5-2018

Genomic Data Mining for Functional Annotation of Human Long Noncoding RNAs

Brian L. Gudenas
Clemson University, bgudena@clemson.edu

Follow this and additional works at: https://tigerprints.clemson.edu/all_dissertations

Recommended Citation
https://tigerprints.clemson.edu/all_dissertations/2146

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.
GENOMIC DATA MINING FOR FUNCTIONAL ANNOTATION OF HUMAN LONG NONCODING RNAs

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Genetics

by
Brian L. Gudenas
May 2018

Accepted by:
Dr. Liangjiang Wang, Committee Chair
Dr. Hong Luo
Dr. Anand Srivastava
Dr. Rajandeep Sekhon
ABSTRACT

Life may have begun in an RNA world, which is supported by the increasingly vital role that RNA has been shown to perform in biological systems. To understand how the genome encodes life, one must look to the transcriptome, the set of all RNA molecules in a cell. The transcriptome illustrates which RNA transcripts are expressed at what times and this orchestrated network of gene expression is responsible for multicellular development. In humans, most genes are noncoding RNAs, meaning that they do not encode proteins. The largest class of noncoding genes are long noncoding RNAs (lncRNAs), RNA transcripts greater in length than 200 nucleotides which lack protein-coding capacity. Some lncRNAs have been shown to be key regulators; however, most lncRNAs are uncharacterized. Therefore, we developed genomic data mining methodologies for lncRNA functional annotation.

Many lncRNAs are brain-specific and their dysregulation is suspected to be involved in neurodevelopmental disorders. Two prevalent brain disorders are intellectual disability (ID) and autism spectrum disorder (ASD), which are genetically heterogeneous with unidentified genetic risk factors. In this study, we created brain developmental gene coexpression networks, for ID and ASD, to identify lncRNAs associated with known disease genes. We found lncRNAs highly co-expressed with ID genes which harbored ID-associated copy number variants (CNVs). To find ASD-associated lncRNAs we identified lncRNAs differentially expressed in the ASD brain and then refined these candidates by filtering for associations with ASD risk genes in a human brain developmental coexpression network. These candidate-ASD associated lncRNAs were associated with the
synaptic transmission and immune response pathways, in addition to residing within ASD-associated CNVs at a high frequency.

The mechanism by which lncRNAs function is partly determined by functional motifs in the RNA transcript sequence. To identify lncRNA motifs, we developed a genetic algorithm capable of finding long motifs and found a motif associated with lncRNA nuclear localization. LncRNA functions are compartmentalized within the cell; therefore, knowledge of lncRNA subcellular localization provides insight into their biological function. We developed a deep learning model that predicts lncRNA subcellular localization from lncRNA transcript sequences. This model obtained high prediction accuracy on lncRNAs with known localizations suggesting that sequence motifs are involved in subcellular localization. In summary, we developed genomic data mining methods for the functional characterization of lncRNAs based on their expression patterns and transcript sequences.
DEDICATION

For my mom and dad, without their lifelong love and support I would have never made it this far.
ACKNOWLEDGMENTS

I want to sincerely thank my research advisor Dr. Wang for his excellent mentorship on bioinformatics, genetics and science. I would not be the scientist I am today without his guidance. I also want to extend gratitude to my graduate committee members, Dr. Luo, Dr. Srivastava and Dr. Sekhon. They have been incredible resources of scientific wisdom which have greatly helped my academic development.
TABLE OF CONTENTS

Page
TITLE PAGE .................................................................................................................... i
ABSTRACT ..................................................................................................................... ii
DEDICATION ................................................................................................................ iv
ACKNOWLEDGMENTS ............................................................................................... v
LIST OF TABLES ........................................................................................................ viii
LIST OF FIGURES ........................................................................................................ ix

CHAPTER
I. LITERATURE REVIEW OF THE FUNCTIONAL ANNOTATION OF
LONG NONCODING RNAs THROUGH GENOMIC DATA MINING ... 1
   Introduction .............................................................................................. 1
   Biological function of IncRNAs .............................................................. 3
   Subcellular localization of IncRNAs ....................................................... 5
   IncRNAs in human disease ................................................................. 8
   Genomic data mining .......................................................................... 10
   Expression-based IncRNA functional annotation .............................. 14
   Sequence-based IncRNA functional annotation ............................... 17
   Concluding remarks ........................................................................... 21

II. GENE COEXPRESSION NETWORKS IN HUMAN BRAIN
DEVELOPMENTAL TRANSCRIPTOMES IMPLICATE THE
ASSOCIATION OF LONG NONCODING RNAs WITH INTELLECTUAL
DISABILITY ............................................................... 22
   Introduction ............................................................................................ 23
   Methods ................................................................................................. 25
   Results and Discussion ......................................................................... 29
   Conclusion ............................................................................................. 38

III. INTEGRATIVE GENOMIC ANALYSES FOR IDENTIFICATION AND
PRIORITIZATION OF LONG NONCODING RNAs ASSOCIATED WITH
AUTISM ........................................................................................................ 40
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Prioritized list of ID-associated lncRNA candidates</td>
<td>37</td>
</tr>
<tr>
<td>5.1</td>
<td>Performance metrics on the test set</td>
<td>86</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Functional themes of lncRNAs</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>LncRNA cellular functions</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Overview of coexpression network analysis</td>
<td>17</td>
</tr>
<tr>
<td>1.4</td>
<td>Deriving sequence-based features using k-mers</td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>ID gene enrichment in coexpression modules</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>Characteristics of ID-gene-enriched coexpression modules</td>
<td>32</td>
</tr>
<tr>
<td>2.3</td>
<td>Coexpression network plot of lncRNAs residing in ID-associated CNVs</td>
<td>36</td>
</tr>
<tr>
<td>3.1</td>
<td>Differentially expressed genes in the ASD cortex</td>
<td>45</td>
</tr>
<tr>
<td>3.2</td>
<td>Enrichment of lncRNAs and ASD genes in brain</td>
<td>48</td>
</tr>
<tr>
<td>3.3</td>
<td>Differential expression in the ASD cortex overlaid</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Characterization of modules enriched for differentially expressed lncRNAs</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>Candidate ASD-associated lncRNA characteristics</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Motif Identification accuracy</td>
<td>69</td>
</tr>
<tr>
<td>4.2</td>
<td>Motif found in nuclear enriched lncRNAs</td>
<td>71</td>
</tr>
<tr>
<td>5.1</td>
<td>Distribution of the lncRNA nuclear to cytosolic transcript ratios</td>
<td>79</td>
</tr>
</tbody>
</table>
List of Figures (Continued)

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>Overview of the DeepLncRNA algorithm</td>
<td>81</td>
</tr>
<tr>
<td>5.3</td>
<td>Model selection based on performance metrics on the validation set</td>
<td>81</td>
</tr>
<tr>
<td>5.4</td>
<td>ROC curve performance comparison on the test set</td>
<td>85</td>
</tr>
<tr>
<td>5.5</td>
<td>Heatmap of lncRNA nuclear to cytosolic transcript ratios</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>across cell types</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>Percent of annotated lncRNAs predicted to localize in the nucleus</td>
<td>89</td>
</tr>
<tr>
<td>5.7</td>
<td>DeepLncRNA predictions on lncRNAs with known subcellular localizations</td>
<td>91</td>
</tr>
</tbody>
</table>
CHAPTER I - LITERATURE REVIEW OF THE FUNCTIONAL ANNOTATION OF LONG NONCODING RNAS THROUGH GENOMIC DATA MINING

1.1 Introduction

The human genome project was a monumental journey to sequence the entire set of human chromosomes which many scientists believed would unlock the secrets of our genome. However, after the completion of the human genome it was discovered that in terms of number of genes, which encoded proteins, humans were somewhere between chickens and grapes (Pertea and Salzberg, 2010). This was a surprise to the scientific community, which preceding the completion of the human genome had estimated around 100,000 human genes only to discover about 22,000 (Pertea and Salzberg, 2010). This relatively small number of genes corresponds to only a few percent of the total human genome, while the rest of the noncoding genome does not encode proteins. This noncoding DNA was referred to as “junk DNA” because of its lack of protein-coding capacity and the presence of noncoding RNAs, transposons, pseudogenes and repetitive regions.

The advent of microarray hybridization technologies allowed the genome-wide detection of noncoding RNAs, which showed the pervasiveness of transcription in the genome. From these genome-wide analyses it is now known that the majority of the genome is actively transcribed (Hangauer, et al., 2013). Why would natural selection favor the transcription, which costs energy, of “junk DNA” with no biological purpose? This question rests on the assumption that junk DNA has no function, which is now known to be invalid. The term “junk DNA” is obsolete after genomic analyses discovered that 80% of the human genome possess biochemical functions (Bernstein, et al., 2012). Genomics research has not only shed light on the dark matter of the genome but also championed for a redefinition of the term gene, due to the vast amount of evidence for functional noncoding RNAs. Genes are no longer always required to encode proteins, thus creating two major
classes of genes, those which encode proteins are protein-coding while those that do not are noncoding RNAs (ncRNAs). This redefinition is of tremendous importance because both noncoding and protein-coding genes are functionally intertwined within the gene network of the genome.

High-throughput RNA-sequencing has largely supplanted microarrays due to the ability to discover novel RNA transcripts, which has resulted in the discovery of tens of thousands of long noncoding RNAs (lncRNAs). LncRNAs are RNA transcripts greater than 200 nucleotides in length, which do not encode proteins. The nucleotide length threshold of 200 nucleotides is largely arbitrary but does serve a key purpose to separate these transcripts from the well-known small noncoding RNAs, such as transfer RNA (tRNA), micro RNA (miRNA) and small nucleolar RNA (snoRNA). A meta-analysis of 7,256 human RNA-sequencing libraries identified 58,648 lncRNA genes, suggesting that 68% of the human transcriptome could be lncRNAs (Iyer, et al., 2015). In total, Iyer et al. identified about 99,000 human genes, which is quite close to the estimate of 100,000 prior to the human genome project, however only approximately 22,000 are protein-coding genes (Iyer, et al., 2015). While this estimate of lncRNAs may be high due to the inclusion of tumor samples, the current GENCODE annotation set (V27), a gold-standard for lncRNA annotation, contains 15,778 lncRNA genes (Harrow, et al., 2006). Therefore, it is evident that lncRNAs are abundant within the human genome with active biological functions.

LncRNAs represent the largest class of noncoding genes with several sub-classes based on their genomic position relative to protein-coding genes. In order of decreasing prevalence, the major lncRNA classifications are long intergenic noncoding RNAs (lincRNAs), antisense lncRNAs (AS-lncRNAs), sense lncRNAs and bidirectional lncRNAs (Derrien, et al., 2012). LncRNAs are most commonly transcribed by RNA polymerase II and generally post-transcriptionally modified just like mRNAs, including 5’ capping, polyadenylation and splicing (Quinn and Chang, 2015).
Indeed, the biogenesis of lncRNAs is suggested to be very similar to mRNAs in most cases. Recently, a key difference between the biogenesis of lncRNAs and mRNAs was discovered, knockouts of the ribonuclease Dicer, responsible for generating miRNAs, resulted in the decreased expression levels of hundreds of lncRNAs, yet not mRNAs (Zheng et al., 2014). This discovery suggests lncRNA biogenesis may function in a regulatory network with miRNAs. Another intriguing difference between mRNAs and lncRNAs is that some lncRNA transcripts possess higher-order structures, like 3′ secondary cloverleaf structures similar to tRNAs. These 3′ secondary structures are cleaved by ribonuclease P to form the mature lncRNA with a 3′ triple helix structure which is predicted to increase transcript stability and facilitate nuclear retention (Quinn and Chang, 2015). While there are many classifications of lncRNAs they generally share common features, including that they are predominantly spliced, expressed at low levels, are tissue-specific and their exonic regions have low levels of interspecies sequence conservation (Derrien, et al., 2012). Furthermore, lncRNA promoters are conserved at a similar level relative to protein-coding genes which suggests that they are positively selected and therefore functionally important (Derrien, et al., 2012).

1.2 Biological function of lncRNAs

The functions of lncRNAs are generally within four major mechanistic themes, which are to act as a signal, decoy, guide or scaffold (Wu, et al., 2013). These different mechanisms can act to regulate other genes at the transcriptional, post-transcriptional, translational, or epigenetic levels. LncRNAs with signaling functions act as a molecular marker to indicate specific biological conditions which then induces a response such as histone modification. An example is the lncRNA Xist which signals X-chromosomal inactivation in females for dosage compensation, by coating the chromosome in Xist
transcripts which signals successive epigenetic modifications such as DNA methylation, histone methylation and histone ubiquitination (Morris, 2016). Decoy lncRNAs function through sequence-based competitive binding of molecules, such as miRNAs which thereby reduces miRNA efficacy. This is commonly observed with lncRNAs acting as miRNA sponges, by binding miRNAs and therefore preventing the translational repression of the miRNAs targets (Geisler and Coller, 2013). Guide lncRNAs bind proteins, such as transcription factors, thereby guiding these complexes to specific genomic loci (Werner and Ruthenburg, 2015). Many lncRNAs function as guides by tethering to chromatin and facilitating the binding of protein complexes such as PRC2 and RNA polymerase II (Werner and Ruthenburg, 2015). The last major lncRNA functional theme, that is currently known, is to act as a scaffold, which mediates the physical interaction between other proteins and ncRNAs, forming ribonucleoprotein complexes. The lncRNA HOTAIR directly facilitates the binding of E3 ubiquitin ligases with multiple substrates for ubiquitination such as Ataxin-1 and Snurportin-1, thereby acting as scaffold for protein ubiquitination (Yoon, et al., 2013). It is important to note that one lncRNA is not confined to a single functional mechanism and can exhibit multiple functions simultaneously.
Figure 1.1 Functional themes of lncRNAs. lncRNAs are shown in purple performing the four main functional themes of lncRNAs. Signaling lncRNAs act in response to a stimulus to induce gene regulation, such as repression, in a spatiotemporal manner. Decoy lncRNAs act as competitive inhibitors, such as miRNA sponges thereby preventing the degradation of the targeted mRNA. Guide lncRNAs bind complexes such as chromatin modifying enzymes and facilitate the targeting to specific genomic loci either in cis or trans. Scaffold lncRNAs act as molecular glue to facilitate the interaction of multiple proteins into a ribonucleoprotein complex. Modified from (Wang and Chang, 2011)

1.3 Subcellular localization of lncRNAs

lncRNAs are not all in the nucleus; some localize to distinct subcellular localizations, such as the chromatin, cytosol, polysomes and exosomes (Kogure, et al., 2013; Heesch, et al., 2014; Chen, 2016). Like proteins, lncRNA functionality is dependent on proper subcellular localization. While many studies suggest that lncRNAs are predominantly enriched in the nucleus, there is countervailing evidence of lncRNA cytoplasmic enrichment (Derrien, et al., 2012; Ulitsky and Bartel, 2013; Heesch, et al., 2014; Werner and Ruthenburg, 2015). However, the factors that govern lncRNA subcellular localization
are mostly unknown. Recently, a nuclear retention motif was identified in the IncRNA BORG through a mutational screen (B. Zhang, et al., 2014). This motif contained a pentamer sequence with two upstream restriction sites and mutations of the motif resulted in loss of nuclear retention (B. Zhang, et al., 2014). Furthermore, the number of copies of this motif present in IncRNAs is correlated with the nuclear to cytoplasmic transcript ratio. The subcellular localization of IncRNAs can reveal insights into their functionality due to the fact that certain IncRNA functionalities are specific to certain cellular locations. Gene regulation by IncRNAs, at the transcriptional, post-transcriptional or epigenetic levels is performed within the nucleus (Sun, et al., 2017). However, translational regulation, binding of miRNAs, subcellular trafficking and processing to produce small ncRNAs are all IncRNA functions exclusive to the cytoplasm (Rashid, et al., 2016). Therefore, predicting the subcellular localization of IncRNAs can provide useful information regarding its biological function.
Figure 1.2 LncRNA cellular functions. (1) LncRNAs can bind chromatin modifying enzymes and facilitate histone modifications such as the trimethylation of histone 3 at lysine 27 (me3K27) inducing gene silencing. (2) LncRNAs are able to form an RNA-DNA triplex which blocks accessibility to gene promoter regions. (3) LncRNAs facilitate the nuclear organization of subcellular structures like nuclear paraspeckles. (4) Alternative splicing can be regulated by lncRNAs ability to bind intronic segments of pre-mRNA. (5) LncRNAs can bind mRNA and ribosomes to regulate translation. (6) LncRNAs are capable of sequestering miRNAs as a miRNA sponge thereby preventing the degradation of the targeted mRNA. (7) LncRNAs can be processed by enzymes such as Dicer to form small ncRNAs like endogenous small interfering RNAs (siRNAs) leading to RNA interference. (8) LncRNAs can function in the subcellular localization of proteins to complexes such as the proteasome. Figure Modified from (Rashid, et al., 2016)
1.4 LncRNAs in human disease

LncRNAs have been implicated in many human diseases. This is not surprising because they are versatile regulators of gene expression with known roles in tissue development. Mutations in lncRNAs can alter their functional efficacy thereby causing aberrant downstream consequences. Interestingly, more than 90% of disease-associated SNPs are found within noncoding regions of the human genome (Maurano, et al., 2012; Ricaño-Ponce and Wijmenga, 2013). Disease-associated single nucleotide variants (SNPs) can alter the expression levels of lncRNAs (Kumar, et al., 2013). In addition, copy-number variants, another major source of disease, have also been shown to change lncRNA expression levels, which has been associated with cancer (Xu, et al., 2017). Aberrant lncRNA expression can dysregulate biological pathways which can be associated with disease.

