LNCRNAKB: A COMPREHENSIVE KNOWLEDGEBASE OF LONG NON-CODING RNAS

by

Fayaz Seifuddin
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

_____ Dr. M. Saleet Jafri, Committee Chair

_____ Dr. Mehdi Pirooznia, Committee Co-Chair

_____ Dr. Haiming Cao, Committee Member

_____ Dr. Huzefa Rangwala, Committee Member

_____ Dr. Iosif Vaisman, Director, School of Systems Biology

_____ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_____ Dr. Ali Andalibi, Interim Dean, College of Science

Date: _____ Summer Semester 2019
George Mason University
Fairfax, VA

lncRNAKB: A Comprehensive Knowledgebase of Long non-coding RNAs

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Fayaz Seifuddin
Master of Science
George Mason University, 2008
Bachelor of Science
Linfield College, 2005

Director: Mehdi Pirooznia, Director
Bioinformatics and Computational Biology Core
National Heart, Lung, and Blood Institute (NHLBI)
National Institutes of Health (NIH)

Summer Semester 2019
George Mason University
Fairfax, VA

DEDICATION

This dissertation is dedicated to my wife and daughter.

ACKNOWLEDGEMENTS

I would like to thank Dr. Mehdi Pirooznia (MP) for his heartwarming support without which none of this would be possible. Words cannot describe his love, dedication and mentorship for eleven great years. MP, I will be expecting to get a cookie, brownie or chocolate chip muffin every day.

I would like to thank Dr. Saleet Jafri for his support, feedback and encouragement as always.

I would like to thank Dr. Haiming Cao for his support and providing me with the direction to begin thinking and understanding about analysis of long non-coding RNAs.

I would like to thank Dr. Komudi Singh for her support and taking up several challenges working with me on many parts of this project.

I would like to thank Abhilash Suresh for his help to execute specific parts of this project as well.

I would like to thank the National Institutes of Health (NIH), the National Heart, Lung, and Blood Institute (NHLBI) and Graduate Partnership Program (GPP) for providing me with the support to attend graduate school at George Mason University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

LNCRNAKB: A COMPREHENSIVE KNOWLEDGEBASE OF LONG NON-CODING RNAS

Fayaz Seifuddin, Ph.D.

George Mason University, 2019

Dissertation Director: Mehdi Pirooznia

High throughput technologies such as next-generation sequencing technologies have allowed the genomic structure to be interrogated at high resolution and scale. That includes long non-coding RNAs (lncRNAs), a class of non-protein-coding transcripts, that range from 200 nucleotides to 100 kb (approximately 10 kb on average). The number of estimated lncRNAs annotations in humans range from 20,000 to 100,000. There are several databases that exist for annotation of human lncRNAs. Most of these databases are available through web-based searchable interfaces. Our objective was to identify current and new lncRNAs databases, download and inspect their latest annotations, integrate this information into a single resource, and create the most comprehensive up-to-date knowledge base that encompasses data from all major resources. Specifically, we provide a "one- stop shop" in which users can search for lncRNAs based on any keywords for e.g. genomic locations, gene names and types. LncRNAs annotations are

commonly used as references for quantifying and identifying differentially expressed genes and transcripts in RNA-seq experiments. We used the Genotype Tissue Expression (GTEx) project RNA-seq data to quantify all the lncRNAs in our knowledge base using 9,425 samples sequenced across 31 solid organ human normal tissues. We performed RNA-seq data analysis using a custom pipeline and created a comprehensive tissue-specific expression body map of human lncRNAs. The sequence-function relationship of lncRNAs is not well understood compared to protein-coding genes whose function can be deduced from primary sequence alone. In addition to understanding and improving the annotations of lncRNAs, we sought to predict and determine molecular, biological and disease functions of lncRNAs. We positionally classified and predicted the coding potential of all lncRNAs using a machine learning approach. Using whole genome sequence (WGS) genotype data from the GTEx project we also identified lncRNAs regulated by genetic variants in *cis*. We performed mRNA-lncRNA co-expression network analysis and identified co-expression gene modules involved in known biological processes thus, deducing the potential function of lncRNAs. Our objective was to functionally annotate and characterize lncRNAs in our knowledge base and provide a comprehensive resource to empower the research community.

CHAPTER 1

INTRODUCTION

Background

Most of the non-protein-coding part of the human genome was considered "junk DNA" in the year 2000 when scientists of the Human Genome Project presented the first rough draft of the human genome sequence (Venter *et al.*, 2001; Lander *et al.*, 2001). This was mainly because of its lack of protein-coding capacity and abundant occurrence of features such as non-coding RNAs, transposons, pseudogenes and repetitive regions. However, that notion is drastically changing with the introduction of high throughput technologies such as microarrays and Next-Generation Sequencing (NGS) that have allowed the non-coding genome to be interrogated at high resolution and scale (You *et al.*, 2017; Hangauer *et al.*, 2013). The Encyclopedia of DNA Elements (ENCODE) project reports that approximately 2% of the genome is protein-coding; however, approximately 80% of all nucleotides are detectably transcribed under some conditions (ENCODE Project Consortium, 2012). The discovery of active transcription of the human genome, coupled with the advances in genomic technologies and research, might be a key element to understand the possible "function" of the inaccurately labelled "junk DNA."

Features of long non-coding RNAs

Long non-coding RNAs (lncRNAs) are a class of non-protein-coding transcripts that range in length from 200 nucleotides/base pairs (bp) to 100 kilobases (kb) (approximately 10 kb on average) (Long non coding RNA biology, 2017). The majority of eukaryotic lncRNAs are produced by RNA polymerase II and capped at the 5′ end similar to protein coding genes (PCGs) (Guttman *et al.*, 2009). LncRNAs may or may not be 3′-end polyadenylated (Long non coding RNA biology, 2017), could undergo splicing and have longer but fewer exons, compared to mRNAs (Derrien *et al.*, 2012). Classes of lncRNAs are usually annotated relative to their position with nearby PCGs (DiStefano, 2018), and include: (1) intergenic lncRNAs or lincRNAs, which are transcribed from regions at least >1 kb from PCGs, (2) bidirectional lncRNAs which are transcribed <1 kb of promoters in opposite direction of protein-coding transcripts, (3) intronic lncRNAs, which are transcribed within introns of PCGs, (4) exonic lncRNAs, which overlap with one or more exons of PCGs, (4) sense lncRNAs, which are transcribed in the same direction of PCGs and overlap with one or more exons or introns of these transcripts and (5) antisense lncRNAs, which are transcribed in the opposite direction of PCGs and overlap with one or more exons or introns of these transcripts.

Many lncRNAs do not show the same pattern of high interspecies conservation as protein coding genes (PCGs) (Hezroni *et al.*, 2015; Cabili *et al.*, 2011; Guttman *et al.*, 2009; Li and Yang, 2017). Sequence conservation is comprised of short, 5′-biased patches of conserved sequence nested in exons (Hezroni *et al.*, 2015). Many studies have reported that lncRNAs have low level of expression (Ponting *et al.*, 2009). However,

lncRNAs have higher tissue-specific expression compared to mRNAs (Cabili *et al.*, 2011; Jiang *et al.*, 2016). Some lncRNAs include short open reading frames (sORFs) and undergo translation, though only a minority of such translation events results in stable and functional peptides (Housman and Ulitsky, 2016; Andrews and Rothnagel, 2014). Due to low sequence conservation and low levels of expression, the knowledge that lncRNAs are merely transcriptional noise is common (Ponjavic *et al.*, 2007).

Biological function, subcellular localization and disease functions of lncRNAs

LncRNAs have been suggested to play diverse and important roles in many fundamental and critical biological processes, including: transcriptional regulation in *cis* or *trans*, post-transcriptional regulation, organization of nuclear domains, regulation of proteins or RNA molecules, epigenetic regulation, organ or tissue development, cell differentiation and apoptosis, cell cycle control, cellular transport, metabolic processes, and chromosome dynamics (Ponting *et al.*, 2009; Kopp and Mendell, 2018).

Predicting the subcellular localization of lncRNAs can provide valuable insights on how they perform many of their biological roles. Initially, lncRNAs were found to be primarily located in the nucleus and chromatin however, recently they have been found in other cellular compartments such as the cytoplasm (Mas-Ponte *et al.*, 2017). Recently, DeepLncRNA, a deep learning algorithm has been developed which predicts lncRNA subcellular localization directly from lncRNA transcript sequences (Gudenas and Wang, 2018).

Many lncRNAs have been associated to human disease (Chen *et al.*, 2013; DiStefano, 2018). Several genetic variants identified using DNA microarrays (Genome Wide Association Studies (GWAS) and NGS technologies (exome and/or whole genome sequencing (WGS)), play an important role in human traits and complex diseases. However, numerous single nucleotide variants (SNVs) and copy number variants (CNVs) fall in the non-coding regions of the human genome (Zhang and Lupski, 2015). Genetic variants associated to disease are known to alter the expression of lncRNAs which successively could also regulate the expression of neighboring PCGs (Tan *et al.*, 2017a; Kumar *et al.*, 2013). Consequently, dysregulation of lncRNAs expression could potentially contribute to a variety of diseases through several biological pathways.

Annotation of lncRNAs

Currently, the number of estimated lncRNAs annotations in humans range from 20,000 to 100,000 (Uszczynska-Ratajczak *et al.*, 2018). Diverse publicly available resources dedicated to annotation of lncRNAs in humans and other species have been developed, which differ in data coverage and quality (Xu *et al.*, 2017; Uszczynska-Ratajczak *et al.*, 2018; Fritah *et al.*, 2014; Xu *et al.*, 2017). Most of these databases are available through web-based searchable interfaces and also provide downloadable lncRNAs annotation files in Gene Transfer Format (GTF) or Gene Feature Format (GFF) (Ma *et al.*, 2019; Chakraborty *et al.*, 2014; Bhartiya *et al.*, 2013). A few of these databases have attempted to integrate annotations from multiple sources and multi-omics data such as expression (occasionally tissue-specific), methylation, variation,

conservation and functional annotation of lncRNAs in humans. However, their annotations and integrations are sometimes outdated, not rigorous, incomprehensive, and incomplete. Frequently used resources of lncRNAs annotation include GENCODEv29 (Frankish *et al.*, 2019; Derrien *et al.*, 2012), CHESS2.1 (Pertea *et al.*, 2018), LNCipedia5.2 (Volders *et al.*, 2015, 2013), NONCODEv5.0 (Fang *et al.*, 2018), FANTOM5.0.v3 (Hon *et al.*, 2017), MiTranscriptomev2 (Iyer *et al.*, 2015) and BIGTranscriptomev1 (You *et al.*, 2017). These resources annotated lncRNAs based on two main approaches: manual or automatic (Uszczynska-Ratajczak *et al.*, 2018). Automatic annotation involves the use of bioinformatics methods such as StringTie (Pertea *et al.*, 2015) and Cufflinks (Trapnell *et al.*, 2012) to reconstruct gene and transcript models based on short sequence reads which is widely used due to the advances in NGS technologies and production of billions of RNA sequences/reads (RNA-seq) (Iyer *et al.*, 2015). Manual annotation involves the creation and curation of gene and transcript models by human annotators based on RNA and protein experimental evidence and defined sets of rules (Frankish *et al.*, 2019; Derrien *et al.*, 2012). After reviewing the frequently used resources, majority of these used a hybrid annotation approach for identifying lncRNAs i.e. a combination of manual and automatic annotations in their pipelines.

The current GENCODE annotation (v29) is widely used as an annotation for PCGs and lncRNAs. GENCODE used a hybrid approach for annotation of genes. It relies heavily on manual annotation where an expert curator interrogates all possible features (sequence, expression data and computational predictions) of a gene and considers all

possible annotations and biotypes for the gene concurrently. Additionally, experimental annotation methods such as Capture Long Seq (CLS) (Lagarde *et al.*, 2017) which captures full length cDNAs, proteomics (Aslam *et al.*, 2017) and RT-PCR (Rio, 2014) augment its annotation pipeline.

CHESS2.1 used billions of short RNA-seq reads from the Genotype Tissue Expression (GTEx) project (GTEx Consortium *et al.*, 2017), which included samples from numerous tissues collected from hundreds of individuals. Using a reference (RefSeq) (O'Leary *et al.*, 2016) guided novel transcriptome assembly approach, CHESS2.1 assembled all samples, merged the results and applied a series of computational filters to exclude transcripts with insufficient evidence. Additional validation was performed based on unmatched mass spectrometry data. CHESS2.1 encompasses all genes from RefSeq and GENCODE however, it adds 224 protein-coding gens and 2,671 lncRNAs based on robust experimental and alignment evidence.

LNCipedia5.2 uses different sources of lncRNA transcripts (Ensembl release 92 (Zerbino *et al.*, 2018), RefSeq – Dec 2014 and NCBI Annotation release 106, FANTOM CAT (Hon *et al.*, 2017)) and attempts to combine these into non-redundant records. It also includes other sources such as LncRNAdb (Amaral *et al.*, 2011) , GENCODE (release 13), NONCODE v4 (Xie *et al.*, 2014) and RNA-Seq data from various resources (Nielsen *et al.*, 2014; Hangauer *et al.*, 2013; Sun *et al.*, 2015; Cabili *et al.*, 2011). For each transcript, LNCipedia5.2 provides information on protein coding potential, locus conservation and published literature available.

NONCODEv5.0 also integrates multiple lncRNA resources (Ensembl, RefSeq, LncRNAdb, LNCipedia, RNA-seq, exosome expression profiles and old versions of NONCODE (Bu *et al.*, 2012; Xie *et al.*, 2014)). However, it contains lncRNA transcripts from multiple species (human, mouse, cow, rat, chimpanzee, gorilla, orangutan, rhesus macaque, opossum, platypus, chicken, zebrafish, fruit fly, Caenorhabditis elegans, yeast, arabidopsis and pig). NONCODEv5.0 provides information on: (i) conservation, (ii) diseases, (iii) lncRNAs and exosome expression profiles, and (iv) lncRNAs and RNA secondary structure.

FANTOM5.0.v3 used a collection of transcript annotations from GENCODE release 19, Human BodyMap 2.0 (Cabili *et al.*, 2011), MiTranscriptome (Iyer *et al.*, 2015), ENCODE and RNA-seq transcript assembly data from 70 FANTOM5 samples. This data was integrated with cap analysis of gene expression (CAGE) data sets (Andersson *et al.*, 2014; Shiraki *et al.*, 2003; FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.*, 2014; Arner *et al.*, 2015) to build an atlas of human lncRNAs with accurate 5' ends. Further characterization of these lncRNAs was performed using epigenomic, genomic and transcriptomic evidence.

MiTranscriptomev2 used 7,256 RNA-seq libraries from tumors, normal tissues and cell lines comprising over 43 Terabytes (Tb) of sequence from 25 independent data sets including ENCODE, The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga) and the Human BodyMap 2.0. It also validated novel lncRNAs (transcripts without coding potential) by searching against a large human proteomics data set derived from benign tissue samples (Kim *et al.*, 2014).

BIGTranscriptomev1 applied a custom transcriptome assembly pipeline, called CAFE to RNA-seq data comprising of 230 billion reads from ENCODE, Human BodyMap 2.0, TCGA and GTEx. Their pipeline claimed to have significantly improved the quality of the resulting transcriptome map by predetermining the orientation of the reads since most of the RNA-seq data used were generated using an unstranded protocol. Additionally, by including information about transcription start sites (CAGE-seq) (FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.*, 2014), cleavage and polyadenylation sites (3P-seq) (Nam *et al.*, 2014; Nam and Bartel, 2012) significantly improved the transcriptome assemblies.

Characterization and classification of lncRNAs

The sequence-function relationship of lncRNAs is not well understood compared to PCGs whose function can be deduced from primary sequence alone (Hezroni *et al.*, 2015). Computational methods for distinguishing between PCGs and lncRNAs can be used to assess the sequence or the evolution of an uncharacterized transcript and predict whether it is likely to encode a protein. These methods use different features for classification of transcripts as coding or non-coding for e.g. the length of ORFs, nucleotide, codon or short word compositions/frequencies (*k*-mers), substitution patterns, the presence of sequences encoding known functional protein domains, similarity to known proteins, conservation or evolution. There are pros and cons for using these features in classification schemes that have to be considered while implementing these methods for e.g. relying on sequence similarities to entries in known protein databases

can be an issue for a putative lncRNA as some databases frequently contain "hypothetical protein" sequences without experimental evidence. In addition, if a lncRNA has sequence similarity to pseudogenes it may contain elements that score highly as potential functional domains or as similar to other proteins, but those elements will typically not reside in a functional ORF. There are several methods for distinguishing between PCGs and lncRNAs (Housman and Ulitsky, 2016). Broadly, these methods can be grouped into machine learning or evolutionary algorithms.

Extensively used algorithms for predicting the coding potential of non-coding sequences are: Coding-Potential Assessment Tool (CPAT) (Wang *et al.*, 2013)), FlExible Extraction of LncRNAs (FEELnc) (Wucher *et al.*, 2017) and PhyloCSF (Lin *et al.*, 2011). CPAT is an alignment-free method that uses a logistic regression framework built with four sequence features: ORF size, ORF coverage defined as the ratio of ORF to transcript lengths, fickett score (combinational effect of nucleotide composition and codon usage bias (Fickett, 1982)) and hexamer score (dependence between adjacent amino acids in proteins (Fickett and Tung, 1992)). Analysis of coding potential using CPAT showed that all four selected features were concordantly higher in coding transcripts and lower in noncoding transcripts. CPAT achieved highest overall accuracy (0.97) when compared with CPC (0.87), PhyloCSF (0.76) and PORTRAIT (Arrial *et al.*, 2009) (0.92).

FEELnc annotates lncRNAs based on a machine learning method, Random Forest (RF) (Breiman, 2001), trained with general features such as multi *k*-mer frequencies, RNA sequence length and ORFs size. It is comprised of three modules: (i) filter, (ii) coding potential, and (iii) classifier. The filter module flags and removes transcripts if

these are monoexonic, short (< 200 base pairs (bp)) and/or overlapping (in sense) exons of the reference annotation, especially protein-coding exons. Using the filtered GTF annotation output file from the filter module, the coding potential module employs RF to calculate a coding potential score (CPS) for each transcript. For the training set, it can use a "known" set/annotation of mRNAs and lncRNAs transcript sequences and build a model using these to calculate CPS of a probable lncRNA sequence. However, due to the lack of a gold standard/known human lncRNAs data set for training, FEELnc implemented three strategies: intergenic, shuffle, and cross-species in the coding potential module. The intergenic approach extracts random intergenic sequences of length *L* from the genome of interest to model species-specific noncoding sequences as the non-coding training set. The shuffle approach builds the training data from random parts of PCGs while preserving a given *k*-mer frequency of the input sequence. The cross-species mode, builds the RF training model based on lncRNAs annotations in other species. To determine an optimal CPS cut-off, FEELnc automatically extracts the CPS that maximizes both sensitivity and specificity based on a 10-fold cross-validation. The CPS is between 0 and 1 where 0 indicates a non-coding RNA and a score close to 1 an mRNA. To classify potential lncRNAs with respect to the localization and the direction of transcription of nearby mRNAs (or other non-coding RNAs) transcripts is a commonly used approach and implemented in the classifier module of FEELnc.

PhyloCSF is based on statistical phylogenetic model comparisons. Two phylogenetic models are created to differentiate coding and non-coding regions. One model represents the evolution of codons in protein-coding genes, and another represents

the evolution of nucleotide triplet sites in non-coding regions. The models have various parameters that can be adjusted based on the genomic regions being classified. When a new sequence needs to be categorized, its probabilities under the coding and non-coding model are calculated. Classification is based on the log-likelihood ratio = coding probability/non-coding probability. Statistical significance is based on a cut-off chosen determined by the distribution of the log-likelihood ratio statistic, or it can be chosen empirically based on classification performance in a test set.

Tissue-specific expression and functional annotation of lncRNAs

Tissue-specificity of lncRNAs expression are an important feature for functionally characterizing lncRNAs. Many lncRNAs show tissue-specific expression patterns, providing important clues about their specific functions (Yang *et al.*, 2018; Cabili *et al.*, 2011; Jiang *et al.*, 2016). Analysis of RNA-seq data derived from 24 human tissues and cell types revealed that the majority of lincRNAs (approximately 80%) exhibit tissue-specific expression patterns, whereas such expression patterns are observed in a much smaller fraction of protein-coding genes (approx. 20%) (Melé *et al.*, 2015; GTEx Consortium *et al.*, 2017). Analysis of RNA-seq data derived across normal human tissues from 16 independent studies showed two classes of lncRNAs; ubiquitously expressed long non-coding RNAs (UE lncRNAs) and tissue-specific lncRNAs (TS lncRNAs) (Jiang *et al.*, 2016). UE lncRNAs were characterized as housekeeping lncRNAs and 12% of these were expressed in all tissues while 2.3% of lncRNAs were expressed in only one tissue (TS lncRNAs). In addition, this study uncovered a range of

features that are specific to UE lncRNAs, including compact gene structure, high conservation, strict combinatorial regulation at transcriptional, post-transcriptional, and epigenetic levels, and strong regulation of enhancers.

LncRNAs are emerging as key regulators of diverse biological processes and diseases. However, the combinatorial effects of these molecules in a specific biological function are poorly understood. Identifying co-expressed protein-coding genes of lncRNAs would provide useful insight into lncRNA functions. Unsupervised clustering methods can be used to extrapolate lncRNAs biological function based on the degree of connections to genes of known function. Using the genes expression profiles, these methods cluster them into group of genes, known as co-expression modules. These modules are then subject to over representation analysis (ORA) with specific pathways for e.g. Gene Ontology (GO) (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2019) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017), which determines the potential function of genes. ORA employs the hypergeometric test (Holmans, 2010) to determine if the overall functional enrichment is different than what would be expected by random chance. The significantly enriched terms can be used as an estimate of annotation for the genes with unknown function for e.g. lncRNAs. This is often referred to as the "guilt by association" principle (Ehsani and Drabløs, 2018a) .

