Large-scale Annotation and Analysis of Allelic Variants and eQTLs for Non-Coding RNAs

Studies have begun to generate large-scale amounts of genotype data in addition to matched functional genomic data (such as RNA-Seq data). These studies have had success in identifying eQTLs or allele-specific variants within a particular study. All of these studies have focused almost exclusively on protein-coding gene annotations. We propose a study focused primarily on non-coding RNAs. We will collect all available data (both public and restricted access) from the Framingham Heart Study (FHS)\cite{25791433}, GTEx\cite{25954001, ENCODE\cite{22955616} and gEUVADIS\cite{24037378} consortia, and we will then uniformly process all the available RNA-Seq data in order to identify both allelic variants as well eQTLs for non-coding RNAs. We will then study the roles that these variant play in the regulation of non-coding RNAs, and will compare our findings with existing knowledge on protein-coding genes.

Our specific aims are as follows:

1. *Construction of an integrated database of allelic variants and eQTLs for non-coding RNAs from large-scale consortia data.* We plan on using our own pipelines to uniformly process all available small and long RNA-Seq datasets for individuals with matching genotype data from the FHS, GTEx, ENCODE and gEUVADIS consortia in order to construct a database of both allele-specific SNVs and eQTLS from all these datasets.

2. *Integrative analysis of allelic variants and eQTLs for non-coding RNAs.* We plan on using this database of allelic variants and eQTLs from Aim 1 to investigate the regulatory roles of SNVs with respect to non-coding RNAs (and in particular miRNAs). Using the genomic locations of SNVs that either exhibit allele-specific expression (ASE) or manifest as eQTLs, we will study their role and compare them with the corresponding regulation of protein-coding genes. We will also study the network behavior of co-expressed allele-specific and eQTL-target genes, and we will compare these with published GWAS results.

Achieving the aims above requires expertise in several areas. Dr. Rozowsky has played a key leadership role in the laboratory of Professor Mark Gerstein in the Department of Molecular Biophysics and Biochemistry at Yale University for the last 14 years. Dr. Rozowsky has extensive experience in analyzing functional genomic data. He is familiar with how the data within these consortia are generated and analyzed, and he is very knowledgeable with the techniques mentioned in this proposal. Dr. Rozowsky had led the lab's efforts in a number of consortia focused on the data generation and processing, as well as integrated analysis of functional genomic data. In particular, Dr. Rozowsky has played a leadership role in all three phases of the ENCODE consortia and the modENCODE consortia, in addition to leading the bioinformatics analysis for the Extracellular RNA Communication Consortia (ERCC), in which he co-chairs the Analysis and RNA-Seq working groups for the consortium. As part of Dr. Rozowsky's role in the lab, he directs the research of a number of undergraduate students, PhD students, postdoctoral fellows as well as associate research scientists.

Innovation

In this proposal, we will perform (for the first time) a large-scale integrative analysis of allele-speciifc variants, as well as eQTLs, for non-coding RNAs using available transcriptome and genotype data from the FHS, GTEx, ENCODE and gEUVARDIS consortia. Integrating these uniformly annotated allele-specific variants and eQTLs for non-coding RNAs, we plan on studying the role of regulation of non-coding RNAs, in contrast to that for protein-coding genes.

Aim 1: Construction of an integrated database of allelic variants and eQTLs for non-coding RNAs from large-scale consortia data

1.1 Preliminary Results

1.1.1 **Collecting and Processing RNA-Seq data**

We have extensive expertise with transcriptome analysis and in developing a wide range of customized tools, as well as building standardized pipelines for analysis and uniform processing of both long and short RNA-seq data. These tools have been evaluated and implemented in several major consortia, including long RNA-seq analysis tools (in mod/ENCODE\cite{22955616, 25164755}) and short RNA-seq pipelines for the analysis of small extracellular RNA-seq data (in the Extracellular RNA Communication Consortium (ERCC) \cite{26320938}).

For general RNA-Seq analysis, we developed an efficient in-house data processing workflow for long RNA-seq data that includes data organization, format conversion, and quality assessment. *RSEQtools* (http://rseqtools.gersteinlab.org/), is a computational package that enables expression quantification of annotated RNAs, as well as identification of splice sites and gene models\cite{21134889}. Comparisons between RNA-Seq samples, and to other genome-wide data, are facilitated in part by our Aggregation and Correlation Toolbox (*ACT*), which is a general purpose tool for comparing genome signal tracks \cite{21349863}. An important challenge in RNA-Seq analysis is detecting unannotated transcription that may be hard to distinguish from noise. Our Database of Annotated Regions with Tools (*DART*) package contains tools for identifying unannotated genomic regions that are enriched for transcription, as well as a framework for storing and querying this information\cite{17567993}. To further investigate newly-discovered transcriptionally active regions, we developed *incRNA*\cite{21177971}, a method that predicts novel ncRNAs using known ncRNAs of various biotypes (which then serves as a gold standard training set). We have also developed specific tools to identify types of transcripts that are difficult to detect using standard analysis pipelines. We recently developed *FusionSeq*, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations \cite{20964841}.