A significant fraction of lncRNAs are specifically expressed in neuronal tissues, therefore we would expect lncRNAs to also be involved in brain disorders (Derrien, et al., 2012). In mammals, lncRNAs have direct roles in neural differentiation and synaptic plasticity (Wu, et al., 2013; Clark and Blackshaw, 2014). Therefore, it is no surprise that lncRNAs have been implicated in neurodegenerative, psychiatric and neurodevelopmental disorders. For example, the lncRNA BACE1-AS increases the stability of the antisense BACE1 mRNA, thereby reinforcing a positive feedback loop increasing beta amyloid levels (Faghihi, et al., 2008). Beta-amyloid peptides can aggregate into extracellular amyloid plaques which are toxic to neurons and can be seen in Alzheimer brains. Two of
the most predominant neurological disorders, intellectual disability (ID) and autism spectrum disorder (ASD), are genetically heterogeneous.

ID and ASD affect up to 3% and 1%, respectively, of the human population causing social, economic and health problems worldwide (Srivastava and Schwartz, 2014). ID is characterized by diminished intellectual capacity and adaptive reasoning, while ASD is recognized by impaired social communications and restrictive or repetitive behavior. Both disorders originate in early childhood and have a large genetic component, yet are genetically heterogeneous with hundreds of implicated risk genes. A large fraction of genes involved in these disorders are associated with the synaptic transmission pathway, supporting the notion that this crucial biological pathway is disrupted or dysregulated in ID and ASD (Verpelli, et al., 2013; De Rubeis, et al., 2014). Moreover, up to 70% of people with ASD also have some form of ID, while 10% of people with ID have some form of ASD (Srivastava and Schwartz, 2014). In most cases of ID or ASD, the exact genetic casual factors of these disorders is unable to be determined, despite being highly genetic disorders as determined by heritability estimates (O’Roak, et al., 2012; Kiser, et al., 2015). However, until recently, only protein-coding genes were studied for involvement in these disorders. Therefore, due to large fraction of cases without an identifiable genetic diagnosis it is probable that the majority of genetic risk factors of ID and ASD reside outside the exome, such as in lncRNAs. Recently, studies have shown involvement of lncRNAs in both disorders (Vondervoort, et al., 2013)

LncRNAs are implicated in many intellectual disability syndromes, such as Coffin-Siris syndrome, Prader-Willi Syndrome, Fragile X-syndrome and Rett syndrome
An ultra-conserved lncRNA named Evf2 was shown to bind the protein BRG1, part of the SWI/SNF nucleosome remodeling complex, and localize to enhancers causing enhancer repression (Cajigas, et al., 2015). Interestingly, mutations causing Coffin-Siris syndrome, a syndromic intellectual disability, were found to reside in the RNA-binding domain of BRG1 thereby implicating the lack of Evf2-induced transcriptional repression in this disorder (Cajigas, et al., 2015). In ASD, hundreds of lncRNAs were found to be differentially expressed in the prefrontal cortex of ASD patients relative to controls (Ziats and Rennert, 2013). Furthermore, the observed differences in lncRNA expression between the prefrontal cortex and cerebellum of ASD patients were significantly lower than between the same brain regions in controls (Ziats and Rennert, 2013). In addition, several specific lncRNAs, such as ST70T1 and PTCHD1AS1 have been associated with ASD through population sequencing studies (Vondervoort, et al., 2013). In a genomic differential expression analysis of ASD leukocytes over 3,000 lncRNAs were found to be differentially expressed, including thirteen lncRNAs associated with synaptic functions (Wang, et al., 2015). These results suggest that lncRNAs are involved in the development of both ID and ASD, which could help identify novel genetic risk factors.

1.5 Genomic data mining

Investigation into biological functions of lncRNAs experimentally, such as gene knockouts, is a highly time-consuming and laborious process which is not easily amenable to parallelization. However, genomic data mining offers a solution to these obstacles. Genomic data mining is the process of utilizing biological datasets to extract hidden
knowledge regarding a specific biological question. Knowledge is gained through the use of data mining algorithms which identify patterns and relationships within the data. Genomic data mining typically consists of three major steps, which are dataset acquisition, data integration and application of data mining algorithms.

Dataset acquisition involves the querying of databases, which are generally publicly available, for biological data which are relevant to the hypothesis at hand. Data integration is the aggregation of diverse or heterogeneous datasets to common formats so that statistical learning can identify generalizable knowledge which is not confounded by variables of individual datasets. Lastly, data mining algorithms, such as coexpression network analysis or deep learning, can be applied to the data for knowledge discovery to help answer a biological question. While still a relatively new discipline, genomic data mining is of monumental importance and will continue to grow in demand proportionally to the vast amount of data being generated. The sequence read archive (SRA), a bioinformatics database for sequencing data, was founded in 2007 and by 2017 already contained over 1000 terabytes of sequencing data. This is only one database, but highlights the need for genomic data mining methodologies to be able to extract knowledge from this explosion of biological data.

A subfield of data mining is machine learning. Machine learning is the creation of models which learn from data, on their own without explicit instructions, to generate predictions for new data instances. Machine learning algorithms use a dataset as input which is composed of data instances, generally referred to as samples in biology, and features that describe these data instances. Therefore the more informative your features
are regarding the problem at hand, the better your prediction results will be. Machine learning is further subdivided into supervised and unsupervised learning. Supervised learning is the process of learning from labeled data in which the response variable, the feature that needs to be predicted, is already known. This allows the learning of generalizable knowledge which can be used to predict the response variable for new data instances with unknown labels. Unsupervised learning is the act of learning from unlabeled data to identify hidden structure within the data for the clustering of data instances into representative groups. Therefore, supervised learning requires a training set with a labelled response variable, a categorical or numeric value, for which to predict, while unsupervised learning does not require labelled data. Both types of machine learning are widely used for the functional annotation of genes.

Supervised machine learning algorithms are used more frequently in genomics due to their ability to predict variables of interest. The ability to make biologically relevant predictions from genomic data has garnered intense interest due to the capability of machine learning algorithms to learn novel complex patterns and structure within complex data. Due to the growing size and complexity of genomic data, machine learning algorithms are needed to discover knowledge in a timely and efficient manner. Two of the most popular machine learning algorithms are the support vector machine (SVM) and random forest (RF). Both of these algorithms are easy to implement due to a small number of parameters, but can achieve high accuracy for both linear and non-linear problem specifications. SVMs and RFs have been used for a diverse array of biological problems from the classification of IncRNAs to the prediction of IncRNA-protein interactions and
ASD-associated IncRNAs (Muppirala, et al., 2011; Cogill and Wang, 2016; Pian, et al., 2016). While SVMs or RFs have achieved superior prediction accuracy in most problems, recently a new set of advanced machine learning algorithms, known as deep learning algorithms have attained first-rate prediction accuracy in complex problems (Ching, et al., 2017).

Deep learning is a term for a group of algorithms utilizing deep artificial neural networks, which mimic how biological brains function through the aggregation of artificial neurons. Artificial neurons are connected to the input features and through randomized transformations, each neuron learns a different representation of the input data. The neurons that are connected to the features, are then connected to another layer of neurons, which learn from the first layer’s output. Therefore, layers of neurons learn from the previous layer and pass on their information to the next layer, meaning each layer learns more complex features of the input data instances. The number of neuronal layers are what give a deep learning model its depth, with each layer transforming its inputs to discover new more advanced features. The discovery of more complex representative features is the main advantage of deep learning techniques over other machine learning algorithms which are restricted to the input features provided. Issues with deep learning include that it requires big datasets to learn generalizable knowledge due to the copious number of parameters that need to be tuned. However, in the age of big data this is becoming less of a concern which is why deep learning will become the next frontier of machine learning in genomics.
In addition to machine learning, evolutionary algorithms, such as the genetic algorithm, are widely used for genomic data mining (Ujjwal, *et al.*, 2011). Genetic algorithms can solve difficult optimization problems due to their ability to stochastically evolve solutions over time, like natural selection. This works by first generating many possible candidate solutions, generally at random, which then evolve towards superior solutions. Genetic algorithms are an iterative process which consists of steps which are continuously repeated over a course of generations, until the best solution can no longer be improved upon. Every solution consists of smaller indivisible pieces, like chromosomes, which in total make up all the properties of the solution. For example, to discover regulatory motifs in the flanking regions of genes, a genetic algorithm was created which used 7-letter substrings of DNA as candidate solutions (Liu, *et al.*, 2004). These candidate motifs were then mutated and mated with each other through crossover for hundreds of generations to find enriched DNA binding motifs. Two major types of genomic data mining for lncRNA functional annotation are discussed below.

### 1.6 Expression-based methods for lncRNA functional annotation

A well-known property of lncRNAs is their tissue and developmental specificity, and thus the expression profile of lncRNAs can be used to help identify its biological function. These methods are especially useful for lncRNAs because they do not encode proteins, therefore the RNA is the functional unit and lncRNA transcript abundances are proportional to their functionality. In contrast, the correlation between a protein and their cognate mRNA abundance has been suggested to be rather poor (Maier, *et al.*, 2009). The
biological functions of lncRNAs can be investigated by examining the differences in expression levels between different groups of samples, such as disease vs control tissues or fetal brains vs adult brains. The two most prominent expression-based approaches are differential gene expression analysis and coexpression network analysis.

Differential gene expression analysis identifies genes which have statistically significant differences in expression levels between two conditions and is commonly used to find genes associated with a disease, tissue-type or experimental treatment. Differential gene expression is a useful method to screen for lncRNAs that may be involved in a condition of interest. However, insights regarding their biological function are limited, although this screening allows for much more efficient further experimentation, such as gene knockouts, to establish the lncRNA’s biological function. For example, differential expression analysis after chemical activation of p53, a tumor suppressor gene, on three different cell lines showed strong up-regulation of an unknown lncRNA, which was subsequently named p53-induced noncoding RNA (PINCR) (Chaudhary, et al., 2017). A gene knockout was performed to functionally characterize PINCR resulting in the discovery that PINCR regulates targets of p53 that manage G1 arrest following DNA damage (Chaudhary, et al., 2017). Differential expression can also reveal lncRNAs which are tissue-specific and may have developmental functions such as the lncRNA, linc-PINT. This lncRNA was found to be up-regulated in radial glial cells of the developing human brain, gene knockouts subsequently suggested this lncRNA functions in cell proliferation (Liu, et al., 2016). While gene knockouts are a valuable experimental tool for lncRNA
functional annotation they are costly and laborious, unlike gene coexpression network analysis.

Gene coexpression network analysis is an unsupervised clustering method which enables the inference of gene’s biological function based on the strength of connections to genes of known function. This method clusters genes by their expression profiles into groups of genes, known as gene modules. These gene modules are then functionally annotated through gene set enrichment analysis which uses a statistical test to check if the overall functional enrichment is different than what would be expected by random chance. Gene coexpression network analysis leverages the biological properties of known genes to gain insight on the uncharacterized genes through a guilt-by-association heuristic. One of the first applications of this technique for the annotation of lncRNA functions was done using microarrays in mice resulting in the predicted annotation of 340 lncRNAs based on coexpression, network characteristics and genomic adjacency (Liao, et al., 2011). These mouse lncRNAs were associated with processes mainly involved in tissue development, cellular transport and metabolic functions (Liao, et al., 2011). Coexpression networks were also used to infer the biological functions of lncRNAs conserved in primates, finding functional associations for fundamental mammalian processes including spermatogenesis, synaptic transmission and placental development (Necsulea, et al., 2014). One concern for expression-based methods for lncRNA functional annotation is that gene expression is dependent on the tissue type, therefore without the relevant dataset the analysis is not possible. However, sequence-based methods for functional annotation are applicable, in
this case, because the lncRNA transcript is the same regardless of what cell type it is expressed in.

**Figure 1.3** Overview of coexpression network analysis. Genes are clustered based on their correlation of gene expression with other genes resulting in gene modules, shown in different colors. Within each gene module, nodes represent genes while edges represent correlations. The length of each edge is inversely proportional to its correlation so the shorter the edge the higher the coexpression. Genes are then annotated based on the functional enrichment of the known genes in the module.

### 1.7 Sequence-based methods for lncRNA functional annotation

LncRNA biological functionality resides in the primary sequence, such as motifs, or from the 3D structure. However, the prediction of lncRNA structure is a research area still in its infancy, mainly due to the paucity of experimentally validated lncRNA structures. Since sequence ultimately dictates structure, the capacity of lncRNAs to bind RNA, DNA and proteins, therefore, is present within the lncRNA primary sequence. While lncRNAs are generally conserved at low levels, functional motifs have been identified in lncRNA sequences, such as miRNA and DNA binding sites. In addition, sequence motifs have been
found which are associated with specific lncRNA higher-order structures such as the AUGC tetraloop motif (Li, et al., 2016). Moreover, lncRNA sequence motifs have been found to be directly involved in their post-transcriptional modifications as well as subcellular localization (Mondal et al., 2015; Zhang et al., 2014). The motifs present in the lncRNA transcript therefore provide insight into its mechanism of action.

