12-2016

Functional Analysis of Human Long Non-coding RNAs and Their Associations with Diseases

Steven Cogill
Clemson University, scogill@g.clemson.edu

FUNCTIONAL ANALYSIS OF HUMAN LONG NON-CODING RNAS AND THEIR ASSOCIATIONS WITH DISEASES

A Dissertation
Presented to
the Graduate School of
Clemson University

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

---

by
Steven Cogill
December 2016

Accepted by:
Dr. Liangjiang Wang, Committee Chair
Dr. Julia Frugoli
Dr. Hong Luo
Dr. Anand Srivastava

ABSTRACT

Within this study, we sought to leverage knowledge from well-characterized protein coding genes to characterize the lesser known long non-coding RNA (lncRNA) genes using computational methods to find functional annotations and disease associations. Functional genome annotation is an essential step to a systems-level view of the human genome. With this knowledge, we can gain a deeper understanding of how humans develop and function, and a better understanding of human disease. LncRNAs are transcripts greater than 200 nucleotides, which do not code for proteins. LncRNAs have been found to regulate development, tissue and cell differentiation, and organ formation. Their dysregulation has been linked to several diseases including autism spectrum disorder (ASD) and cancer. While a great deal of research has been dedicated to protein-coding genes, the relatively recently discovered lncRNA genes have yet to be characterized. LncRNA function is tied closely to when and where they are expressed. Co-expression network analysis offer a means of functional annotation of uncharacterized genes through a "guilt by association" approach. We have constructed two co-expression networks using known disease-associated protein-coding genes and lncRNA genes. Through clustering of the networks, gene set enrichment analysis, and centrality measures, we found enrichment for disease association and functions as well as identified high-confidence lncRNA disease gene targets. We present a novel approach to the identification of disease state associations by demonstrating genes that are associated with the same disease states share patterns that can be discerned from transcriptomes of healthy tissues. Using a machine learning algorithm, we built a model to classify ASD

versus non-ASD genes using their expression profiles from healthy developing human brain tissues. Feature selection during the model-building process also identified critical temporospatial points for the determination of ASD genes. We constructed a webserver tool for the prioritization of genes for ASD association. The webserver tool has a database containing prioritization and co-expression information for nearly every gene in the human genome.

DEDICATION

I dedicate my dissertation to my parents, friends, and partner. Without their support and affection, this work would not be possible.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Wang for his guidance and support in my research project. I am thankful that he has shared his wealth of knowledge in bioinformatics and academic science with me. I would like to thank Dr. Luo for the opportunity to learn benchtop skills in his lab, Dr. Frugoli for her career and writing advice, and Dr. Srivastava for sharing his human genetics knowledge. I would also like to thank my lab mates, Jose Guevara and Brian Gudenas, for their support. Finally, I would like to thank the faculty, staff, and students at the Clemson University Department of Genetics and Biochemistry and the Greenwood Genetic Center for providing me with the opportunity to pursue my PhD studies.

TABLE OF CONTENTS

Page

Table of Contents (Continued)

Page

LIST OF TABLES

LIST OF FIGURES

List of Figures (Continued) Page

CHAPTER I – LITERATURE REVIEW OF THE FUNCTIONAL ANNOTATION OF HUMAN LONG NON-CODING RNAS USING COMPUTATIONAL METHODS

1.1 Introduction

Biological data has expanded both in quantity and complexity from next-generation sequencing and high-throughput methods. Most noteworthy is the global aspect of data. We have begun to work in 'omics', looking at systems as a whole, and data at this scale and complexity diminishes the ability of the scientist alone to efficiently and effectively discern actionable knowledge. An example would be the study of BReast Cancer susceptibility gene 1 (BRCA1). Mutations in BRCA1 were discovered in 1990 to be associated with families at high risk for breast and ovarian cancers (Hall *et al.,* 1990). Given its impact on cancer research and relatively early discovery, it has been intensely studied for nearly 30 years (Scalia-Wilbur *et al.*, 2016). Currently there are 12,812 articles pertaining to BRCA1 in Pubmed (2011, accessed on 8/30/16). A scientist with a specific question may be able to find a relevant article or a review which can answer their query or provide an overview of BRCA 1, but a broad query such as identifying a comprehensive list of potential binding partners for BRCA1 and discerning interesting shared characteristics amongst the binding partners would require a comprehensive view of the available literature. While an individual scientist may not be able to read the 12,812 articles, bioinformatics data mining techniques such as text mining and sentiment analysis could provide a solution.