CHAPTER 2

lncRNAKB: A comprehensive knowledgebase of long non-coding RNAs (lncRNAs)

ABSTRACT

Motivation: There are several databases that exist for annotation of human long non-coding RNAs (lncRNAs) that contain between 20,000 to 100,000 entries. These databases contain unique and overlapping lncRNAs that have been identified by next generation sequencing (NGS) methods. The information on lncRNAs provided in these databases is not rigorous thus, making it difficult to understand their molecular and cellular functions. Consequently, there is a need to systematically and carefully combine these annotations, create a non-redundant resource and provide valuable functional information on lncRNAs.

Results: We have created the long non-coding RNA knowledgebase (lncRNAKB) by methodically integrating six widely used lncRNAs annotation databases (CHESS2.1, FANTOM5.0.v3, LNCipedia5.2, NONCODEv5.0, MiTranscriptomev2 and BIGTranscriptomev1). We present an annotation of a large number of unique lncRNAs (*n*=77,199). The lncRNAKB incorporates coding potential, classification/localization with respect to messenger RNAs (mRNAs), gene expression, tissue-specificity scores, expression quantitative trait loci (eQTL)-regulated lncRNA genes, phylogenetic

conservation and functional characterization to identify co-expressed mRNAs that would provide potential understanding on lncRNAs function. A machine learning approach was used to calculate the coding potential scores and classify the lncRNAs in the lncRNAKB annotation database as putative lncRNAs or mRNAs. Gene expression data of 9,074 RNA-seq samples, collected from the Genotype Tissue Expression (GTEx) project was used to provide tissue-specific expression profiles and tissue-specificity scores in 31 solid organ human normal tissues. Using whole genome sequence (WGS) genotype data of 652 subjects and tissue-specific gene expression data from the GTEx project we calculated *cis*-eQTLs in all tissues. We calculated and compared evolutionarily exon conservation between lncRNAs and protein-coding genes (PCGs) using an alignment of 30 vertebrate species. We used Weighted Gene Co-expression Network Analysis (WGCNA) to identify co-expression modules encompassing lncRNA-mRNA pairs that were subjected to enrichment analysis using Gene Ontology (GO) pathways to identify meaningful biological processes that lncRNAs could be potentially involved in and created dynamic Cytoscape networks for exploration and visualization. All components are provided in a user-friendly web interface.

Availability: lncRNAKB is available at http://www.lncrnakb.org

Contact: fayaz.seifuddin@nih.gov

Supplementary Information: Supplementary data are available online.

INTRODUCTION

Most of the non-protein-coding part of the human genome was considered "junk DNA" in the year 2000 when scientists of the Human Genome Project presented the first rough draft of the human genome sequence (Venter *et al.*, 2001; Lander *et al.*, 2001). However, that notion is drastically changing with the introduction of high throughput technologies such as Next-Generation Sequencing (NGS) that have allowed the non-coding genome to be interrogated at high resolution and scale (You *et al.*, 2017). The Encyclopedia of DNA Elements (ENCODE) project reports that approximately 2% of the genome is protein-coding; however, approximately 80% of all nucleotides are detectably transcribed under some conditions (ENCODE Project Consortium, 2012). Long non-coding RNAs (lncRNAs) are a class of non-protein-coding transcripts that range in length from 200 nucleotides/base pairs (bp) to 100 kilobases (kb) (approximately 10 kb on average) (Long non coding RNA biology, 2017). Currently, the number of estimated lncRNAs annotations in humans range from 20,000 to 100,000 (Uszczynska-Ratajczak *et al.*, 2018).

LncRNAs have been suggested to play diverse and important roles in many fundamental and critical biological processes, including transcriptional and post-transcriptional regulation, epigenetic regulation, organ or tissue development, cell differentiation and apoptosis, cell cycle control, cellular transport, metabolic processes and chromosome dynamics (Ponting *et al.*, 2009; Kopp and Mendell, 2018). Many lncRNAs do not show the same pattern of high interspecies conservation as protein coding genes (PCGs) (Hezroni *et al.*, 2015; Cabili *et al.*, 2011; Guttman *et al.*, 2009; Li

and Yang, 2017). Sequence conservation is comprised of short, 5′-biased patches of conserved sequence nested in exons (Hezroni *et al.*, 2015). Many studies have reported that lncRNAs have low level of expression (Ponting *et al.*, 2009). However, lncRNAs have higher tissue-specific expression compared to mRNAs (Cabili et al., 2011; Jiang et al., 2016). Some lncRNAs include short open reading frames (sORFs) and undergo translation, though only a minority of such translation events results in stable and functional peptides (Housman and Ulitsky, 2016; Andrews and Rothnagel, 2014).

Recently, diverse publicly available resources dedicated to annotation of lncRNAs in humans and other species have been developed, which differ in data coverage and quality (Xu *et al.*, 2017; Uszczynska-Ratajczak *et al.*, 2018; Fritah *et al.*, 2014; Xu *et al.*, 2017). Most of these databases are available through web-based searchable interfaces and also provide downloadable lncRNAs annotation files in Gene Transfer Format (GTF) or Gene Feature Format (GFF) (Ma *et al.*, 2019; Chakraborty *et al.*, 2014; Bhartiya *et al.*, 2013). A few of these databases have attempted to integrate annotations from multiple sources and multi-omics data such as expression (occasionally tissue-specific), methylation, variation, conservation and functional annotation of lncRNAs in humans. However, their annotations and integrations are sometimes outdated, not rigorous, incomprehensive, and incomplete.

Frequently used resources of lncRNAs annotation include GENCODEv29 (Frankish *et al.*, 2019; Derrien *et al.*, 2012), CHESS2.1 (Pertea *et al.*, 2018), LNCipedia5.2 (Volders *et al.*, 2015, 2013), NONCODEv5.0 (Fang *et al.*, 2018), FANTOM5.0.v3 (Hon *et al.*, 2017), MiTranscriptomev2 (Iyer *et al.*, 2015) and

BIGTranscriptomev1 (You *et al.*, 2017). These resources annotated lncRNAs based on two main approaches: manual or automatic (Uszczynska-Ratajczak *et al.*, 2018). Automatic annotation involves the use of bioinformatics methods such as StringTie (Pertea *et al.*, 2015) and Cufflinks (Trapnell *et al.*, 2012) to reconstruct gene and transcript models based on short sequence reads which is widely used due to the advances in NGS technologies and production of billions of RNA sequences/reads (RNA-seq) (Iyer *et al.*, 2015). Manual annotation involves the creation and curation of gene and transcript models by human annotators based on RNA and protein experimental evidence and defined sets of rules (Derrien *et al.*, 2012).

We developed lncRNAKB (https://www.lncrnakb.org) which carefully combines the frequently used lncRNAs annotation resources mentioned above using a cumulative stepwise intersection method. Our method of integration compares the annotations thoroughly, discarding redundant and ambiguous lncRNAs records. In addition, the cumulative approach accounts for the large overlap between the lncRNAs annotation databases. The lncRNAKB provides a comprehensive downloadable, searchable and viewable (via the UCSC Genome Browser) (Casper *et al.*, 2018) GTF annotation file of human PCGs and a large number of lncRNAs (n=77,199) that can be used by researchers to quantify their RNA-seq expression data for lncRNAs discovery. Using the method of FEELnc (FlExible Extraction of LncRNAs) (Wucher *et al.*, 2017), we filtered and classified the lncRNAs with respect to overlap with mRNAs, thus, providing additional categorization of lncRNAs. Furthermore, with FEELnc we calculated the coding potential of the aforementioned lncRNAs based on a Random Forest (RF) (Breiman, 2001) model

trained with general features such as multi *k*-mer frequencies and presence of ORFs. To significantly enrich and improve lncRNAs' annotations in the lncRNAKB to support function inference we implemented the latest analysis pipeline and analyzed the largest and most comprehensive tissue-specific RNA-Seq data available through the Genotype Tissue Expression (GTEx) project (GTEx Consortium *et al.*, 2017) and created a tissue-specific expression body map of human lncRNAs. We calculated tissue-specificity scores across all the genes in the lncRNAKB which elucidated the tissue-specificity of lncRNAs compared to mRNAs. In addition, we calculated expression quantitative trait Loci (eQTL)-regulated lncRNA genes using the GTEx expression with whole genome sequencing (WGS) genotype data and created a tissue-specific eQTL body map of human lncRNAs in the lncRNAKB. We also calculated and compared the conservation scores (derived from an alignment of 30 vertebrate species) (Casper *et al.*, 2018) for all PCGs and lncRNAs (exon-level) in the lncRNAKB, which can be viewed and downloaded from the website. Genome-wide gene expression data (raw counts and normalized), eQTL results and tissue-specificity scores across all tissues are freely available for visualization and download on the lncRNAKB website. Furthermore, to predict lncRNA functions, we used the guilt by association principle (Ehsani and Drabløs, 2018a) and Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008) method analyzing lncRNAs-mRNAs co-expression patterns in a tissue-specific manner. We performed enrichment analysis of annotation terms for all co-expression modules identified and showed the functional pathways associated with lncRNAs, creating a tissue-specific body map of functionally annotated lncRNAs. All pathway results files are

freely available to download. Moreover, for each module we have created a dynamic network figure on the website to view the strength of connections between the highest ranking mRNAs and lncRNAs by pathway. The lncRNAKB is highly beneficial because it integrates multi-omics data with the aim to significantly enrich and improve lncRNAs' annotations to support functional implications.

MATERIALS AND METHODS

Data sources and collection

We identified lncRNAs databases by conducting a broadly cast literature search of the PubMed database through February 28th, 2019 with the following keyword algorithm: *(lncrna or long noncoding or long non-coding rna or noncoding) and (annotation or function or database).* A total of 13,412 articles were returned filtered by humans species and published within the past five years sorted by the best match criteria. These were manually reviewed by looking at their titles, abstracts, keywords, and full text as needed to identify those that reported on lncRNAs annotations, databases and function. We further searched the references of these articles to identify any other articles that were potentially missed by the initial PubMed search. In addition to the literature search, we reviewed the frequently used and updated lncRNABlog (https://www.lncrnablog.com/) Database section which provides a summary of up-to-date published lncRNAs resources. We also consulted with clinicians and researchers in the field to identify lncRNAs data sources that are widely used. After the review, we chose to integrate six main sources of existing well-known and widely used human lncRNAs annotation databases in the

knowledgebase, including: CHESS2.1 (Pertea *et al.*, 2018), LNCipedia5.2 (Volders *et al.*, 2013, 2015), NONCODEv5.0 (Bu *et al.*, 2012; Xie *et al.*, 2014; Fang *et al.*, 2018), FANTOM5.0.v3 (Hon *et al.*, 2017), MiTranscriptomev2 (Iyer *et al.*, 2015) and BIGTranscriptomev1 (You *et al.*, 2017). Table 2.1 summarizes significant details for each data source after evaluation and the lncRNAs related-information that will be publicly and freely available via our knowledgebase after integration.

Table 2.1: Summary of lncRNAs annotation databases.
Summary of lncRNAs annotation databases that have been integrated into the lncRNAKB and the types of data included in each resource.

	CHESS2.1	**LNCipedia5.2**	**NONCODEv5.0**	**FANTOM5.0.v3**	**MiTranscriptomev2**	**BIGTranscriptomev1**	**lncRNAKB**
Annotation source file name	chess2.1.gff	lncipedia_5_2_hc_hg38.gtf	NONCODEv5_human_hg38_lncRNA.gtf	FANTOM_CAT.lv3_robust.only_lncRNA.gtf	mitranscriptome.hg19.v2.gtf	BIGTranscriptome_lncRNA_catalog.hg19.gtf	**lncRNAKB_hg38_v6.gtf**
Website	http://ccb.jhu.edu/chess/	https://lncipedia.org/info	http://www.noncode.org/index.php	http://fantom.gsc.riken.jp/5/	http://mitranscriptome.org/	http://bhyou.dothome.co.kr/	**https://www.lncrnakb.org**
Reference [PMID]	Pertea *et al.*, 2018 [30486838]	Volders *et al.*, 2013, 2015, 2019 [30371849]	Fang *et al.*, 2018 [29140524]	Hon *et al.*, 2017 [28241135]	Iyer *et al.*, 2015 [25599403]	You *et al.*, 2017 [28396519]	**Under preparation**
Genome reference build/version	hg38	hg19, hg38	hg19, hg38	hg19	hg19	hg19	**hg38**
Annotation method/source	GENCODE (release 25 and 27), RefSeq, FANTOM5, RNA-seq (GTEx-phs000424.v6.p1 in May of 2016), transcript assembly, mass spectrometry (validation)	Ensembl, RNA-seq (Human BodyMap 2.0 lincRNAs), LncRNAdb, GENCODE (release 13), RefSeq-Dec2014, RefSeq-NCBI (release 106), Nielsen *et al.,* Hangauer *et al.,* NONCODE, Sun and Gadad *et al.,* 2015, FANTOM CAT	Ensembl, RefSeq, lncRNAdb, LNCipedia, RNA-seq (Human BodyMap 2.0 lincRNAs), Exosome Expression Profile, old versions of NONCODE	GENCODE (release 19), Human BodyMap 2.0, miTranscriptome, ENCODE and an RNA-seq assembly from 70 FANTOM5 samples, cap analysis of gene expression (CAGE) data	7,256 RNA sequencing (RNA-seq) libraries from tumors, normal tissues and cell lines comprising over 43 Tb of sequence from 25 independent studies including ENCODE, TCGA, Human BodyMap 2.0, proteomics (validation)	ENCODE, TCGA, GTEx, Human BodyMap 2.0, the Human Protein Atlas, GENCODE (release 19), RefSeq, PacBio	**CHESS2.1, LNCipedia5.2, NONCODEv5.0, FANTOM5.0.V3, MiTranscriptomev2, BIGTranscriptomev1**

	CHESS2.1	LNCipedia5.2	NONCODEv5.0	FANTOM5.0.v3	MiTranscriptomev2	BIGTranscriptomev1	lncRNAKB
Number of lncRNAs (genes)	18,887	56,946	96,308	27,871	63,505	14,090	**77,199**
Number of lncRNAs (transcripts)	56,927	127,802	172,216	89,833	175,259	26,591	**224,286**
Number of lncRNAs (exons)	159,891	357,620	429,240	251,201	539,840	87,316	**611,340**
Number of protein-coding genes	22,883	-	-	-	-	-	**22,518**
Tissue-specific Expression/score	yes	no	yes	yes	yes	yes	**yes**
Tissue-specific Expression Quantitative Trait Loci (eQTLs)	no	no	no	no	no	no	**yes**
UCSC genome browser/Custom Genome Browser Track	no	yes	yes	yes	yes	no	**yes**
External Gene information/links (Gene Cards or RefSeq or Ensembl or UCSC)	yes	yes	yes	yes	yes	yes	**yes**
Coding potential prediction/score	no	yes	no	yes	yes	yes	**yes**

	CHESS2.1	LNCipedia5.2	NONCODEv5.0	FANTOM5.0.v3	MiTranscriptomev2	BIGTranscriptomev1	lncRNAKB
Conservation information	no	yes	yes	yes	yes	no	**yes**
Gene-level functional annotation, mRNA co-expression, pathway enrichment analysis	no	no	yes	yes	yes	no	**yes**

Data integration

We downloaded lncRNAs annotation files in gene transfer format (GTF) or gene feature format (GFF) (https://useast.ensembl.org/info/website/upload/gff.html#moreinfo) from all six annotation databases (links in Table 2.1) identified by the comprehensive analytical review above. To streamline the data integration step all the GTF or GFF annotations were parsed to the same format using the following steps: (i) if necessary, we updated the coordinates of annotation using the UCSC liftOver tool (Casper *et al.*, 2018) from hg19 to hg38 (latest genome build), and (ii) for each chromosome, we split the gene and transcript records into individual files named by chromosome, strand, start and stop base pair locations. Each gene block file contained the transcripts information and the transcript block file contained the exons information. In cases where the annotation file did not have any genes information (only containing transcripts or exons records) we used the gene ids in the transcripts or exons records to get the first and last exon, then manually created a gene entry using the base pair locations of the first exon (as gene start), of the last exon (as gene stop), and transcript strand to represent the gene strand. We also removed redundant records from all annotation files.

Using CHESS2.1 as the reference annotation database (containing both protein-coding and lncRNAs genes) we used a cumulative stepwise intersection method to merge it with the rest of the five lncRNAs annotation databases in this order: (i) FANTOM5.0.v3, (ii) LNCipedia5.2, (iii) NONCODEv5.0, (iv) MiTranscriptomev2 and (v) BIGTranscriptomev1 at the genes and transcripts levels. The order of intersection was arbitrary. Figure 2.1 illustrates the cumulative stepwise intersection method for two

databases as an example, D1 (CHESS2.1) in blue and D2 (FANTOM5.0.v3-lncRNAs only) in green. At the gene level (top panel), we only kept those genes from D2 that had full overlap, were completely enclosed within D1 genes, or no overlap with D1 genes on the same strand (results of intersection shown in orange). We discarded genes in D2 that had partial overlap with D1 genes (marked in red X) because we did not want to re-define boundaries of genes in the reference annotation database.

For genes that intersected, we compared these individually at the transcript level (D1 and D2 transcripts shown in blue and green with smaller bars, introns and exons as compared to genes, bottom panel, respectively). For each gene, we compared the D2 transcripts starts and stops with the intersected gene boundary (dashed gray lines) because we observed that in some databases the transcript boundaries exceeded the gene boundaries. By this rule, we removed several transcripts (marked with a red X) that were probably incorrectly assigned to genes. In addition, if a transcript in D2 had partial overlap with transcripts in D1, we incorporated that transcript (marked with red ticks). If a transcript in D2 had no overlap with any transcripts in D1, we added the transcript including all the exons to the gene record accordingly. If a transcript in D2 was exactly identical to any transcript in D1 we did not add it. For genes with no overlap in D1, we added all the transcripts and corresponding exons to the merged annotation as a lncRNA entry (marked in red ticks).

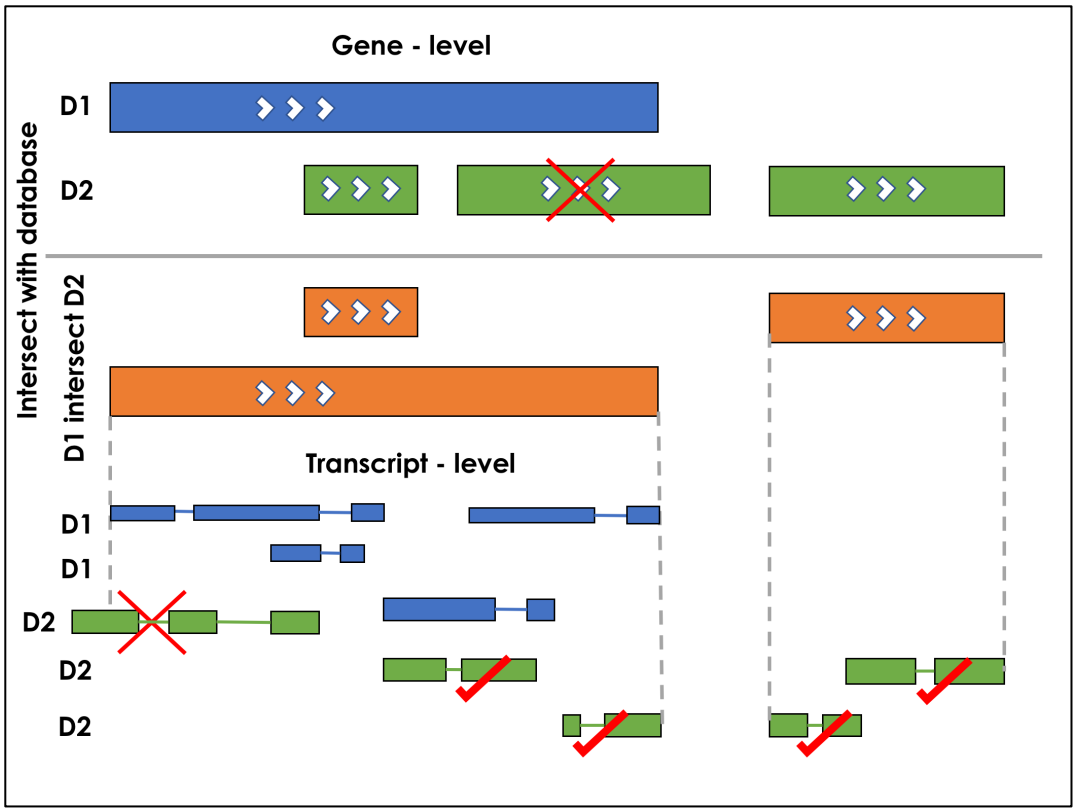


Figure 2.1: Illustration showing the stepwise intersection of two annotation databases. Illustration showing the stepwise intersection of two annotation databases D1 (CHESS2.1) (blue) and D2 (FANTOM5.0.v3-lncRNAs only) (green) at the gene and transcript levels. The genes are shows as solid rectangles and the transcripts are shown with exons and introns. The white arrows show the direction/strand in which the gene is transcribed. The orange bars show the results of the intersection (D1 intersect D2) at the gene level. The red X marks show transcripts that were not incorporated into the merged annotation and vice versa for the red ticks. D3 (LNCipedia5.2), D4 (NONCODEv5.0), D5 (MiTranscriptomev2) and D6 (BIGTranscriptomev1) were merged using the same cumulative stepwise intersection method (see Methods: Data integration).