LncRNA transcript sequences can be utilized as inputs into genomic data mining algorithms to identify motifs or make functional predictions. Almost all machine learning algorithms require a fixed number of input features which creates a challenge due to lncRNA nucleotide sequences of variable length. A common way around this issue is to transform the sequences into k-mers of all possible substrings of a fixed length $k$. The number of k-mers grows exponentially with the selection of $k$, therefore most sequence-based methods using nucleotides as inputs use 4-mers, which equals 256 unique DNA 4-mers. Larger k-mers are more informative but as the number of features increases, so should the number of data instances in your training data, to prevent overfitting. K-mers can be highly informative features because they can contain known functional motifs without having to know the motif beforehand. However, k-mers contain only local sequence information and cannot include long-range sequence interactions, unlike structural features.
Figure 1.4 Deriving sequence-based features using k-mers. K-mers are all possible substrings of length $K$ in a sequence, which are counted to form a K-mer matrix. This method allows the derivation of a fixed number of features from sequences of variable lengths.

LncRNA sequences can also be used as inputs into structural prediction algorithms that predict secondary structure conformations, which can then be used to derive structural associated features (Lorenz, et al., 2011). These structurally derived features commonly include vectors of hydrogen bonding, minimum free energy secondary structures, RNA loops and the number of possible structural conformations (Lorenz, et al., 2011).
LongTarget, a prediction model for lncRNA DNA-binding motifs, was developed based on traditional and hoogstein base-paring features which can predict DNA binding sites in lncRNA sequences that form RNA-DNA triplexes (He, et al., 2018). RNA-RNA interactions have also been predicted using RNA sequences to infer the minimum free energy secondary structure of the interacting RNAs using a genetic algorithm (Montaseri, et al., 2014). While predicting structural features using lncRNA sequences can be informative, lncRNA structural predictions are highly variable based on the algorithm used due to the infancy of our knowledge regarding lncRNA secondary structure.

The interaction of lncRNAs with proteins is a well-known source of lncRNA functionality, whether it is guiding chromatin modification enzymes or acting as molecular scaffold. Several machine learning models have been developed to predict lncRNA protein interactions from their sequences using k-mer based approaches. Using lncRNA and protein sequence pairs to derive k-mers, the method RPISeq predicts lncRNA-protein interactions with an accuracy of 80%, utilizing a support vector machine and random forest (Muppirala, et al., 2011). LncPro predicts lncRNA-protein interactions using a structure-based approach, by first deriving structure-based features from the lncRNA and protein sequences, which attains prediction accuracy similar to RPISeq (Lu, et al., 2013). The structural features derived include vectors of hydrogen bonding, minimum energy secondary structures and vectors of Van der Waal’s which are then transformed using a Fourier transform to attain features of fixed length (Lu, et al., 2013). However, through the addition of a stacked denoising autoencoder, a type of deep neural network, the IPminer method was able to achieve an accuracy of 89% for lncRNA-protein interactions using
only k-mers (Pan, et al., 2016). This accuracy is superior to that of RPISeq and the structural based LncPro, suggesting that deep learning with k-mers is an optimal approach for the prediction of lncRNA-protein interactions. In summary, lncRNA sequences appear highly informative regarding their binding potential with DNA, RNA and proteins which can provide insights into their biological function.

1.8 Concluding remarks

In this chapter, we have explored the functional roles of lncRNAs and their importance in gene regulation especially in context of human neurodevelopmental disorders such as ID and ASD. Furthermore, we briefly examined some genomic data mining methods for lncRNA functional annotation based on their expression or primary sequences. In chapter 2, brain developmental coexpression networks are used to identify lncRNAs associated biological pathways dysregulated in intellectual disability. In chapter 3, lncRNAs are found to be differentially expressed in the ASD brain that are also highly co-expressed with ASD risk genes in neuronal development. In chapter 4, a genetic algorithm is developed that finds long sequence motifs in groups of lncRNA sequences. In chapter 5, a deep neural network is constructed that predicts lncRNA subcellular localization directly from transcript sequences.
CHAPTER II - GENE COEXPRESSION NETWORKS IN HUMAN BRAIN DEVELOPMENTAL TRANSCRIPTOMES IMPLICATE THE ASSOCIATION OF LONG NONCODING RNAs WITH INTELLECTUAL DISABILITY

Brian L. Gudenas and Liangjiang Wang
Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA.
Published: Bioinformatics and Biology Insights, 9, (2015)

Abstract

The advent of next-generation sequencing for genetic diagnoses of complex developmental disorders such as intellectual disability (ID) has facilitated the identification of hundreds of predisposing genetic variants. However, there still exists a vast gap in our knowledge of causal genetic factors for ID as evidenced by low diagnostic yield of genetic screening, in which identifiable genetic causes are not found for the majority of ID cases. Most methods of genetic screening focus on protein-coding genes, however, noncoding RNAs may outnumber protein-coding genes and play important roles in brain development. Long noncoding RNAs (LncRNAs) specifically have been shown to be enriched in the brain and have diverse roles in gene regulation at the transcriptional and post-transcriptional levels. LncRNAs are a vastly uncharacterized group of noncoding genes, which could function in brain development and harbor ID predisposing genetic variants. We analyzed LncRNAs for coexpression with known ID genes and affected biological pathways within a weighted gene coexpression network derived from RNA-sequencing data spanning human brain development. Several ID-associated gene modules were found to be enriched for LncRNAs, known ID genes and biological pathways dysregulated in ID. Utilizing a list of de novo and pathogenic copy number variants (CNVs), a major risk factor detected in ID probands,
we identified lncRNAs overlapping these genetic structural variants. By integrating our results, we have made a prioritized list of lncRNAs associated with ID based on a coexpression network in the developing brain and genetic structural variants found in ID probands.

2.1 Introduction

Intellectual disability (ID) is a developmental brain disorder characterized by diminished intellectual function and adaptive behaviors, with an estimated prevalence in the population between 1 and 3% (Leonard and Wen, 2002). Currently, in most cases of ID an identifiable genetic cause is still unclear (Kaufman, et al., 2010). However, known ID genes are predominantly involved in synaptic function such as cytoskeletal reorganization and synaptic plasticity (Verpelli, et al., 2013). Disruptions in synapse function likely cause a cascade of detrimental effects persisting into brain development, indicating the necessity of the precise spatiotemporal gene expression required for normal brain development. Genetic studies on ID have almost exclusively focused on variants in protein-coding genes, such as copy number variants (CNVs) and single nucleotide variants (SNVs). However, in human cells, the majority of RNA transcripts may not encode proteins, suggesting the need to expand the search for causal factors of ID beyond protein-coding genes (Yoon, et al., 2014). In this study, we have examined over 4,000 long noncoding RNAs (lncRNAs) to assess their potential association with ID through the integration of RNA-sequencing data and genetic structural variants detected in ID affected individuals.
LncRNAs are longer than 200 nucleotides with diverse emerging regulatory mechanisms; while some lncRNAs may encode small peptides, lncRNAs are vastly untranslated (Slavoff, et al., 2013; Lander, 2014; Li, et al., 2014). Functionally, lncRNAs have been shown to be involved in transcriptional and post-transcriptional regulation, in addition to roles in epigenetic mechanisms (Schaukowitch and Kim, 2014). Specifically in the brain, lncRNAs have been shown to be involved in neural differentiation and synaptic plasticity (Wu, et al., 2013; Iyengar, et al., 2014; Schaukowitch and Kim, 2014). CNVs in the genomic regions of these lncRNAs likely cause disruptive effects through alteration of gene copy number, thereby leading to aberrant expression and possible downstream effects. The identification of lncRNAs involved in neuronal and developmental processes which are also affected by ID-predisposing CNVs may lead to the identification of novel ID genes.

Considering the genetic heterogeneity of intellectual disability and largely uncharted molecular roles of lncRNAs, we chose to construct a gene coexpression network using RNA-seq data from the developing brain. This analysis facilitates the functional annotation of uncharacterized lncRNAs by clustering genes based on correlations of expression levels across brain developmental stages. Weighted gene coexpression network analysis (WGCNA) is a well-established method for biological data mining (Langfelder and Horvath, 2008). In previous studies, WGCNA has been used to elucidate convergent molecular pathways, specific brain regions and developmental periods associated with ASD, illustrating the functionality of coexpression networks for complex developmental disorders (Voineagu, et al., 2011; Parikshak, et al., 2013; Liu, et al., 2014; Sugathan, et al.,
Within a coexpression network, a module represents a group of correlated genes, based on expression profiles, which likely share genetic regulation and/or biological function. Thus, by clustering genes into coexpression modules, the biological function of a lncRNA may be inferred from the Gene Ontology enrichment of the known genes in the module and the degree to which the lncRNA correlates within the module. In this study, the gene coexpression network is based on a comprehensive dataset of human brain developmental transcriptomes. We compiled a list of known ID genes from multiple sources, and used them to identify potential ID-associated lncRNAs in the coexpression network. Moreover, we used a list of CNVs identified in a large cohort of probands with ID to identify lncRNAs residing within the CNVs. This approach has facilitated the prioritization of candidate ID-associated lncRNAs based on the developing brain gene coexpression network seeded with known ID genes and disruptive genetic variants found in ID probands.

2.2 Methods

Brain developmental transcriptome data

The BrainSpan developmental transcriptome dataset contains RNA-seq expression profiles summarized to gene-level reads per kilobase million mapped reads (RPKM) with GENCODE 10 annotations (Gardiner, 2015; Greenwood Genetic Center, 2015). Only samples less than or equal to 3 post-natal years and from the neocortex were used in this study. This resulted in 210 RNA-seq samples derived from 28 different individuals across 11 regions within the neocortex. LncRNAs were related to the genes of the developmental
transcriptome by using the GENCODE v22 long noncoding RNA annotations. Genes were variance-filtered by removing the lowest 25% of genes based on standard deviation. The gene-level RPKM values were then normalized by using the log2 (RPKM+1) for all further analysis.

Gene lists

The full ID gene list was compiled through the combination of three curated gene sets, the “ID all” gene set by Parikshak et al, the XLID gene panel and the ID gene database (Parikshak, et al., 2013; Gardiner, 2015; Greenwood Genetic Center, 2015) The ASD gene list was obtained from the SFARI Human Gene Autism Database (Basu, et al., 2009). The “ID only” gene list was created by removing genes from the full ID gene list that are also classified as ASD genes. The “ID & ASD” gene set contains all overlapping genes from the “ID all” and ASD gene list. The “XLID” gene set represents the XLID diagnostic panel (Greenwood Genetic Center, 2015). All gene lists are provided in Additional File A-1.

Weighted gene coexpression network analysis

Signed weighted gene coexpression network analysis (WGCNA) was performed in R, version 3.2.0, utilizing the WGCNA, v1.46, R package (Langfelder and Horvath, 2008; R Core Team, 2015). Readers interested in the mathematical derivations of the technique are recommended to view the excellent WGCNA theory website (http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/). Traditional coexpression networks are created by filtering a symmetric correlation matrix by a hard-threshold which likely results in many false negatives due to the arbitrary cutoff of the threshold. WGCNA
does not suffer from this pitfall due to utilization of a soft-threshold which simply emphasizes high correlations. In addition, this soft threshold enables the coexpression network to approximate scale-free topology, an inherent property of biological networks (Albert, 2005). First, a correlation matrix, also known as a symmetric adjacency matrix, was made by calculating the biweight midcorrelation, a robust alternative to the Pearson correlation coefficient, between all gene pairs. This adjacency matrix was then raised to a soft threshold power of 10 to achieve a scale free topology. The topological overlap measure (TOM) is computed for all genes by taking into account direct pairwise correlations as well as shared correlations between other genes. Gene modules were formed by unsupervised clustering of genes of the hierarchical cluster tree based on the threshold of dissimilarity, 1-TOM. The minimum module size was set at 50 genes with the module merging cut height set at 0.20. The WGCNA results from this study are provided in Additional File A-2.

Modular gene set enrichment analysis

For each gene set, we determined overrepresentation within each module using the Fischer’s exact test. All resulting p-values from each gene set were adjusted by the false discovery rate method, Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). We required an adjusted p-value of <0.05 and odds ratio of >1 to classify modular enrichment of a specific gene list. Heatmaps of the −log10(adjusted P-value) were created with the gplots R package (Warnes, et al., 2009). Results from the ID gene enrichment for all modules are provided in Additional File A-3.
CNV detection

CNV genomic coordinates were obtained from 213 idiopathic ID subjects using array CGH analysis (Qiao, et al., 2013). We extracted the genomic coordinates of all lncRNAs, from the human Ensembl assembly NCBI 37 using the R package “biomaRt” (Durinck, et al., 2009). We performed genomic liftover, using the “rtracklayer” R package for conversion to assembly NCBI 36 coordinates, to match the coordinates of the CNV dataset (Lawrence, et al., 2009). All CNV overlaps were quantified using the “GenomicRanges” R package (Lawrence, et al., 2013).

LncRNA candidate prioritization

The network plot was constructed by selecting the top two lncRNAs with the highest modular membership values within each module that was statistically enriched for the “ID Only” gene list. Modular membership is defined as the correlation of a gene's expression profile with the eigengene of the module. Next, for each of the chosen lncRNAs we selected the top three most highly correlated genes based on connectivity. The network plot was generated using Cytoscape v3.1.1 package by visualizing all biweight midcorrelations above 0.65 (Shannon, et al., 2003). LncRNA prioritization was performed by ranking all lncRNAs that overlapped a de novo or DECIPHER CNV by the maximal absolute Pearson correlation coefficient to all genes in the full ID gene list. The full ranked lncRNA list is available in Additional File A-4, including genomic location and type of CNV overlap.
2.3 Results and Discussion

Coexpression network analysis and identification of gene modules enriched with known ID genes and lncRNAs

For weighted gene coexpression network analysis we utilized a comprehensive developmental transcriptome dataset, which contains RNA-seq data from 210 neocortical samples during early brain development (from 8 weeks post-conception to 3 post-natal years). This developmental transcriptome dataset contains more than 50,000 genes, including over 9,000 currently classified as lncRNAs. To reduce the search space for ID-associated lncRNAs, we chose to remove genes thought not to be involved in brain development by filtering out genes with the lowest variance across the different stages during brain development. After removing the bottom quartile of genes based on variance, we retained a total of 39,000 genes, 6,000 of which are classified as lncRNAs. In addition, we utilized high thresholds for module construction for further filtering, resulting in a final gene coexpression network comprised of 26,030 genes including 4,070 lncRNAs, which are distributed amongst 16 gene modules labeled by colors.

Next, we asked if the known ID genes converge onto specific coexpression modules. Since coexpression implies shared function and/or regulation, ID-gene-enriched modules would suggest that the genes in these modules are potentially involved in ID-associated biological pathways. We compiled a comprehensive list of known ID genes (Additional File A-1) by combining gene lists from three curated sources, a diagnostic X-linked ID gene panel, the ID gene database project and the ID gene list from a recent publication (Parikshak, et al., 2013; Gardiner, 2015; Greenwood Genetic Center, 2015).
Identification of ID-related modules also requires taking into account autism spectrum disorders (ASD), which are present in up to 20% of ID cases (Kaufman, et al., 2010). Utilizing the ASD gene list from the Simons Foundation Autism Research Initiative (SFARI) Human Gene Database, we found our compiled ID gene list had a 20% overlap with the ASD gene list (Basu, et al., 2009). This is not surprising because of the shared genetic components between the disorders, such as the synaptic plasticity and transmission pathways (Srivastava and Schwartz, 2014). To assess the enrichment of non-syndromic ID, we created an ID-specific gene set, called “ID only”, by removing genes present in both the ID and ASD gene lists, and the overlapping genes are referred to as “ID & ASD”.