The research environment has become data rich, information poor. The origins of the phrase "data rich, information poor" or otherwise known as the DRIP syndrome is difficult to determine. One of the first mentions of data rich information poor in biological research context is from Williamson (1987) which roughly coincided with the first international medical informatics conference hosted in 1985 (Sewell and Thede, 2012). The phrase itself is in reference to the low cost of data production and the lack of return on this accumulation of data. Return comes in the form of knowledge which gives the data clarity and allows for decision making and predictions. Data in itself is meaningless, and applicable knowledge has to be extracted from it using data mining techniques (Obeidat *et al*., 2015). This idea became very apparent with widespread use of microarray technologies in the late 90s before there was a suitable infrastructure to handle the data deluge (Schulze and Downward, 2000). Data has begun to accumulate faster than it can be analyzed (Schatz *et al*., 2010), which has led to the rise of cloud computing, bioinformatics cores, and the incorporation of bioinformatics curricula at research universities (Dai *et al*., 2012; Lewitter *et al*, 2009, Welch *et al*., 2014).

The earliest work in bioinformatics was focused on databases. Following the advent of protein sequencing by Tuppy and Sanger (1951), the Protein Data Bank was formed in 1972 to archive the new sequences (Bernstein *et al*., 1977). With the advent of the internet, next generation sequencing, and high-throughput methodologies, the number of publically available databases has ballooned to a recent listing of 1,552 databases (Fernandez-Suarez *et al*., 2014). Biological databases can be categorized by the type of data contained within and how the data was obtained. Biological database data types are

DNA, RNA, protein, expression, pathway, disease, nomenclature, literature, and standard/ontology (Zou *et al*., 2015). Databases which act as repositories for experimentally derived results are considered primary databases, whereas databases which house data from analysis on this primary data are considered secondary databases (Mewes *et al.*, 2010). An example of a primary database is GenBank, which is a comprehensive collection of all publically available DNA sequences (Benson *et al*., 2013). An example of a secondary database is the UniProt knowledgebase, which contains annotated protein entries (Bateman *et al.*, 2015). The structuring and archiving of information into accessible formats is an active area of research. One of the challenges that lie in biological database research is structuring data from unstructured data. This is particularly difficult with the more traditional relational databases that store data in interconnected formally-defined tables. Novel approaches are being developed to address the issue. An example is the approach of Lysenko *et al.* (2016) who applied graph theory (see section 1.3) to the structuring of the database. Databases are so much an active area of bioinformatics research that the journal *Nucleic Acids Research* releases an annual database issue covering new releases and advances in the field.

Bioinformatics is a multidisciplinary field that stands at the cross-section of mathematics, statistics, biology, and computer science. The emergence of bioinformatics as a major discipline incorporating data mining amongst other tools followed closely with the publishing of the human genome (Lander *et al*., 2001). Indeed, a formal definition was not proposed until 2001 when Luscombe *et al*. (2001) describe it as: "Conceptualizing biology in terms of macromolecules (in the sense of physical-

chemistry) and then applying ‘informatics’ techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale”. Although the definition remains apt, Bioinformatics is pluralistic, and can be further broken down into three sub-disciplines. The first is the development of algorithms which relies heavily on mathematics. Commonly, this field tries to reduce the complexity of existing algorithms, and one common area this is applied to is in image processing. An example is the study done by Zhao *et al.* (2016) where they developed a variation on principal component analysis for faster processing of cryo-electron microscopy images. The second sub-discipline is the analysis and interpretation of biological data which may also be considered computational biology and relies on biological training. An example of this type of study would be co-expression network analysis (see section 1.3). The third sub-discipline is the development of tools, which relies heavily on computer science. An example of this type of study would be the development of a disease gene prioritization systems (see section 1.5). Additionally, bioinformatics studies may show overlap between these different sub-disciplines. The definition of a bioinformatician is also pluralistic as well. Many would emphasize computational aspects over the biological aspects and vice versa. A recent survey by Bartlet et al. (2016) amongst bioinformaticians in the United Kingdom found that there was a large cultural divide amongst the different disciplines of bioinformatics and that the backgrounds of key members varied greatly as well. Although it is evolving, bioinformatics is a field that has allowed for extraction of useful knowledge from the data deluge.