We recently developed the *extra-cellular RNA processing toolkit*, *exceRpt* \cite{manuscript in preparation} (http://github.gersteinlab.org/exceRpt/), a set of tools and a pipeline designed for comprehensive analysis of small RNA-seq datasets: read preprocessing, filtering and alignment, biotype abundance estimation, visualization and quality assessment. It is specifically designed to handle technical issues that are often characteristic of small RNA-seq samples, such as those obtained from extra-cellular preparations. In addition, the software is perfectly capable of processing data from more standard cellular preparations and long RNA-seq data. The *exceRpt* pipeline is used for uniform processing of hundreds of RNA-seq datasets submitted to the exRNA Atlas (http://exrna-atlas.org/) repository.

1.1.2 **Building Personal Genomes and Imputation**

Current human genome annotations are based on the reference genome and, as such, do not provide an accurate representation for the large genomic diversity of the human population. We have developed approaches and tools to incorporate personal variation data into the reference genome sequence producing the individual's personal diploid genome sequence and matching annotation. For personal genome construction, we have developed a computational tool, vcf2diploid \cite{21811232}. This tool integrates an individual's genomic variation data (SNVs, indels, and SVs) into the reference genome. Phase information of heterozygous variants is also incorporated, producing maternal and paternal haplotypes. Chain files generated by the program can be used to account for coordinate offsets between the individual's parental haplotypes and the original reference genomic sequence. The versatility to convert between reference and personal genome coordinates allows mapping of annotated genomic regions (e. g., gene or peak coordinates for RNA-seq and ChIP-seq, respectively) between the genomes using available tools, such as the UCSC LiftOver tool {Rhead:2010if}.

We have previously constructed the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878 (also known as GM12878). We have made this assembly available as a resource at alleleseq.gersteinlab.org and have been updating it as new versions of the human reference genome, genomic annotations, and NA12878 genetic variation data are released. Furthermore, the availability of a computational tool enables the construction of personal genomes in a high-throughput fashion, as demonstrated in a recent publication\cite{27089393} in which we built 382 personal genomes using the variant call sets from the 1000 Genomes Project.

1.1.3 **Uniform calling of ASE & eQTLs and constructing a database of annotations**

We have extensive experience in developing large databases of QTLs and allelic sites. AlleleSeq \cite{21811232} is a tool developed specifically for the detection of allelic sites, including those associated with gene expression and transcription factor binding using RNA-seq and ChIP-seq datasets. Using AlleleSeq, we have spearheaded and published allele-specific analyses as part of our work in several major consortia, including ENCODE and the 1000 Genomes Project \cite{22955616, 26432245}. Overall, we found a substantial number of genomic elements associated with ASB and ASE [24]. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression [6] (Fig. 1). Furthermore, our AlleleSeq tool (which is available online) provides lists of detected allelic variants, and the personal diploid genome and transcriptome of NA12878 [36]. We continually update AlleleSeq, and the resource is being used by the scientific community, as evidenced by citations for our tool in the published literature \cite{22955620,22955619}

We have developed techniques to link eQTLs to phenotype data by quantifying the amount of information necessary to identify an individual in order to demonstrate how the information security of the given dataset may be compromised[42]. We used eQTLs from the expression datasets generated by the GEUVADIS project\cite{24037378} and the genotype dataset from the 1000 Genomes Project. Considering this as a test case, we developed statistical formalisms for quantifying the leakage of information that enables the identification of specific individuals in genotype and phenotype datasets with use of QTLs.