By mapping the gene sets to the developmental coexpression network, we have identified 2 out of 8 modules that are enriched for “ID only” genes without also being enriched for “ID & ASD” genes or ASD genes (Figure 2.1 and Additional File A-2). These two ID-specific modules are labeled as magenta and turquoise. The highest scoring term of biological processes from Gene Ontology enrichment analysis for the turquoise module is immune response (p-value < 0.001), whereas, the magenta module does not show any significant functional term enrichment (Figure 2.2B). Interestingly, the blue, black and purple modules are enriched for all four ID gene sets, in addition to the ASD gene list. These modules are likely involved in core synaptic and regulatory pathways which are affected in both ID and ASD. Gene Ontology functional analysis finds enrichment for transcriptional regulation, synaptic transmission and protein localization, respectively (Figure 2.2B). Furthermore, the blue, black, brown and turquoise modules possess a
significant amount of lncRNAs (Additional File A-3), and thus are interesting modules to examine the relationships between lncRNAs and known ID genes.

**Figure 2.1** ID gene enrichment in coexpression modules. Overrepresentation of known ID genes as a heat map displaying the \(–\log_{10}\) (adjusted P-values) of the Fischer’s exact test. Odds ratio values are overlaid onto the heat map if they are greater than one. Only modules with an adjusted p-value < 0.05 in at least one gene set are displayed resulting in 12 out of the 16 modules shown.
Figure 2.2 Characteristics of ID-gene-enriched coexpression modules. A) Gene expression of lncRNAs from 2 months post-conception up to one post-natal year, grouped by module, was normalized and plotted across developmental time using a scatter smoothing function. The red line represents birth and the shaded adjacent regions of the trend lines represent 95% confidence intervals. B) Modular gene enrichment was analyzed using DAVID and the top 5 Gene Ontology terms of biological processes were displayed. The red vertical lines represent the significance threshold of \( p \)-value= 0.01.

Besides modular functional annotation, we also examined the developmental expression patterns of the co-expressed lncRNAs in these modules, which may be used to infer biological function (Figure 2.2A). Interestingly, the two modules, turquoise and
magenta, which are enriched specifically for non-syndromic ID genes, show a very similar developmental expression trajectory. LncRNAs of both modules become highly expressed in the late fetal period, a developmental period known for axonal and dendritic outgrowth (Andersen, 2003). In contrast, the developmental expression profiles of the blue, black and purple modules enriched for all ID gene sets show a mirrored developmental trend in which they are down-regulated in the late fetal period and rise during the early post-natal stage known for synaptic overproduction before the pruning stage (Andersen, 2003). The blue module is enriched for transcriptional regulation, a known biological role of lncRNAs (Additional file A-3). In addition, both the blue and black modules are also enriched for synaptic transmission (both p-values < 0.00001). Interestingly, the expression patterns of the modules enriched solely for ID genes and the modules enriched for all ID gene sets appear to mirror each other, in terms of up- and down-regulation (Figure 2.2A). The fact that these modules are enriched for ID genes, lncRNAs and known pathways in ID disorders suggests that these modules are likely essential contributors to normal cognitive development.

**LncRNAs overlap with ID-associated CNVs**

Next, we asked if lncRNAs could reside within CNVs and if those lncRNAs were clustered into ID gene-enriched modules. To answer this question, we examined the genomic overlaps of lncRNAs with CNVs observed in idiopathic ID from 213 probands (Qiao, et al., 2013). Four subtypes of CNVs, including de novo, familial, common in ID cases and common in controls, were experimentally classified (Qiao, et al., 2013). In addition, a fifth subtype of known pathogenic CNVs associated with ID was added as a
positive control from DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources) (http://decipher.sanger.ac.uk/). We found that the lncRNAs within the turquoise, blue and brown modules had the highest number of genomic overlaps with both the de novo CNVs and known pathogenic CNVs from DECIPHER. These three modules had over three times the amount of ID-predisposing CNVs than the gray module although the gray module is over two times the size of any other module. The gray module contains genes that failed to merge with any other modules due to low topological overlap. Thus, the gray module could effectively serve as a randomized control for modular CNV overlap analysis.

Prioritization of candidate lncRNAs associated with ID

Focusing on modules which possessed features of ID gene enrichment and CNV overlap, we asked whether lncRNAs residing in likely pathogenic CNVs were strongly co-expressed with known ID genes. We examined the coexpression network neighborhood of the top CNV-harboring lncRNAs based on modular membership, of all ID enriched modules, along with each lncRNAs most highly correlated genes (Figure 2.3). The blue gene cluster contains two known ID genes, MBD5 and MEF2C, both of which are highly co-expressed with lncRNAs residing in DECIPHER pathogenic CNVs. MBD5 is required for methyl-CpG-binding specificity to methylated DNA, and haploinsufficiency of MBD5 is associated with intellectual disability (Williams, et al., 2009). Moreover, the expression of the blue gene cluster is negatively correlated with developmental time, suggesting its involvement in early neurodevelopmental processes. It is possible that these highly co-expressed lncRNAs may be involved in the transcriptional regulation of ID genes, such as
MBD5 and MEF2C. In addition, disruptive CNVs affecting lncRNAs involved in transcriptional regulation in the early fetal period could cause detrimental effects lasting throughout neurodevelopment. The lncRNA CTC-467M3.1 is highly correlated with MEF2C (Pearson’s correlation coefficient of 0.915). CTC-467M3.1 is located on the antisense strand relative to MEF2C, suggesting the possibility that CTC-467M3 might be involved in the cis-regulation of MEF2C. Natural antisense transcripts (NATs) have been shown to be involved in altering gene expression of their protein-coding counterparts, typically by suppression at the epigenetic level, but their roles in transcriptional activation and alternative splicing have also been observed (Vadaie and Morris, 2013). Thus, our approach has found that lncRNAs can be highly co-expressed with known ID genes and also overlap possibly pathogenic CNVs. These findings have allowed us to prioritize a list of potential ID-associated lncRNAs for further analysis as candidates for novel ID genes (Table 2.1). These lncRNAs reside within ID-predisposing CNVs and are ranked by the highest correlation to known ID genes. Notably, the lncRNA CTC-467M3.1, is among the top 15 ranked lncRNAs.
Figure 2.3 Coexpression network plot of lncRNAs residing in ID-associated CNVs. Network nodes represent genes, with the color being representative of the expression correlation to developmental months. Red nodes show expression negatively correlated with developmental months while green indicates positive correlation. The colored border of the node indicates the module to which that gene belongs.

<table>
<thead>
<tr>
<th>LncRNA</th>
<th>MODULE</th>
<th>CORRELATION TO DEV AGE</th>
<th>CNV TYPE</th>
<th>CORRELATION TO ID GENE</th>
<th>ID GENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC004019.13</td>
<td>Brown</td>
<td>0.80</td>
<td>DECIPHER</td>
<td>0.975</td>
<td>ALDH4A1</td>
</tr>
<tr>
<td>RP5-1059M17.1</td>
<td>Brown</td>
<td>0.57</td>
<td>De Novo</td>
<td>0.960</td>
<td>PTHLH</td>
</tr>
<tr>
<td>RP11-466F17.1</td>
<td>Blue</td>
<td>−0.59</td>
<td>DECIPHER</td>
<td>0.957</td>
<td>SATB2</td>
</tr>
<tr>
<td>AC017095.1</td>
<td>Yellow</td>
<td>−0.53</td>
<td>DECIPHER</td>
<td>0.956</td>
<td>SATB2</td>
</tr>
<tr>
<td>RP11-2E17.1</td>
<td>Brown</td>
<td>0.85</td>
<td>DECIPHER</td>
<td>0.938</td>
<td>ALDH4A1</td>
</tr>
<tr>
<td>AC099508.1</td>
<td>Blue</td>
<td>−0.79</td>
<td>De Novo</td>
<td>0.935</td>
<td>DCX</td>
</tr>
<tr>
<td>ATP13A4-AS1</td>
<td>Brown</td>
<td>0.75</td>
<td>DECIPHER</td>
<td>0.929</td>
<td>KANK1</td>
</tr>
<tr>
<td>AP00039.3</td>
<td>Brown</td>
<td>0.78</td>
<td>De Novo</td>
<td>0.929</td>
<td>ASPA</td>
</tr>
<tr>
<td>RP1-163G9.1</td>
<td>Turquoise</td>
<td>0.74</td>
<td>De Novo</td>
<td>0.928</td>
<td>KCNJ10</td>
</tr>
<tr>
<td>CTD-2210P24.4</td>
<td>Brown</td>
<td>0.79</td>
<td>DECIPHER</td>
<td>0.926</td>
<td>ALDH4A1</td>
</tr>
<tr>
<td>RP11-1026M7.2</td>
<td>Turquoise</td>
<td>0.57</td>
<td>De Novo</td>
<td>0.928</td>
<td>CACNA1F</td>
</tr>
<tr>
<td>RP11-588H23.3</td>
<td>Turquoise</td>
<td>0.76</td>
<td>DECIPHER</td>
<td>0.921</td>
<td>CA2</td>
</tr>
<tr>
<td>AC007096.1</td>
<td>Greenyellow</td>
<td>−0.65</td>
<td>De Novo</td>
<td>0.919</td>
<td>USP9X</td>
</tr>
<tr>
<td>CTC-467M3.1</td>
<td>Purple</td>
<td>−0.62</td>
<td>De Novo</td>
<td>0.915</td>
<td>MEF2C</td>
</tr>
</tbody>
</table>

Table 2.1 Prioritized list of ID-associated lncRNA candidates. LncRNAs overlapping pathogenic CNVs are ranked based on maximum correlation with known ID genes. The lncRNA module assignment and correlation of expression to developmental age are also provided.

Our results show the high coexpression between lncRNAs and known ID genes, suggesting the association of lncRNAs with ID. This is the first time that an assessment of associations between lncRNAs and ID has been performed on a genomic scale. However, there have been a few specific lncRNAs associated with ID such as the lncRNA Evf2, also known as DLX6-AS1, which is involved in a negative feedback loop of active chromatin remodeling leading to transcriptional repression (Cajigas, et al., 2015). Interestingly, mutations identified in Coffin-Siris syndrome (CSS), an intellectual disability, were found to localize to components of the chromatin remodeling complex, such as DLX1, which
upregulates DLX6-AS1 (Cajigas, et al., 2015). In our brain gene coexpression network, DLX6-AS1 belongs to the blue module, which is enriched for transcriptional regulation. Interestingly, DLX6-AS1 possesses the highest correlation (Pearson’s correlation coefficient = 0.87) with DLX2, which has been shown to be functionally redundant to DLX1 (Petryniak, et al., 2007).

2.4 Conclusion

In this study, we have identified potential ID-associated IncRNAs based on coexpression with known ID genes. Some coexpression modules enriched for known ID genes are also enriched for IncRNAs. The IncRNAs in these modules show specific developmental expression patterns in the brain. We have observed two distinct expression patterns of IncRNAs in ID-gene-enriched modules, showing inverse relationships most noticeably with regards to the mid to late fetal period. The coexpression modules show high level of connectivity between IncRNAs and known ID genes, affected pathways and developmental periods. Moreover, we have identified IncRNAs residing within de novo and pathogenic CNVs, which are major risk factors in ID. We have shown that IncRNAs overlapping the CNVs are also highly co-expressed with known ID genes. For instance, we have identified IncRNAs that show strong connections with MDB5 and MEF2C within the brain gene coexpression network. Finally, we have prioritized the IncRNAs overlapping ID-associated CNVs based on their coexpression to known ID genes in the developing brain. This prioritized list was constructed through ranking IncRNAs by their maximal correlation to known ID genes. The IncRNAs selected in this study can serve as a starting point for a new direction of inquiry for expanding upon the causal genetic factors of ID.
Acknowledgment

This work is supported, in part, by CSREES/USDA under project number SC-1000675 and a grant from the Self Regional Healthcare Foundation.
CHAPTER III - INTEGRATIVE GENOMIC ANALYSES FOR
IDENTIFICATION AND PRIORITIZATION OF LONG NONCODING RNAs
ASSOCIATED WITH AUTISM

Brian L. Gudenas¹, Anand K. Srivastava², Liangjiang Wang¹*
¹Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina, United States of America
²J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, South Carolina, United States of America
Published: PloS one, 12, 5, (2017)

Abstract

Genetic studies have identified many risk loci for autism spectrum disorder (ASD) although causal factors in the majority of cases are still unknown. Currently, known ASD risk genes are all protein-coding genes; however, the vast majority of transcripts in humans are noncoding RNAs (ncRNAs) which do not encode proteins. Recently, long noncoding RNAs (lncRNAs) were shown to be highly expressed in the human brain and be crucial for normal brain development. We utilize a novel computational pipeline for the integration of various genomic datasets to identify lncRNAs associated with ASD. This method utilizes differential gene expression patterns in affected tissues in conjunction with gene coexpression networks in tissue matched non-affected samples. We analyzed RNA-seq data from the cortical brain tissues from ASD cases and controls to identify lncRNAs differentially expressed in ASD. We derived a gene coexpression network from an independent human brain developmental transcriptome and detected a convergence of the differentially expressed lncRNAs and known ASD risk genes into specific coexpression modules. Coexpression network analysis facilitates the discovery of associations between previously uncharacterized lncRNAs with known ASD risk genes, affected molecular
pathways and at-risk developmental time points. In addition, we show that some of these lncRNAs have a high degree of overlap with major CNVs detected in ASD sequencing studies. By utilizing this integrative approach comprised of differential expression analysis in affected tissues and connectivity metrics from a developmental coexpression network, we have prioritized a set of candidate ASD-associated lncRNAs. The identification of lncRNAs as novel ASD susceptibility genes could help explain the genetic pathogenesis of ASD.

3.1 Introduction

Autism spectrum disorder (ASD) is a group of highly heritable genetic neurodevelopmental disorders characterized by impaired social communications with an estimated prevalence of 1 out of 68 births in 2010 (CDC, 2014). ASD risk genes include hundreds of protein-coding genes most commonly affected by copy number variants (CNVs) which can perturb gene expression; however, each known ASD risk gene only accounts for less than a few percent of ASD cases (Leblond, et al., 2014; Yu, et al., 2015). Most of these ASD risk genes function in several biological pathways, such as synaptic transmission, transcriptional regulation, immune response and chromatin remodeling (Garbett, et al., 2008; Voineagu and Eapen, 2013; De Rubeis, et al., 2014; Ander, et al., 2015). ASD is highly heterogeneous and there is still unaccounted genetic risk factors that likely resides outside protein-coding regions such as in regulatory noncoding RNAs (ncRNAs).

Long noncoding RNAs (lncRNAs) are transcripts greater in length than 200 nucleotides, which do not encode proteins. LncRNAs have been shown to be involved in a diverse array of neurodevelopmental functions such as brain development, neural
differentiation and synaptic plasticity (Wu, et al., 2013; Iyengar, et al., 2014; Schaukowitch and Kim, 2014; Briggs, et al., 2015). LncRNAs can also have epigenetic functions by interacting with chromatin re-modeling complexes to facilitate gene silencing or activation (Nie, et al., 2012; Schaukowitch and Kim, 2014). In addition, IncRNAs function in brain development, contributing to increased cognitive function and neuronal tissue specification (Clark and Blackshaw, 2014). Other than epigenetic functions, studies have shown lncRNA regulatory actions through diverse mechanisms such as multi-protein scaffolding, transcriptional interference, post-transcriptional modification and miRNA blocking (Geisler and Coller, 2013). The functionalities of IncRNAs to negatively or positively affect gene expression at the transcriptional, post-transcriptional, translational and epigenetic levels exhibit their regulatory versatility. In addition, because IncRNAs are highly tissue-specific and highly expressed in the human brain, IncRNAs are likely involved in complex neurodevelopmental disorders such as ASD (Wilkinson and Campbell, 2013; Wu, et al., 2013; Ziats and Rennert, 2013; Clark and Blackshaw, 2014; Wang, et al., 2015).