1.2 Long non-coding RNAs

The evolution and quantity of lncRNAs in the human genome

While only 2% of the human genome is made up of protein-coding regions, it is estimated that nearly 75% of the entire genome is transcribed (Djebali *et al*., 2012). When the genome was first sequenced, the genome size and the number of protein coding genes showed little to no correlation with organismal complexity. It has since been found that the complexity of an organism is actually more closely associated with the number of non-coding RNA genes (Taft *et al*., 2007, Necsulea *et al*., 2014).

Non-coding RNA genes are transcribed but not translated. Non-coding RNAs lack an open reading frame (ORF) with coding potential. Determining the coding potential of a transcript is a multi-step process. The GENCODE consortium (Harrow *et al*., 2012), whose lncRNA annotations were used for a majority of the studies in this dissertation, first compare the transcript to known sequences using the Basic Local Alignment Search Tool n (BLASTn) (Altschul *et al*., 1990) to cluster the transcripts and then compare the clusters to existing non-coding RNA families in RFAM (Nawrocki *et al.*, 2015). Next they determine the length of the longest potential ORF if one is present. If the length of the ORF is greater than 35% of the transcript length then the transcript is determined to have coding potential. They also look for homology between potential proteins coded within the ORF and any known Protein families in Pfam (Finn, R. 2016). The last aspect of coding potential that is considered by the consortium is codon substitution frequency within the potential ORF. The PhyloCSF method performs multiple sequence alignments and measures conservation based upon the frequencies of synonymous codon

substitutions, conservative amino acid substitutions, and missense and non-sense substitutions, and this has been demonstrated as being effective in determining coding potential as non-coding sequences have lower conservation (Lin *et al*., 2011). Most groups which attempt to identify lncRNAs within the human genome employ a methodology similar to that of the GENCODE consortium for determining the coding potential of ORFs within transcripts. For example, the lncRNA gene identification study of Iyer *et al.* (2015) used the Coding Potential Assessment Tool (CPAT) to determine coding potential. CPAT employs a logistic regression model accounting for ORF size, Fickett Testcode Statistic (Fickett, 1982), and hexamer usage bias, which are similar features to those accounted for in the GENCODE method (Wang *et al.*, 2013). Another interesting method of note for determining coding potential is ribosome profiling. It has been found that ribosomes can potentially bind long non-coding RNAs (lncRNAs), but this does not lead to translation (Guttman *et al*., 2013). The ribosome profiling can distinguish between a coding and a non-coding transcript based upon the sharpness of ribosome release e.g. coding transcripts release once they reach the stop codon where as non-coding transcripts have a much more variable release point (Guttman *et al*., 2013).

Until the early 90s, the functions of RNAs were relegated to messenger RNAs (mRNAs), which is the intermediate between DNA and proteins, and the housekeeping RNAs such as transfer RNAs and ribosomal RNAs which are constitutively expressed and help maintain the base functionality of the cell (Yang *et al.*, 2016). Non-coding RNAs are classified based upon their size with small non-coding RNAs (sncRNA) being less than 200 nucleotides in length, and lncRNAs being greater than 200 nucleotides in