Architecture of the database

The 3-tier server architecture model containing data, logic and presentation tiers has been implemented as shown in Figure 2.2. The popular MySQL open source relational database management system (RDBMS) has been employed for the data tier,

expanded with a NoSQL document storage. NoSQL document storage is a JSON-based (JavaScript Object Notation) data structure format and as such has a flexible dynamic structure with no schema constraints which makes it suitable for literature and document storage. The MySQL RDBMS (version 8.0) is ideal for data indexing and a powerful query system for relational data. The logic tier is responsible for the communication between the user queries from the presentation tier and fetching the outcome from the data tier, as well as data integration from MySQL and NoSQL data sources. The presentation tier contains several modules based on AJAX (Asynchronous JavaScript and XML), jQuery (JavaScript Query system version 3.3.1 - https://jquery.com/), and the PHP server-side scripting language (version 7.1.18.), as well as the CSS (Cascading Style Sheets) code to describe how HTML elements are to be displayed on user side web interface. JQuery and AJAX have the advantage of asynchronous background calls to the logic tier, native JSON parsing, and dynamic rendering of the browser display, which makes the data retrieval system perform more efficiently. The Web server is hosted on a CentOS 7 operating system using an Apache (2.4.33) web server. The user interface is functional across major web-browsers such as Chrome, Safari, and Firefox on Linux, Mac, iOS, Android, and Windows OS platforms.

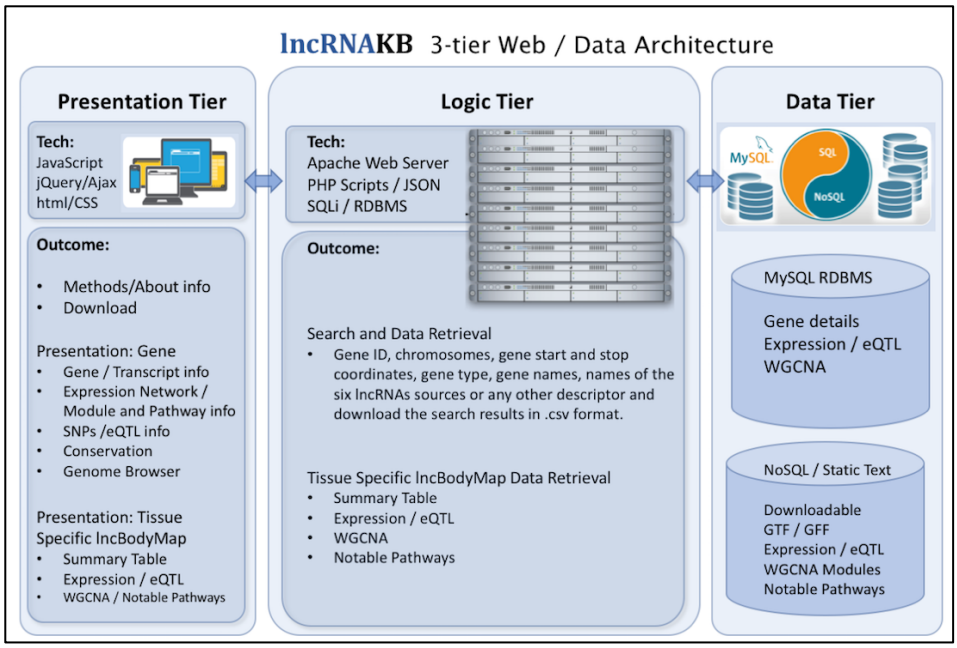


Figure 2.2: Schema of the web/database segment of the lncRNAKB.

Classification/Annotation and coding potential of lncRNAs using Random Forest

We used FEELnc (FlExible Extraction of LncRNAs) (Wucher *et al.*, 2017) to classify/annotate and calculate the coding potential of lncRNAs in the lncRNAKB. FEELnc annotates lncRNAs based on a machine learning method, Random Forest (RF) (Breiman, 2001), trained with general features such as multi *k*-mer frequencies, RNA sequence length and open reading frames (ORFs) size. It is comprised of three modules: (i) filter, (ii) coding potential, and (iii) classifier.

Filter The filter module flags and removes transcripts overlapping (in sense) exons of the reference annotation and especially protein-coding exons. We used the GENCODEv29 (Derrien *et al.*, 2012) GTF file as the reference annotation to get an estimate of the number of transcripts from lncRNAKB overlapping with protein_coding

transcripts (transcript_biotype=protein_coding). We used GENCODE as the reference annotation for protein-coding genes because it is comprehensively and manually curated. We arbitrarily set the minimal fraction out of the candidate lncRNAs size to be considered for overlap to be excluded as 0.75 (> 75% overlap) to retain many lncRNAs transcripts. Transcripts < 200 base pairs (bp) long were filtered out and monoexonic transcripts were retained.

Coding potential We used the filtered GTF annotation output file from the filter module and calculated a coding potential score (CPS) for each transcript using the coding potential module. Due to the lack of a gold standard/known human lncRNAs data set for training, we used the "intergenic" mode in the module. This approach extracts random intergenic sequences of length *L* from the genome of interest to model species-specific noncoding sequences as the non-coding training set. We used the human reference genome FASTA file (hg38) and the GENCODE GTF file as the reference annotation. To get the best training set of known mRNAs, we used transcript_biotype=protein_coding and transcript_status=KNOWN for the RF model. We used the default values for the *k*-mer sizes, number of trees and ORF type. To determine an optimal CPS cut-off, FEELnc automatically extracts the CPS that maximizes both sensitivity and specificity based on a 10-fold cross-validation. The CPS was between 0 and 1 where 0 indicates a non-coding RNA and a score close to 1 an mRNA.

Classifier To classify potential lncRNAs with respect to the localization and the direction of transcription of nearby mRNAs (or other non-coding RNAs) transcripts as shown in Supplementary Figure 2.1, we used the classifier module. We used the final set

of lncRNAs transcripts output from the coding potential module and classified them using the GENCODEv29 GTF file as the reference annotation. A sliding window size around each lncRNA was used to check for possible overlap with nearest reference transcripts. We used a minimum and maximum window size of 10 kilobase (kb) and 100kb respectively. The classification method reported all interactions within the defined window and established a best partner transcript using certain rules.

Tissue-specific expression profiling and expression quantitative trait loci (eQTLs)

Expression profiling We analyzed the largest and most comprehensive tissue-specific RNA-seq data available through the Genotype Tissue Expression (GTEx) project (GTEx Consortium *et al.*, 2017) to create a tissue-specific expression body map of human lncRNAs across all the genes in the GTF annotation file from lncRNAKB. We downloaded raw paired-end RNA-seq data (FASTQ files – GTEx Release v7) from the dbGap portal (study_id=phs000424.v7.p2) of 31 solid organ human normal tissues. For each solid tissue, quality control of paired-end reads were assessed using FastQC tools (Andrews), adapter sequences and low-quality bases were trimmed using Trimmomatic (Bolger *et al.*, 2014) and aligned to latest version of the human reference genome (H. sapiens, GRCh38) using the latest version of HISAT2 (Kim *et al.*, 2015), which is a splice-aware aligner that maps reads to the reference. Using uniquely aligned reads to the human genome, gene-level expression (raw read counts) were generated with the featureCounts software (Liao *et al.*, 2014), which assigns reads to features in a fast and parallelizable framework. After visualizing the distribution of uniquely mapped paired-

end reads assigned to genes across all the GTEx samples we chose to exclude samples with < 10^6 reads assigned to genes. In addition, there were samples with data that we could not map or download. We normalized the raw read counts to Transcripts Per Kilobase Million (TPM) (Wagner *et al.*, 2012) (https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/). For each gene in the lncRNAKB annotation database, we created a box plot distribution to visualize its tissue-specific expression pattern across all tissues.

Tissue-specificity scores In addition to gene expression visualization, we calculated two tissue-specificity metrics (Tau and Preferential Expression Measure (PEM)) (Kryuchkova-Mostacci and Robinson-Rechavi, 2017; Russ and Futschik, 2010) using the normalized TPM expression values across all genes and tissues (no filter applied). Tau was calculated as follows (Yanai *et al.*, 2005): $\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1}$; $\hat{x}_i = \frac{x_i}{\underset{1\leq i\leq n}{max(x_i)}}$ where (x_i= expression of the gene in tissue i, n= number of tissues). Tau summarizes in a single number whether a gene is tissue-specific or ubiquitously expressed across all tissues. PEM was calculated as follows (Huminiecki *et al.*, 2003): $PEM = log_{10}\left(\frac{\sum_{i=1}^{n} s_i}{s_i} * \frac{x_i}{\sum_{i=1}^{n} x_i}\right)$ where (x_i= expression of the gene in tissue i, s_i= summary of the expression of all genes in tissue i, n= number of tissues). PEM shows for each tissue separately how specific the gene is to that tissue. The PEM scores the expression of a gene in a given tissue in relation to its average expression across all other genes and tissues. To compute Tau and PEM, we calculated and used the average expression across all replicates for each gene by tissue. All genes that were not expressed in at least one tissue were removed

from the analysis. For comparison purposes with Tau, for each gene, we used the maximum specificity value of PEM across all tissues (normalized between 0 to 1, using the maximum across all genes).

Principal Component Analysis To explore gene expression similarity between tissues and across GTEx samples as well as summarize lncRNAs tissue-specific expression we performed a principal component analysis (PCA) (Son *et al.*, 2018). We used the normalized TPM expression values, transformed by taking the $log_2(TPM + 1)$, across all lncRNAs (n = 77,199) and tissues (n = 31) (no filters applied). We used the prcomp package in R (Team, 2012).

Genotype file processing We also downloaded whole genome sequence (WGS) data in blood-derived DNA samples (Variant Call Format (VCF) file – GTEx Release v7) from the dbGap portal (study_id=phs000424.v7.p2) to conduct tissue-specific expression quantitative trait loci (eQTL) analysis. We created an eQTL body map of human lncRNAs, across all the genes in the GTF annotation file from lncRNAKB. We preprocessed the VCF file using the following steps with a combination of PLINKv1.9 (Chang *et al.*, 2015; Purcell and Chang) vcfv0.1.15 (Danecek *et al.*, 2011) and bcfv1.9 tools (Narasimhan *et al.*, 2016): (i) removed indels; (ii) excluded missing and multi-allelic variants; (iii) selected "FILTER == 'PASS'" variants; (iv) excluded variants with minor allele frequency (MAF) < 5%; (v) updated the coordinates of single nucleotide polymorphisms (SNPs) using the UCSC liftOver tool (Casper *et al.*, 2018) from hg19 to hg38 (latest genome build); (vi) changed the SNPs IDs to dbSNP (Sherry *et al.*, 2001) rsID using dbSNP Build 151; (vii) converted to bed, bim and fam format. For each solid

tissue, we only selected subjects that had both WGS data and gene expression data. We generated a subset of the VCF files by tissue and re-calculated the MAF to exclude variants with MAF < 5%. After converting to ped and map format, we ran principal component analysis (PCA) on each tissue to get a set of genotype covariates using eigensoftv6.1.4 (Price *et al.*, 2006; Patterson *et al.*, 2006).

eQTL analysis For each solid tissue, we implemented a two-step filtering approach using the raw gene expression counts matrix quantified across all the genes in the lncRNAKB GTF annotation file. First, we filtered genes based on the normalized TPM expression values, keeping genes with TPM > 0.50 in at least 20% of the samples. Second, we filtered genes based on the raw counts, keeping genes with counts > 2 in at least 20% of the samples. The edgeR (Robinson *et al.*, 2010) package in R (Team, 2012) was used to process the filtered read counts into log2 counts per million (log2CPM) and the limma-voom R package (Ritchie *et al.*, 2015) (Law *et al.*, 2014) was used to normalize the data between samples using trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010). The expression files were then sorted by gene start and stop, compressed with BGZIP and indexed with TABIX (Li, 2011). Only tissues with > 80 samples were included in the *cis*-eQTL analysis. In the eQTL analysis, we included the first five principal components (PCs) that explained the most variation in the genotype data by looking at their scree plots by tissue. Sex was also included as a covariate. Within each tissue, *cis*-eQTLs were identified by linear regression, as implemented in FastQTLv2.0 (threaded option) (Ongen *et al.*, 2016), adjusting for the five PCs and sex. We restricted our search to variants within 1 Megabase (Mb) of the transcription start site

(TSS) of each gene and in the tissue of analysis. To evaluate the significance of the most highly associated variant per gene we used the adaptive permutations option in FastQTL between 1000 and 10000 permutations. Once we obtained the permutation p-values for all the genes, we accounted for multiple testing to determine the significant *cis*-eQTLs. We used the Benjamini and Hochberg correction method (Haynes, 2013) to calculate the false discovery rate (FDR) in R statistical programming language (R) (Team, 2012). For each tissue, all *cis*-eQTL results were visualized using a manhattan plot created using the qqman package in R (Turner, 2014).

Conservation Analysis

Conservation of exons between protein-coding genes and lncRNAs in the lncRNAKB annotation database was analyzed using the bigWigAverageOverBed (Pohl and Beato, 2014) and the cons30way (hg38) track (Siepel *et al.*, 2005) both downloaded from the UCSC genome browser. This track shows multiple alignments of 30 vertebrate species and measurements of evolutionary conservation using two methods (phastCons and phyloP (Cooper *et al.*, 2005)) from the PHAST package (Hubisz *et al.*, 2011) for all thirty species. The multiple alignments were generated using multiz (Blanchette *et al.*, 2004) and other tools in the UCSC/Penn State Bioinformatics comparative genomics alignment pipeline. An exon-level BED file was created using the lncRNAKB GTF annotation file separately for protein-coding genes and lncRNAs. We merged overlapping exons within transcripts to avoid counting conservation scores of overlapping base pairs more than once. For each exon, the bigWigAverageOverBed

function calculates the average conservation score across all base pairs. Using boxplots we visualized and compared the average conservation score differences between lncRNAs and protein-coding exons.

Functional characterization of lncRNAs using a network-based approach

We used the "guilt by association" principle to functionally characterize lncRNAs in the lncRNAKB across all 31 solid organ human normal tissues in the GTEx data (Ehsani and Drabløs, 2018a). This method is widely used to identify well-annotated genes that seem to be involved in some of the same processes as a given un-annotated gene. It is based on comparison of gene expression profiles between lncRNAs and mRNAs using metrics such as Pearson or Spearman correlation, applying a specific cut-off, and performing enrichment analysis of annotation terms in the most highly ranked mRNAs. The significantly enriched terms can be used as an estimate of annotation for the lncRNAs.

Using the filtered log2CPM and TMM normalized gene expression data (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs), subsection: eQTL analysis), we used the weighted gene co-expression network analysis (WGCNA) approach (Langfelder and Horvath, 2008) as implemented in the Co-Expression Modules identification Tool (CEMiTool) package in R (Russo *et al.*, 2018) to identify modules of lncRNA-mRNA pairs that are co-expressed and therefore likely work in concert to carry out various biological functions. Additionally, we filtered the gene expression data by log2CPM > 2 in at least 20% of the samples to avoid random

correlations. In the CEMiTool, default parameters were mostly used with the following exceptions: (i) Pearson method was used for calculating the correlation coefficients, (ii) the network type used was unsigned, (iii) no filter was used for the expression data, (iv) applied Variance Stabilizing Transformation (VST) and the correlation threshold for merging similar modules were set to 0.90. After identifying co-expressed gene modules, we performed over-representation analysis (ORA) by module based on the hypergeometric test (Yu *et al.*, 2012) that can be used to reveal if a set of co-expressed genes is enriched for genes belonging to known pathways or functions. We used Gene Ontology (GO) pathways (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2019; Gene Ontology Consortium, 2015) to check for overrepresentation of genes and determined the most significant module functions based on pathways FDR q-value $\leq$ 0.05 (Storey, 2002). The background set used for the pathway enrichment analysis was genes represented across all GO pathways. To visualize the interactions between the genes in each co-expression module, we output the top 25 most notable pathways across all modules and the entire module adjacency/correlation matrix (correlations $>$ 0.20), available for downloading. We filtered the module adjacency matrices based on correlations $>$ 0.20 across all genes in each pathway, and created a JSON file (one per pathway) to produce interactive networks using Cytoscape v3.6.0 JavaScript modules (Shannon *et al.*, 2003). This will give users the ability to visualize the potential biological functionality of lncRNAs in lncRNAKB across all 31 solid organ human normal tissues.

RESULTS

Database Content

Figure 2.3 illustrates the results of the lncRNAKB that are discussed in detail in each section below. Table 2.1 summarizes the number of lncRNAs features (genes *n*=77,199, transcripts *n*= 224,286 and exons *n*= 611,340) in the lncRNAKB GTF annotation file.

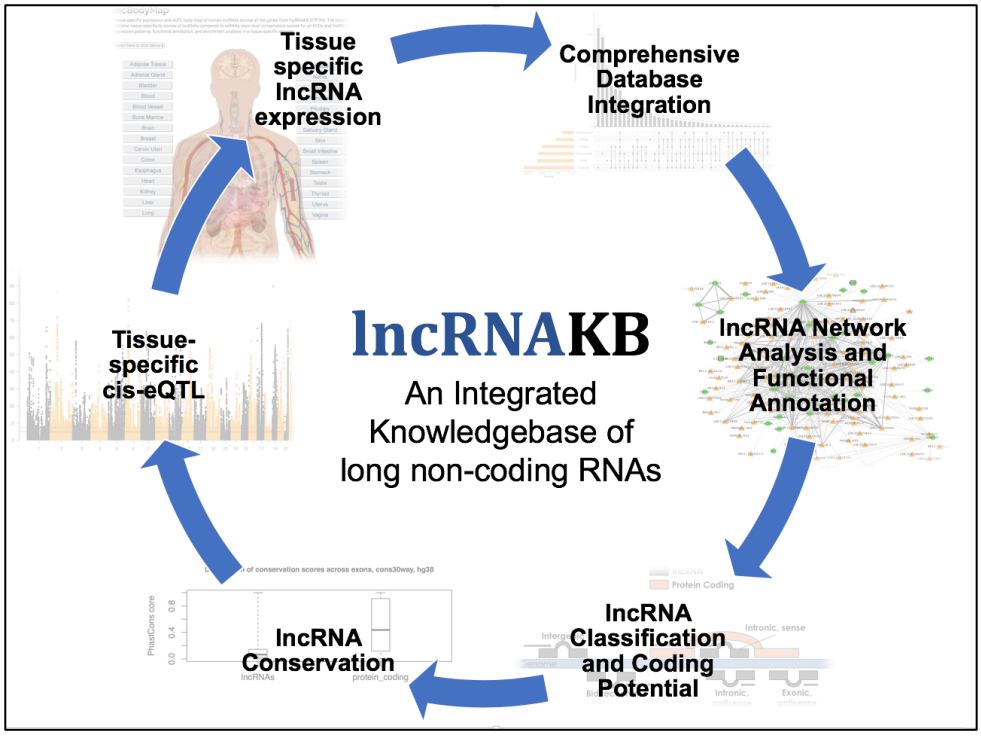


Figure 2.3: Overview of specific components of the lncRNAKB.
All components of the lncRNAKB which provide valuable information on lncRNAs and are freely available for viewing and downloading on the web resource.

Downloadable, searchable and viewable annotation database in gene transfer format (GTF)

The final merged GTF annotation in lncRNAKB has 99,717 genes, 530,947 transcripts, 3,513,069 exons (include both PCGs and lncRNAs) and is freely available to download from the website. We generated a UCSC Genome Browser custom track of the lncRNAKB GTF annotation file. The UCSC Genome Browser is accessible within the lncRNAKB website via a html IFRAME to view the lncRNAKB GTF annotation file in combination with other tracks on the Genome Browser. On the website, users can search the lncRNAKB annotation database by ID, chromosomes, gene start and stop coordinates, gene type, gene names, names of the six lncRNAs sources or any other descriptor and download the search results in .csv format.

Table 2.2 shows the results of the cumulative stepwise intersection method across the six lncRNAs annotation databases compared to the reference (CHESS2.1) at the gene level. NONCODEv5.0 and MiTranscriptomev2 added 20,700 and 15,164 genes respectively, which was a substantial contribution. While CHESS2.1 already incorporated data from FANTOM5.0.v3, based on the cumulative stepwise intersection method we added additional 7,157 genes from FANTOM5.0.v3. LNCipedia5.2 on the other hand added 10,506 genes. We arbitrarily chose the order of the annotation databases for intersection therefore. The last source, BIGTranscriptomev1 contributed only 333 genes which indicates that there was extensive overlap with other annotation databases.

To be specific, after merging all six annotation databases, we define “overlap” (between one or more sources for a gene) if transcripts (with exons) were added to an

overlapping gene (within or fully overlapping) or a new gene (with transcript and exons) was added to the annotation entirely. Figure 2.4 illustrates the overlap at the gene level after the cumulative stepwise intersection method was applied across all six sources. 5,295 genes overlapped between all six sources. In addition, there was considerable overlap between different annotation databases. All of LNCipedia5.2 genes overlapped with one or more of the annotation databases. NONCODEv5.0 added the highest number of non-overlapping genes (*n*=16,080) followed by MiTranscriptomev2 (*n*=14,620). BIGTranscriptomev1 added only 333 unique gene entries due to its overlap with genes in the other databases. CHESS2.1 was used as the reference annotation database and contains protein-coding (*n*=20,352) and lncRNAs genes (*n*=18,897). However, from Figure 2.4, we observed that the number of non-overlapping genes added from CHESS2.1 is 9,595, which could possibly indicate that we added non-coding transcripts from overlapping lncRNAs in other annotation databases to the protein-coding genes. Supplementary Table 2.1a and 2.2b shows the number of transcripts and the sources of annotation databases at gene level for protein-coding genes between CHESS2.1 and lncRNAKB, respectively. Comparing Supplementary Table2.1a and Supplementary Table2.2b showed that the number of transcript entries for the protein coding genes in lncRNAKB was much higher than that in chess (approximately 40,330 more transcript entries in lncRNAKB compared to CHESS2.1). This suggests that a good proportion of the lncRNAs transcripts (~15%) overlap with or fall within the boundary of protein coding genes. Supplementary Table 2.2a and 2.2b shows the number of transcripts and the sources of annotation databases at gene level for non-coding genes between

CHESS2.1 and lncRNAKB, respectively. By comparing all 4 tables, we show that we have effectively added numerous non-coding genes (*n*=53,941) and non-coding transcripts (*n*=207,681) from different lncRNAs annotation databases.

Table 2.2: Results of the cumulative stepwise intersection method.
Results of the cumulative stepwise intersection method across the six lncRNAs annotation databases compared to the reference (CHESS2.1) at the gene level.