Microarray gene expression profiling of the ASD cortex indicated that the number of differentially expressed lncRNAs between the prefrontal cortex and cerebellum in ASD brains was lower than between the same brain regions in controls (Ziats and Rennert, 2013). This paradigm has also been observed in other studies in the ASD cortex with regards to mRNA expression and differentially methylated regions (Voineagu, et al., 2011; Nardone, et al., 2014). These studies suggest that ASD may be caused by aberrant neurodevelopment which could dysregulate neuronal tissue-specification. In a differential
expression analysis of ASD leukocytes, more IncRNAs were found to be differentially expressed than mRNAs, including thirteen IncRNAs associated with synaptic functions (Wang, et al., 2015). Genetic lesions of these IncRNAs, such as by CNVs, can impair gene expression and/or regulation which could have downstream regulatory consequences affecting neurodevelopment (Vincent, et al., 2010; Lin, et al., 2011; Vondervoort, et al., 2013; Zhubi, et al., 2014).

Our goal is to find differentially expressed IncRNAs in the ASD cortex and then identify which IncRNAs are also highly co-expressed with known ASD risk genes and ASD-affected biological pathways in neurodevelopment. The reasoning for using coexpression networks is that IncRNAs are vastly functionally uncharacterized, therefore, highly correlated gene and IncRNA expression patterns across developmental time imply shared biological function and/or regulation. Coexpression allows us to refine our list of candidate ASD-associated IncRNAs through gained functional insights utilizing this expression-based guilt-by-association heuristic. We have utilized an integrative approach for identifying ASD-associated IncRNAs, by analyzing differentially expressed IncRNAs in the ASD cortex and mapping them onto a brain developmental gene coexpression network. This approach, despite the genetic heterogeneity of ASD, facilitates the identification of ASD-associated IncRNAs by leveraging the information of ASD and non-ASD developmental cortex transcriptomes.
3.2 Results and Discussion

Differential expression of lncRNAs in the ASD cortex

We speculated that genes differentially expressed in the ASD cortex would be informative for identifying ASD-associated lncRNAs because the human cortex has been implicated in ASD pathophysiology by multiple transcriptomic studies (Parikshak, et al., 2013; Willsey, et al., 2013). Therefore, we re-analyzed RNA-seq data from the ASD cortex from a previous independent study which focused on differential splicing of protein-coding genes, yet did not analyze lncRNAs (Voineagu, et al., 2011). We found 1602 differentially expressed genes (FDR adjusted p-value < 0.05; \(|\log_2 \text{fold change}| \geq 1\)) (Figure 3.1). Furthermore, genes significantly down-regulated in the ASD cortex were enriched for biological processes related to synaptic function such as chemical synaptic transmission and synaptic signaling (p-values < .001) (Additional File A-5). The up-regulated genes were enriched for biological functions such as immune system process, cell surface receptor signaling pathway and response to cytokines (p-values < .001) (Additional File A-5). These results are in concordance with previous findings from ASD brain gene expression studies, where genes functioning in the synaptic transmission pathway were down-regulated, while genes involved in immune response were upregulated (Voineagu and Eapen, 2013). Furthermore, known ASD risk genes, curated by the Simons Foundation Autism Research Initiative (SFARI), were enriched within the differentially expressed genes (p-value < .001) (Additional File A-5) (Basu, et al., 2009). Thus, differentially expressed genes in the ASD cortex appear to be representative of known ASD
pathophysiology based on enrichments of biological pathways dysregulated in ASD and overrepresentation of known ASD risk genes.

![Differentially Expressed Genes in the ASD brain](image)

**Figure 3.1** Differentially expressed genes in the ASD cortex. Volcano plot displaying genes differentially expressed in the ASD cortex. A gene was required to have an absolute value of log₂ fold change greater than or equal to one and an adjusted p-value less than 0.05 to be considered differentially expressed.

Interestingly, we detected 263 lncRNAs differentially expressed between ASD and control cortical brain samples (FDR adjusted p-value < 0.05; |Log₂ fold change| ≥ 1) (Additional File A-7). Almost half of these differentially expressed lncRNAs were from intergenic regions (45%), with most of the remaining lncRNAs antisense to protein-coding genes (41%) (Additional File A-8). Next, we identified the nearest neighboring gene to each lncRNA since lncRNAs can have cis-regulatory mechanisms; remarkably, 5 of these
lncRNAs are antisense to known ASD risk genes such as RAPGEF4, DLX6, STXBP5, KLC2 and DMXL2 (Additional File A-8).

Next, we asked if these lncRNAs are specifically expressed in the human brain relative to other tissue types, which would suggest brain-specific biological functions. We extracted RNA-seq data from the Genotype-Tissue Expression project for over 40 different human tissues, each containing over 50 samples, and plotted the median expression of the differentially expressed lncRNAs for each tissue (Lonsdale, et al., 2013) (Additional Figure B-1). The majority of these lncRNAs are highly expressed in brain tissues relative to other tissue types (Additional Figure B-1). Furthermore, we found that the tissue type with the highest average expression for these lncRNAs is the brain cortex, suggesting that these lncRNAs perform cortex associated biological functions (Additional Figure B-1). We now have a list of lncRNAs differentially expressed in the ASD cortex, of which approximately 50% of the lncRNAs have a fractional expression level greater than 50% in the human brain suggesting tissue-specificity (Additional File A-8). We further refine the list of candidate lncRNAs through coexpression network analysis.

*Gene coexpression network analysis indicates that differentially expressed lncRNAs are involved in biological processes dysregulated in ASD*

We built genome-wide gene coexpression networks by utilizing the BrainSpan developmental transcriptome dataset, which consists of brain samples from eight weeks post-conception up to 40 years of age (Brainspan, 2013). First, we extracted all samples within cortical brain regions and then filtered out lowly expressed genes. This pre-processing resulted in a final RNA-seq dataset consisting of 352 cortical brain samples and
26,188 genes, of which 127 out of the 263 differentially expressed lncRNAs were present. Next, we used this refined RNA-seq dataset for signed weighted gene coexpression network analysis (WGCNA) (Langfelder and Horvath, 2008) which identified 33 gene coexpression modules (Additional File A-7). These coexpression modules symbolize groups of genes with similar developmental expression profiles through cortical development. Measuring the coexpression of randomly sampled groups of genes of equal size to each module shows that these coexpression modules are all significantly co-expressed (Additional Figure B-2).

Next, we asked if there were modules enriched for both differentially expressed lncRNAs and known ASD risk genes. To assess gene enrichment of ASD risk genes within modules, we utilized two independent lists of ASD risk genes. The list referred to as SFARI has been curated by the Simons Foundation Autism Research Initiative (SFARI) and these genes are scored based on the degree and strength of evidence for implications in ASD (Basu, et al., 2009). To avoid any bias in the SFARI gene set, which is manually curated, we also utilized a gene list known as the ME16 module which was identified in an independent unsupervised genome-wide coexpression study in brain tissue (Parikshak, et al., 2013). The ME16 gene list was shown to be enriched for genes with rare de novo genetic variants in ASD probands and a gene list known as “asdM12”, which contains genes aberrantly expressed in the ASD cortex (Voineagu, et al., 2011; Parikshak, et al., 2013). The differentially expressed lncRNAs show statistical enrichment in three modules (Blue, Brown and Black) (Figure 3.2). Interestingly, the Blue module is also enriched for SFARI ASD risk genes and ME16 genes (Fig 3.2). The co-enrichment of two ASD gene
sets and lncRNAs differentially expressed in the ASD cortex within the same developmental brain coexpression module suggests that the Blue module and the lncRNAs within it, are likely functionally involved in ASD pathogenesis.

**Figure 3.2** Enrichment of lncRNAs and ASD genes in brain developmental coexpression modules. Heatmap showing module based enrichment of gene lists, “DE LncRNAs” are lncRNAs differentially expressed in the ASD cortex, “SFARI ASD” are known ASD risk genes and “ME16” is an ASD-associated gene coexpression module identified in an independent study. Enrichment of gene lists was determined by a Fischer’s exact test requiring the FDR-adjusted p-value < 0.05 and an Odds Ratio > 1. Only modules containing at least 1 differentially expressed lncRNAs are shown.

Next, we asked if the identified coexpression modules were dysregulated in ASD by assessing their average differential expression in the ASD cortex of each module. Since we have identified less than two thousand differentially expressed genes in the ASD cortex
and there are over 26,000 genes in our developmental coexpression network, the vast majority of genes in the network have an ASD cortical fold change of zero (log2 fold change). Therefore, modules which show a statistically significant average ASD fold-change which deviates from zero likely represent biological pathways dysregulated in ASD. When overlaying the differential expression fold changes calculated from the ASD cortex onto the coexpression modules, 13 out of the 33 modules were found to be significantly differentially expressed on average compared to randomly sampled gene sets of the same size (Figure 3.3). Interestingly, all modules enriched for lncRNAs are on average differentially expressed in ASD, with the Blue module showing down-regulation while the Brown and Black modules are up-regulated (Figure 3.3). This suggests common mechanisms dysregulating these ASD-associated gene networks. Next, we examine the functional enrichments of these ASD-associated modules.
Figure 3.3 Differential expression in the ASD cortex overlaid onto developmental coexpression modules. Average log₂ fold changes of genes differentially expressed in the ASD cortex were overlaid onto the coexpression modules formed using the BrainSpan Developmental Transcriptome. Any genes that failed to reach significance had their log₂ fold changes set to 0. The red circle within each bar plot is the average log₂ fold change of 10,000 random gene samplings of equal size to the respective module. Significance of differential expression compared to the permuted distribution (FDR < 0.05) is denoted by a black asterisk adjacent to a module's respective bar plot.

We functionally characterized all modules enriched for differentially expressed lncRNAs by performing Gene Ontology Enrichment Analysis and visualizing their developmental expression pattern in the human cortex (Figure 3.4). The Blue module’s top three enriched biological processes are synaptic signaling, chemical synaptic transmission and anterograde trans-synaptic signaling (p-values < .001) (Figure 3.4). The synaptic transmission pathway is a well-known biological process dysregulated in ASD from gene
expression and genome-wide association studies (Voineagu, et al., 2011; Berg and Geschwind, 2012; Parikshak, et al., 2013; Voineagu and Eapen, 2013; De Rubeis, et al., 2014; Srivastava and Schwartz, 2014). Moreover, the expression of the genes within the Blue module show a positive correlation with developmental time in the cortex (Pearson’s correlation coefficient, $R^2 = 0.55$) (Figure 3.4), possibly coinciding with major cortical development (Parikshak, et al., 2013). The Brown and Black modules, which are only enriched for differentially expressed lncRNAs, have functional enrichments for immune response and lipid transport, respectively (Figure 3.4). Perturbations in the immune system as well as the transport of fatty acids has also been associated with ASD (Shimamoto, et al., 2014; Nazeen, et al., 2016). Remarkably, the differentially expressed lncRNAs are enriched in modules which have all been functionally linked with ASD, the synaptic transmission, immune response and lipid transport pathways (Qiu, et al., 2006; Garbett, et al., 2008; Voineagu and Eapen, 2013; De Rubeis, et al., 2014; Shimamoto, et al., 2014).
Figure 3.4 Characterization of modules enriched for differentially expressed lncRNAs. Gene ontology functional enrichments were performed for each module and adjusted for multiple comparisons (FDR < 0.05). The scatterplots show modular developmental expression profiles based on a modules eigengene (1st principal component) through developmental time, months PC means months post-conception (2 months post-conception to 1 post-natal year), with the blue vertical line demarcating birth. The trend line of each scatterplot is derived from a locally weighted scatterplot smoothing function.

Prioritization of candidate ASD-associated lncRNAs

The gene ontology enrichments and developmental expression trajectories (Figure 3.4) are representative of entire coexpression modules, however the differentially expressed lncRNAs represent a minority of the total genes within the modules. To assess the direct relationships between lncRNAs and ASD risk genes we examined the coexpression solely between lncRNAs and ASD risk genes compared to random
permutations. In both ASD gene sets, SFARI and ME16, we find statistically significant (p-values < .0001) summed correlations directly between the differentially expressed lncRNAs and ASD gene sets relative to randomly sampled gene sets (Additional figure B-3). These results further suggest that the identified lncRNAs are involved in the similar convergent biological processes dysregulated in ASD.

All the evidence presented thus far has been expression-based; therefore, we speculated that the integration of genetic mutational data, such as ASD-associated CNVs would be beneficial for the ranking of candidate ASD-associated lncRNAs. We incorporated a list of 5,030 major ASD-associated CNVs curated by SFARI from ASD genetic sequencing studies and calculated overlaps for all genes in the developmental network. To prioritize differentially expressed lncRNAs, we ranked the lncRNAs based on their module assignment to prioritize lncRNAs in modules enriched for ASD risk genes followed by ranking lncRNAs within the same module by their total overlaps with ASD-associated CNVs. Interestingly, the highest ranked lncRNA, HTR5A-AS1, is highly brain specific and its most highly correlated gene in the network is AGBL4, a known ASD risk gene which is also down-regulated in the ASD cortex (Pearson correlation = 0.98) (Additional file A-8). A summary of the genomic characteristics and coexpression network results for the candidate ASD-associated lncRNAs identified here, grouped by module, are presented in (Figure 3.5) including the total ASD CNV-lncRNA overlaps detected. In summary, we observe the largest amount of the DE lncRNAs in the blue, black and brown modules (Figure 3.5A), where the lncRNAs in the blue module are heavily down-regulated in the ASD cortex while the lncRNAs in the black and brown modules are up-regulated
Furthermore, the lncRNAs in the blue and black modules due on average exhibit tissue-specificity to the human brain relative to all tissue types (Figure 3.5C). In addition, the lncRNAs of the blue module show the highest level of ASD CNV overlaps (Figure 3.5D), suggesting these lncRNAs may be affected by CNVs in ASD.
Figure 3.5 Candidate ASD-associated IncRNA characteristics.
(A) Module assignment for DE IncRNAs, only modules with at least 3 IncRNAs are displayed, in addition we provide the module function which is the highest scoring gene ontology biological process for the whole module. (B) Average log2 fold change, from the ASD cortex, of the DE IncRNAs in each module. (C) Average fractional expression levels in the brain of the DE IncRNAs in each module. Fractional brain expression for each IncRNA is calculated from the Genotype Tissue expression project RNA-seq data, as the total expression in brain tissues divided by the sum of expression across all tissue-types (Lonsdale, et al., 2013). The red line at 50% represents the threshold for tissue-specificity as defined by Ayupe et al (Ayupe, et al., 2015). (D) Total overlaps between ASD CNVs and DE IncRNAs in each module.

For each differentially expressed IncRNA we also identified its most highly co-expressed protein-coding gene within the developmental network. Remarkably, fourteen candidate ASD-associated IncRNAs most highly co-expressed genes, in a network of over 21,000 genes, are known ASD risk genes (Additional file A-8). This result indicates these IncRNAs are likely involved in the molecular function and/or regulation of these specific ASD risk genes. However, further experimental studies will be needed to decipher the true relationship between these candidate ASD-associated IncRNAs and their highly co-expressed ASD risk genes. This prioritized list of ASD-associated IncRNAs can assist geneticists by providing high-quality novel experimental targets to further elucidate ASD pathogenesis.

