Reading Between the Genes: Computational Models to Discover Function from Noncoding DNA

Yves A. Lussier[†], Joanne Berghout, Francesca Vitali, Kenneth S. Ramos
Center for Biomedical Informatics and Biostatistics,
The Center for Applied Genetic and Genomic Medicine,
BIO5 Institute, UA Cancer Center, and Dept of Medicine; University of Arizona
1657 E Helen St, Tucson, AZ 85719, USA
Emails: yves@email.arizona.edu, jberghout@email.arizona.edu, francescavitali@email.arizona.edu, ksramos@email.arizona.edu

Maricel Kann
Dept of Biological Sciences; University of Maryland, Baltimore County,
1000 Hilltop Circle Baltimore, MD 21250 United States
Email: mkann@umbc.edu

Jason H. Moore
Department of Biostatistics, Epidemiology, and Informatics; University of Pennsylvania,
3700 Hamilton Walk, Philadelphia, PA, 19104, USA
Email: jhmoore@exchange.upenn.edu

Noncoding DNA - once called "junk" has revealed itself to be full of function. Technology development has allowed researchers to gather genome-scale data pointing towards complex regulatory regions, expression and function of noncoding RNA genes, and conserved elements. Variation in these regions has been tied to variation in biological function and human disease. This PSB session tackles the problem of handling, analyzing and interpreting the data relating to variation in and interactions between noncoding regions through computational biology. We feature an invited speaker to how variation in transcription factor coding sequences impacts on sequence preference, along with submitted papers that span graph based methods, integrative analyses, machine learning, and dimension reduction to explore questions of basic biology, cancer, diabetes, and clinical relevance.

Keywords: non-coding DNA, intergenic, LncRNA, microRNA, sncRNA, miRNA, piRNA, LINEs, LINE1, repetitive elements

[†] Work partially supported by This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, HL132532, CA023074, 1UG3OD023171, 1R01AG053589-01A1, 1S10RR029030)

1. Introduction

The majority of the human genome is comprised of noncoding DNA. Estimating the percentage of noncoding DNA that comprises functional elements is somewhat controversial, depending heavily on definitions used, specific assays, and cell type or other biological contexts explored. However, we can all agree that it's not zero. Unlike protein-coding sequences where a biologically active function can be relatively readily assigned through standard experimental techniques, most noncoding sequences largely remain as functional black boxes. Even for those noncoding sequence variants confidently linked to variation in a biological trait (i.e. via GWAS), the mechanism of precisely how they exert an effect on the phenotype often remains unclear. Identifying the potential and most relevant relationships or functions in the absence of an a priori biological hypothesis requires the intersection of big data, computing, and creativity.

Noncoding DNA includes RNA genes (miRNA, lncRNA, piRNA), regulatory regions (transcription factor binding sites, eQTL-associated SNPs, promoters, enhancers, insulators), epigenetic mark associated regions, repetitive elements (LINEs, transposons, Alu elements, telomeres), pseudogenes, and structural elements among others. With this diversity of potential function, largely incomplete annotation, and substantial degree of sequence variation between individuals, defining even common canonical motifs at the resting state is challenging. Recognizing these needs, several international efforts such as ENCODE, GTEX, NIH Epigenomics Roadmap, and the International Epigenome Consortium have been established. These researchers use high throughput technologies and systematic approaches to start developing regulatory maps and begin learning about "genomic grammar", or, the rules that govern meaning as a cell or protein "reads" through DNA. With these international projects along with many other academic laboratories generating vast quantities of genome-scale data from sequence, expression, ChIP-Seq, CHIA-PET, ATAC-Seq and other new technologies [1], there also emerges a need to develop methods for appropriate data handling, integration, and analysis.

Thus, there arises a unique opportunity for computational biologists to identify network and systems properties of noncoding DNA, linking evidence from biochemical assays, genetics, and evolutionary biology with other datasets. These can predict downstream biology, functional convergence, and impacted mechanisms that ultimately lead to disease. In addition to novel basic science insights, understanding the mechanisms perturbed by variation in noncoding DNA can implicate new pathophysiological mediators and unveil new therapeutic targets. The computational tools that have been created by researchers to handle these data have been complex, innovative, and some great work has been done by data-generating scientists and research parasites alike, leading to unexpected discovery and new research questions. As a result, we believe that a PSB session devoted to the topic of non-coding DNA would be timely, interesting and yield some excellent paper submissions.

2. Sessions summary

2.1. *Invited Talk*

We have the privilege to have the participation of a guest speaker, Dr. Martha Bulyk from the Division of Genetics at Harvard, also an Associate Member of the Broad Institute. Dr Bulyk has made a name for herself with work at interface between DNA sequence and protein binding specificity to explore cis regulatory motifs. She invented a high throughput method for detecting transcription factor binding preferences via a novel protein binding microarray, and developed integrated computational approaches to interpret these data, with an eye towards TF binding site clustering, combinatorial co-occurrences, and cross-species conservation.