length (Kapranov *et al*., 2007). The theory of lncRNAs as regulators was proposed in 1961 by Jacob and Monod (Jacob and Monod, 1961; Kung *et al.*, 2013). The first discovery in 1990 of an lncRNA with a regulatory role was that of H19, which regulates in a cis fashion the expression of insulin like growth factor 2 and plays a role in embryonic development (Brannan *et al.*, 1990; Gabory *et al.*, 2006). This was followed closely in 1992 by the discovery of X-inactive specific transcript (Xist) which is key in the inactivation of the X chromosome (Brockdorff *et al.*, 1992; Brown *et al.*, 1992). With the discovery of the regulatory role lncRNAs, it was hypothesized that mRNAs may serve dual functions (Karapetyan *et al*., 2013). In this instance, genes whose transcripts undergo alternative splicing could code for both messenger RNAs (mRNAs), which are translated to proteins, and lncRNAs serving an alternative function. While there have been recent examples, there are currently very few examples of genes demonstrating this behavior (Karapetyan *et al*., 2013). While the roles of sncRNAs have been well characterized over the last 10 years, lncRNAs remain poorly characterized (Clerget *et al*., 2015; Xu *et al*., 2016).

LncRNAs share many similarities to mRNAs in that they can have a 5’ cap and 3’ polyadenylation, undergo alternative splicing, and are transcribed by RNA polymerase II (Ulitsky and Bartel, 2013). Currently the GENCODE project, which seeks to identify all genes within the human genome, lists 15,941 lncRNA genes (Harrow *et al*., 2012). However, it is believed that lncRNA genes are more numerous than protein coding genes, and a recent computational study by Iyer *et al*. (2015) detected 58,648 lncRNA genes. Discrepancies in the estimated number of lncRNA genes arises from differences in

methodologies. GENCODE is widely used and considered by some to be the standard of gene annotations for the human genome. Examples supporting this claim are its use in the building of the BrainSpan dataset (Hawrylycz *et al*., 2012) and its use in the University of California Santa Cruz genome browser (Kent *et al.,* 2002). Given its use as a resource and need for highest accuracy, the GENCODE consortium is conservative in its estimates, and they employ a manual curation through the HAVANA group which looks at the genome itself rather than mapping the transcripts as outlined above (Harrow *et al*., 2012). Another discrepancy is the amount of RNA seq data that is produced. While Iyer *et al.* (2015) produced the largest human lncRNA discovery list to date, they also used ~100 fold greater RNA seq data than previous studies that included data from tumor tissues and cancer cell lines. Another key difference between lncRNA discovery studies is the handling of single exon lncRNAs. Earlier versions of GENCODE did not include them within their list of lncRNAs due to their unreliability in accurately being mapped and determined definitively to be non-coding, but Iyer *et al*., included them within their study.

LncRNAs are classified by their genomic location into four categories: sense overlapping, intronic sense, bi-directional promoter, antisense, and intergenic or intervening (Figure 1.1) (Ma *et al*., 2013). Intergenic lncRNAs otherwise known as long intergenic/intervening RNAs (lincRNAs), as their name implies, lie between protein coding genes and comprise 7,539 of the 15,767 (48%) lncRNAs listed in GENCODE (Harrow *et al*., 2012). Sense overlapping lncRNAs contain a protein coding gene within their coding region on the same strand. Antisense lncRNAs overlap exon or introns of

protein-coding genes on the opposing strand. Intronic sense lncRNAs are located within the introns of protein-coding genes on the same strand. Bidirectional promoter lncRNAs overlap the promoter of a protein-coding gene on the opposing strand.