3.3 Conclusions

Utilizing differential expression analysis in affected tissues coupled with an independent developmental gene coexpression network, we have identified a list of candidate ASD-associated IncRNAs. These IncRNAs are differentially expressed in the ASD cortex, highly expressed in brain and cortical tissues and also co-expressed with ASD
risk genes in the developing cortex. We have identified a coexpression module enriched for both differentially expressed lncRNAs and ASD risk genes; this module is functionally enriched for the synaptic signaling and transmission pathways. In addition, two modules were enriched solely for differentially expressed lncRNAs, which were functionally enriched for the immune response and lipid transport pathways, biological processes suspected to be dysregulated in ASD. Furthermore, all modules enriched for lncRNAs are on average significantly differentially expressed in the ASD cortex, with the synaptic module showing down-regulation and the immune and lipid transport modules being up-regulated. Finally, we identified 14 lncRNAs whose most highly co-expressed genes in the entire network are known ASD risk genes suggesting direct functional associations. The convergence of the independent genomic results presented here suggest an association between these previously uncharacterized lncRNAs and ASD in the human brain. These lncRNAs can serve as prioritized candidate ASD risk genes to expedite future genetic studies researching ASD pathogenesis.

3.4 Methods

RNA-seq data analysis

We extracted raw RNA-sequencing data from human cortical tissues for three ASD cases and controls, each with two technical replicates, from a published study (GEO accession GSE30573) (Voineagu, et al., 2011). Transcript abundances were quantified from RNA-seq reads which were mapped to the human transcriptome version GRCh38 with gene annotations from ENSEMBL v86, using Salmon (v0.7.1) (Patro, et al., 2017).
Transcript counts were then imported into R (version 3.3.2) using tximport (R package, v1.2.0) (Soneson, et al., 2015). All subsequent analyses were performed in R (version 3.3.2) on a 64-bit Windows 7 system.

**Differential expression analysis**

Differential expression was performed based on the difference in gene counts between ASD cases and controls using DESeq2 (R package, v1.10.1) (Love, et al., 2014). DESeq2 estimates differential expression using a negative binomial model which have been shown to be reduce false positives compared to other methods (Z. H. Zhang, et al., 2014). A gene was considered differentially expressed if it had an FDR-adjusted p-value less than 0.05 and an absolute log2 fold change greater than or equal to one.

**Tissue-specific expression of lncRNAs**

We downloaded the tissue-specific RNA-seq dataset, in which gene-level median RPKM abundances are reported for each tissue type, from the Genotype-Tissue Expression (GTEx) program (Lonsdale, et al., 2013). We filtered the data to remove any tissue types with less than 50 samples. For each differentially expressed lncRNA, FPKM values were Z-score normalized across the tissues then plotted in a heatmap using the gplots (R package, v3.0.1) (Warnes, et al., 2009). Hierarchical clustering was performed on the tissues and all brain tissues were highlighted blue using a color key. Fractional brain expression is calculated by the summation of all expression values in brain tissues divided by the total expression for all tissue-types multiplied by 100 to get a relative percent.
Gene lists

The “DE lncRNA” gene list is composed of all significant differentially expressed genes which have lncRNA biotypes according to Ensembl (v84). The ASD gene list was extracted from AutDB, the SFARI human gene database (Basu, et al., 2009). We utilized the SFARI “Gene Score”, which categorizes ASD risk genes based on evidence for implication in ASD, to filter the ASD risk genes. The “SFARI” gene list is only composed of genes with evidence levels 1-5 (high evidence – minimal evidence). We obtained M16 from an independent genomics study which applied WGCNA to the BrainSpan developmental RNA-seq dataset to identify modules in the developing brain enriched for ASD risk genes (Parikshak, et al., 2013).

Cortical development expression data

We downloaded the BrainSpan Developmental transcriptome Gencode v10 dataset (summarized to gene-level reads) from http://www.brainspan.org/. Only samples from cortical brain regions were used in our analysis and we used a variance filter to remove the bottom two quartiles of genes based on across sample variance. This preprocessing resulted in an RNA-seq dataset with 352 cortical samples and 26,188 genes which was used as input for gene coexpression network analysis.

Coexpression network analysis

Coexpression networks were built using WGCNA (R package, v1.51) (Langfelder and Horvath, 2008). The biweight midcorrelation, a correlation metric more robust to outliers than the Pearson correlation, was used to calculate correlations between all gene pairs. Afterwards, a signed weighted network was created using a soft-threshold power of
12 to approximate network scale-free topology. Next, the topological overlap was calculated between all gene pairs. The topological overlap metric represents not only pair-wise relationships, but relationships between cliques of genes and is therefore advantageous for clustering genes over pair-wise methods. To identify coexpression modules a hierarchical cluster tree is created based on the topological overlap matrix, with modules representing distinct branches of the dendrogram. Modules are formed from the result of a dynamic tree cutting algorithm (Langfelder and Horvath, 2008). All parameters used for network construction are included in (Additional file A-5).

**Network validation**

To validate that all modules were co-expressed above what would be expected by random chance, we performed a coexpression permutation test. This was done by assessing the average correlation of randomly sampled gene sets, which were equal in size to the modules derived previously, 10,000 times. Comparing the distributions of the average biweight mid-correlation of randomly sampled gene sets with the coexpression of each module, we found all modules to be significantly co-expressed above random chance (modules, p < 1x10^-4).

Next, we asked if the up- and down-regulated genes that we identified in the ASD brain segregated into the developmental modules created using BrainSpan (Brainspan, 2013). We performed another permutation test where we calculated the average log₂ fold change of 10,000 randomly created gene sets equal in size to each module and then compared these distributions to each module’s average log₂ fold change in the ASD brain.

**Gene set enrichment analysis**
Overrepresentation of gene lists within modules was calculated using a one-sided Fisher exact test to assess gene list enrichment. All p-values, from all gene sets and modules, were adjusted for multiple testing using the False Discovery Rate method (Benjamini and Hochberg, 1995). We required an odds ratio > 1 and an adjusted p-value < 0.05 to claim a gene set is enriched within a module. The –log10(p-values) was then plotted in a heatmap using the gplots (R package, v3.0.1) (Warnes, et al., 2009).

**Module characterization**

Gene Ontology enrichment analysis in each lncRNA-enriched module was characterized using the GOstats (R Package, v2.36), reported biological processes were required to have an FDR-adjusted p-value < 0.05 (Falcon and Gentleman, 2007). Module eigengenes, representing module developmental trajectories, from 8 weeks post-conception to 1 post-natal year were plotted and fitted with a locally weighted scatterplot smoothing function using ggplot2 (R package, v2.1.0) (Wickham, 2009).

**LncRNA-ASD risk gene coexpression analysis**

The summation of all biweight midcorrelations between differentially expressed lncRNAs with SFARI or ME16 genes was calculated. Next, the summed correlation between lncRNAs and randomly sampled gene sets, of equal size to the respective ASD gene list was calculated 10,000 times. This resulted in a permuted normal distribution from which a p-value was derived for the actual summed correlation between the lncRNAs and ASD gene sets.

**CNV analysis**
CNV summary data was downloaded from SFARI and filtered to only retain CNVs with a report class of “Major” (Abrahams, et al., 2013). Using the cytoband locus we converted all CNVs into hg38 genomic coordinates according to the hg38 cytoband coordinates from UCSC (Additional file A-5). Overlaps were quantified for each gene if its genomic coordinates overlapped the CNV range using the GenomicRanges (R package, v1.22.4) (Lawrence, et al., 2013).
CHAPTER IV - A GENETIC ALGORITHM FOR FINDING DISCRIMINATIVE FUNCTIONAL MOTIFS IN LONG NONCODING RNAS

Brian L. Gudenas¹, Anand K. Srivastava², Liangjiang Wang¹*
¹Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina, United States of America
²Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA
*Corresponding author: liangjw@clemson.edu
Published: Bioinformatics Research and Applications, 13th International Symposium proceedings, (2017)

Abstract

Long noncoding RNAs (lncRNAs), each with >200 nucleotides in length, constitute a large portion of the human transcriptome. Although recent studies indicate that lncRNAs play key roles in gene regulation, development and disease, the RNA functional motifs are still poorly understood. Most of the existing algorithms for motif finding are severely limited in scalability with regards to sequence and motif size. In this study, we propose a novel genetic algorithm for discriminative motif identification capable of handling large input sequences and motif sizes by utilizing genetic operators to learn and evolve in response to the input sequences. We utilize our method on long noncoding RNA (lncRNA) transcripts as a test case to identify functional motifs associated with subcellular localization. Our methodology shows high accuracy and the ability to identify functional motifs associated with subcellular localization in lncRNAs, which recapitulates a previous experimental study.
4.1 Introduction

The identification and subsequent functional annotation of short reoccurring motifs within molecular sequences has been integral for the field of genetics. Generally, the first step regarding functional annotation of a novel protein or RNA sequence is the identification of known functional motifs within the primary sequence. Once identified, these motifs allow the functional inference of the previously uncharacterized sequence. Functional motifs are generally identified assuming a random uniform background nucleotide model, however the nucleotide sequence within a gene are not randomly distributed, but have structure; therefore, the utilization of real sequences as a background set is advantageous. The use of a positive and negative sequence set is known as discriminative motif identification. Popular algorithms for discriminative motif discovery include discriminative regular expression motif elicitation (DREME) (Bailey, 2011). However, the DREME algorithm suggests input sequences are less than 500 nucleotides and that the motif width is less than or equal to 8 nucleotides.

Recently, tens of thousands of long noncoding RNAs (lncRNAs) have been discovered in primates, the vast majority of which are functionally uncharacterized. LncRNAs are poorly conserved across species and can perform a myriad of diverse functions, adding to the complexity of their functional annotation. Unlike mRNAs, lncRNAs can localize in many different places within the cell, which can provide insights into their functionality. Localization motifs have been identified in lncRNA transcripts which regulate subcellular localization. Thus it may be possible to identify motifs within lncRNA transcripts, genome-
wide, associated with subcellular localization, providing a valuable first step in the functional annotation of human IncRNAs. One method to do this would be to identify two distinct sets of IncRNAs, one enriched in the cytoplasm and the other enriched in the nucleus, followed by finding motifs overrepresented in one set of transcripts but not the other. However, there are tens of thousands of human IncRNAs, which have a median transcript length of 592 bp and could possibly contain long functional motifs (Derrien, et al., 2012). Due to these issues and limitations, previous approaches are unsuitable for our purposes of the identification of functional motifs in full length IncRNAs genome-wide.

Genetic algorithms, which mimic biological evolution to stochastically evolve a population of solutions over time, have been utilized for motif finding previously, such as MDGA (Che, et al., 2005). MDGA represents a solution as a vector of indices which indicate the starting position of the motif in each sequence and therefore fail to utilize all the information present in a sequence, such as multiple motif occurrences. Furthermore, all the methods mentioned previously were developed for the identification of transcription factors binding sites, which are small motifs and are contained in very short sequences, such as ChIP-Seq peaks (< 100 nucleotides). We propose to develop a novel genetic algorithm (GA) for discriminative motif discovery to identify long functional motifs in full-length IncRNA transcripts, a use-case previous methods are incapable of.
4.2 Methods

Representation and Population Initialization

We represent a solution as a position weight matrix (PWM) of length \( w \), which is therefore not dependent on the input sequence size. A fixed number of individuals is initialized to derive a population by creating randomized PWMs utilizing the conjugate dirilecht distribution.

Fitness

The fitness function must identify similar sequence motifs in the positive set which are underrepresented in the negative set. A useful metric for the identification of informative similar sequences is the information content (\( IC \)), the total information content of a PWM \( M \) is as follows (Shannon, 1948).

\[
IC(M) = \sum_{i=1}^{w} \sum_{\beta \in \{A,C,G,T\}} M_{\beta i} \log_2 \frac{M_{\beta i}}{p_\beta}
\]

Where \( W \) is the motif width, \( M_{\beta i} \) is the frequency of nucleotide \( \beta \) of column \( i \) and \( p_\beta \) is the background probability of nucleotide \( \beta \). We chose to augment this popular metric to incorporate our goal of minimizing the matches of the PWM in the negative sequence set. We define the fitness score to be the information content of the PWM divided by the total matches of the PWM in a small random subset \( S^-_\beta \) of negative set \( S^- \).

\[
Fitness = \frac{IC(M)}{\sum_{S^-_\beta} \sum_{\beta \in \{A,C,G,T\}} M_{\beta i} p_\beta |S^-_\beta|}
\]
Where $\sigma_M(S)$ equals the match score of motif $M$ in sequence set $S$ and $|S_\bar{\mu}|$ is the total sequence length scanned, used as a normalization factor. Using a sliding-window across each sequence we calculate a score as the log-likelihood PWM score divided by the maximal PWM score.

*Selection*

During each iteration, selection occurs to determine which solutions survive into the next generation. A solution’s probability of survival is approximately proportional to their fitness score. Linear-rank selection then occurs with replacement to generate a new population of solutions.

*Crossover and Mutation*

Crossover occurs by randomly selecting two parent individuals which are then recombined to create two novel children solutions. Therefore, crossover allows the recombination of solutions to further explore the solution space. To avoid the positional bias of the traditionally used 1-point crossover, we chose to utilize random uniform crossover. In random uniform crossover, we randomly choose a crossover number $c$ between $(w+1: w-1)$, then we randomly select $c$ columns of the parental solution PWMs to be switched. The resulting children solutions then replace the initial parental solutions used for crossover.

Mutation occurs by randomly selecting solutions which are then altered stochastically to allow further exploration of the search space, thereby avoiding local maxima. For each
solution $\mathbf{M}$ selected for mutation, we randomly select a fixed proportion $\beta$ of the positive sequence set $\mathbf{S}^+$, then we score the solution across all possible windows of the sequences in $\mathbf{S}_\beta^+$. Based on the maximal scoring position of $\mathbf{M}$ in each sequence of $\mathbf{S}_\beta^+$ we then update the PWM to form $\mathbf{M}'$ using the alignment of the selected positions.

**Implementation**

The program was implemented in the R statistical language utilizing the framework from the R package GA [7,8]. The pseudo-code is as follows:

1. Set Input Parameters:
   
   $\mathbf{S} =$ Full sequence set ordered by positive/negative
   
   $\mathbf{W} =$ Motif Width; $\mathbf{P} =$ Population size; $i = 0$
   
   $\mathbf{R} =$ Number of generations to terminate if best solution not improved

2. INITIALIZE Population: Randomly create $\mathbf{P}$ PWMs

3. Fitness EVALUATION: $\mathbf{BEST} = \max(\text{Fitness}(\mathbf{P}))$

4. Genetic Operators: While ($i < \mathbf{R}$)
   
   ELITISM: Save top 5% of fittest solutions for next generation
   
   SELECTION: Linear-rank selection to create new $\mathbf{P}$
   
   Random uniform CROSSOVER of random subset of $\mathbf{P}$
   
   MUTATE random subset of $\mathbf{P}$
   
   Evaluate fitness of new $\mathbf{P}$
   
   TERMINATION: If ( $\max(\text{fitness}(\mathbf{P})) == \mathbf{BEST}$ ) {
4.3 Results

To evaluate our genetic algorithm (GA) we first begin with synthetic data in order to obtain complete control over the sequence attributes. We created a set $S^+$ of 100 independent and identically distributed sequences over a range of lengths with a single randomly implanted motif of length 15, each with 3 random mutations per sequence. Our negative sequence set $S^-$ is simply a dinucleotide shuffle of $S^+$. To assess the sensitivity of the GA we evaluate its performance over a large range of different sequence lengths $N$, because as $N$ increases the noise to signal ratio increases (Figure 4.1).
Motif similarity is represented by the log-likelihood PWM score divided by the maximal log-likelihood PWM score, averaged over 11 trials for each of the different sequence lengths. The red line is the threshold for a significant match as defined by Hansen (Hansen, et al., 2012).