[1]The original number of genes in the sources shown here are slightly less than the actual downloaded GTF/GFF annotation files (Table 2.1) because we removed redundant genes and transcripts records (see Materials and Methods: Data integration).

CHESS2.1	**lncRNAKB_hg38_v2.gtf**	**Source**	**[1]Original number of genes in source after applying the lncRNAKB redundancy filter**
46,421	45,857	CHESS2.1	46,421
	7,157	FANTOM5.0.v3	21,457
	10,506	LNCipedia5.2	40,188
	20,700	NONCODEv5.0	63,055

		15,164	MiTranscriptomev2	45,282
		333	BIGTranscriptomev1	13,525
Total	**46,421**	**99,717**		**229,928**

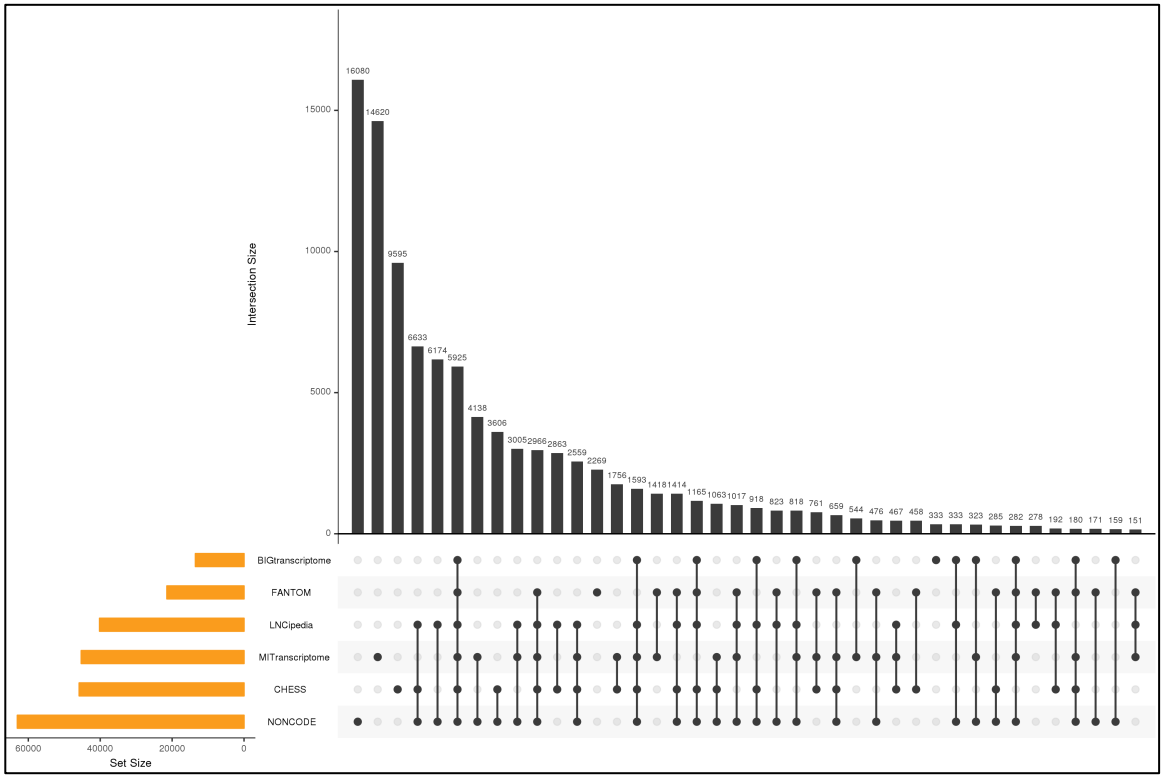


Figure 2.4: Upset plot showing the overlap of all six lncRNAs annotation databases. Upset plot showing the overlap of all six lncRNAs annotation databases at the gene level, after the cumulative stepwise intersection method across. The orange bars indicate the total number of genes in each source before merging. The black bars indicate the total number of genes present within a database or shared between databases indicated by black dots present below the x-axis of the plot. Genes uniquely contributed by a single database would be represented as a single dot that horizontally aligns with the respective database. Black dots connected by lines indicate the number of databases that share the genes represented in the bar plot.

Classification/annotation and coding potential of lncRNAs using Random Forest

The final merged GTF annotation in lncRNAKB has 99,717 genes, 530,947 transcripts, 3,513,069 exons (include both PCGs and lncRNAs). After executing the FEELnc filter module on the lncRNAKB GTF, the output had 96,539 genes, 311,241 transcripts and 1,200,236 exons that were not filtered (transcripts > 200 bp long and < 75% overlap with protein-coding transcripts) and considered to be "candidate lncRNAs." The coding potential score (CPS) cut-off determined by the Random Forest (RF) classification (see Methods: Classification/Annotation and coding potential of lncRNAs using Random Forest: Coding Potential) on the training data was 0.434 (separating protein-coding (mRNAs) versus lncRNAs transcripts). Based on this cut-off, 83,190 genes, 219,324 transcripts, 622,122 exons were classified as lncRNAs and 31,402 genes, 91,845 transcripts, 577,978 exons as protein-coding. The classification module categorized 141,394 lncRNAs transcripts as GENIC (when the lncRNA transcript overlaps an mRNA/protein-coding transcript from the reference annotation file) and 50,540 as INTERGENIC (lincRNAs). Table 2.3 summarizes the results of the classifier module with a breakdown of interactions between the two types of lncRNAs and their partner mRNAs/protein-coding transcripts. The lincRNAs are, on average 23kb away from their mRNA partner.

Table 2.3: Summary of classification of lncRNAs transcripts.
Summary of classification of lncRNAs transcripts with respect to the localization and the direction of transcription of proximal RNA transcripts. The legend below explains the categories in detail:

[1]GENIC: when the lncRNA gene overlaps an RNA gene from the reference annotation file
[2]INTERGENIC (lincRNA): otherwise

GENIC type :
Then exonic or intronic locations:
[1a]Overlapping subtype: the lncRNA partially overlaps the RNA partner transcript
[1b]Containing subtype: the lncRNA contains the RNA partner transcript
[1c]Nested subtype: the lncRNA is contained in the RNA partner transcript

INTERGENIC type:
[2a]Divergent subtype: the lncRNA is transcribed in head to head orientation with RNA partner transcript

- Then upstream or downstream locations

[2b]Convergent subtype: the lncRNA is oriented in tail to tail with orientation with RNA partner transcript

- Then upstream or downstream locations

[2c]Same_strand subtype: the lncRNA is transcribed in the same orientation with RNA partner transcript

- Then upstream or downstream locations

	[1]GENIC			
	[1a]Overlapping	**[1b]Containing**	**[1c]Nested**	**Total**
Antisense Exonic	9,326	1,816	3,552	**14,694**
Antisense Intronic	1,302	1,284	8,330	**10,916**
Sense Exonic	29,942	42,160	29,087	**101,189**
Sense Intronic	327	994	13,274	**14,595**
Total	**40,897**	**46,254**	**54,243**	**141,394**
	[2]INTERGENIC			
	[2a]Convergent	**[2b]Divergent**	**[2c]Same_Strand**	**Total**
Upstream	-	14,930	13,408	**26,470**
Downstream	11,540	-	10,662	**24,070**
Total	**11,540**	**14,930**	**24,070**	**50,540**

Tissue-specific expression profiling and expression quantitative trait loci (eQTLs)

Expression profiling Supplementary Table 2.3 shows the number of RNA-seq samples we analyzed across 31 solid organ human normal tissues from GTEx (*n*=9,425). Supplementary Table 2.4 shows the summary statistics of alignment (total number of paired-end reads, total number of uniquely aligned paired-end reads, unique and overall alignment rate) across all samples analyzed by tissue. Supplementary Table 2.5 shows the summary statistics of quantification (total gene count, total number of uniquely aligned paired-end reads used for quantification, total number of uniquely aligned paired-end reads assigned to genes and proportion of successfully assigned paired-end reads to genes) across all RNA-seq samples analyzed by tissue. Supplementary Figure 2.2 shows the distribution of uniquely aligned paired-end reads assigned to genes across all samples. Bars highlighted in red show the numbers of samples with $< 10^6$ reads assigned to genes (*n*=351) that were excluded from further analysis. In the lncRNAKB, users can visualize the normalized gene expression levels (TPM) across 31 solid organ human normal tissues by searching for any gene. Figure 2.5 shows an example box plot distribution of gene *NPPB* (natriuretic peptide B) for visualization. *NPPB* had a Tau (overall) and PEM score (top five highest positive tissue-specificity score in the heart tissue) of 1 and 1.49 respectively (see subsection: Tissue-specificity Scores). It functions as a cardiac hormone and plays a key role in cardiac homeostasis. A high concentration of this protein in the bloodstream is indicative of heart failure. Even though *NPPB* is categorized as a PCG, it has three transcript isoforms that are characterized as lncRNAs. Users can download these boxplots by gene and download genome-wide gene expression matrices (raw counts

and TPM) in text format across all 31 solid organ human normal tissues in the lncRNAKB.

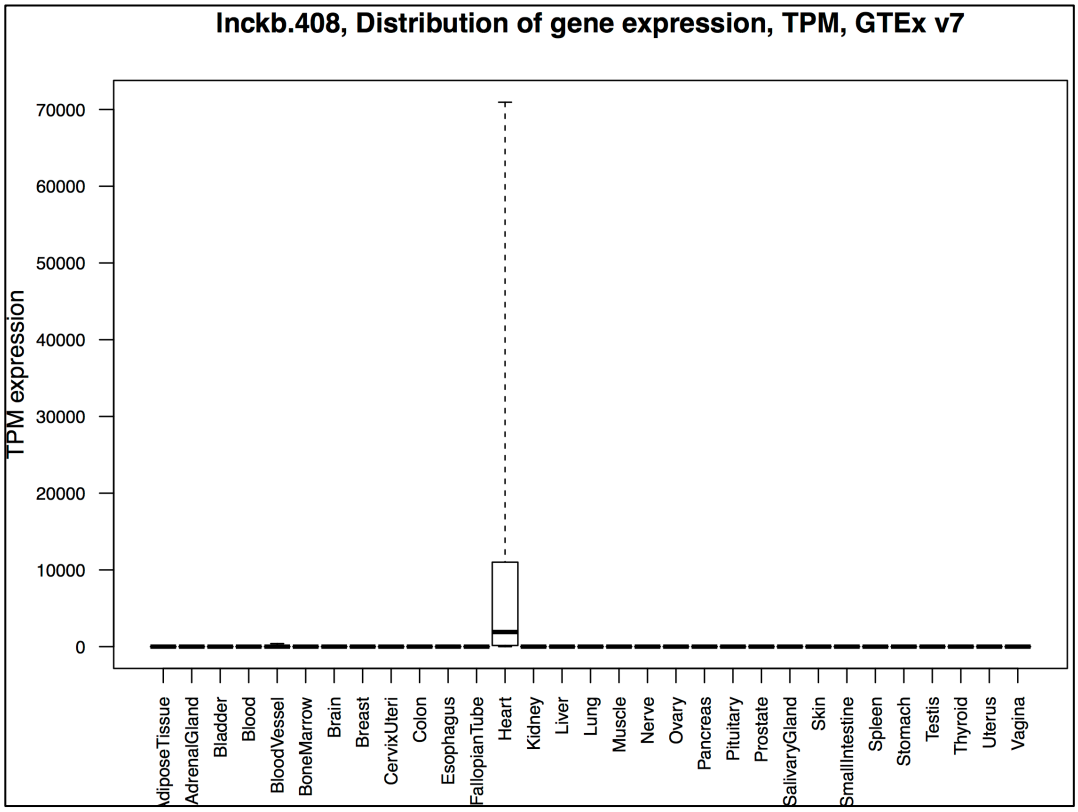


Figure 2.5: Gene expression box plot distribution.
Gene expression box plot distribution of gene *NPPB* (natriuretic peptide B). The x-axis represents the 31 solid organ human normal tissues from GTEx and y-axis is the TPM expression. *NPPB* was ranked among the top five heart-specific genes.

Tissue-specificity scores Figure 2.6 shows the density distribution of tissue-specificity metrics (Tau and Preferential Expression Measure (PEM)) across protein-coding genes (PCGs) and lncRNAs in the lncRNAKB annotation database as a comparison. The tissue-specificity scores vary from 0 to 1, where 0 means broadly expressed, and 1 is specific. Figure 2.6 displays the maximum specificity value of PEM

among all tissues while Tau is calculated and displayed across all tissue (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): subsection: Tissue-specificity scores). Overall, Figure 2.6 shows that lncRNAs have higher tissue-specificity compared to PCGs and the possibility that lncRNA tissue-specificity could be driven by individual tissues. We created a clickable human body map highlighting 31 solid organ human normal tissues from GTEx, in which users can click on any tissue and rank the genes by PEM score to get a potential list of tissue-specific lncRNAs. Supplementary Figure 2.3 shows the density distributions of PEM scores across PCGs and lncRNAs as a comparison in the lncRNAKB annotation database by tissue. It reports a positive value for genes over-expressed and a negative value for genes under-expressed in the specific tissue, respectively. Supplementary Figure 2.3 shows that the tissue-specificity and gene expression between PCGs and lncRNAs varies among tissues but, most of the lncRNAs have similar expression patterns compared to PCGs. Supplementary Figure 2.3 shows that individually lncRNAs might be not very specific to one tissue but, overall (as shown in Figure 2.6) these are tissue-specific. All tissue-specific figures and score files are available for viewing and downloading individually on the lncRNAKB website in tissue-specific pages.

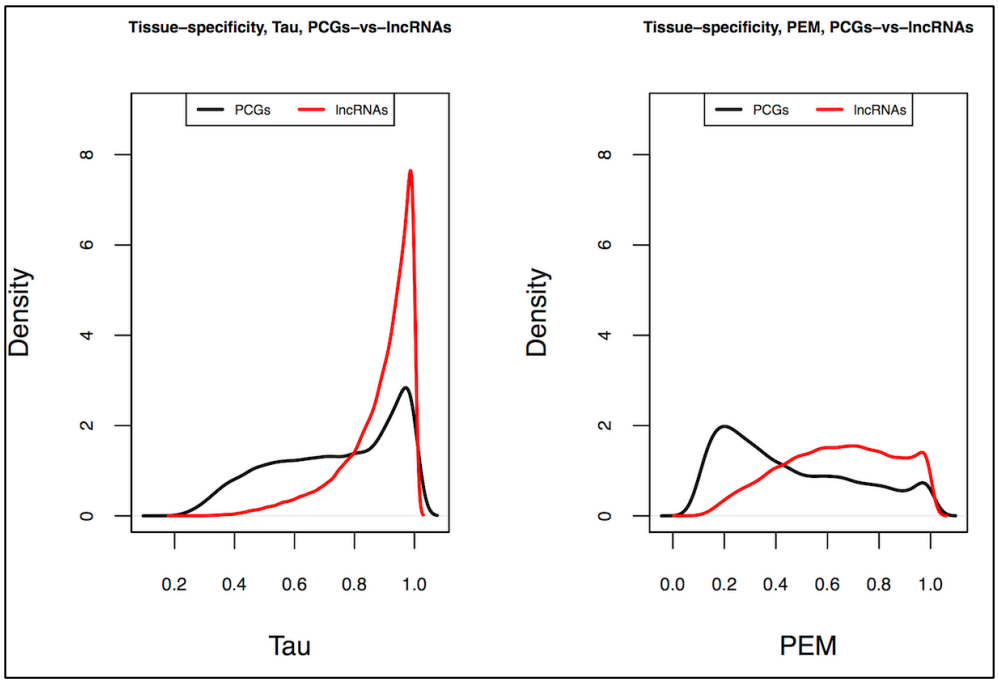


Figure 2.6: Distribution of tissue-specificity scores.
Distribution of tissue-specificity scores (Tau [left] and PEM [right]) with data for RNA-seq of 31 solid organ human normal tissues from GTEx across protein-coding genes (PCGs) and lncRNAs in the lncRNAKB as a comparison. The tissue-specificity scores varies from 0 to 1, where 0 means broadly expressed, and 1 is specific. Graph created with density function from R, which computes kernel density estimates.

Principal Component Analysis Figure 2.7 shows the results of PCA using $log_2(TPM + 1)$ transformed lncRNAs expression data across all tissues in the lncRNAKB. Tissues show a characteristic transcriptional signature, as revealed by PCA of lncRNA expression. The separation is between nonsolid (blood) and solid tissues and, within solid tissues, brain and testis are the most distinct. This is an additional confirmation that lncRNAs are tissue-specific.

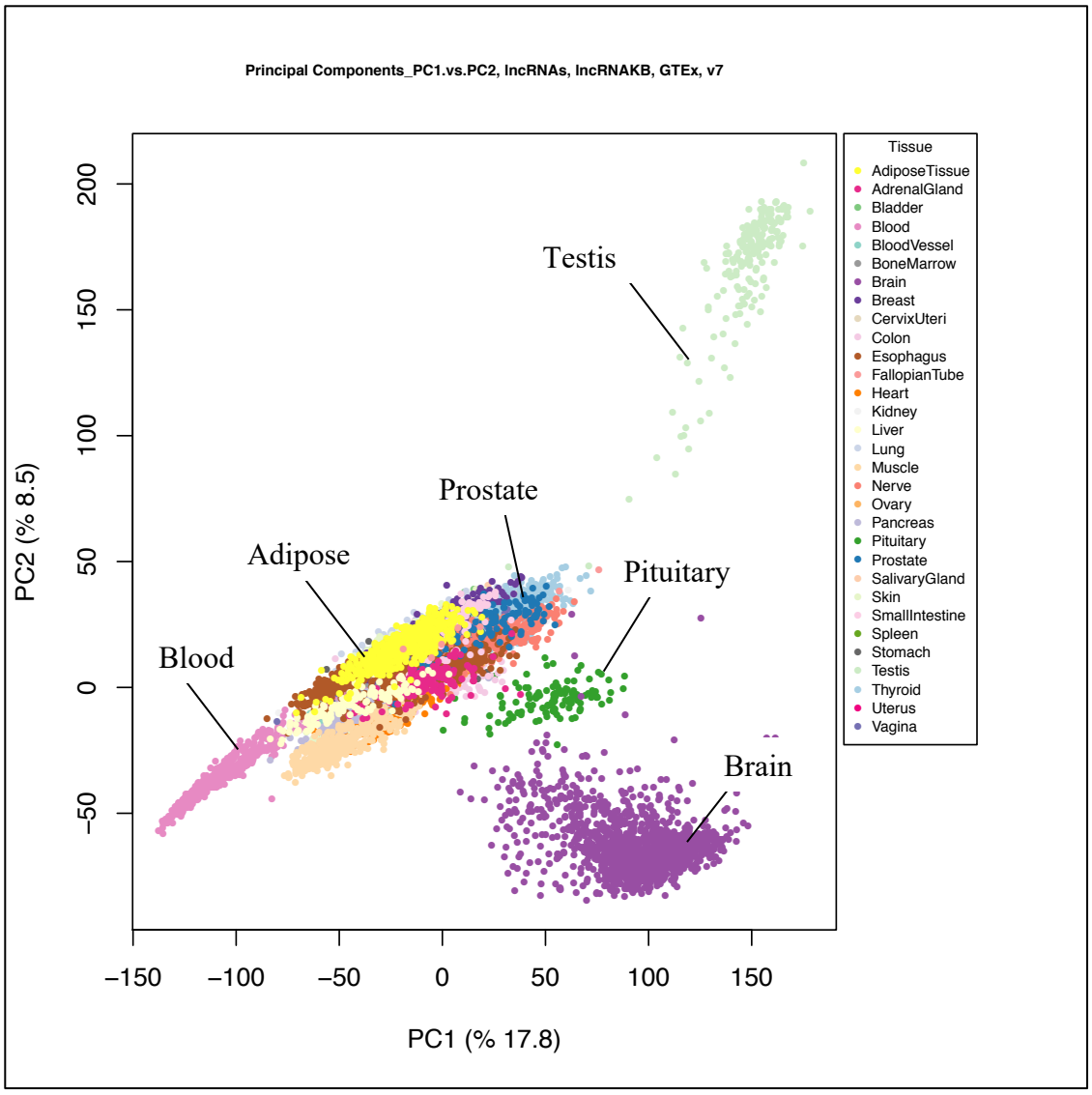


Figure 2.7: Principal Component Analysis of GTEx samples using lncRNA expression. Principal Component Analysis (PCA) of GTEx samples based on lncRNA expression. PCA of all samples based on expression levels of lncRNAs ($\boldsymbol{log_2(TPM+1)}$ transformed). Expression of lncRNAs alone also recapitulates tissue types.

eQTL analysis Table 2.4 summarizes the results of the *cis*-eQTL analysis. The number of RNA-seq samples with WGS data across 31 solid organ human normal tissues

from GTEx were (*n*=5,502). 25 tissues had > 80 samples (*n*=5,393) with WGS and included in the *cis*-eQTL analysis. After pre-processing the WGS VCF file (initially with 50,862,464 variants) across all samples (*n*=652), 5,835,187 SNPs were leftover for the *cis*-eQTL analysis (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): Genotype File Processing). For each tissue, Table 2.4 shows the number of samples (stratified by sex), the number of SNPs available after preprocessing, the number of genes that met the TPM threshold criteria from the RNA-seq data (PCG and lncRNAs), the total number of SNP-gene pairs that were tested within 1 Mb of the transcription start site (TSS) of each gene and the number of top *cis*-eQTL genes that met the permutation p-value <= 0.05 threshold after the FastQTLv2.0 adaptive permutations approach (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): eQTL Analysis). In the lncRNAKB, users can visualize the *cis*-eQTL results by tissue via a manhattan plot. Figure 2.8 shows an example plot from the heart tissue. All figures are available for viewing and download individually on the lncRNAKB website in tissue-specific pages. Users can also download genome-wide compressed *cis*-eQTL results files (text format) by tissue (all SNP-gene pairs and top SNP-gene pairs generated via permutation).