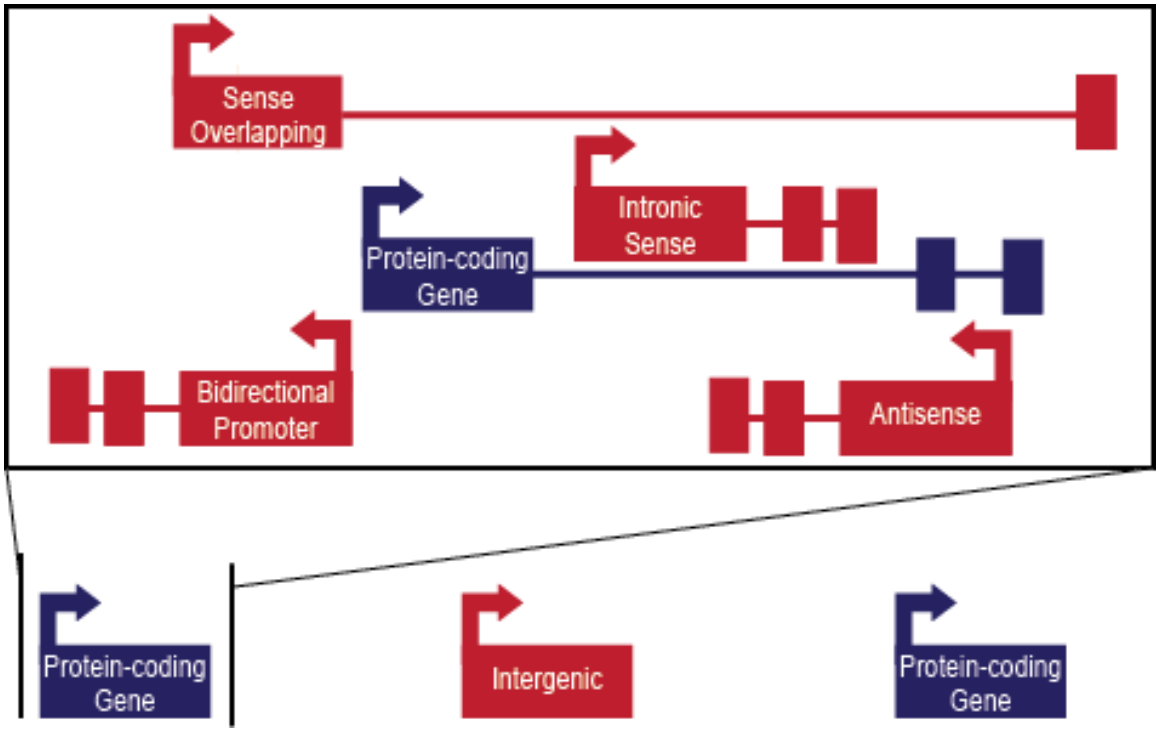


Figure 1.1 Positional classification of lncRNA genes. Protein-coding genes are shown in blue and lncRNA genes are shown in red. For each gene, the arrows indicate strand placement, the thick lines indicate exons, and the thin lines indicate introns. The box is an expanded view of the region surrounding the protein coding gene. Adapted from Derrien *et al*. (2012).

In a landmark study, Nesculea *et al*. (2014) demonstrated that evolutionarily, lncRNA genes have poor interspecies conservation across exons in comparison to protein-coding genes, but their splice sites and promoter regions are more conserved than protein-coding genes. They also found that lncRNAs have higher time and tissue specific expression, and lncRNAs that are found to be conserved from lower organisms have been

primarily associated with embryogenesis based on the higher likelihood of their promoter regions to contain HOX transcription factor binding sites. Collectively, this indicates a rapid evolution of lncRNAs and implicates the more recent lncRNAs in the formation of the complex organs in higher organisms. In perhaps one of the most complex organs, the human brain, lncRNAs have shown elevated expression relative to other tissues (Derrien *et al*., 2012).

Functional Annotation

After the identification of the numerous lncRNAs within the genome, the next step in gaining a deeper understanding of their role is to determine their functions (Figure 1.2). A genome annotation for a gene is a description of the gene, its product, which can be either RNA or protein, and the function of that product (Koonin and Galperin, 2003). The description is further refined by gene ontology, which seeks to apply a structured vocabulary to biological processes, cellular components, and molecular functions associated with a given gene. This concept was proposed by the Gene Consortium in 1998 for model organism databases (Ashburner *et al*., 2000). The hierarchies have been adopted and expanded on by multiple term enrichment software packages to allow for more fields such as disease associations and keywords (see more in Section 1.3). The general process of functional gene annotation is first the identification of the gene within the genome. This is followed by *in silico* annotation which seeks to associate a putative function for the gene product. The third step of the process is the experimental validation of the annotation. While automated annotation cannot account for all caveats and can display inconsistencies (Devos and Valencia, 2001; Schnoes *et al*., 2009; Brenner, 1999),

the putative functional assignment provided by the computational analysis can guide the experimental aspect as well as prioritize targets warranting further study (Koonin and Galperin, 2003). In addition, manual annotation errors can still arise and in larger datasets, they tend to persist longer (Percudani *et al*., 2013).