**Differentially Localized LncRNAs**

Lastly, we wanted to identify possible functional motifs in LncRNAs which are differentially localized within the cell. For this test case we used published results from which LncRNA transcript abundances were quantified from fractionated cellular compartments, either from the nucleus or the cytoplasm (Derrien, et al., 2012). We extracted the sequences for all LncRNAs quantified between the nuclear and cytoplasmic fractions, using sequences enriched in the nucleus as our positive sequence set and the cytoplasmic LncRNAs as the negative set. This data resulted in a set of 1749 LncRNA
transcripts with a total sequence length of 2.42 megabases, of which 981 sequences are enriched in the nucleus while 768 are enriched in the cytoplasm. Using the nuclear transcripts as our positive set and the cytoplasmic transcripts as our negative set we ran the GA on this large dataset to identify 14-mers associated with subcellular localization (Figure 4.2).

Notably, the best motif identified contains a core pentamer RNA motif previously identified to dictate nuclear localization of IncRNAs (B. Zhang, et al., 2014). The pentamer RNA sequence motif was found to be AGCCC with the restriction sites of (G or C) at -3 and (T or A) at -8, which the motif we identified contains all of, except the -8 restriction site (B. Zhang, et al., 2014). Next, we calculated the total matches of the identified motif in each sequence set, finding 3,133 instances of the motif in the nuclear set and only 1,871 in the negative set. Furthermore, the counts of this motif in each IncRNA shows a small yet significant correlation with the nuclear/cytoplasm FPKM ratio (Pearson’s Correlation Coefficient = 0.14, p-value = 4.31×10⁻⁹).
Motif found in nuclear enriched lncRNAs. The output of the genetic algorithm ran with the nuclear enriched lncRNA transcripts as the positive set and the cytoplasmic lncRNAs as the negative set. The first plot shows the fitness per generation. The second figure is the sequence logo of the best motif found.

4.4 Conclusion

Based on our preliminary observations we have shown that our genetic algorithm is capable of identifying discriminative motifs in large sequence sets, such as sequence sets containing thousands of lncRNA transcripts. We demonstrated that our algorithm can achieve high accuracy in synthetic tests despite a high noise to signal ratio. In addition, utilizing entire lncRNA transcripts derived from the transcript quantification of fractionated cells we have identified a motif enriched in nuclear transcripts. Remarkably, the motif identified recapitulates an experimentally identified lncRNA localization motif identified in an independent study (B. Zhang, et al., 2014). The counts of this motif in the lncRNA transcripts also shows a significant positive correlation with subcellular
localization. This preliminary work shows that it is possible to computationally identify functional motifs in previously uncharacterized IncRNAs.

Our method is capable of handling large input sequence sets as well as identifying arbitrarily large motifs. In addition, to speed up the motif identification process over a range of motif widths (\(w\)), solutions from the final population can be used as a seed population for identifying motifs of size \(w+1\). This procedure can be done iteratively and will speed up the motif identification procedure because the initial population will already contain informative seed motifs to be expanded upon. Furthermore, the genetic algorithm framework could likely be improved upon through augmentation of the genetic operators.
CHAPTER V - PREDICTION OF LncRNA SUBCELLULAR LOCALIZATION WITH DEEP LEARNING FROM SEQUENCE FEATURES

Brian L. Gudenas and Liangjiang Wang*

Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA
*Corresponding author: liangjw@clemson.edu

Abstract

Long noncoding RNAs are involved in biological processes throughout the cell including the nucleus, chromatin and cytosol. However, most lncRNAs remain unannotated and functional annotation of lncRNAs is difficult due to their low conservation and their tissue and developmentally specific expression. LncRNA subcellular localization is highly informative regarding its biological function, although it is difficult to discover because no prediction methods currently exist. While protein subcellular localization prediction is a well-established research field, lncRNA localization prediction is a novel research problem. We developed DeepLncRNA, a deep learning algorithm which predicts lncRNA subcellular localization directly from transcript sequences. We identified differentially localized lncRNAs using 93 strand-specific RNA-seq samples, from multiple cell types, which underwent fractionation to isolate the nucleus and cytosol prior to RNA-seq. Sequence-based features were extracted from the lncRNAs to train the model, achieving high prediction accuracy, demonstrating that primary sequence motifs are a major driving force in the subcellular localization of lncRNAs. DeepLncRNA obtained an accuracy of 73.3%, sensitivity of 80.7%, specificity of 66.2% and area under the receiver operating characteristic curve of 0.812.

5.1 Introduction
The inner workings of the cell are orchestrated by complex interactions between the products of DNA, both noncoding RNAs and proteins. This idea has superseded the view that proteins and their corresponding messenger RNA (mRNA) are solely responsible for cellular function. Noncoding RNAs are now known to be an integral functional component of genetic regulation, and are involved in crucial roles such as the regulation of gene expression. The most prevalent and one of the most functionally diverse classes of noncoding RNAs are the long noncoding RNAs (lncRNAs).

LncRNAs are large RNA transcripts which do not encode proteins and are estimated to outnumber protein-coding genes within the human genome (Iyer, et al., 2015). However, lncRNAs are poorly conserved at the sequence level, which makes functional annotation difficult. LncRNAs perform a diverse repertoire of essential molecular functions, in many different subcellular locations (Geisler and Coller, 2013). However, determining the functional role of lncRNAs experimentally is highly time-consuming and laborious. Like proteins, lncRNA functionality is dependent on proper subcellular localization. LncRNA transcripts can localize in many different places within the cell, including the chromatin, nucleus, cytoplasm and exosomes (Heesch, et al., 2014; Morris, 2016). Knowing the localization patterns of lncRNAs allows the inference of their biological functional. Therefore, the possibility to learn where any given lncRNA localizes would provide valuable information regarding its biological function as well as the RNA localization mechanism.

LncRNA subcellular localization is likely dependent on many factors, including sequence and structural motifs which can facilitate binding to proteins involved in
localization (Goff and Rinn, 2015). Identification of structural motifs in IncRNAs is currently problematic both experimentally and computationally due to the high-level of complexity of intra-molecular organization that IncRNAs can exhibit (Yan, et al., 2016). However, sequence motifs in IncRNAs associated with subcellular localization have been identified such as the pentamer motif AGCCC which is highly associated with IncRNA nuclear localization (B. Zhang, et al., 2014). Therefore, it is evident that motifs in the IncRNA primary sequence are involved in IncRNA subcellular localization. While, obtaining IncRNA structural data is difficult, IncRNA transcript sequences are readily available.

Protein subcellular localization has been an active research area for decades and many localization motifs have been identified. These localization motifs either reside in the primary sequence, such as the N-terminal signal peptide associated with the secretory pathway, or within the 3D protein structure, such as DNA-binding domains in nuclear proteins. A well-known method for protein subcellular localization prediction is MultiLoc, a support vector machine (SVM) which used sequence-derived features to achieve an average accuracy of 75% (Höglund, et al., 2006). DeepLoc, a deep learning algorithm, recently achieved an accuracy of 91% on the same data set used by MultiLoc (Almagro Armenteros, et al., 2017). However, the proteins in this dataset have been found to be highly homologous and therefore might provide an overly optimistic model evaluation. Using a more comprehensive dataset of proteins which localize to ten different subcellular locations, DeepLoc achieved an accuracy of 77%, while MultiLoc2, an upgraded version of MultiLoc, only achieved an accuracy of 55% (Almagro Armenteros, et al., 2017).
Sequence-based features thus appear to be highly informative for protein subcellular localization and deep learning attains exceptional accuracy in comparison to other machine learning algorithms. Despite the well-established knowledge regarding protein localization prediction, we know relatively nothing about the prediction of lncRNA localization.

Our goal is to learn a model that predicts lncRNA subcellular localization directly from lncRNA nucleotide sequences. We have chosen to utilize a deep neural network (DNN), which have shown promise in many bioinformatics applications such as the annotation of noncoding variants and identification of enhancers (Quang, et al., 2015; Kim, et al., 2016). Deep learning methods, such as DNNs, avoid the need to manually craft informative features and instead automatically learn high-level features through the iterative aggregation of features in each layer of the network. Since nuclear retention motifs have already been found in nucleus-localized lncRNAs, differences in sequence composition between distinct nuclear and cytosolic lncRNAs are expected (B. Zhang, et al., 2014). In this study, we performed used binary classification to learn how to discriminate between differentially localized nuclear and cytosolic lncRNAs. Our task is to predict the subcellular localization of lncRNAs based on their transcript sequence, therefore we named our algorithm DeepLncRNA, an acronym for “Deep Learning of Nuclear Classification of long noncoding RNAs”. We trained our model on the sequences of differentially localized lncRNAs, which are either enriched in the nucleus or the cytosol. DeepLncRNA scans the lncRNA sequence, computing a range of k-mer frequencies and protein-binding motifs which are then used to predict the lncRNA localization.
Sequence-based features were extracted from lncRNA transcript sequences, suggesting that our methodology could be easily applied to uncharacterized lncRNAs. LncRNAs are lowly conserved between species, and a large fraction of human lncRNAs are even primate-specific (Necsulea, et al., 2014; Washietl, et al., 2014). Therefore, our model could be applicable to lncRNAs in closely related primates such as the chimpanzee or bonobo. This study represents a first step in lncRNA subcellular localization prediction which will be a valuable resource for the functional annotation of this large, diverse and not yet fully understood class of noncoding genes.

5.2 Methods

Datasets

We analyzed paired-end strand-specific RNA-sequencing data from human cell lines from the ENCODE project (Bernstein, et al., 2012). Samples underwent cellular fractionation, to separate either the nucleus or cytosol, prior to RNA-seq. In total, we acquired 93 RNA-seq profiles from 14 human immortalized cell lines, of which 45 were from the cytosol and 48 from the nucleus. All cell lines were required to contain at least two samples from each cellular fraction. Samples underwent different RNA library protocols such as poly(A)+ (n = 62), total RNA (n = 8) or poly(A)- (n = 23). Using the total RNA and poly(A)- library protocols in addition to the standard poly(A)+ samples allows a complete transcriptomic analysis of lncRNAs, which are not all polyadenylated. All sample metadata, as well as transcriptome alignment rates are displayed in (Additional file A-9).

Raw RNA-seq reads were mapped to the human transcriptome and quantified using Kallisto (v0.43.1), (Bray, et al., 2016). In total, ~ 6 billion reads were aligned to the human
transcriptome. Differential transcript expression analysis between the nuclear and cytosolic fractions for each cell type was performed using Sleuth (R package, v0.29.0) which was shown to be superior to other methods at identifying differentially expressed transcripts (Pimentel, et al., 2016). If multiple RNA library protocols were used for a single cell type then we added this as a covariate when testing for differential transcript expression.

Identification of Differentially Localized Human LncRNAs

We performed differential transcript expression to quantify the differences in lncRNA transcript abundances between the nuclear and cytosolic cellular fractions for each cell type. We aggregated the log2 fold-change values for each lncRNA across all cell-types using a weighted average. Computing the nuclear to cytosolic log2 fold change allowed the examination of the distribution of lncRNA subcellular localization for over 18000 lncRNA transcripts (Figure 5.1). In agreement with previous studies, we found lncRNAs to be predominantly enriched in the nucleus (Derrien, et al., 2012; Djebali, et al., 2012). However, we do detect a large portion of lncRNAs (n = 4380) with transcript abundances higher in the cytosol than the nucleus (Figure 5.1). Part of the nuclear skew of this distribution is likely explained by the fact that all lncRNAs, regardless of destination, must originate in the nucleus through the act of transcription. Furthermore, once transcribed the export of lncRNAs from the nucleus to the cytoplasm must take some amount of time due to the export mechanism, such as assembly of ribonucleoprotein complexes and recruitment of exporters (Köhler and Hurt, 2007). Due to these two factors we expect the median lncRNA nuclear to cytosol transcript ratio to be greater than zero and indeed the median log2 fold-change was 1.6. Therefore, since our distribution is not centered at zero,
like a standard differential expression test, we must adjust the commonly used symmetric log2 fold-change threshold to classify differential expression. To account for the nuclear skew of transcript ratios we selected new log2 fold-change thresholds, corresponding to the first and fourth quartile, to signify differential localization (cytosolic < 0, nuclear > 2.8). Applying these fold-change thresholds to our data resulted in a balanced dataset of 4380 cytosolic lncRNAs and 4298 nuclear lncRNAs. The dataset was then split into a training, validation and testing set using a randomized 70/15/15 percent split.

Figure 5.1 Distribution of the lncRNA nuclear to cytosolic transcript ratios. A histogram showing the log2 fold-change ratios for lncRNA transcripts (n = 18,068) detected across all cell types. Colored bars indicate differentially localized lncRNAs which passed fold-
change thresholds (Cytosolic < 0; Nuclear > 2.8) resulting in a training set of 4380 cytosolic lncRNAs and 4298 nuclear lncRNAs.

**Extraction of Sequence Features from LncRNAs**

To derive sequence-based features of uniform length from transcript sequences of variable length we counted k-mers. Using the sequences of the differentially localized lncRNAs we computed a k-mer frequency matrix, containing the frequency of all possible oligonucleotides for $k$ equal to two through five resulting in ($4^2+4^3+4^4+4^5$) = 1360 k-mer features. In addition, the genomic loci of lncRNAs are known to be important regarding their functionality which is why lncRNAs are classified based on their genomic context such as, intergenic, antisense or sense lncRNAs (Ma, et al., 2013). Therefore, we added additional features representing these major lncRNA subtypes based on the transcript annotations from ENSEMBL. We also added the chromosomal location of the lncRNA to further capture any effects of its genomic location. Lastly, the binding of RNA by proteins represents a possible mechanism in which lncRNAs may be localized. Therefore, we added features representing the presence of known RNA-binding motifs which were obtained from the CISBP—RNA database (Ray, et al., 2013). Matches were counted using a sliding-window approach, and a match was scored if the sub-sequence obtained a log-likelihood position weight matrix (PWM) score greater than 80% of the maximal PWM score (Andersen, et al., 2008). In total, we obtained 1582 sequence-based features which are the inputs for DeepLncRNA (Figure 5.2).
Figure 5.2 Overview of the DeepLncRNA algorithm
Deep Neural Network Model

DeepLncRNA is a feed-forward multi-layer deep neural network. The architecture used consists of one input layer, three hidden layers using the rectified linear unit activation function and a softmax output layer. Hidden layer dropout was applied which randomly masks half of the connections in each layer during training of the DNN which reduces the propensity for overfitting. Input dropout was also applied which randomly masks some of the hidden units in each layer to increase the generalizability of the model. Furthermore, regularization was applied using the L1 and L2 weight penalties to the cost function. All model parameter values were selected using a random search over all possible parameter combinations seeking to minimize the misclassification rate on the validation set (Additional File A-10). DeepLncRNA was trained with stochastic gradient descent using the backpropagation algorithm which adjusts network weights by minimizing the error between the response variable and the predicted output. DeepLncRNA was built using the h2o R package (H2O.ai, 2017).

