRISPR-Cas technologies provide a powerful approach to map regulatory elements to phenotypes of interest, but computational methods that effectively model these experimental approaches for different CRISPR technologies are not readily available. Here we present CRISPR-SURF, a deconvolution framework to identify functional regulatory regions in the genome from data generated by CRISPR-Cas nuclease, CRISPR interference (CRISPRi), or CRISPR activation (CRISPRa) tiling screens. We validated CRISPR-SURF on previously published and new data, identifying both experimentally validated and new potential regulatory elements. With CRISPR tiling screens now being increasingly used to elucidate the regulatory architecture of the non-coding genome, CRISPR-SURF provides a generalizable and accessible solution for the discovery of regulatory elements. (6)

Delineation and annotation of the human regulatory landscape across 400+ cell types and states

Wouter Meuleman, *Investigator, Altius Institute for Biomedical Sciences*

The human genome encodes vast numbers of non-coding elements whose combined actuation patterns reflect regulatory processes across cellular states and conditions. Despite large-scale technology development for interrogating non-coding parts of the genome, pragmatic annotated high-resolution maps of regulatory regions and their inter-cell type dynamics have been lacking. To address this issue, we applied a joint experimental and computational approach, integrating 733 deeply sequenced DNase I hypersensitivity assays spanning more than 400 distinct human cell types and states. These data enable a systematic and principled approach to studying regulatory architecture and dynamics on a global scale. We define a common coordinate system for regulatory DNA marked by DNase I hypersensitive sites, encompassing over 3 million elements defined and annotated with unprecedented resolution and detail. Through systematic analysis of the dynamics of these regulatory regions across cell types and states, we derive a collection of Regulatory Components, providing a novel multi-component annotation of the human regulome. Using admixtures of multiple components, we show that it is possible to decompose biological features of cell and tissue samples and define the extent to which individual regulatory elements contribute to broader cellular regulatory programs. These previously unappreciated features allow us to characterize the functional properties of genes and pathways. For instance, based solely on their regulatory landscape, we readily identify genes coding for lineage-specifying factors. Moreover, we associate specific regulatory structures with distinct binding site motifs, as well as with gene expression patterns across cell types. Moreover, our Regulatory Components provide a fundamentally new framework for understanding how disease-associated variation maps to genome function, not otherwise appreciated. Taken together, through integrative analysis across hundreds of cell types and states, we provide a novel multi-component annotation of the human regulatory landscape. Our Regulatory Components are predictive for functional and regulatory characteristics

of genes, pathways and genetic variants. As such, they open up new horizons on the architecture of human genome regulation and function.

Genetically explainable non-coding RNA expression by SNVs

Lana Garmire, *Associate Professor, Molecular Biosciences and Bioengineering, U of Michigan*

Long intergenic non-coding RNAs have been shown to play important roles in cancer. However, because lincRNAs are a relatively new class of RNAs compared to protein-coding mRNAs, the mutational landscape of lincRNAs and the impact of mutations on lincRNA expression are not extensively studied. We comprehensively characterize expressed somatic nucleotide variants within lincRNAs using 6118 primary tumor samples from 12 cancer RNA-Seq datasets in TCGA. Due to uncertainty of somatic or germline mutations from analyzing un-paired RNA-Seq data alone, we first build a highly accurate machine-learning model (AUC 0.987) to discriminate somatic variants from germline variants within lincRNAs, using a subset of samples that have both exome-seq and RNA-seq data. We use this model to predict highly confident eSNVs (expressed SNVs) and found that they are especially enriched in chr2p11.2, chr14q32.33, chr22q11.22 and chr3q29 regions. To understand the effect of molecular features on lincRNA somatic eSNVs, we build another model (AUC 0.72) and identify molecular features that are strongly associated with lincRNA mutations, including copy number variation, conservation, substitution type and histone marker features. Finally, we prioritize the lincRNAs by their eSNV influence, and propose a short list of genetically affected lincRNAs to be validated by experimental studies.

Interpreting genetic variants by gene regulatory network

Yong Wang, *Professor, Institute of Applied Mathematics, Academy of Mathematics and Systems Science & National Ctr for Mathematics and Interdisciplinary Science, Chinese Academy of Science*

Interpreting genetic variance (including SNP and structural variants) is the key to precision health. Most of these variants will affect disease risk, response to drugs or other traits such as height in a tissue or condition-specific way. How can we figure out which variants affect the function and regulation of genes in which condition? We propose to use gene regulatory network to integrating omics data and interpret genetic variants. Particularly, we will discuss the models and algorithms to organize, analyze, model, and integrate the genetic variant, DNA accessibility data, transcriptional data, and functional genomic regions together. We believe that the integrative paradigm on chromatin and expression levels will eventually help us to understand the information flow in cell and will influence research directions across many fields.

References

1. A. Battle, CD Brown, BE Engelhardt, SB Montgomery, *Nature.* **550**(7675), 204-13 (2017).
2. CA Sloan, ET Chan, JM Davidson, et al., *Nucleic Acids Research.* **44**(D1), D726-32 (2016).
3. JB Wright, NE Sanjaya, *Trends in Genetics.* **32**(9), 526-9 (2016)
4. M Amorim, S Salta, R Henrique, C Jeronimo *J Transl Med.* **14**, 265 (2016).
5. C Viner, J Johnson, N Walker, et al., *bioRxiv (preprint).* doi.org/10.1101/043794 (2016)
6. JY Hsu, CP Fulco, MA Cole, et al., *bioRxiv (preprint).* doi.org/10.1101/345850, (2018)