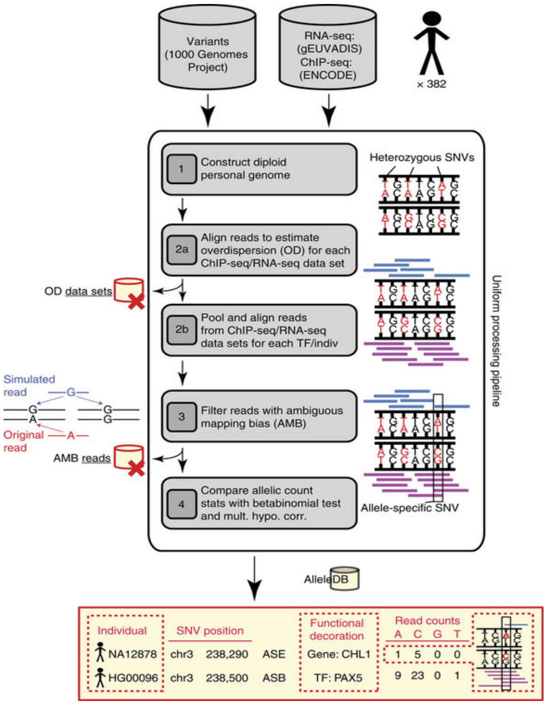
Variants (1000 Genomes Project)
RNA-seq: (gEUVADIS) ChIP-seq: (ENCODE)
× 382
1
Construct diploid personal genome
Heterozygous SNVs
2a
Align reads to estimate overdispersion (OD) for each ChIP-seq/RNA-seq data set
OD data sets
2b
Pool and align reads from ChIP-seq/RNA-seq data sets for each TF/indiv
Simulated read
G
A
G
G
Original read
A
3
Filter reads with ambiguous mapping bias (AMB)
AMB reads
4
Compare allelic count stats with betabinomial test and mult. hypo. corr.
Allele-specific SNV
Uniform processing pipeline
AlleleDB
Individual
NA12878
HG00096
SNV position
chr3 238,290 ASE
chr3 238,500 ASB
Functional decoration
Gene: CHL1
TF: PAX5
Read counts
A C G T
1 5 0 0
9 23 0 1

Figure 1. Workflow for uniform processing of functional genomics data from hundreds of individuals, assessment of allele-specific expression and binding events and construction of AlleleDB

Recently, we have further developed AlleleSeq and applied the new version to 1,139 RNA-seq and ChIP-seq datasets for 382 cell lines in the 1000 Genomes Project. For each cell line, we harmonized and aggregated multiple RNA-seq and ChIP-seq datasets separately, and then uniformly reprocessed each using the updated AlleleSeq. This allowed us to annotate the 1000 Genomes Project SNP catalog with allelic information. We constructed a database, AlleleDB, to house all the results. The database can be queried for specific genomic regions and visualized as a track in the UCSC browser [13], as well as other visualizers such as the Integrated Genomics Viewer [14]. It may also be downloaded as flat files for downstream analyses for users that are more advanced in bioinformatics training. We continue to maintain and update AlleleSeq as a publicly available resource. It has been utilized considerably by the scientific community, as indicated by the number of citations and publications using our data and tool.

1.2 Proposed Research

1.2.1 **Collecting and Processing RNA-Seq data**

We will collect RNA-seq data from the Framingham Heart Study (FHS)\cite{25791433}, GTEx\cite{25954001, ENCODE and gEUVADIS. For each sample, we will uniformly process the RNA-seq data for both long RNAs and small RNAs. Specifically, we will apply the exceRpt pipeline to process small RNA-seq data in order to quantify the expression levels for all annotated small RNAs. We will apply the *ENCODE long RNA-Seq pipeline* to uniformly to quantify the expression levels of all long non-coding RNAs. We will use the latest version of the GENCODE annotation \cite{22955987} as the default set of annotations, which includes the latest miRNA annotation from miRbase\cite{XXX}.

1.2.2 **Building Personal Genomes and Imputation**

Using personal genomes allows us to account for differences due to the impact of variants (SNPs, indels and SVs) between individuals, as well as between haplotypes of the same individual. This improves mappability of functional assay reads, and it alleviates reference bias in the identification and analyses of allele-specific events.

HOW

We will leverage our current collection of 382 personal genomes of the 1000 Genome Project individuals with matching functional datasets from the gEUVADIS\cite{24037378} and ENCODE\cite{22955616} consortia. We will extend these sets of personal genomes by integrating additional available variant data from FHS, GTEx, ENCODE, gEUVADIS, as well as other sources, in order to construct personal genomes for each genotyped individual, which will be used for further analyses. For individuals that are only partially genotyped using SNP arrays, we will employ imputation-based methods to predict genome-wide sets of variants for these individuals. We will assess and develop approaches to alleviate the impact of heterogeneity between different variant call approaches used in different studies (i.e., array-based imputed vs. WGS-based calls, depth of sequencing, sequencing technologies, etc.)