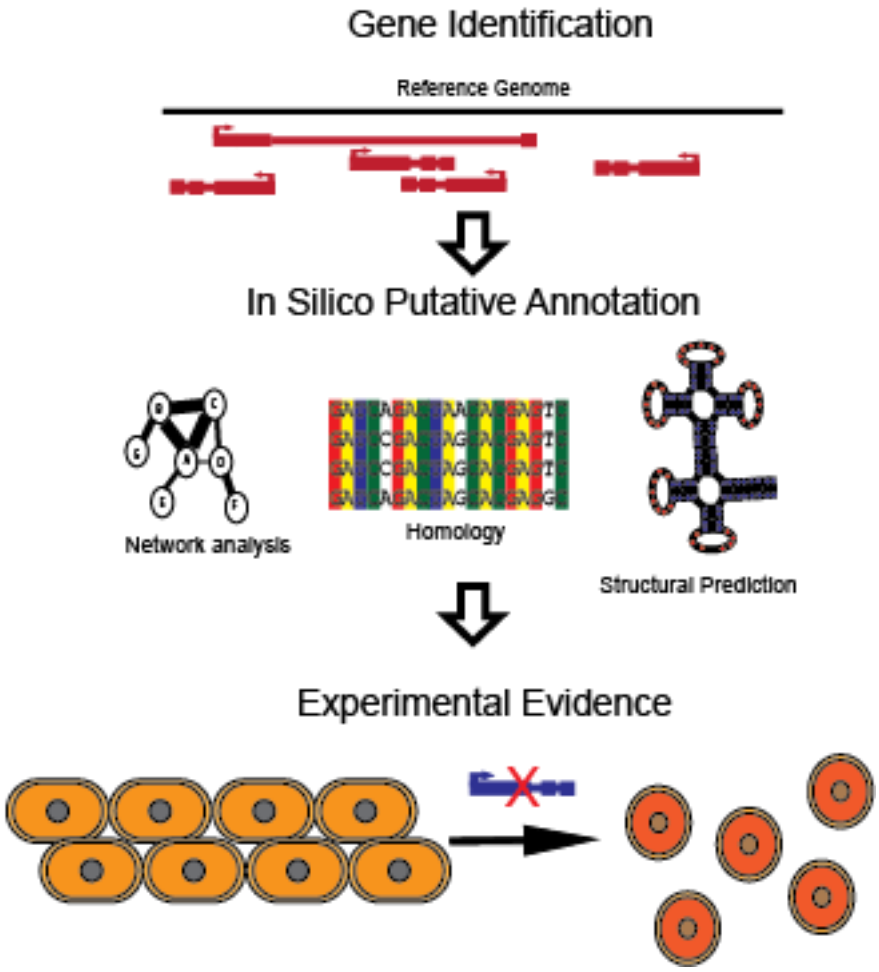


Figure 1.2 Schematic of the functional annotation process. This is a step by step process of the functional annotation of genes. The first step shows the mapping of transcripts to a reference genome. The second steps indicates some of the potential computational methods that are used to annotate the gene. The final step is the experimental validation of function. An example of this being the effect of the knockout of a gene, which affects the morphology of the cell.

Functional annotation is the precursor step to understanding how genes interact on a systems level, and lncRNA genes present interesting challenges in terms of annotation. The first is that although the number of lncRNA genes is comparable if not greater than protein coding genes (see above), very few of the lncRNA genes lack functional annotation (Laurent *et al.*, 2015). The lncRNA database, which derives its lncRNA entries from literature searches, only has 184 entries for human lncRNAs (Quek *et al.*, 2015). With such a small fraction of lncRNAs with known function, comparative analysis provides little insight into function.

The second challenge is that, in contrast to sncRNAs whose function are closely related to their sequence (Clerget *et al*., 2015), lncRNA function can be dependent on sequence or derived from structure. For example the function of lncRNA referred to as highly upregulated in liver cancer (HULC) is dependent upon its structure. It has been found to have a competitive endogenous function in that it binds and sequesters the miRNA, miR-372 (Wang *et al*., 2010). The well-studied HOX transcript antisense RNA (HOTAIR) has been shown to have structurally dependent binding to polycomb repressive complex 2 (PCR2) and lysine-specific histone demythylase 1A (LSD1) (Huang *et al*., 2014). These are just two examples of the dichotomy of function demonstrated by lncRNAs.