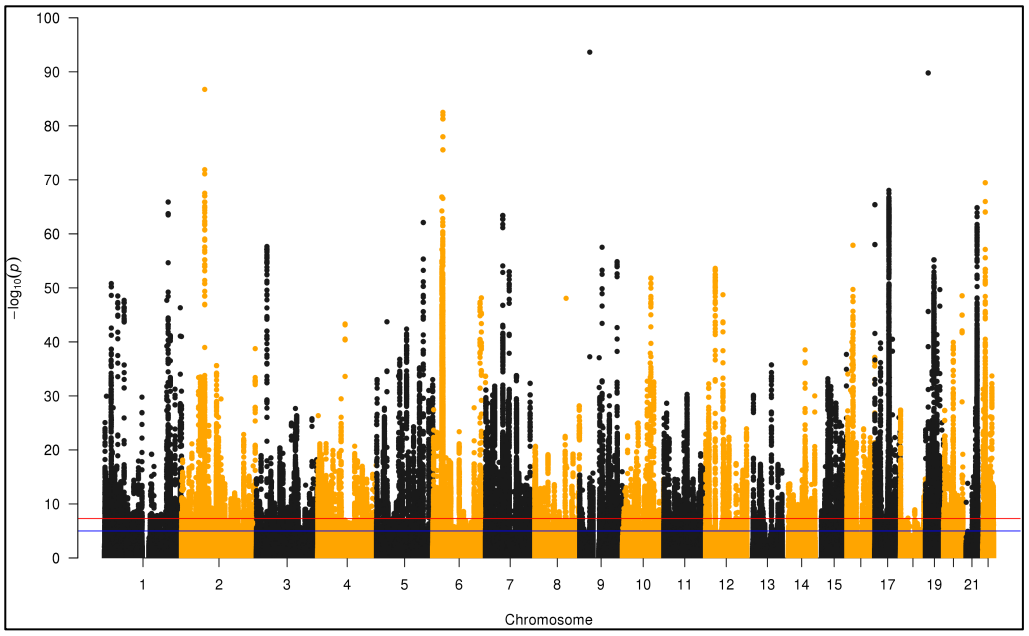


Figure 2.8: Manhattan plot illustrating the results of the cis-eQTL analyses.
Manhattan plot illustrating the results of the *cis*-eQTL analysis from the heart tissue. The x-axis are the chromosomes and each dot on the y-axis represents the *cis*-eQTL -log10 (p-values) of the SNP-gene pairs that were tested within 1 Mb of the TSS of each gene.

Table 2.4: Summary results of the *cis*-eQTL results available from the lncRNAKB.
Summary results of the *cis*-eQTL results available from the lncRNAKB. Tissues with < 80 samples are shown here but, were excluded from the analysis.

Tissue	Number_of_RNA_seq_samples_with_WGS	Number_of_Males	Number_of_Females	Number_of_SNPs_with_MAF_greater_than_0.05	Total_number_of_genes_passed_filter	Total_number_of_PCGs	Total_number_of_lncRNAs	Total_SNP_gene_pairs_eQTLs	Total_SNP_gene_pairs_with_permutation_pvalue_less_than_0.05
Adipose_Tissue	363	220	143	5,952,169	27,029	15,175	11,854	54,871,184	5,766
Adrenal_Gland	146	82	64	5,886,806	25,943	14,973	10,970	51,879,876	4,077
Bladder	9	4	5	5,462,615	28,695	15,597	13,098	-	-
Blood	356	226	130	5,953,536	18,412	11,788	6,624	37,414,178	2,877
Blood_Vessel	378	241	137	5,963,536	25,614	14,770	10,844	51,947,442	5,854
Bone_Marrow	-	-	-	-	-	-	-	-	-
Brain	170	116	54	5,857,467	31,339	16,148	15,191	62,844,553	3,488
Breast	184	102	82	5,901,708	28,839	15,680	13,159	58,130,064	4,267
Cervix_Uteri	8	0	8	5,522,234	28,706	15,649	13,057	-	-
Colon	250	148	102	5,907,992	28,297	15,781	12,516	57,063,773	4,767
Esophagus	353	221	132	5,941,386	26,803	15,439	11,364	54,314,052	4,815
Heart	251	163	88	5,913,705	24,959	14,788	10,171	50,153,256	4,375
Kidney	29	23	6	5,742,588	28,917	15,726	13,191	-	-
Liver	118	77	41	5,871,833	23,846	14,204	9,642	47,689,780	2,759
Lung	274	182	92	5,926,605	29,045	15,744	13,301	58,884,074	5,461
Muscle	359	220	139	5,962,131	22,042	13,558	8,484	44,548,539	4,454
Nerve	268	174	94	5,941,274	29,326	15,472	13,854	59,363,204	7,416
Ovary	99	0	99	5,873,449	27,292	14,845	12,447	54,588,663	3,466
Pancreas	167	98	69	5,905,087	23,569	14,210	9,359	47,408,959	#N/A
Pituitary	108	76	32	5,814,865	30,586	15,848	14,738	60,707,019	3,949
Prostate	101	0	101	5,810,666	30,373	15,931	14,442	60,377,553	#N/A
Salivary_Gland	63	43	20	5,771,591	28,409	15,679	12,730	-	-
Skin	442	278	164	5,966,760	27,316	15,442	11,874	55,698,051	6,210
Small_Intestine	90	54	36	5,777,092	30,046	15,950	14,096	59,426,622	2,987

Spleen	108	62	46	5,874,443	28,284	14,969	13,315	56,914,604	4,743
Stomach	182	104	78	5,890,077	26,974	15,530	11,444	54,242,450	3,804
Testis	171	0	171	5,875,543	47,909	17,777	30,132	98,376,057	8,951
Thyroid	286	183	103	5,941,584	29,715	15,604	14,111	60,217,108	7,611
Uterus	82	0	82	5,795,583	28,175	15,166	13,009	55,748,102	3,037
Vagina	87	0	87	5,837,620	28,423	15,629	12,794	56,861,978	2,865

Conservation Analysis

Figure 2.9 shows the two box plot distributions of exon sequence conservation scores comparing protein-coding and lncRNAs in the lncRNAKB annotation database. Overall, it shows that exons in the protein-coding genes have higher mean sequence conservation scores compared to lncRNAs. Users can download the conservation scores across exons on the lncRNAKB website stratified by protein-coding genes and lncRNAs. In addition, on the gene display page we will present the conservation scores for exons in that gene.

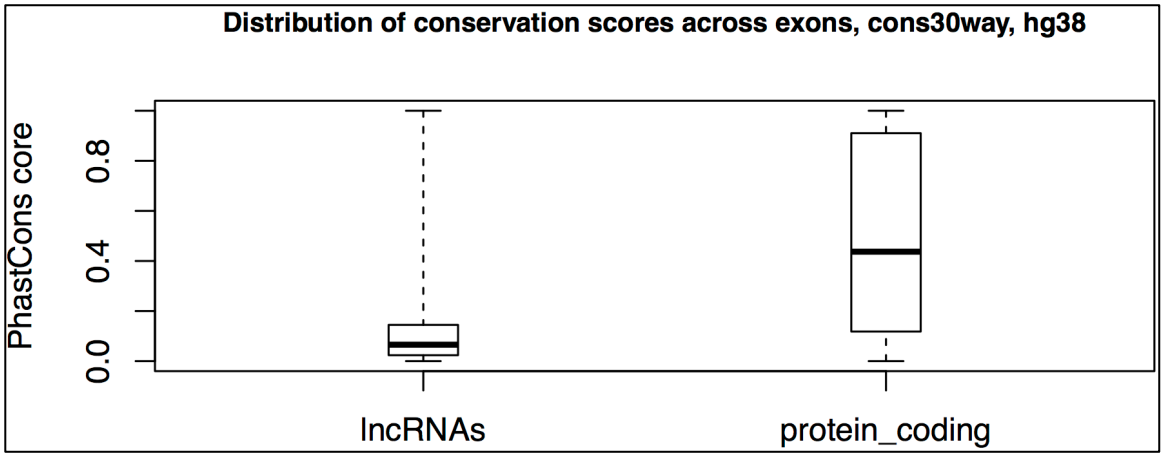


Figure 2.9: Distribution of mean PhastCons exon sequence conservation scores. Distribution of mean PhastCons exon sequence conservation scores across lncRNAs and protein-coding genes in the lncRNAKB.

Functional characterization of lncRNAs using a network-based approach

Supplementary Table 2.6 summarizes the results of the WGCNA analysis across the 31 solid organ human normal tissues using the GTEx RNA-seq data. After

filtering genes with low expression in the lncRNAKB annotation database by tissue (See Methods: Functional characterization of lncRNAs using a network-based approach), on average, we had gene expression data across approximately 14,699 protein-coding genes and 3,389 lncRNAs per tissue. We identified 1,208 lncRNA-mRNA co-expression modules across all tissues (on average approximately 43 modules per tissue). On average, across all tissues, each module had approximately 487 genes including 92 lncRNAs, indicating favorable co-expression of lncRNAs with PCGs. Supplementary Table 2.7 summarizes the results of the over-representation analysis (ORA) based on the hypergeometric test using the Gene Ontology (GO) pathways across all the modules identified in the 31 solid organ human normal tissues. It displays the number of GO pathways tested, number of pathways with p-value ≤ 0.05 and FDR q-value ≤ 0.05 in all modules by tissue. On average, across all modules, each tissue had approximately 10,849 and 2,592 pathways (out of approximately 83,240 pathways that were tested on average across all modules per tissue) with p-value ≤ 0.05 and q-value ≤ 0.05 respectively, indicating significant enrichment of biological processes within these modules.

Supplementary Table 2.8 shows the results of WGCNA in heart tissue for all lncRNA-mRNA co-expression modules identified. There were 61 modules identified in the heart using gene expression data across 16,882 protein-coding genes and 2,762 lncRNAs. Supplementary Table 2.8 separates the number of genes and lncRNAs in each module, representing the size of each. It displays a list of lncRNAs and top 20 hub genes (genes with highest connectivity) in each module. Supplementary Table 2.9 shows the results of all GO pathways enriched in the heart tissue by module. There were several

significant pathways identified (q-value <= 0.05) with many of these involved in heart related biological processes. Figure 2.10 highlights the network figure created using Cytoscape for module M2 identified in the heart tissue. This module is involved in heart-specific processes such as heart growth, development and contraction. The network has 148 genes (34 protein-coding and 106 lncRNAs) after filtering the adjacency matrix with correlations < 0.20 and "heart development" specific pathways/genes. The orange triangles and green circles/nodes represent lncRNAs and mRNAs respectively. The thickness of the edges highlights the connectivity (degree) between nodes. The relatively strong connections of several lncRNAs to PCGs in this network hypothetically shows that these could be potentially involved in the same heart development specific biological processes.

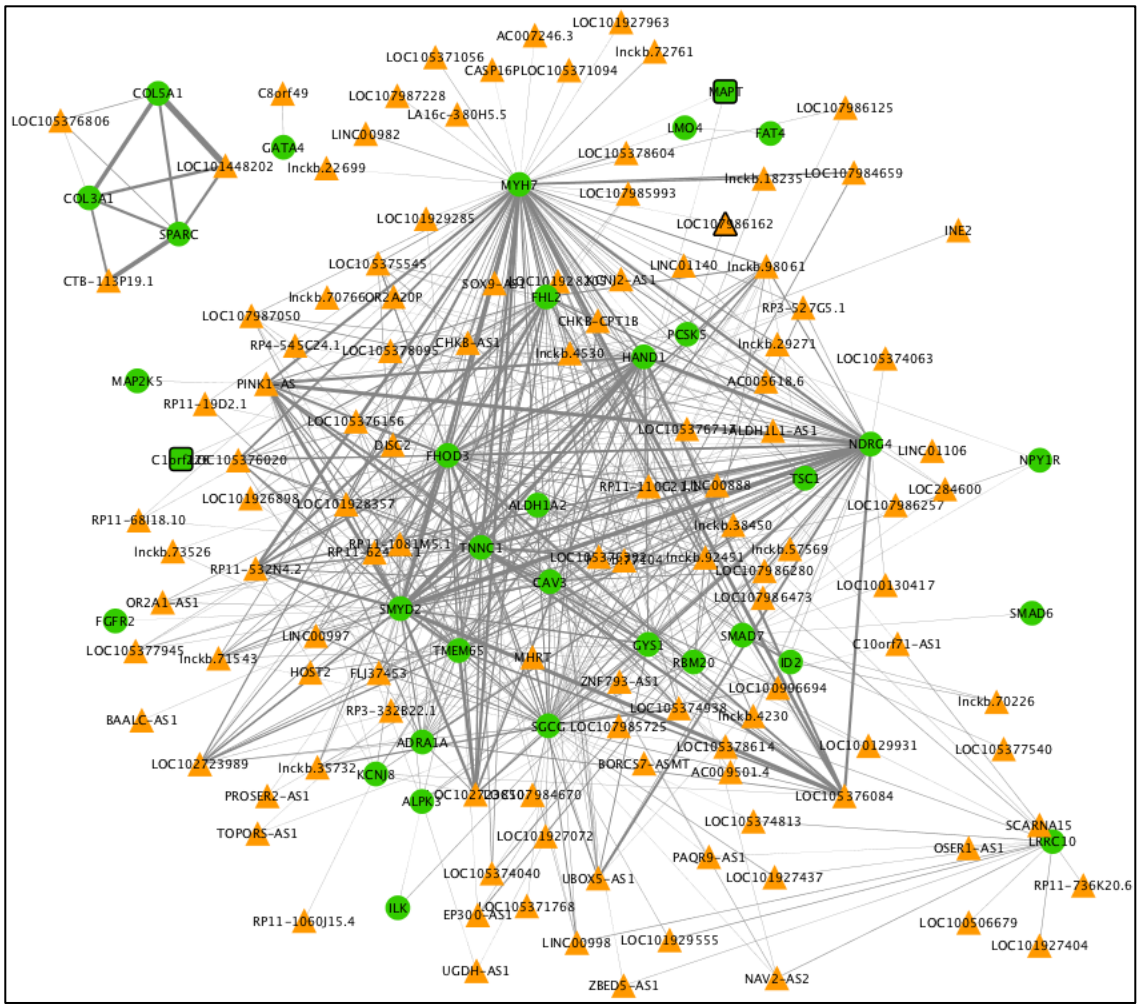


Figure 2.10: Cytoscape network for lncRNA-mRNA co-expression module in the heart. Cytoscape network for lncRNA-mRNA co-expression Module 2 (M2) in the heart identified using WGCNA. The network was filtered for heart development genes (n=148) and correlations > 0.20. Orange triangles and green circles/nodes represent lncRNAs and PCGs respectively. The density of gray lines/edges represents the strength of the connection between genes.

On the lncRNAKB website, for each tissue, users can view and download the WGCNA results (similar to Supplementary Table 2.8) and module enrichment results for all GO pathways (similar to Supplementary Table 2.9) as comma separated (.csv) files. In addition, for notable pathways in a module, users can visualize and download the

corresponding network figures checking for connections of lncRNAs with relevant mRNAs involved in known biological processes.

DISCUSSION AND FUTURE DIRECTIONS

There is a large volume of transcriptomics data publicly available and currently being produced at an unprecedented rate. Novel transcripts assembly using RNA-seq data is a method that generates thousands of new transcripts that need to be characterized. Several of these novel transcripts are categorized as lncRNAs. A few lncRNAs annotation databases and knowledgebases have materialized to store and collect relevant data, hypothesizing that this information can be used by biologists to determine the functional repertoire of these transcripts. Generally, lncRNA knowledgebases are constructed using one or both of these methods: (i) transcriptome assembly using public RNA-seq data (several hundred to thousands of samples) or (ii) combining existing annotation databases. A few of these databases attempt to integrate annotations from multiple sources and go beyond that by integrating multi-omics data such as expression (occasionally tissue-specific), methylation, variation, conservation and functional annotation of lncRNAs in humans. However, from our review of existing resources, those annotations and integrations are sometimes outdated, not rigorous, incomprehensive, and incomplete.

We have created the lncRNAKB, a well-structured research tool that delivers valuable data on human lncRNAs, which can be used for functional molecular studies or development of methods for classification and annotation. These important features are

central points of the lncRNAKB: (i) carefully integrating six extensively used lncRNAs annotation databases using a cumulative step-wise intersection method, (ii) filtering lncRNAs based on their genomic positions relative to known PCGs, (iii) annotating and classifying lncRNAs based on a machine learning method using features such as *k*-mer frequencies and ORF length, (iv) creating a tissue-specific expression body map of lncRNAs, (v) calculating tissue-specificity scores of lncRNAs compared to mRNAs, (vi) creating a tissue-specific eQTL body map of lncRNAs, (vii) calculating exon level conservation scores for all PCGs and lncRNAs, (viii) using the guilt by association principle and WGCNA method to analyze lncRNAs-mRNAs co-expression patterns in a tissue-specific manner, (ix) creating a tissue-specific body map of functionally annotated lncRNAs using enrichment analysis of annotation terms for all co-expression modules and (x) implementing a web resource providing organized and easy-to-follow navigation for users to view and download all content related to lncRNAs. Additionally, when extensive new lncRNAs annotation databases emerge, we can incorporate these into lncRNAKB.

CHAPTER 3

Annotation and functional characterization of lncRNAs using integrative GWAS, eQTL and network analysis in lncRNAKB: a case-study in the context of heart diseases

ABSTRACT

Advent of high-throughput sequencing technologies and development of efficient tools to analyze big data has led to the surprising discovery that only ~2% of the human genome is protein coding and that a majority of the genome though transcribed, falls into the non-protein coding transcripts category. LncRNAs are non-protein coding transcripts that are longer than 200bp (base pairs). Several groups of researchers have used manual and automated techniques to identify and annotate lncRNAs in the human genome. However, questions remain about the purpose and function of these lncRNAs and whether they play a critical role in normal cell function and/or in disease. Here we present a pipeline and case study in heart tissue that uses publicly available RNA-seq and genome-wide association (GWAS) data to functionally characterize 77,199 lncRNAs available in the lncRNAKB built by step-wise integration of six commonly used lncRNA human annotation databases. We performed tissue-specific expression quantitative trait loci (*cis*-eQTL) analysis and overlaid these to GWAS summary data from the UK

Biobank on several heart diseases, to identify subsets of single nucleotide polymorphisms (SNPs) in lncRNAs that may have pleiotropic association between gene expression and disease phenotype using Summary-data Mendelian Randomization (SMR) analysis. We constructed heart-specific protein coding-lncRNA co-expression networks to functionally characterize the cellular processes that the lncRNAs may be involved in using Weighted Gene Co-expression Network Analysis (WGCNA) and overlapped the SMR prioritized lncRNAs on "notable" pathways. These analyses will provide insight into the underlying biology of lncRNAs in cell function and heart disease. Using our pipeline, we were able to find heart specific lncRNAs and creating co-expression networks sheds light on the function of lncRNAs in heart disease and how they might contribute to gene expression and pathology.

INTRODUCTION

Long non-coding RNAs (lncRNAs) are a class of non-protein-coding transcripts that range from 200 nucleotides to 100 kb (approximately 10 kb on average) (Long non coding RNA biology, 2017). The majority of eukaryotic lncRNAs are produced by RNA polymerase II and capped at the 5′ end similar to protein coding genes (PCGs) (Guttman *et al.*, 2009). LncRNAs may or may not be 3′-end polyadenylated (Long non coding RNA biology, 2017), could undergo splicing and have longer but fewer exons, compared to mRNAs (Derrien *et al.*, 2012). Classes of lncRNAs are usually annotated relative to their position with nearby PCGs (DiStefano, 2018), and include: (1) intergenic lncRNAs or lincRNAs, which are transcribed from regions at least >1 kb from PCGs, (2)

bidirectional lncRNAs which are transcribed <1 kb of promoters in opposite direction of protein-coding transcripts, (3) intronic lncRNAs, which are transcribed within introns of PCGs, (4) exonic lncRNAs, which overlap with one or more exons of PCGs, (4) sense lncRNAs, which are transcribed in the same direction of PCGs and overlap with one or more exons or introns of these transcripts and (5) antisense lncRNAs, which are transcribed in the opposite direction of PCGs and overlap with one or more exons or introns of these transcripts.

Many lncRNAs do not show the same pattern of high interspecies conservation as protein coding genes (PCGs) (Hezroni *et al.*, 2015; Cabili *et al.*, 2011; Guttman *et al.*, 2009; Li and Yang, 2017). Sequence conservation is comprised of short, 5′-biased patches of conserved sequence nested in exons (Hezroni *et al.*, 2015). Many studies have reported that lncRNAs have low level of expression (Ponting *et al.*, 2009). However, lncRNAs have higher tissue-specific expression compared to mRNAs (Cabili *et al.*, 2011; Jiang *et al.*, 2016). Some lncRNAs include short open reading frames (sORFs) and undergo translation, though only a minority of such translation events results in stable and functional peptides (Housman and Ulitsky, 2016; Andrews and Rothnagel, 2014). Due to low sequence conservation and low levels of expression, the knowledge that lncRNAs are merely transcriptional noise is common (Palazzo and Lee, 2015). Owing to the advances in high throughput sequencing technologies there have been thousands of lncRNAs that have been annotated across human tissues however, only a small proportion have been functionally characterized (Palazzo and Lee, 2015). Individual lncRNAs have been suggested to carry out a variety of functions, including

transcriptional regulation in *cis* (e.g., *XIST*) (Clemson *et al.*, 1996) or *trans* (e.g., *Fendrr*) (Grote *et al.*, 2013), organization of nuclear domains, and regulation of proteins or RNA molecules (Kopp and Mendell, 2018). In addition, there has been increased evidence suggesting that dysregulation of lncRNAs is involved in many diseases (Chen *et al.*, 2013; Wapinski and Chang, 2011). Despite the fact that only a minority of lncRNAs have been adequately functionally characterized, there is no agreement on the transcriptional regulatory mechanisms by which lncRNAs might perform (Bassett *et al.*, 2014; Palazzo and Lee, 2015).