1.2.3 **Uniform calling of ASE & eQTLs and construction of an annotations database**

To identify SNVs associated with allele specific expression (ASE) for both short and long non-coding RNAs (using our AlleleSeq and AlleleDB pipelines \cite{21811232,27089393}), we will uniformly process the available long and small RNA-Seq reads from FHS, GTEx, ENCODE, gEUVADIS data, in conjunction with the matching personal genomes constructed for these individuals.

Similarly, we will apply a uniform pipeline to call tissue-specific and cross-tissue eQTLs across all datasets. We will first use PEER factors to normalize the mapping counts generated by exceRpt, which accounts for hidden and known covariates, when available \cite{22343431,20463871}. eQTL calls will be generated using Matrix eQTL \cite{22492648}. We will call *cis* and *trans* eQTLs separately controlling for factors such as poor mappability, cross mapping and repetitive elements as in \cite{bioRxiv:074419}. Finally, we will apply fine mapping methods, such as CAVIAR, to propose causal variants by using the correlation structure of the genetic variants induced by linkage disequilibrium \cite{25104515}.

To construct a database for the annotations, we will enlarge the framework of AlleleDB\cite{27089393} to include ASE variants as well as eQTLs called using these datasets. The database can be queried for specific genomic regions and visualized as a track in the UCSC browser \cite{12045153} or the Integrated Genomics Viewer \cite{21221095}. It may also be downloaded as flat files for more detailed downstream analyses. Datasets that are restricted access via dbGap will not be included for public download and will only be privately accessible.

Aim 2: Integrative analysis of allelic variants and eQTLs for non-coding RNAs

2.1 Preliminary Results

2.1.1 **Interpreting mechanistic roles of ASEs and eQTLs for non-coding RNAs and associated protein-coding genes**

Extracellular RNA (exRNA) and cellular miRNA have received growing attention in recent years. A number of studies have recently focused on these classes of miRNAs by taking advantage of the ease with which they may be extracted from human blood samples\cite{25791433,27123852,27112789}. A recently-published\cite{25791433} dataset of microRNA-eQTLs ("miR-eQTLs") in human adult whole-blood samples (taken from the FHS) includes 5,269 cis-miR-eQTLs for 76 miRNAs. As a first step toward investigating the potential mechanistic roles of these cis-eQTLs, we have performed several analyses using these data to elucidate the relationships between miRNA expression heritability, genomic distance between the associated variants and the associated miRNAs, and metrics for eQTL strength (as measured by effect sizes, t-statistics, and significance values). Thus far, our existing results are largely consistent with intuition -- we find that stronger eQTLs (i.e., those that result in more pronounced changes in miRNA expression values) are positively correlated with significance values and miRNA expression heritability, and they we find them to be negatively correlated with genomic distance.

To further investigate the potential mechanistic roles of eQTLs in human whole-blood, we have also carried out a preliminary round of analyses with the objective of determining whether cis-miR-eQTLs may act to silence target genes. We have merged the SNVs within the above-discussed cis-miR-eQTLs with SNVs associated with eQTLs in whole-blood samples of the GTEx dataset\cite{25954001}. We have found that SNVs that positively influence miRNA expression values tend to negatively influence protein-coding gene expression values (and vice-versa). Our next steps entail interrogating these miRNA-gene pairs for direct interactions (see section 2.2.1 within Proposed Research).

2.1.2 **Network analysis of co-expressed allele-specific and eQTL-target genes and relating ASE SNVs and eQTLs with published GWAS results**

We have been at the forefront of efforts to map the large-scale structure of gene expression networks in human and model organisms through our involvement in the ENCODE and modENCODE consortia\cite{22955619}\cite{21177976}. Integrated analysis of regulatory networks, including protein-coding genes and miRNAs in human and *C. Elegans*, have revealed extensive large-scale hierarchical organization, differential enrichment of network motifs across hierarchical levels, and distinct preferences for TF binding at proximal and distal regions. More recently, we have developed an approach to identify regulatory network motifs at a finer level of granularity using logical functions, and have shown these to interact differentially within the hierarchical structure of Yeast and Human networks \cite{25884877}.

We have further introduced methods for analyzing transcriptional architecture by using co-expression networks to identify modules of co-expressed genes. We have developed a novel cross-species multi-layer network framework, OrthoClust, for analyzing the co-expression networks in an integrated fashion by utilizing the orthology relationships of genes between species \cite{25249401}. OrthoClust revealed conserved modules across human, worm and fly that are important for development. Identifying such modules revealed extensive correspondences between various developmental stages in worm and fly \cite{25164755}. Recently, we have shown that it is also possible to reveal large-scale expression patterns over time by identifying dynamical interactions between co-expression modules, revealing conserved dynamical interactions in worm and fly development \cite{27760135}.