The annotation of lncRNAs is further complicated by a lack of conservation. While protein-coding genes show high levels of conservation in their sequences, lncRNAs only demonstrate small and difficult to discern ultra-conserved regions (Johnsson *et al*., 2014). These considerations rule out traditional similarity methods that

are employed to determine function. However similarity methods such as network analysis can be applied between the expression patterns of well-studied protein-coding genes and lncRNA genes.

The molecular mechanisms of lncRNAs

Although many lncRNAs have been discovered, there various functions are still being determined (Kung *et al*., 2013). To provide organization to the known functions of lncRNAs, Wang and Chang (2011) proposed four archetypes of molecular mechanisms employed by lncRNAs: signal, decoy, guide, and scaffold (Figure 1.3), and here I use these archetypes to further describe the function of lncRNAs

As mentioned previously, lncRNA have highly specific temporal spatial expression patterns in comparison to protein-coding genes. As signals, lncRNAs can respond to stimuli and initiate biological processes or provide feedback as to the current state of the cell. As an example of an lncRNA initiating a biological process, Xist is activated during development to inactivate the X chromosome (Brockdorff *et al.*, 1992; Brown *et al.,* 1992). As an example of cell state feedback, prostate cancer antigen 3 (PCA3) is a lncRNA that is only expressed in prostate cancer cells (Bussenmakers *et al*., 1999; Hessels *et al*., 2003). These links to biological processes and cell state based on their expression allow us an avenue into the functional annotation of lncRNAs through shared expression patterns with the well-characterized protein coding genes. Another mechanism of lncRNAs is as a decoy. This again points to the dichotomy of function of lncRNAs based on either sequence or structure as decoys bind and attenuate the function

of their target. An example of sequence based decoy function are the competitive endogenous long non-coding RNAs which share binding sites with miRNAs. These lncRNAs essentially act as a sponge to 'soak up' through binding and attenuate the function of miRNAs (Thompson and Dinger, 2016). This role is critical and in dysregulation can lead to disease states. Recently, Liu *et al*. (2016) discovered that the lncRNA Ras suppressor protein 1 pseudogene 2 (RSR1P2) competitively binds the micro RNA let-7a in a process that promotes cervical cancer. An example of a structure based decoy is the lncRNA Growth arrest-specific 5, which binds to and represses the glucocorticoid receptor (Kino *et al.*, 2010). LncRNAs classified as guides effectively direct protein functions to designated targets. The previously mentioned HOTAIR is an example of an lncRNA which employs this mechanism. It binds both PRC2 and LSD1 which are chromatin modifying enzymes and targets them to HOX gene loci to repress their expression (Huang *et al*., 2014). The final mechanism archetype is scaffold. Given the large number of potential secondary structures for lncRNAs (Wan, *et al.*, 2014) and the 6%-8% frequency of RNA binding proteins (Jingna *et al.*, 2015), lncRNAs are particularly effective in the formation of protein complexes. While HOTAIR also employs this mechanism archetype, probably the most well-known example would be ribosomal RNAs. The full diversity of lncRNA function remains to be determined, which is why lncRNAs are currently one of the most active areas of research.