We have developed a comprehensive lncRNAs annotation database called the long non-coding RNA knowledgebase (lncRNAKB – http://www.lncrnakb.org) in which we systematically integrated six lncRNAs annotation databases and generated a large number of unique lncRNAs (*n*=77,199). There are many components of the lncRNAKB, generated across 31 solid human normal tissues using RNA-seq (9,074 samples) and Whole Genome Sequence (WGS) (652 samples) genotype data, obtained from the Genotype Tissue Expression (GTEx Release v7) project (GTEx Consortium *et al.*, 2017). Some of these include: (i) tissue-specific gene expression profiles, (ii) tissue-specific expression quantitative trait loci (*cis*-eQTLs) and (iii) tissue-specific Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008) analyzing lncRNAs-mRNAs co-expression patterns, followed by Gene Ontology (GO) pathways enrichment analysis (Gene Ontology Consortium, 2015; The Gene Ontology Consortium, 2019; Ashburner *et al.*, 2000). All these data are publicly available for viewing and downloading through the web resource.

In light of availability of these data on lncRNAs from the lncRNAKB, we sought out to perform a multi-omics integration analysis to build a reasonable case for assigning probable functions of lncRNAs. We focus our integrative analysis on the heart tissue and present a case-study to identify lncRNAs that could be potentially involved in transcriptional regulatory mechanisms possibly leading to heart diseases. There have been several genome-wide association (GWAS) studies that have identified many genetic variants associated with heart disorders such as cardiovascular disease, coronary heart disease, congenital heart disease, cardiomyopathy, heart failure, acute myocardial infarction, atherosclerotic heart disease, atrial fibrillation and chronic ischemic heart disease (Buniello *et al.*, 2019). However, the mechanism by which these genetic variants exert their effects on complex diseases/traits is generally unknown because of the complex linkage disequilibrium (LD) between them and the causal mutations (Raychaudhuri, 2011). In addition, GWAS have shown that majority of these variants are located in non-coding regions enriched for long intergenic non-coding RNAs (lincRNAs) (Cabili *et al.*, 2011). Several eQTL studies have been implemented across many human tissues and cell types showing that the regulation in expression of PCGs is mediated by GWAS/trait associated genetic variants in *cis* (*cis*-eQTLs) (Nicolae *et al.*, 2010). Additionally, lincRNAs-eQTLs studies have also shown that there are a substantial number of *cis*-eQTLs in lincRNAs and demonstrated that the genetic regulation of lincRNA expression is independent of the regulation of neighboring PCGs (Popadin *et al.*, 2013). It has also been shown that the transcription of *cis*-eQTLs in lincRNAs that are associated with traits, contributes to chromosomal architecture and thereby the regulation

of nearby trait-associated PCGs expression levels (Tan *et al.*, 2017b). Significant enrichment of *cis*-eQTLs in lincRNAs suggests that many Single Nucleotide Polymorphism (SNP)-trait associations could perhaps act through gene expression of lincRNAs (i.e. SNP → Gene expression (PCGs and/or lincRNAs) → trait) (Hernandez *et al.*, 2012; Porcu *et al.*, 2018). This relationship can be investigated in studies of traits if SNP and gene expression data are available from the same samples. However, it is rare for transcriptomic studies examining complex diseases/traits to have supplementary genotype data and large sample sizes due to shortage of genomic material and cost thus, underpowered. In contrast, GWAS of traits alone have large sample sizes but, do not have accompanying gene expression data. Therefore, overall detection of eQTLs and those with small effect sizes is hindered. Considering the limitations stated above, there have been development of methods in prioritizing causal genes at GWAS loci (Veturi and Ritchie, 2018; Schaefer *et al.*, 2018).

Transcriptome-wide association studies (TWAS) have been previously employed to integrate GWAS and eQTL data to prioritize genes that are associated with traits (Gusev *et al.*, 2016). The general idea for a TWAS is to use a known gene expression and SNP dataset (eQTLs) as a reference panel to determine the gene-trait association from GWAS. There are three steps for conducting a TWAS: (i) the reference expression panel (for e.g. GTEx) is used to perform a *cis*-eQTL analysis with gene expression data, searching for variants within a specified distance for e.g. 1 Megabase (Mb) of the transcription start site (TSS) of each gene and in the tissue of analysis which is considered your training data, (ii) the training data is used to predict/impute the gene

expression of thousands of individuals in a GWAS study, and (iii) the predicted/imputed gene expression measures are statistically associated to trait. The predicted/imputed gene expression estimates can be conceptualized as a linear model of genotypes with weights (strength of correlation between SNPs and gene expression from the training data) while accounting for LD among SNPs. For steps (ii) and (iii), two approaches could be used: (a) If individual-level genotype data is available from the GWAS study, expression prediction may be performed sequentially using the effect sizes from the reference panel and measure the association between predicted expression and a trait for e.g. using PrediXcan2 (Gamazon *et al.*, 2015), and/or (b) if only summary-level data is available from the GWAS study, the SNP-trait standardized effect sizes can be used directly (weighted linear combinations) while accounting for LD among SNPs to estimate association between predicted expression and a trait for e.g. using Fusion (Gusev *et al.*, 2016) and S-PrediXcan (Barbeira *et al.*, 2018).

Mendelian Randomization (MR) is a method that can be used to test the effect of exposure (gene expression) on an outcome (trait) through an instrumental variable (genetic variant/SNP) (Porcu *et al.*, 2018; Smith and Ebrahim, 2003). Therefore, MR can be used to prioritize causal genes at GWAS loci, if the gene expression and outcome (trait) data on the same samples are available. MR analysis can be conducted using a two-stage method. It comprises of two regression steps: the first-step is the regression of the exposure (gene expression) on the instrumental variable (genetic variant/SNP), and the second-step is the regression of the outcome (trait) on the fitted values of the exposure from the first step. Since the outcome is continuous, the causal effect of the risk factor in

the outcome can be estimated using a 2-stage least squares approach (Burgess *et al.*, 2017). GWAS data with large sample sizes that include both gene expression and outcome (trait) data on the same samples are rare. However, many GWAS meta-analysis and eQTL studies with large cohorts and summary-level results (for e.g. effect sizes or test statistics) are publicly available (GTEx Consortium *et al.*, 2017; Bycroft *et al.*, 2018). Summary Mendelian Randomization analysis (SMR) (Zhu *et al.*, 2016) is a method that prioritizes genes that are targeted by genetic variants/SNPs in GWAS of complex diseases. It combines (using MR concepts) summary-level data from two-samples for e.g. independent GWAS and data from eQTL studies to identify pleiotropic association between the expression level of a gene (exposure) and a trait (outcome). Pleiotropic association is when the causal variant affects both gene expression and trait. The test statistic from SMR is an approximate χ^2 interpreted as the effect of gene expression on outcome (trait) free of non-genetic confounders. It is also possible that there are two distinct causal variants in LD with each other, one influencing gene expression and the other the outcome (trait) which is of less interest in prioritizing genes with SMR. In pleiotropic associations, the effect sizes of gene expression on outcome (trait) would be approximately similar for genetic variants/SNPs in LD with each other in the *cis*-eQTL region. Therefore, testing for heterogeneity in the effect sizes of gene expression on outcome (trait), estimated for the SNPs in the cis-eQTL region would be equivalent to testing against the null hypothesis that there is a single causal variant affecting both the gene expression and outcome (trait). SMR has developed a method to test for heterogeneity in dependent instruments (HEIDI).

To better understand the contribution of lncRNAs expressed in heart tissue and prioritize these based on the effects of genetic variation/SNPs on lncRNAs expression in heart diseases, we performed a systematic *cis*-eQTL study and SMR analysis using seven heart disease related GWAS (myocardial infarction, atrial fibrillation, atherosclerosis, cardiomyopathy, chronic heart disease, heart failure and obesity) obtained from the UK Biobank (Bycroft *et al.*, 2018). In addition, to further understand the potential transcriptional regulatory mechanisms by which these prioritized lncRNAs may perform, we overlap the WGCNA lncRNA-mRNA co-expression gene modules information identified in the heart tissue (results available on lncRNAKB) against the prioritized lncRNAs (across all seven GWAS) in any module enriched for heart specific GO processes. Therefore, using these approaches we demonstrate the relevance of data available on lncRNAKB and one of the numerous ways that biologists can use these datasets to understand lncRNAs and augment their research on lncRNAs discovery.

MATERIALS AND METHODS

Heart-specific tissue expression and expression quantitative trait loci (eQTLs) data:

We downloaded the heart-specific transcriptome and eQTLs datasets from the lncRNAKB (http://www.lncrnakb.org). The lncRNAKB contains RNA-seq expression profiles summarized at gene-level in raw read counts and normalized to Transcripts Per Kilobase Million (TPM) (Wagner *et al.*, 2012) (https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/) using the GTEx project (Release v7) RNA-seq data and the gene transfer format (GTF) annotation file

(https://useast.ensembl.org/info/website/upload/gff.html#moreinfo) from lncRNAKB. After excluding samples with < 10^6 reads assigned to genes, there were 430 samples quantified across 99,717 genes (77,199 lncRNAs and 22,518 PCGs) in the heart tissue. Details of the RNA-seq data analysis are outlined in Chapter two (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): subsection: Expression profiling).

Using the TPM normalized expression matrix across all genes and tissues (no filter applied) we calculated a tissue-specificity metric called Preferential Expression Measure (PEM) (Kryuchkova-Mostacci and Robinson-Rechavi, 2017). PEM shows how specific a gene is to the heart tissue. The PEM scores the expression of a gene in the heart tissue in relation to its average expression across all other genes and tissues. To compute PEM, we calculated and used the average expression across all replicates for each gene by tissue. All genes that were not expressed in at least one tissue were removed from the analysis. Details of the tissue-specificity score analysis are outlined in Chapter two (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): subsection: Tissue-specificity scores).

The lncRNAKB also contains heart-specific *cis*-eQTL data generated using the methods outlined in Chapter two (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): subsection: eQTL analysis). Briefly, 50,153,256 SNP-gene pairs (*cis*-eQTLs) were calculated using 251 RNA-seq samples with blood-derived WGS genotype data (163 males:88 females), 5,913,705 SNPs, and 24,959 genes (10,171 lncRNAs:14,788 PCGs). We filtered the original genotype data

(50,862,464 variants) across all samples (*n=652*), from the GTEx project using standard measures i.e. excluding variants that were indels, missing, multiallelic, "FAIL" and < 5% minor allele frequency (MAF). We subset the genotype data for 251 heart samples and filtered SNPs < 5% MAF as an additional step. Details of the genotype file processing are outlined in Chapter two (see Methods: Tissue-specific expression profiling and expression quantitative trait loci (eQTLs): subsection: Genotype file processing). Genes were retained if TPM > 0.50 in at least 20% of the samples. Furthermore, only genes with counts > 2 in at least 20% of samples were kept. Log2 counts per million (log2CPM) and the limma-voom R package (Ritchie *et al.*, 2015) (Law *et al.*, 2014) was used to normalize the data between samples using trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010). The *cis*-eQTLs were identified by linear regression, as implemented in FastQTLv2.0 (threaded option) (Ongen *et al.*, 2016), adjusting for the five PCs and sex within 1 Megabase (Mb) of the transcription start site (TSS) of each gene. The *cis*-eQTL results were visualized using a manhattan plot created using the qqman package in R (Turner, 2014). To evaluate the significance of the most highly associated variant per gene we used the adaptive permutations option in FastQTL between 1000 and 10000 permutations. After permutation, 4,365 unique genes (1,913 lncRNAs) had a *cis*-eQTL with permutation p-value ≤ 0.05 which were used for the multi-omics integrative analysis with GWAS data using SMR.

Genome-wide association studies (GWAS) in heart diseases:

Table 3.1 shows the seven heart diseases related GWAS summary data that were selected from the UK Biobank (Bycroft *et al.*, 2018) for the integrative analysis. Details of the methods for generating the GWAS data are outlined in the UK Biobank manuscript and website (http://www.nealelab.is/uk-biobank). GWAS summary data used for this case study did not include hundreds of thousands of samples as desired. However, as a proof of concept, we illustrated the integration process of GWAS, eQTL and network analysis to functionally annotate lncRNAs thus, highlighting potential use of datasets available on thousands of lncRNAs (across numerous tissues) in the lncRNAKB using considerably larger GWAS summary data obtained via meta-analysis.

Table 3.1: Descriptive summary of the GWAS summary data for SMR analysis.
Descriptive summary of the GWAS summary data used for the SMR analysis
n = sample size, n_{case} = number of cases, $n_{control}$ = number of controls, m_{SNP} = number of SNPs

GWAS (UK Biobank) - Disease	n_{case}	$n_{control}$	m_{SNP}	**GWAS summary data download link**
Chronic Heart Disease	8,755	328,444	11,400,324	https://www.dropbox.com/s/rdz3f8ind6mqlcp/40001_I259.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O 40001_I259.gwas.imputed_v3.both_sexes.tsv.bgz
Myocardial Infarction	3,927	333,272	11,400,324	https://www.dropbox.com/s/53ksig4hhxkfjh6/40001_I219.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O 40001_I219.gwas.imputed_v3.both_sexes.tsv.bgz
Atrial Fibrillation	3,818	333,381	11,400,324	https://www.dropbox.com/s/uluig1qhvz7veek/I48.gwas.imputed_v3.both_sexes.ts

				v.bgz?dl=0 -O I48.gwas.imputed_v3.both_sexes.tsv.bgz
Heart Failure	617	336,582	11,400,324	https://www.dropbox.com/s/ruhz8ihfqviok49/I50.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O I50.gwas.imputed_v3.both_sexes.tsv.bgz
Atherosclerosis Heart Disease	340	336,859	11,400,324	https://www.dropbox.com/s/vojnvx9zc4lodo1/40001_I251.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O 40001_I251.gwas.imputed_v3.both_sexes.tsv.bgz
Cardiomyopathy	285	336,914	11,400,324	https://www.dropbox.com/s/x164s4tlb4vjffe/I42.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O I42.gwas.imputed_v3.both_sexes.tsv.bgz
Obesity	217	336,982	11,400,324	https://www.dropbox.com/s/o5gfs1ny01n50za/E66.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0 -O E66.gwas.imputed_v3.both_sexes.tsv.bgz

Summary Mendelian Randomization (SMR) integrating GWAS and eQTL data for lncRNA candidate prioritization in heart diseases:

To test for the relationship between SNPs, gene expression, and a trait we applied the SMR method across the whole genome using summary data from seven GWAS of heart diseases (see Methods: Genome-wide association studies (GWAS) in heart diseases) and heart eQTLs (see Methods: Heart tissue expression and expression quantitative trait loci (eQTLs) data). In a traditional MR analysis, consider z as an instrumental variable (genetic variant/SNP), x as the exposure (gene expression) and y as the outcome (trait). MR tests whether the effect of z on y is mediated by x (i.e. SNP (z) $\rightarrow$ Gene expression (PCGs and/or lincRNAs) (x) $\rightarrow$ trait (y)) using the estimates of x on

z, $beta_{zx}$ and estimates of y on z, $beta_{zy}$ thus, $beta_{xy} = beta_{zy}/beta_{zx}$ derived using a 2-stage least squares approach (Burgess *et al.*, 2017). MR analysis requires SNP, gene expression and trait data to be measured on the same sample. However, it is not feasible to collect GWAS and transcriptomic data on hundreds of thousands of samples in the same study. SMR uses a two-sample MR approach to derive an approximate χ^2 test statistic defined by the following equation $T_{SMR} = \frac{z_{zy}^2 z_{zx}^2}{z_{zy}^2 + z_{zx}^2}$ where z_{zy} and z_{zx} (equation 5, online methods) in (Zhu *et al.*, 2016) are the summary statistics from the GWAS and *cis*-eQTL study, respectively (Zhu *et al.*, 2016). The SMR test can possibly detect three modes of association between gene expression and trait through SNPs: (i) causality ($z \rightarrow x \rightarrow y$), (ii) pleiotropy ($z \rightarrow x$ and $z \rightarrow y$) or (iii) linkage ($z_1 \rightarrow x$ and $z_2 \rightarrow y$), z_1 and z_2 are in LD. Since SMR uses only the top *cis*-eQTL that is strongly associated with gene expression as an instrument (z) it is unable to distinguish between causality and pleiotropy. As mentioned in the introduction, that LD between two causal variants, one influencing gene expression and the other influencing trait is of less interest in SMR. The test for heterogeneity in dependent instruments (HEIDI), developed in SMR, considers that $beta_{xy}$ i.e. the estimated effect sizes of gene expression on outcome (trait) are expected to be equal at any of the *cis*SNPs to the top associated *cis*-eQTL due to LD between them if there is only a single causal variant affecting both gene expression and outcome (trait). LD is estimated from the reference sample used to calculate the *cis*-eQTLs. Genes with a HEIDI p-value < 0.05 indicate signs of heterogeneity and are not considered for prioritization.

SMR was run using all the estimate of SNP effects on the traits from summary GWAS related to heart disease. The summary GTEx heart eQTL data consisted of 9,391,776 *cis*-eQTLs SNPs for 4,365 genes, selected using a *cis*-eQTL permutation p-value ≤ 0.05. Only *cis*-eQTLs with p-value ≤ 5 x 10^{-4} (475,150 *cis*-eQTLs) were included in the SMR analysis. In the SMR test, the top associated *cis*-eQTL for each gene was used as the instrument. Default values were used for the HEIDI test. The significance level for the SMR test was defined as nominal SMR p-value (p_{SMR}) ≤ 0.05. Genes with HEIDI p-value (p_{HEIDI}) ≥ 0.05 were retained as prioritized showing no evidence of heterogeneity.

Heart-specific functional characterization of lncRNAs using a network-based approach:

Unsupervised clustering methods can be used to extrapolate lncRNAs biological function based on the degree of connections to genes of known function. Using the genes expression profiles, these methods cluster them into group of genes, known as co-expression modules. These modules are then subject to over representation analysis (ORA) with specific pathways for e.g. Gene Ontology (GO) (Gene Ontology Consortium, 2015; The Gene Ontology Consortium, 2019) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017), which determines the potential function of genes. ORA employs the hypergeometric test (Holmans, 2010) to determine if the overall functional enrichment is different than what would be expected by random chance. The significantly enriched terms can be used as an estimate of annotation for the genes with

unknown function for e.g. lncRNAs (Ehsani and Drabløs, 2018b). This is often referred to as the "guilt by association" principle.

The co-expressed gene modules were identified using the Co-Expression Modules identification Tool (CEMiTool) (Russo *et al.*, 2018). This is a slight modification of the Weighted Gene-Coexpression Network Analysis (WGCNA) (Langfelder and Horvath, 2008) method. Excellent details of the WGCNA theory are available from this website (https://horvath.genetics.ucla.edu/coexpressionnetwork/). The goal of WGCNA is to identify modules of genes that are co-expressed in an experiment and therefore likely work in concert to carry out various biological functions. Briefly, pairwise Pearson correlation coefficients are calculated between all gene expression levels in the analysis. The resulting correlation matrix is transformed into a so-called adjacency matrix of connection strengths using a power function with the formula $((1 + correlation)/2)^{\beta}$ where the parameter of this function is selected so that the resulting network best satisfies a scale-free topology. This adjacency matrix is then converted into a topological overlap matrix (TOM), and then modules of densely interconnected genes are derived from the network by hierarchical clustering of the topological overlap. The main difference between the algorithm in CEMiTool and WGCNA is that WGCNA provides a function named "pickSoftThreshold" that can automatically select the β value; however, the CEMiTool has created an alternative algorithm that improves the automatic selection of the β value, allowing for more reliable and consistent results.

The lncRNAKB contains heart-specific WGCNA results generated using the methods outlined in Chapter two (see Methods: Tissue-specific expression profiling and

expression quantitative trait loci (eQTLs): Functional characterization of lncRNAs using a network-based approach). Briefly, there were 61 gene co-expression modules identified in the heart using expression data from the GTEx project across 16,882 protein-coding genes and 2,762 lncRNAs. On average there were approximately 276 genes including 45 lncRNAs per module. Network plots were constructed by selecting "notable" pathways based on the GO pathways enrichment results. The network plots were generated using Cytoscape v3.6.0 (Shannon *et al.*, 2003) package by visualizing edges with correlation > 0.20. LncRNAs that were prioritized by SMR analysis for each GWAS were overlaid on top of these networks to potentially elucidate their functional roles in heart disease.

RESULTS

Heart-specific tissue expression and expression quantitative trait loci (eQTLs):

Figure 3.1 illustrates the density distributions of PEM scores across PCGs (n = 20,220) and lncRNAs (n = 72,083) in the heart tissue. It reports a positive value for genes over-expressed and a negative value for genes under-expressed, respectively. Figure 3.1 shows that most of the lncRNAs have similar expression patterns compared to PCGs in the heart. Supplementary Table 3.1 shows PEM scores across all genes in the lncRNAKB annotation database in heart tissue. Figure 3.2 shows the gene expression distribution across all tissues for top five heart-specific lncRNAs (selected by highest PEM scores). Figure 3.2 illustrates that the expression values of all five lncRNAs are higher in heart compared to other tissues suggesting heart-specificity. Figure 3.3 shows a manhattan plot

that illustrates the genome-wide results of 50,153,256 *cis*-eQTLs analyzed in the heart tissue. 4,365 genes (1,913 lncRNAs) had a *cis*-eQTL with an adaptive permutation p-value ≤ 0.05.