Evaluation Criteria

In this work, we develop a DeepLncRNA to identify lncRNAs to be enriched in the nucleus (positive class) or cytosol (negative class). We use the common machine learning evaluation metrics such as accuracy, sensitivity, specificity and Matthews correlation coefficient respectively defined below.

\[ TP = \text{True Positive} \]
\[ TN = \text{True Negative} \]
\[ FP = \text{False Positive} \]
\[ FN = \text{False Negative} \]
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]
\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]
\[
\text{Specificity} = \frac{TN}{TN + FP}
\]
\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

5.3 Results

To evaluate the performance of DeepLncRNA we compared it to other advanced machine learning algorithms. We compared DeepLncRNA with a random forest (RF) and support vector machine (SVM) (Figure 5.3). Based on all three measures, accuracy, specificity and sensitivity DeepLncRNA achieved better performance. The ability to abstract complex non-linear features does appear to enhance the performance of DeepLncRNA compared with the other machine learning algorithms. Interestingly, the specificity of every model is more than 5% lower than its sensitivity. Specificity, known as the true negative rate, represents the ability to correctly identify cytosolic lncRNAs, suggesting that it is more difficult to predict cytosolic localization than nuclear localization.
Figure 5.3 Model selection based on performance metrics on the validation set. The performance metrics of DeepLncRNA, random forest (RF) and radial support vector machine (SVM) on the validation set.

Model parameters were selected based on the maximization of accuracy on the validation set. Since DeepLncRNA has more parameters than either the random forest or support vector machine it is possibly an over-optimistic evaluation of its accuracy. Therefore, we generated ROC curves on the unseen test set for all three models (Figure 5.4). The ROC curve shows DeepLncRNA has the highest discriminatory power between the nuclear and cytosolic lncRNAs. Furthermore, we compared DeepLncRNA to the other machine learning models using a range of performance metrics and found DeepLncRNA achieved better performance on every metric except specificity (Table 5.1). While
DeepLncRNA obtained a specificity 4.5% lower than that of a random forest, its sensitivity is 8.1% higher. Based on the more comprehensive metrics such as accuracy, AUC and MCC, we conclude that DeepLncRNA is the optimal model for the prediction of lncRNA subcellular localization.

Figure 5.4 ROC curve performance comparison on the test set.
Table 5.1 Performance metrics on the test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.716</td>
<td>0.726</td>
<td><strong>0.707</strong></td>
<td>0.788</td>
<td>0.433</td>
</tr>
<tr>
<td>SVM</td>
<td>0.713</td>
<td>0.721</td>
<td>0.706</td>
<td>0.795</td>
<td>0.427</td>
</tr>
<tr>
<td>DeepLncRNA</td>
<td><strong>0.733</strong></td>
<td><strong>0.807</strong></td>
<td>0.662</td>
<td><strong>0.812</strong></td>
<td><strong>0.473</strong></td>
</tr>
</tbody>
</table>

To show that DeepLncRNA can be applied to lncRNAs in cell types other than the ones used for training, we examined the role that cell type has on lncRNA subcellular localization. Different cell types have distinct gene expression profiles which means different abundances of the export machinery, such as exporter proteins, which could be needed for specific lncRNAs to exit the nucleus. Therefore, we visualized the conservation of lncRNA subcellular localization across all cell types used in this study (Figure 5.5). Despite the vast differences in tissue types, lncRNA subcellular localization appears highly conserved across cell type. Since the subcellular localization of a lncRNA is not dependent on cell-type, our model is applicable to all human lncRNAs. However, for a small number of lncRNAs there are changes in subcellular localization between certain cell types. This suggests that it may be beneficial to add cell type specific features in the future for the prediction of lncRNA subcellular localization.
Figure 5.5 Heatmap of IncRNA nuclear to cytosolic transcript ratios across cell types. Each bar is a IncRNA transcript colored according to its nuclear to cytosolic log$_2$ fold-change (L2FC) in the respective cell type, white bars indicate the IncRNA was not detected in that cell type. Cell types were then clustered based on their IncRNA localization patterns.

To examine the subcellular localization properties of different subcategories of IncRNAs we used DeepLncRNA to predict the subcellular localization of all annotated human IncRNAs, excluding any IncRNAs in our training set. In total, we predicted the localization
of over 20,000 lncRNAs which we then grouped by gene biotype and evaluated based on the proportion which localize to the nucleus (Figure 5.6). Intriguingly, we observed drastically different proportions of nuclear localization between lncRNA biotypes. Most notably, sense intronic lncRNAs, which reside in the intron of a protein-coding gene, are almost entirely predicted to be enriched in the nucleus. In fact, sense overlapping lncRNAs which can share exons with protein-coding genes are also predicted to be highly nuclear. Thus, both types of sense lncRNAs appear to be highly nuclear which may suggest they predominantly function in the cis-regulation of their embedded protein-coding gene. Almost half of antisense lncRNAs are predicted to be enriched in the cytosol. This is compatible with the fact that many antisense lncRNAs are known to increase the stability of their cognate mRNA by protection from miRNA in the cytoplasm (Rashid, et al., 2016).
Figure 5.6 Percent of annotated lncRNAs predicted to localize in the nucleus. DeepLncRNA predictions of the localization of all annotated lncRNAs grouped by lncRNA biotype. Each bar represents the total percent of lncRNAs in that biotype that are predicted to be localized in the nucleus. The red vertical line represents the boundary between a predominantly cytosolic enriched or nuclear enriched biotype.

Next, we compared the predictions of DeepLncRNA with experimental results from RNA profiling studies of lncRNA subcellular localization. Several lncRNAs have already had their subcellular localization studied through experimental approaches such as fluorescent in situ hybridization of RNA (Cabili, et al., 2015). From the current literature we curated a list of twenty-one lncRNAs with known subcellular localizations, including three lncRNAs which were found to be dual-localized in both subcellular fractions.
(Additional file A-11). However, many of these differentially localized lncRNAs were present in our dataset, therefore, we removed all of them from the training and validation set and recreated DeepLncRNA using the exact same parameters originally used. We then used the new version of DeepLncRNA to predict the subcellular localization of these lncRNAs which have had their localization experimentally tested yet have never been seen by our model (Figure 5.7). DeepLncRNA correctly predicted 8 out of 9 nuclear lncRNAs and 6 out 9 cytoplasmic lncRNAs, based on greater than 50% probability for their respective fraction. Despite not training on dual-localized lncRNAs, DeepLncRNA predicted all three dual-localized lncRNAs are present in the cytoplasm which is correct. The nuclear lncRNA BORG and the cytoplasmic linc-p21 included, are mouse lncRNAs and DeepLncRNA correctly predicted the nuclear retention of BORG (Figure 5.7). These results suggest DeepLncRNA learned generalizability sequence-based features which can predict the lncRNA subcellular localization of novel lncRNAs.
Figure 5.7 DeepLncRNA predictions on lncRNAs with known subcellular localizations. A stacked bar plot showing the percent of lncRNA transcripts predicted to localize to a specific subcellular fraction. LncRNA gene names colored by (red, black and blue) represent nuclear, dual-localized and cytoplasmic lncRNAs, respectively, identified in experimental studies (Additional File A-11).

5.4 Conclusion

In conclusion, we developed DeepLncRNA, a deep learning algorithm which predicts lncRNA subcellular localization directly from lncRNA transcript sequences. DeepLncRNA obtained better accuracy relative to other state-of-the-art machine learning
algorithms and represents a major advancement in lncRNA subcellular localization prediction. The high accuracy of DeepLncRNA indicates that lncRNA primary sequence motifs play a large role in subcellular localization. We predicted the subcellular localization of all annotated human lncRNAs, finding different biotypes possess distinct subcellular localization properties. DeepLncRNA also correctly predicted the localization of more than 75% of a manually curated list of lncRNAs with experimentally validated localizations. In the future, lncRNA subcellular localization prediction will enable the examination of the role that disease-associated point mutations and copy-number variants have on lncRNA function. Since DeepLncRNA is a novel model to predict lncRNA subcellular localization and the number of lncRNAs is expanding we expect DeepLncRNA to play a pivotal role in the functional annotation of lncRNAs.
CHAPTER VI – CONCLUSIONS

In this study, we used genomic data mining methods for the functional annotation of human lncRNAs. Methodologies were developed using either expression-based or sequence-based approaches. For expression-based approaches, we selected two major neurodevelopmental disorders as test cases to identify disease-associated lncRNAs. LncRNA expression was evaluated using a gene coexpression network through the integration of RNA-seq datasets, curated disease-associated gene lists and disease-specific mutational data. First, we curated a list of high-confidence intellectual disability (ID) genes and built a genome-wide coexpression network from RNA-seq data of the developing human brain. We detected a gene module, associated with the immune response, which was specific to ID and the lncRNAs in this module were highly expressed during the mid-to-late fetal period of neurodevelopment. Furthermore, many of these ID-associated lncRNAs were validated through integration of independent mutational data, suggesting that they harbor ID-associated CNVs.

To identify candidate autism spectrum disorder (ASD) associated lncRNAs, we first identified lncRNAs differentially expressed in the ASD cortex, finding hundreds of differentially expressed lncRNAs. To prioritize the lncRNA candidates, we used a brain developmental coexpression network and identified which lncRNAs were highly co-expressed with ASD risk genes. This refined list of candidate ASD-associated lncRNAs was then validated based on ASD mutational data, suggesting that these lncRNAs were present in CNVs at a high frequency. Consequently, our final high-confidence ASD-associated lncRNAs were differentially expressed in the ASD brain, co-expressed with
ASD risk genes in neurodevelopment and present in ASD-associated CNVs higher than expected by chance. Coexpression network analysis is a useful methodology for the functional annotation of lncRNAs because it allows the inference of lncRNA function, based on a “guilt by association” heuristic, with genes of known function. Furthermore, we have demonstrated that the approach can be augmented by the integration of different datasets into the coexpression network such as disease gene lists, mutational data and differential expression data. In the future, coexpression methodologies can be improved upon from the integration of more data types such as methylation data, miRNA data, SNV data and more disease-affected RNA-seq data.

Sequence-based approaches were also developed to discover new knowledge in lncRNA transcript sequences. Due to the limitations of current motif finders, such as small motif sizes and input sequence limits, we developed a genetic algorithm to find discriminative functional motifs. This algorithm scans two sets of lncRNAs and finds a position weight matrix which is enriched in one set yet not the other. Our preliminary results show this algorithm can find large motifs, such as 15-mers, even when multiple random positions are mutated in the test set. Furthermore, using a publicly available list of differentially localized lncRNAs we found a motif that was experimentally associated with nuclear localization of lncRNAs. Future work can likely improve the genetic algorithm performance from the incorporation of more sophisticated motif representation schemes such as hidden markov models which could capture more sequential information.

LncRNA’s biological function is heavily dependent on their subcellular localization because certain functions are compartmentalized to specific cellular locations.
We have constructed a deep learning model, DeepLncRNA, which predicts the nuclear or cytosolic enrichment of a lncRNA directly from the transcript sequence. DeepLncRNA was trained using more than six thousand differentially localized lncRNAs which were identified from differential expression analysis of fractionated samples. DeepLncRNA obtained high accuracy on the unseen test dataset and a curated high-confidence list of known differentially localized lncRNAs. Furthermore, when DeepLncRNA was used to predict the localization of all annotated lncRNAs, it found large differences in localization trends between different lncRNA biotypes. While protein subcellular localization prediction is a well-established field, lncRNA subcellular localization is a novel research area which will greatly aid in the functional annotation of lncRNAs. In the future as experimental data accumulates, more advanced sequence-based features can be added to improve prediction accuracy, such as post-transcriptional modification motifs and structure-associated motifs. Furthermore, RNA-seq data from fractionation techniques to isolate the chromatin, nucleoplasm, ribosome and exosome fractions could be incorporated into the training data to change the model into a multi-class predictor which should improve discriminative power.

In the future, there are many paths toward more sophisticated functional characterization of lncRNAs. For expression-based methods of lncRNA functional annotation, the interaction of lncRNAs and protein-coding genes is only one component of their regulatory function. Another known aspect of lncRNA functionality is their interaction and regulation with small noncoding RNAs, such as miRNAs. However, most RNA-seq protocols do not profile small RNAs like miRNAs due to either the RNA
extraction technique used or a size selection screening based on sequence length. Therefore, coexpression network analysis is missing an important piece of the puzzle for the characterization of IncRNA function. The integration of miRNAs into coexpression networks will allow the identification of miRNA- IncRNA interactions which could function as regulatory control mechanisms for the gene expression of other genes. For example, if a miRNA targets the mRNA of a disease gene, reducing its expression, and a IncRNA is found to sequester that miRNA, then it could be said that this IncRNA is indirectly regulating the disease gene. These type of complex functional relationships can only be discovered if all RNA species are adequately profiled, small and large, prior to coexpression network analysis.

Sequence-based methods for IncRNA functional annotation will also see great improvements in the future, due to the accumulation of more data and the application of sophisticated deep learning methodologies. The utilization of k-mers leave much to be desired due to their omission of long-range sequential interactions and positional information. Recurrent neural networks are being used to learn both short and long-range sequential patterns in sequences of arbitrary length. These types of algorithms will be superior in the functional characterization of IncRNAs due to the learning of both sequential information, which motifs are present, as well as positional patterns, where are the motifs located. The motifs identified by these models will provide great insight into the biological function of IncRNAs by enabling the clustering of IncRNAs into functional families based on short and long-range motif characteristics.
As the number of IncRNAs appears to keep growing, data mining approaches will become invaluable tools for the functional annotation of IncRNAs. Currently, expression and sequence-based methods offer the best means of understanding IncRNA functionality. Integrating both approaches will likely be the future of genomic data mining for IncRNA functional annotation. Furthermore, as experimentally derived 3D structures of IncRNAs accumulates structure-based approaches will likely fill in some of the missing pieces of deciphering IncRNA functionality. Once we comprehend the functional transcriptome we will began to unlock the secrets of disease and biological development.
Appendix A – Additional Files

Additional File A-1 ID gene lists.

Additional File A-2 ID WGCNA results.

Additional File A-3 Modular ID gene enrichment.

Additional File A-4 Ranked ID lncRNAs.

Additional File A-5 Characteristics of genes differentially expressed in ASD.

Additional File A-6 R markdown document containing source code and exploratory ASD data analysis.

Additional File A-7 ASD Weighted gene coexpression network analysis results.

Additional File A-8 Prioritized ASD-associated lncRNAs.

Additional File A-9 Subcellular localization RNA-seq sample metadata

Additional File A-10 DeepLncRNA parameter optimization

Additional File A-11 Manually curated lncRNAs with known localizations
Figure B-1 Tissue specificity of differentially expressed lncRNAs. The heatmap displays lncRNAs median FPKM abundance for each human tissue type. LncRNAs were Z-score normalized across tissue types, and then hierarchical clustering was performed on the tissue types. The blue boxes adjacent to the hierarchical cluster tree denote if the tissue type is derived from the brain.
Figure B-2 Coexpression of developmental gene modules. The bar plot shows the average biweight midcorrelation of all gene modules in the brain developmental network. The red circle within each bar represents the average biweight midcorrelation of 10,000 randomly selected gene sets of equal size to the genes within the respective module.
Figure B-3 Coexpression of ASD gene sets with differentially expressed lncRNAs. The histograms display the summed biweight midcorrelation between ASD gene sets and 10,000 randomly selected gene sets of equal size to the differentially expressed lncRNAs. The red vertical line represents the sum of the biweight midcorrelation between an ASD gene set with the differentially expressed lncRNAs. P-values were calculated based on the difference between the actual summed correlation (red line) and the permuted normal distribution and adjusted for multiple comparisons (adjusted p-values < 0.001).
REFERENCES


H2O.ai (2017) h2o: R Interface for H2O. Available at: ttps://cran.r-project.org/package=h2o.


Kumar, V. *et al.* (2013) Human Disease-Associated Genetic Variation Impacts Large


Voineagu, I. et al. (2011) Transcriptomic analysis of autistic brain reveals convergent