2.2 Proposed Research

2.2.1 **Interpreting mechanistic roles of ASE SNVs and eQTLs for non-coding RNAs and associated protein-coding genes. Integrative analyses to elucidate eQTL mechanisms through intermediary miRNAs**

We will combine data on cis- and trans-eQTLs linked to miRNAs (so-called "mir-QTLs") with eQTLs for protein-coding genes as well as SNVs associated with ASE. This integration and downstream analysis will be motivated by a simple model in which the SNV of an eQTL linked to a specific miRNA may also function as a causative eQTL variant linked to the expression of protein-coding genes. The mechanism by which these “shared eQTLs” act may be one in which the SNV directly influences miRNA expression, and that miRNA may act to degrade the transcript of a target gene, thereby reducing its measured expression levels (see Fig. 2).

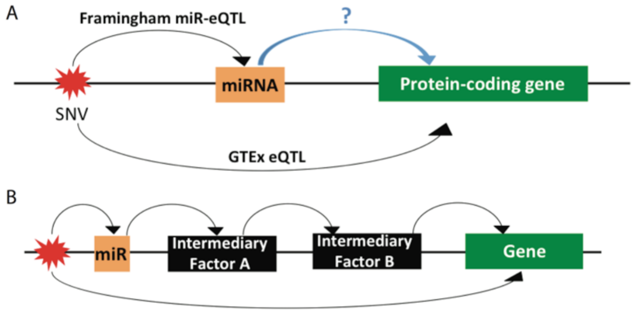

Figure 2: A) A proposed mechanism of broker miRNAs, wherein eQTLs detected for miRNAs in human whole blood samples are matched to the SNVs for eQTLs in GTEx. B) A scenario in which a miRNA and a protein-coding gene share an SNV as an eQTL, but the means by which the miRNA influences protein-coding gene expression is one of an indirect cascade of effects involving intermediary regulatory genes.

TR4PCE

We will first select the mirQTLs and allele-specific miRNAs from the datasets that we will uniformly process in Aim 1, as well as previously determined eQTLs for protein-coding genes from these same studies. We will distinguish between cases in which the miRNA lies within the target gene itself, and those in which the miRNA lies outside of the target gene. We will investigate cases in which SNVs may act in an allele-specific manner on the miRNAs (by definition, according the mechanism proposed above, the manner in which SNVs act to regulate the expression of the protein-coding genes would be trans-regulatory effects -- the miRNA would be a diffusible factor regulating the transcript levels of the target gene).

We emphasize that, in restricting the search for eQTLs to only those genes which are targets of particular miRNAs, greater statistical power may be achieved in identifying trans-eQTLs for protein-coding genes (experimentally-validated target genes for specific miRNAs will be obtained from TarBase 6.0 \cite{22135297}). We can perform similar analyses for other classes of non-coding RNAs.

2.2.2 **Network analysis of co-expressed allele-specific and eQTL-target genes and relating ASE SNVs and eQTLs with published GWAS results**

We will build cross-tissue and tissue-specific gene co-expression modules at both the lncRNAs and miRNA levels using all available data. Specifically we will use the FHS data both to look for modules that are differentially expressed in patients with heart disease, and for modules that are enriched in ASNVs (ASE SNVs) and eQTLs present in the FHS disease genotypes. ASNVs and eQTLs associated with such modules may be considered putative causative variants for the disease, and we will develop statistical techniques that utilize the modular network structure to increase the power to detect such variants. We will adapt Orthoclust to identify co-expression modules across and within tissues \cite{25249401}, and will additionally look for network motifs at the module level, and will then test for enrichment of FHS ASNVs and eQTLs within these motifs.

We will further investigate the potential of recently introduced methods based on Bayesian structured sparsity \cite{27467526}\cite{Zhao JMLR 2016; 17(196):1-47}, which provide a unified framework for jointly identifying eQTLs and co-expression modules, drawing on the increased statistical strength provided by treating eQTL identification as a structured multivariate regression problem \cite{ArXiv:1512.02306}. We will adapt existing structured methods to incorporate both miRNA and long-RNA modules and their interactions by incorporating prior information based on known miRNA targets (using TarBase as above), and to incorporate both allele-specific SNV and eQTL traits, while coherently modeling their expected interactions when affecting the same genes.

Finally, we will compare the genetic variants associated with FHS eQTLS and ASNVs with known disease-associated variants from GWAS studies (such as CVD), and test for association with similar genes, and for enrichment in similar co-expression modules. The contrasting approaches above will allow us to compare the consistency of our findings using modules identified by contrasting methods.

mORE