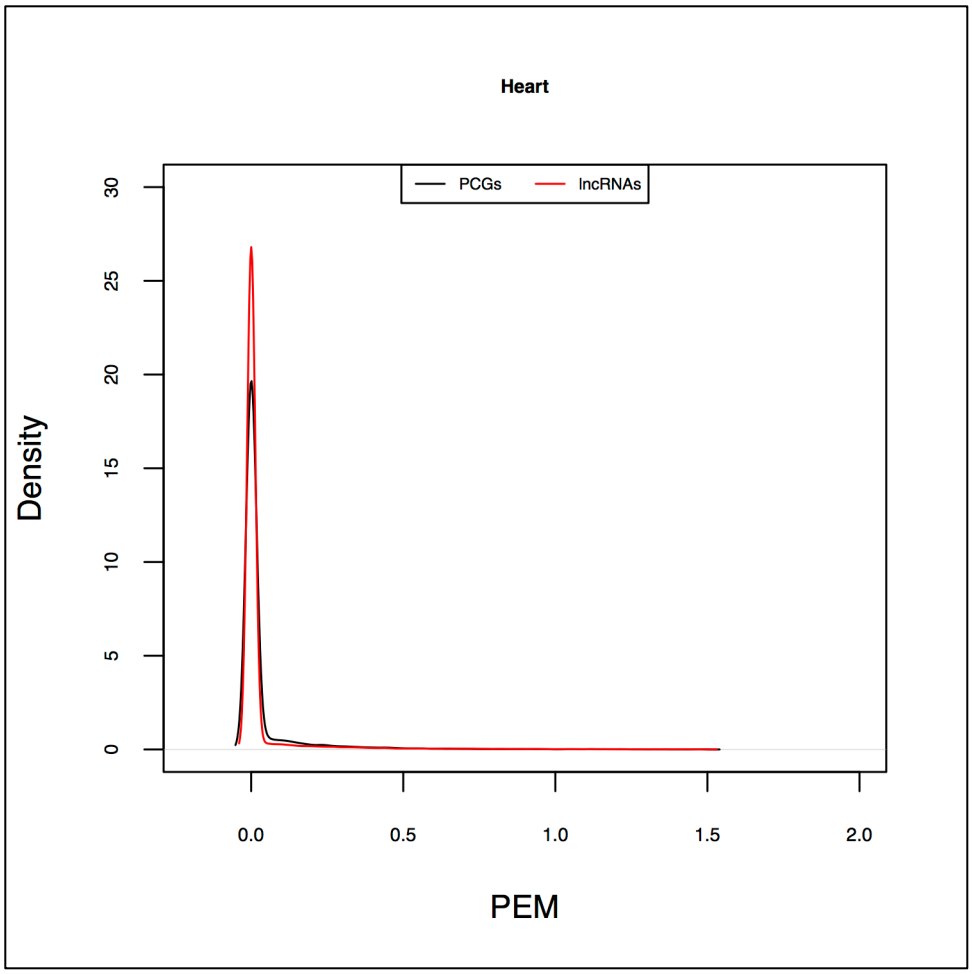


Figure 3.1: Distribution of PEM tissue-specificity scores in heart tissue.
Distribution of Preferential Expression Measure (PEM) tissue-specificity scores calculated with RNA-seq data of heart tissue (430 samples) from GTEx across PCGs (n = 20,220) and lncRNAs (n = 72,083) in the lncRNAKB as a comparison. A positive or negative value indicates genes are over-expressed or under-expressed, respectively. Graph created with density function from R, which computes kernel density estimates.

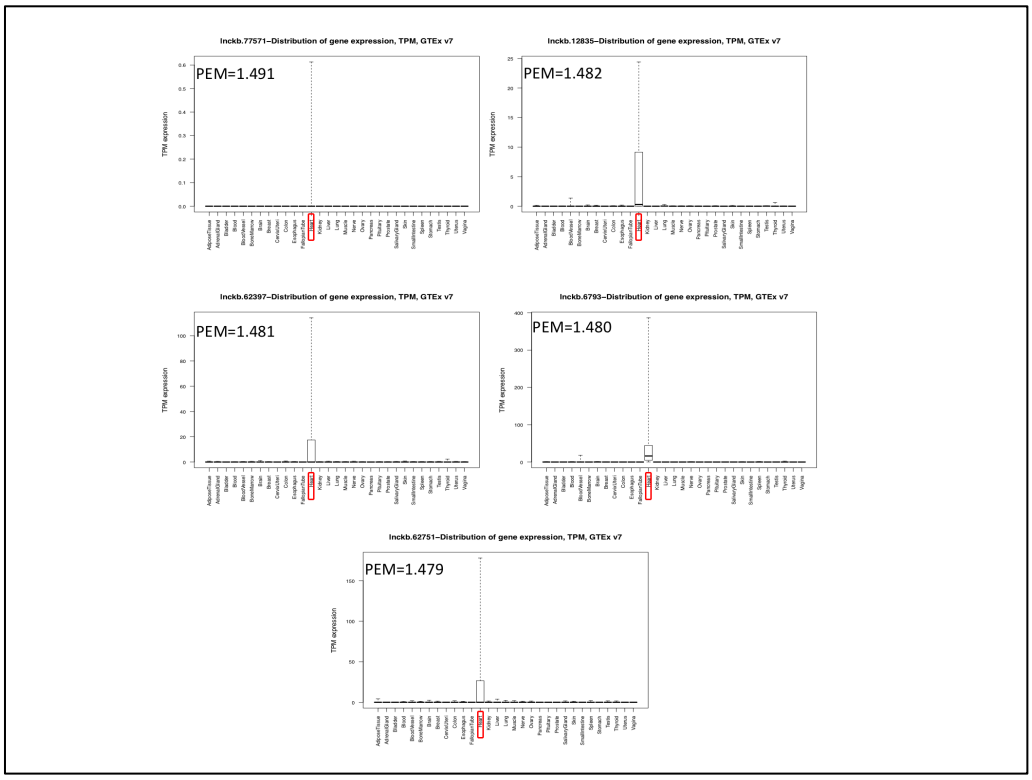


Figure 3.2: Gene expression box plot distribution of the top five lncRNAs with the highest heart-specific PEM scores.
Gene expression box plot distribution of the top five lncRNAs with the highest heart-specific PEM scores. The x-axis represents the 31 solid organ human normal tissues from GTEx and y-axis is the TPM expression. The heart tissue is highlighted using a solid red box and the PEM scores are shown on the left-hand side of individual plots.

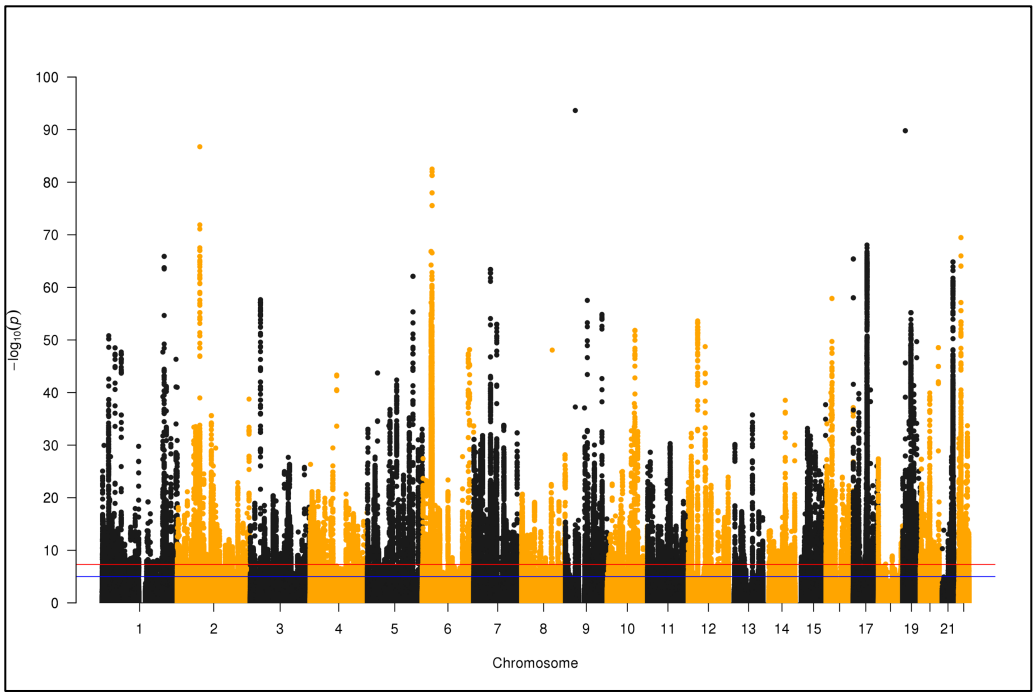


Figure 3.3: Manhattan plot illustrating the results of the cis-eQTL analyses from the heart tissue.
Manhattan plot illustrating the results of the *cis*-eQTL analyses from the heart tissue. The x-axis are the chromosomes and each dot on the y-axis represents the *cis*-eQTL -$\log_{10}$ (p-values) of the SNP-gene pairs that were tested within 1 Mb of the TSS of each gene.

SMR prioritized lncRNA candidates in heart diseases:

Figure 3.4 shows manhattan plots to visualize the results of SMR analysis across seven GWAS related to heart disease. In total, we identified 1,054 genes (out of 4,365 genes), tagged by 859 SNPs/*cis*-eQTLs, at the nominal significance level ($p_{SMR} \leq 0.05$) for the seven heart diseases (Figure 3.4 and Supplementary Table 3.2). Table 3.2 summarizes the results of the SMR analysis across each GWAS. Table 3.2 shows that in addition to PCGs, SMR also prioritized several lncRNAs.

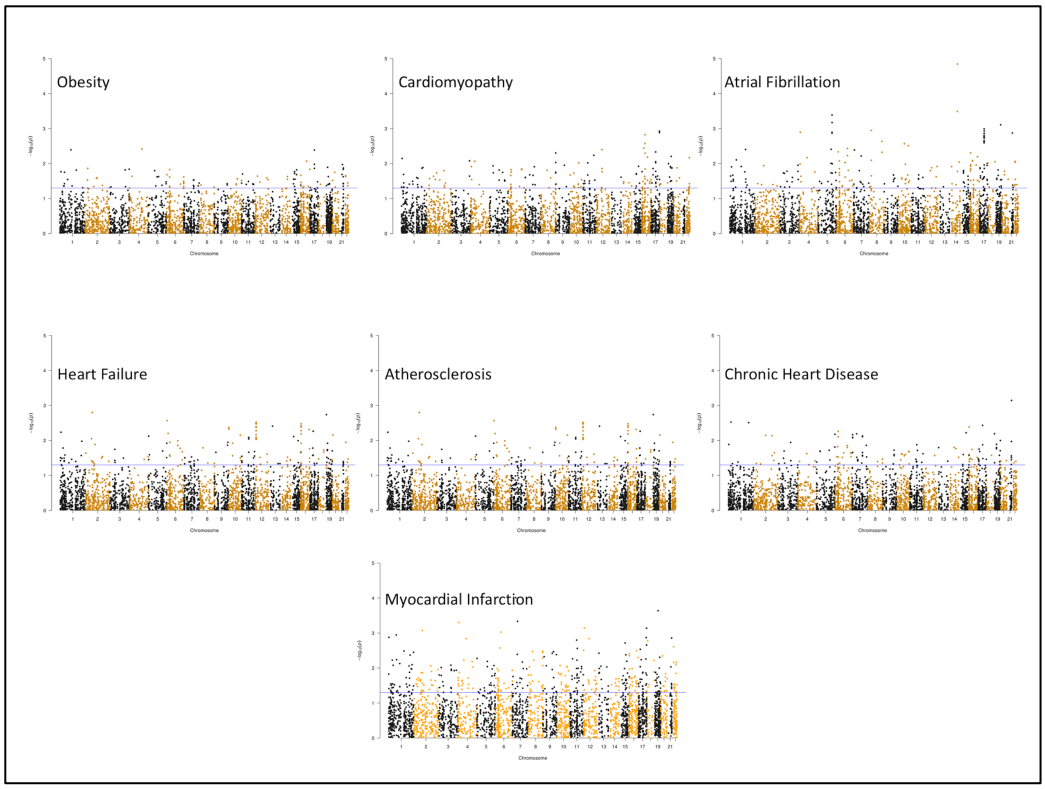


Figure 3.4: Manhattan plots illustrating the results of the SMR analysis across seven GWAS related to heart disease.
Manhattan plots illustrating the results of the SMR analysis across seven GWAS related to heart disease. The x-axis are the chromosomes and each dot on the y-axis represents the gene-based SMR -$\log_{10}$ ($\boldsymbol{p_{SMR}}$), indicating pleiotropic association between the expression level of a gene (exposure) and a trait (outcome) i.e. when the causal variant affects both gene expression and trait. Genes above the blue line with SMR p-value ($\boldsymbol{p_{SMR}} \leq 0.05$) and HEIDI p-value ($\boldsymbol{p_{HEIDI}} \geq 0.05$) were considered as SMR prioritized/significant.

Table 3.2: Number of genes (PCGs and lncRNAs) prioritized by SMR.
Number of genes (PCGs and lncRNAs) prioritized by SMR ($\boldsymbol{p_{SMR}} \leq 0.05$) for seven heart diseases.

GWAS (UK Biobank) - Trait	Number of genes (PCGs) passing SMR test	Number of genes (lncRNAs) passing SMR test
Chronic Heart Disease	77	59

Myocardial Infarction	47	45
Atrial Fibrillation	104	92
Heart Failure	83	65
Atherosclerosis Heart Disease	63	84
Cardiomyopathy	79	69
Obesity	51	57

Heart-specific functional characterization of lncRNAs using a network-based approach:

For weighted gene co-expression network analysis we utilized a comprehensive heart transcriptome dataset, which contains RNA-seq data from the GTEx project with 251 samples (including WGS genotype data) across 16,882 PCGs and 2,762 lncRNAs. 61 co-expression gene modules were identified (see Chapter two: Supplementary Table 2.8). On average there were approximately 276 genes including 45 lncRNAs per module. The largest module (M1) had 3,532 genes including 595 lncRNAs, whereas one module (M60) had no lncRNAs. On average, across all modules, 13,830 and 2,391 GO pathways (out of 104,175 pathways that were tested across all modules) had a p-value ≤ 0.05 and q-value ≤ 0.05 respectively, indicating significant enrichment of biological processes within these modules (see Chapter two: Supplementary Table 2.9). Table 3.3 highlights four "notable" pathways (in three modules) that were chosen based on the following criteria: (i) q-value ≤ 0.05, (ii) the number of genes that overlap in each pathway (Count), (iii) the number of lncRNAs in each module, and (iv) their biological relevance to heart tissue. The heart development process has 466 genes whose specific outcome is the progression

of the heart over time, from its formation to the mature structure thus, related to cardiac development. The lipid catabolic pathway has 247 genes resulting in lipid breakdown, lipid catabolism, lipid degradation, lipolysis and multicellular organism lipid catabolic process thus, related to cholesterol metabolism. The muscle system process has 282 genes and involved in muscle physiology thus, could affect the heart muscle largely.

Table 3.3: Four "notable" GO pathways enriched in three co-expression gene modules. Four "notable" GO pathways enriched (ORA) in three co-expression gene modules (WGCNA) identified using GTEx RNA-seq data (n = 251, with genotype data) across 16,882 PCGs and 2,762 lncRNAs in the heart tissue.

Module	**GO_Pathway_ID**	**pvalue**	**qvalue**	**Count_of_genes_in_pathway**	**number_gene s_in_module**	**number_lncrnas_in_module**
M1	GO_HEART_DEVELOPMENT	3.93E-05	0.0080	93	3532	595
M2	GO_HEART_DEVELOPMENT	0.0002	0.0144	54	1955	428
M2	GO_LIPID_CATABOLIC_PROCESS	1.85E-07	4.85E-05	41	1955	428
M6	GO_MUSCLE_SYSTEM_PROCESS	9.44E-06	0.0031	18	512	84

Figure 3.5 illustrates the Cytoscape network plots that were generated for the four "notable" pathways (in three modules) shown in Table 3.3. LncRNAs that were prioritized by SMR analysis ($p_{SMR} \leq 0.05$) for each GWAS were overlaid on top of these networks to potentially elucidate their functional roles in heart diseases.

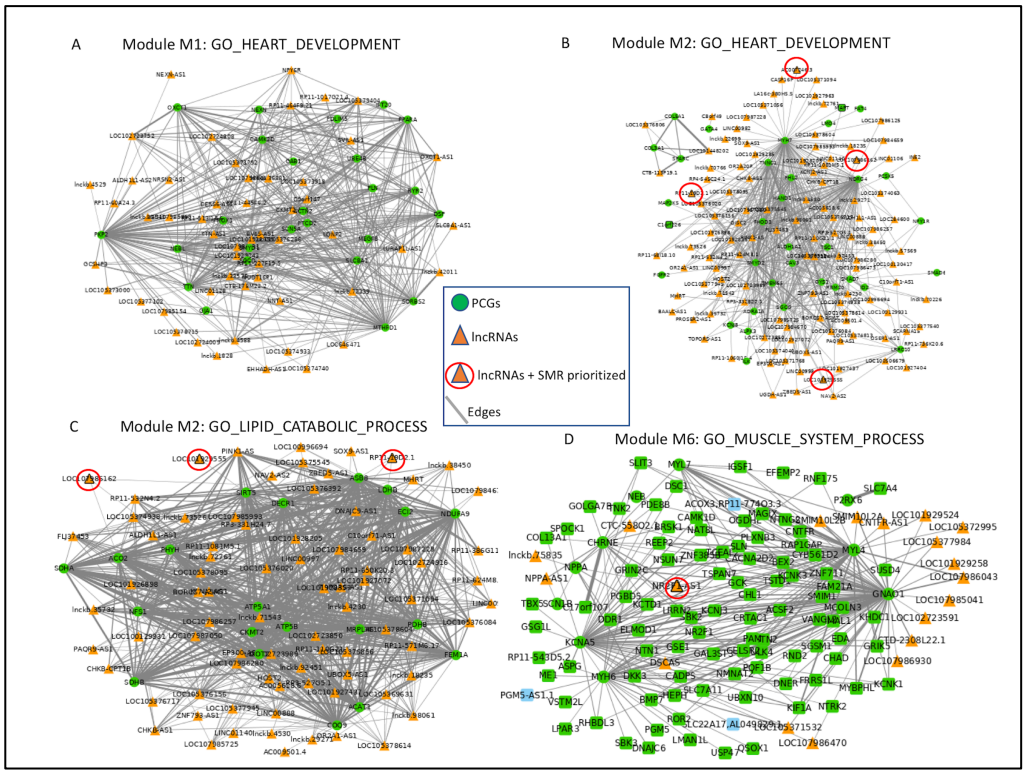


Figure 3.5: Cytoscape networks for four "notable" GO pathways in the heart enriched in three co-expression gene modules.
Cytoscape networks for four "notable" GO pathways in the heart enriched (ORA) in three co-expression gene modules (WGCNA). The networks were filtered based on correlations > 0.20. Orange triangles and green circles/nodes represent lncRNAs and PCGs respectively. Orange triangles with a red circle represent SMR prioritized ($\boldsymbol{p_{SMR}} \leq$ 0.05) lncRNAs. The density of gray lines/edges represents the strength of the connection between genes.

Table 3.4 shows the number of lncRNAs that were SMR prioritized (Table 3.2 and Supplementary Table 3.2) overlaid on top of the four "notable" networks/pathways as a summary. First, Table 3.4 demonstrates that there were several lncRNAs within each network with correlations > 0.20 thus, indicating robust mRNA-lncRNA co-expression. Second, there were a couple of lncRNAs that were SMR prioritized ($p_{SMR} \leq 0.05$) in three out of four "notable" networks/pathways using GWAS summary data from seven heart diseases: myocardial infarction, atrial fibrillation, atherosclerosis, cardiomyopathy, chronic heart disease, heart failure and obesity obtained from the UK Biobank (Bycroft *et al.*, 2018). Supplementary Table 3.3, 3.4, 3.5, and 3.6 provides individual details for each network (mainly gene ids and gene type) and flags lncRNAs that were SMR prioritized.

On further review of PCGs and lncRNAs in these "notable" networks we have successfully identified potential genes which are known to play a role in cardiac related diseases for e.g. in co-expression gene module M2 (heart development), we observed a tight co-expression of several lncRNAs with *MYH7* (myosin heavy chain 7), which encodes molecular motor proteins for heart contraction. *MYH7* has been shown to be regulated by a cluster of lncRNAs (Han *et al.*, 2014). We also observed *MHRT* (Myosin Heavy Chain Associated RNA Transcript), a lncRNA known to have a cardioprotective role by acting as a decoy to the *BRG1* (brahma-related gene 1) (Hermans-Beijnsberger *et al.*, 2018). *MHRT* is co-expressed with *SGCG* (sarcoglycan gamma) which maintains the structure of muscle tissue and is highly expressed in heart tissue. In co-expression gene module M2 (lipid catabolic process), we observed a lncRNA *RP11-532N4.2* whose expression has been shown to be dysregulated following ischemia (Saddic *et al.*, 2017).

We also observed *SIRT5* (sirtuin 5) which is an important regulator of heart function. Mutations in the *SIRT5* gene promoter have been associated with acute myocardial infarction (Chen *et al.*, 2018). In co-expression gene module M6 (muscle system process) we observed *NPPA-AS1* which is a lncRNA antisense to *NPPA* (Natriuretic Peptide A) and mutations in it are likely to be associated with atrial fibrillation. There have also been a few reviews in describing the potential of lncRNAs involvement to cardiac biology in humans and mice (Frank *et al.*, 2016; Scheuermann and Boyer, 2013; Hu *et al.*, 2018; Hobuß *et al.*, 2019).

Table 3.4: Details of Cytoscape networks for the four "notable" networks/pathways. Details of each Cytoscape network for the four "notable" networks/pathways (in three modules) chosen. Disease represents the seven heart diseases which were used for the SMR analysis. The number of lncRNAs SMR prioritized had $\boldsymbol{p_{SMR}} \leq 0.05$.