2.1.1. *Her talk is entitled "Survey of coding variation in human transcription factors reveals prevalent DNA binding changes"*

> Invited talk abstract: Sequencing of exomes and genomes has revealed abundant genetic variation affecting the coding sequences of human transcription factors (TFs), but the consequences of such variation remain largely unexplored. We developed a computational, structure-based approach to evaluate TF coding variants for their impact on DNA-binding activity and used universal protein binding microarrays (PBM) to assay sequence-specific DNA-binding activity across reference and variant alleles found in individuals of diverse ancestries and families with Mendelian diseases. We found variants that affect DNA-binding affinity or specificity and identified thousands of rare alleles likely to alter the DNA-binding activity of human sequence-specific TFs. Altered sequence preferences correlated with changes in genomic TF occupancy (ChIP-Seq peaks) and gene expression of the associated target genes. Our results suggest that most individuals have unique repertoires of TF DNA-binding activities, which may contribute to phenotypic variation.

2.2. *Papers*

2.2.1. *Evaluating relationships between pseudogenes and genes: from pseudogene evolution to their functional potentials*

Johnson et al. developed a novel approach integrating graph analysis, sequence alignment and functional analysis to classify pseudogene-gene relationships. They identified ~15,000 pseudogenes from the Human GENCODE release 24 using cufflinks gffread. Every pseudogene was then aligned to the best consensus sequence of ~3200 gene families using pairwise ClustalW. Finally, they mined the resulting network of pseudogene-gene edges jointly with gene-Gene Ontology terms to impute the former function of poorly characterized pseudogenes. These observations extend on previous work [2], and may lead to new insight on families of pseudogenes and to characterize the potential function of new pseudogenes in cancer.

2.2.2. *Convergent downstream candidate mechanisms of independent intergenic polymorphisms between co-classified diseases implicate epistasis among noncoding elements*

Han et al. characterize convergent downstream candidate mechanisms of distinct intergenic SNPs across distinct diseases within the same clinical classification by conducting an integrative analysis of four networks: disease class to disease annotations, GWAS disease-SNP associations, eQTL SNP-mRNA associations, and Gene Ontology (GO) gene-mechanisms annotations. At FDR≤5%, they prioritize 167 intergenic SNPs, 14 classes, 230 mRNAs, and 134 GO mechanisms. They expand on their previous study [3] and observe that co-classified SNP were more likely to be prioritized in the same mechanisms as compared to those of distinct classes (odds ratio ~3.8). The SNPs prioritized to the same GO mechanisms were also enriched in regions bound to the same/interacting transcription factors and/or interacting in long-range chromatin interactions suggestive of epistasis (odds ratio ~ 2,500). Such network of genetic epistasis associated to candidate biological mechanisms has the potential to reposition medications that target proteins within downstream mechanisms of intergenic SNPs associated to disease risks, the latter generally considered undruggable before this study.

2.2.3. *Pan-cancer analysis of expressed single nucleotide variants in long intergenic non-coding RNAs*

Ching et al. analyzed 6118 primary tumor samples of 12 distinct cancers from TCGA and detected 94,700 somatic mutations in lincRNAs, and 15.5 million germline variants in lincRNAs. They further built machine-learning models to impute sensitive regions to mutations and variants of lincRNAs, suggesting that some nucleotide positions within lincRNA are more likely to gain somatic mutations than other positions. Non-coding but not trivial, the authors extend our understanding of polymorphisms and mutations in lincRNAs, increasing their contributions to "non-coding but not trivial" lincRNAs [4].

2.2.4. *Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 diabetes GWAS*

Manduchi et al. utilized evidence for enhancer-promoter interactions from functional genomics data in order to build biological filters to narrow down the search space for two-way Single Nucleotide Polymorphism (SNP) interactions in Type 2 Diabetes (T2D) Genome Wide Association Studies (GWAS). They confirmed the validity of the method by identifying a statistically significant pairs of type 2 diabetes SNPs in statistical epistasis [5] and replicated in an independent datasets. Their framework, that accounts for epistasis, expands substantially on GWAS analyses as the current paradigm consists on linear additive analyses of GWAS.

References

1. Ward LD, Kellis M: Interpreting noncoding genetic variation in complex traits and human disease**.** *Nat Biotechnol* 2012, **30:**1095-1106.

2. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JM, Li Y, Menzies A, Mudie L, Ramakrishna M, et al: Processed pseudogenes acquired somatically during cancer development. *Nat Commun* 2014, **5:**3644.
3. Li H, Achour I, Bastarache L, Berghout J, Gardeux V, Li J, Lee Y, Pesce L, Yang X, Ramos KS, et al: Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions**.** *NPJ Genom Med* 2016, **1**.
4. Ching T, Masaki J, Weirather J, Garmire LX: Non-coding yet non-trivial: a review on the computational genomics of lincRNAs**.** *BioData Min* 2015, **8:**44.
5. Moore JH, Williams SM: Epistasis and its implications for personal genetics**.** *Am J Hum Genet* 2009, **85:**309-320.