Module	**GO_Pathway_ID**	**Count_of_genes_correlation_greater_than_0.2**	**number_of_PCGs_in_network**	**number_of_lncRNAs_in_network**	**number_of_lncRNAs_SMR_prioritized**	**Disease**
M1	GO_HEART_DEVELOPMENT	78	25	46	0	-
M2	GO_HEART_DEVELOPMENT	148	34	106	4	Atherosclerosis, Myocardial Infarction, Atrial

						Fibrillation
M2	GO_LIPID_CATABOLIC_PROCESS	101	20	75	3	Atherosclerosis, Myocardial Infarction, Atrial Fibrillation
M6	GO_MUSCLE_SYSTEM_PROCESS	126	104	18	1	Cardiomyopathy

DISCUSSION

To date, there are hundreds of thousands of novel lncRNA transcripts that are being annotated in the human genome. Consequently, there has been some effort in assigning function to them based on various criteria such as: (i) expression levels, (ii) splicing, (iii) conservation, and (iv) experimental evidence. However, lncRNAs are known to have low levels of expression and low conservation thus, challenging to study experimentally. In an effort to combine widely used lncRNAs annotation databases and collate valuable information on lncRNAs in a systematic manner we have developed the lncRNAKB. There are several types of data sets available on lncRNAKB across 77,199 lncRNAs and 31solid organ human normal tissues that could be used to evaluate

lncRNAs function on a case-by-case basis. Here we show an example pipeline in the human heart tissue of how researchers could use lncRNAKB to provisionally label lncRNAs as functional and prioritize them for further experimental studies. The main components of the pipeline include: (i) identifying genetic variants/SNPs within a specific window that regulate the expression of lncRNAs (*cis*-eQTLs) in a tissue-specific manner since tissue enriched lncRNAs are an ideal starting point to search for human lncRNAs that are functional in a pathophysiological setting, (ii) overlapping GWAS summary data in a trait-relevant tissue with *cis*-eQTL summary data, to identify subsets of SNPs in lncRNAs that may have pleiotropic association between gene expression and disease phenotype, (iii) identifying tissue-specific modules of gene co-expression between mRNAs and lncRNAs to assign potential function to lncRNAs due to correlation with mRNAs of known function determined by modular pathway enrichment analysis with known biological processes, and (v) overlapping the subset of lncRNAs prioritized with GWAS and *cis*-eQTL colocalization within "notable" networks/pathways. This pipeline provides a way to gain insight into the function of numerous lncRNAs that we have added in the lncRNAKB and we can characterize numerous lncRNAs that may play a significant role in pathways related to many diseases/traits across several human tissues.

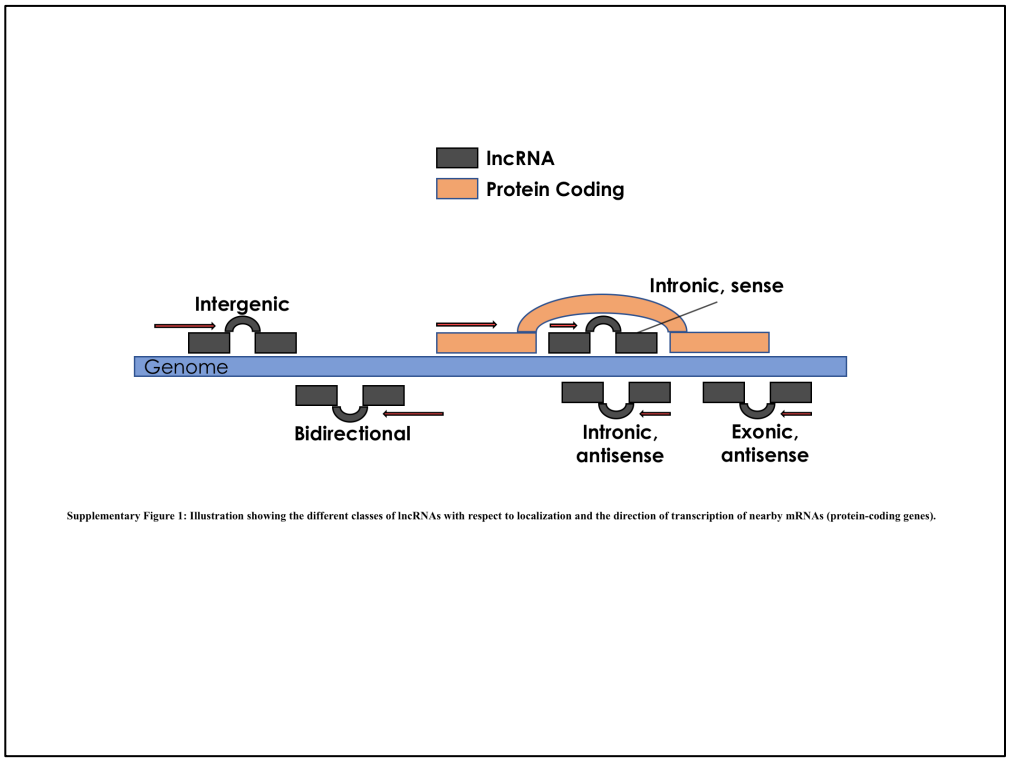
CONCLUSION AND FUTURE DIRECTIONS

There is a large volume of transcriptomics data publicly available and currently being produced at an unprecedented rate. Novel transcripts assembly using RNA-seq data is a method that generates thousands of new transcripts that need to be characterized. Several of these novel transcripts have been categorized as lncRNAs. While these data support the presence of lncRNAs in cell and tissue specific manner, bigger questions surrounding the purpose and functionality of lncRNAs in human biology remain. Several lncRNA databases have tried to address some of these questions, however, there is clearly a need to integrate the lncRNA annotation between databases to create a non-redundant list of well-annotated lncRNA entries that could be utilized by biologists to pursue research in this area.

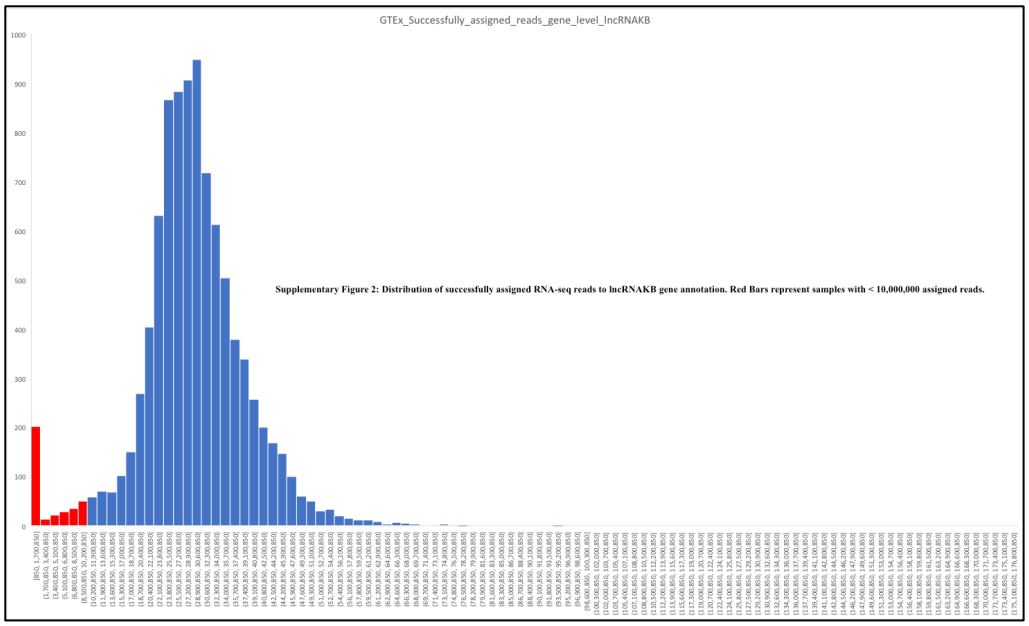
To address this need, we have created the lncRNAKB, a well-structured research tool that delivers valuable data on human lncRNAs, which can be used for data exploration and hypothesis building purpose. Briefly, the lncRNAKB is the end-product of systematic step-wise integration of six widely used lncRNA databases that resulted in a total non-redundant 99,717 genes entries that were accompanied with 530,947 transcripts and 3,513,069 exon entries. All the annotated lncRNAs can be browsed at http://www.lncrnakb.org. This web-resource also provides a comprehensive list of information that researchers can access for every lncRNA entry. This includes viewing

and downloading coding potential, conservation score, tissue specific expression information, and tissue specificity score for any gene of interest from the website. In addition to the gene-level information, we have also created a lncRNA body map where we have utilized the gene expression information and created a tissue specific gene expression and network pages that can be browsed by researchers and the information downloaded as per their research needs. This information includes gene expression count and TPM matrix for all the tissues, eQTL results and lncRNA-mRNA co-expression clusters/modules and the pathway enrichment results of respective modules. Put-together, the above-described features presented in the lncRNAKB web-resource will provide a comprehensive set of information that could be used by biologist interested in pursuing research in lncRNAs in their tissue or biological process of interest.

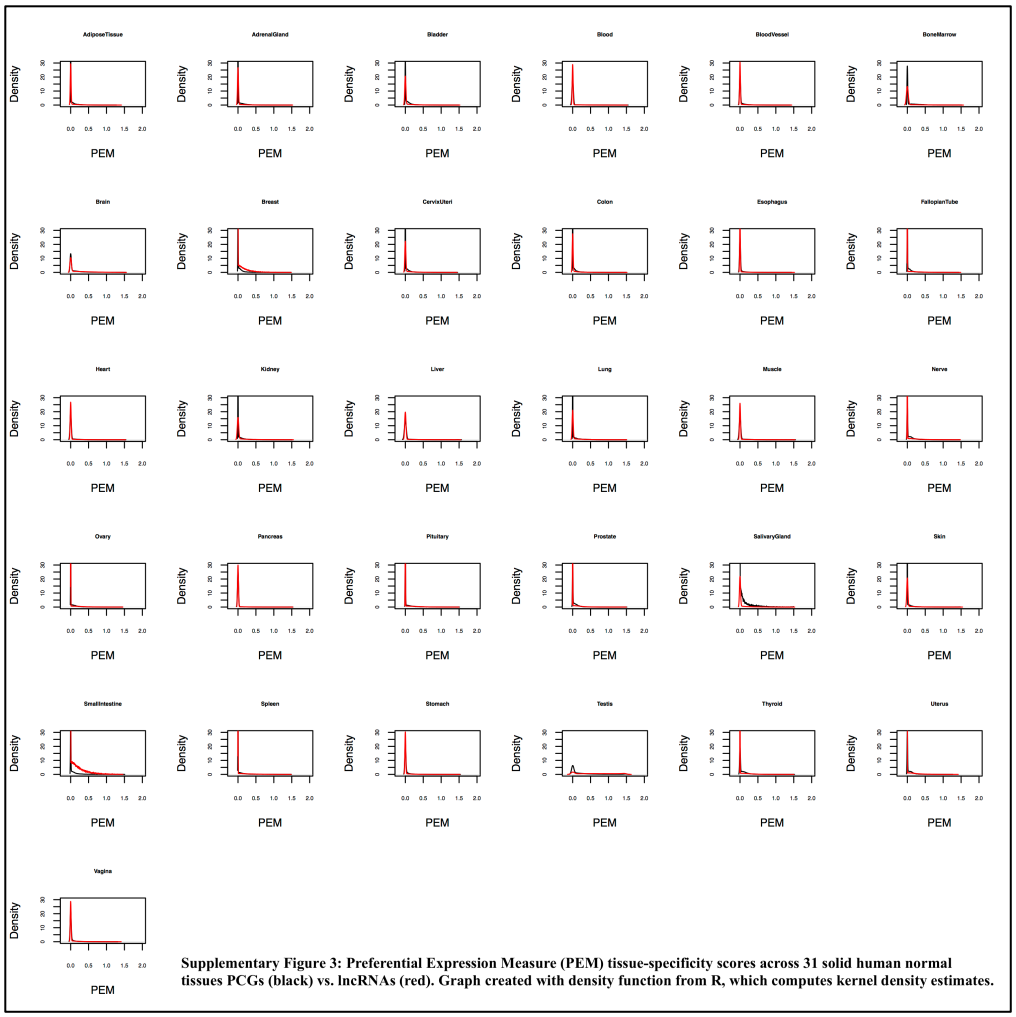
APPENDIX



Supplementary Figure 2.1: Illustration showing the different classes of lncRNAs with respect to localization and the direction of transcription of nearby mRNAs (protein-coding genes).



Supplementary Figure 2.2: Distribution of successfully assigned RNA-seq reads to lncRNAKB gene annotation. Red bars represent samples with < 10,000,000 assigned reads.



Supplementary Figure 3: Preferential Expression Measure (PEM) tissue-specificity scores across 31 solid human normal tissues PCGs (black) vs. lncRNAs (red). Graph created with density function from R, which computes kernel density estimates.

Supplementary Figure 2.3: Preferential Expression Measure (PEM) tissue-specificity scores across 31 solid organ human normal tissues PCGs (black) vs. lncRNAs (red). Graph created with density function from R, which computes kernel density estimates.

REFERENCES

Amaral,P.P. *et al.* (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146-151.

Andersson,R. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.

Andrews,S. FastQC a quality control tool for high throughput sequence data.

Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.

Arner,E. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.

Arrial,R.T. *et al.* (2009) Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics*, **10**, 239.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Aslam,B. *et al.* (2017) Proteomics: Technologies and Their Applications. *J Chromatogr Sci*, **55**, 182–196.

Barbeira,A.N. *et al.* (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*, **9**, 1825.

Bassett,A.R. *et al.* (2014) Considerations when investigating lncRNA function in vivo. *Elife*, **3**, e03058.

Bhartiya,D. *et al.* (2013) lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database (Oxford)*, **2013**, bat034.

Blanchette,M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Breiman,L. (2001) Machine Learning. 45:5.

Bu,D. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210-215.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Burgess,S. *et al.* (2017) A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*, **26**, 2333–2355.

Bycroft,C. *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.
Cabili,M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
Casper,J. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.
Chakraborty,S. *et al.* (2014) LncRBase: an enriched resource for lncRNA information. *PLoS ONE*, **9**, e108010.
Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
Chen,G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983-986.
Chen,L. *et al.* (2018) Functional genetic variants in the SIRT5 gene promoter in acute myocardial infarction. *Gene*, **675**, 233–239.
Clemson,C.M. *et al.* (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J. Cell Biol.*, **132**, 259–275.
Cooper,G.M. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
Derrien,T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
DiStefano,J.K. (2018) The Emerging Role of Long Noncoding RNAs in Human Disease. In, DiStefano,J.K. (ed), *Disease Gene Identification*. Springer New York, New York, NY, pp. 91–110.
Ehsani,R. and Drabløs,F. (2018a) Measures of co-expression for improved function prediction of long non-coding RNAs. *BMC Bioinformatics*, **19**.
Ehsani,R. and Drabløs,F. (2018b) Measures of co-expression for improved function prediction of long non-coding RNAs. *BMC Bioinformatics*, **19**.
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
Fang,S. *et al.* (2018) NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.*, **46**, D308–D314.
FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
Frank,S. *et al.* (2016) A lncRNA Perspective into (Re)Building the Heart. *Frontiers in Cell and Developmental Biology*, **4**.

Frankish,A. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, **47**, D766–D773.
Fritah,S. *et al.* (2014) Databases for lncRNAs: a comparative evaluation of emerging tools. *RNA*, **20**, 1655–1665.
Gamazon,E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.
Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049-1056.
Grote,P. *et al.* (2013) The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell*, **24**, 206–214.
GTEx Consortium *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
Gudenas,B.L. and Wang,L. (2018) Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features. *Sci Rep*, **8**, 16385.
Gusev,A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.
Guttman,M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
Han,P. *et al.* (2014) A long noncoding RNA protects the heart from pathological hypertrophy. *Nature*, **514**, 102–106.
Hangauer,M.J. *et al.* (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.*, **9**, e1003569.
Haynes,W. (2013) Benjamini–Hochberg Method. In, Dubitzky,W. *et al.* (eds), *Encyclopedia of Systems Biology*. Springer New York, New York, NY, pp. 78–78.
Hermans-Beijnsberger,S. *et al.* (2018) Long non-coding RNAs in the failing heart and vasculature. *Noncoding RNA Res*, **3**, 118–130.
Hernandez,D.G. *et al.* (2012) Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol. Dis.*, **47**, 20–28.
Hezroni,H. *et al.* (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*, **11**, 1110–1122.
Hobuß,L. *et al.* (2019) Long Non-coding RNAs: At the Heart of Cardiac Dysfunction? *Frontiers in Physiology*, **10**, 30.
Holmans,P. (2010) Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.*, **72**, 141–179.
Hon,C.-C. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, **543**, 199–204.
Housman,G. and Ulitsky,I. (2016) Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim. Biophys. Acta*, **1859**, 31–40.
Hu,G. *et al.* (2018) Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. *Oncotarget*, **9**, 18648–18663.

Hubisz,M.J. *et al.* (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinformatics*, **12**, 41–51.
Huminiecki,L. *et al.* (2003) Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics*, **4**, 31.
Iyer,M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
Jiang,C. *et al.* (2016) Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. *Oncotarget*, **7**, 7120–7133.
Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
Kim,M.-S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.
Kopp,F. and Mendell,J.T. (2018) Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*, **172**, 393–407.
Kryuchkova-Mostacci,N. and Robinson-Rechavi,M. (2017) A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinformatics*, **18**, 205–214.
Kumar,V. *et al.* (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.*, **9**, e1003201.
Lagarde,J. *et al.* (2017) High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.*, **49**, 1731–1740.
Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
Law,C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**, R29.
Li,D. and Yang,M.Q. (2017) Identification and characterization of conserved lncRNAs in human and rat brain. *BMC Bioinformatics*, **18**, 489.
Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.
Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
Lin,M.F. *et al.* (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275-282.
Long non coding RNA biology (2017) Springer Berlin Heidelberg, New York, NY.
Ma,L. *et al.* (2019) LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.*, **47**, D128–D134.
Mas-Ponte,D. *et al.* (2017) LncATLAS database for subcellular localization of long noncoding RNAs. *RNA*, **23**, 1080–1087.
Melé,M. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.
Nam,J.-W. *et al.* (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell*, **53**, 1031–1043.

Nam,J.-W. and Bartel,D.P. (2012) Long noncoding RNAs in C. elegans. *Genome Res.*, **22**, 2529–2540.
Narasimhan,V. *et al.* (2016) BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.
Nicolae,D.L. *et al.* (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**, e1000888.
Nielsen,M.M. *et al.* (2014) Identification of expressed and conserved human noncoding RNAs. *RNA*, **20**, 236–251.
O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733-745.
Ongen,H. *et al.* (2016) Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, **32**, 1479–1485.
Palazzo,A.F. and Lee,E.S. (2015) Non-coding RNA: what is functional and what is junk? *Front Genet*, **6**, 2.
Patterson,N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
Pertea,M. *et al.* (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, **19**, 208.
Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
Pohl,A. and Beato,M. (2014) bwtool: a tool for bigWig files. *Bioinformatics*, **30**, 1618–1619.
Ponjavic,J. *et al.* (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**, 556–565.
Ponting,C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
Popadin,K. *et al.* (2013) Genetic and epigenetic regulation of human lincRNA gene expression. *Am. J. Hum. Genet.*, **93**, 1015–1026.
Porcu,E. *et al.* (2018) Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits Genetics.
Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
Purcell,S. and Chang,C. PLINK 1.9.
Raychaudhuri,S. (2011) Mapping rare and common causal alleles for complex human diseases. *Cell*, **147**, 57–69.
Rio,D.C. (2014) Reverse transcription-polymerase chain reaction. *Cold Spring Harb Protoc*, **2014**, 1207–1216.
Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.
Russ,J. and Futschik,M.E. (2010) Comparison and consolidation of microarray data sets of human tissue expression. *BMC Genomics*, **11**, 305.
Russo,P.S.T. *et al.* (2018) CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics*, **19**, 56.
Saddic,L.A. *et al.* (2017) The Long Noncoding RNA Landscape of the Ischemic Human Left Ventricle. *Circ Cardiovasc Genet*, **10**.
Schaefer,R.J. *et al.* (2018) Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. *Plant Cell*, **30**, 2922–2942.
Scheuermann,J.C. and Boyer,L.A. (2013) Getting to the heart of the matter: long non-coding RNAs in cardiac development and disease. *EMBO J.*, **32**, 1805–1816.
Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
Shiraki,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15776–15781.
Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
Smith,G.D. and Ebrahim,S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, **32**, 1–22.
Son,K. *et al.* (2018) A Simple Guideline to Assess the Characteristics of RNA-Seq Data. *BioMed Research International*, **2018**, 1–9.
Storey,J.D. (2002) A direct approach to false discovery rates. *J.R. Statist. Soc. B*, **64**, 479–498.
Sun,M. *et al.* (2015) Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Mol. Cell*, **59**, 698–711.
Tan,J.Y. *et al.* (2017a) cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture. *Cell Rep*, **18**, 2280–2288.
Tan,J.Y. *et al.* (2017b) cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture. *Cell Rep*, **18**, 2280–2288.
Team,R.C. (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.
The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
Trapnell,C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562–578.

Turner,S.D. (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*.
Uszczynska-Ratajczak,B. *et al.* (2018) Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.*
Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
Veturi,Y. and Ritchie,M.D. (2018) How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? *Pac Symp Biocomput*, **23**, 228–239.
Volders,P.-J. *et al.* (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, D174-180.
Volders,P.-J. *et al.* (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*, **41**, D246-251.
Wagner,G.P. *et al.* (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.
Wang,L. *et al.* (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.
Wucher,V. *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.*, **45**, e57.
Xie,C. *et al.* (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, D98-103.
Xu,J. *et al.* (2017) A comprehensive overview of lncRNA annotation resources. *Brief. Bioinformatics*, **18**, 236–249.
Yanai,I. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.
Yang,R.Y. *et al.* (2018) A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation.
You,B.-H. *et al.* (2017) High-confidence coding and noncoding transcriptome maps. *Genome Res.*, **27**, 1050–1062.
Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
Zerbino,D.R. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet.*, **24**, R102-110.
Zhu,Z. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.

BIOGRAPHY

Fayaz Seifuddin received his Bachelor of Sciences in Computer Science from Linfield College in 2005. He received his Master of Sciences in Bioinformatics and Computational Biology from George Mason University in 2008.