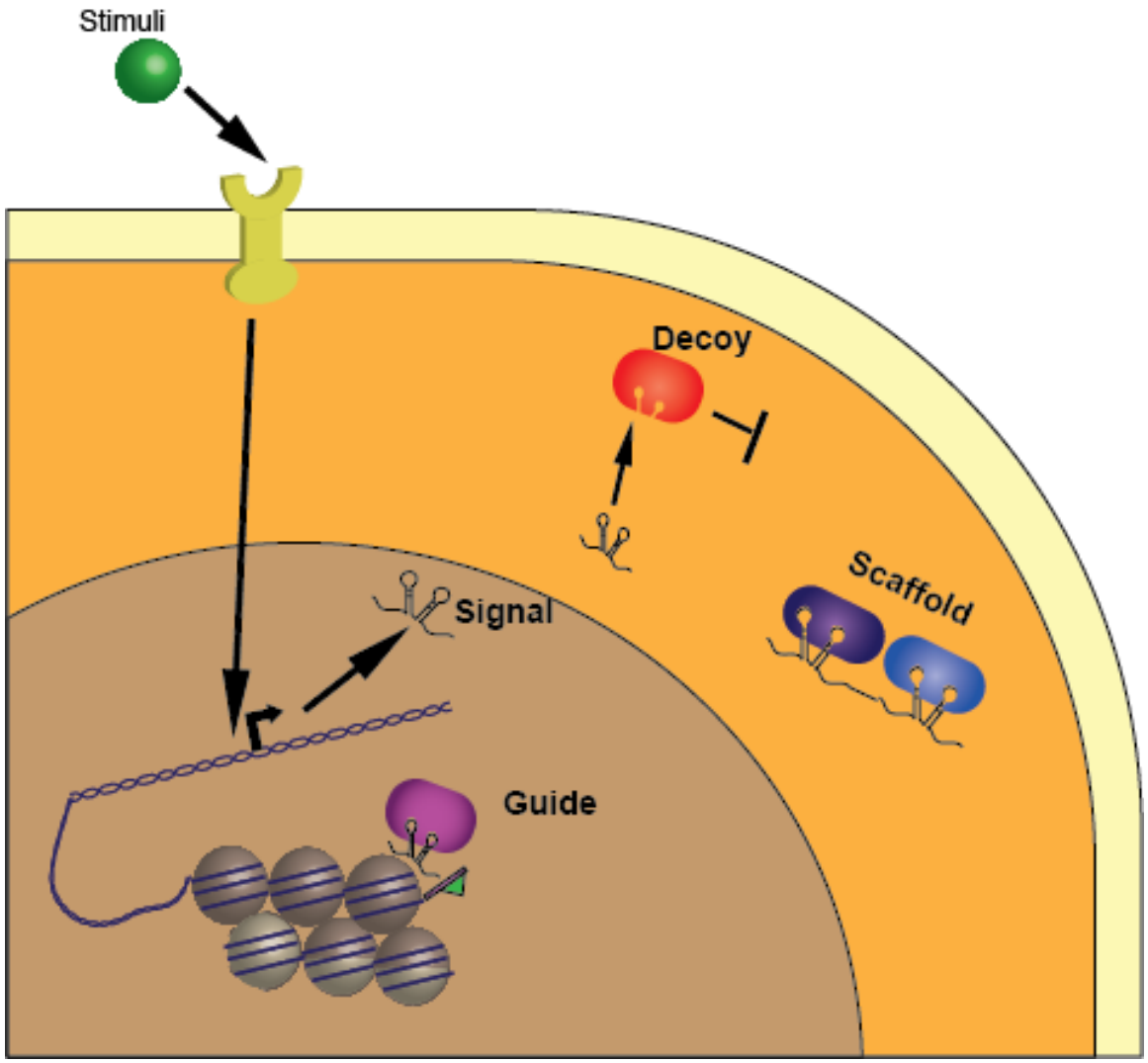


Figure 1.3 Schematic of the molecular mechanism archetypes for lncRNAs. The diagram shows lncRNAs acting as signals, decoys, guides, and scaffolds within both the cytoplasm and nucleus of the cell. Adapted from Wang and Chang (2011).

LncRNAs in human disease

Given characterization of lncRNAs above, it is not surprising that lncRNAs are associated with multiple diseases. The lncRNADisease database contains entries for experimentally determined lncRNAs associated with disease based upon literature

searches (Chen *et al.*, 2013). They currently have 478 lncRNA entries and they have found associations with 166 diseases. Although it may be reflective of the degree to which diseases receive focus, lncRNAs have been found to have particularly critical roles in cardiac diseases and cancers. One example of a critical lncRNA in cardiac disease is tie-1 AS regulates the expression of tyrosine kinase with immunoglobulin-like and EGF-like domains 1 (tie-1). Tie-1 helps maintain cell junctions and when overexpressed, tie-1 leads to impairment of vascular development (Li *et al.*, 2010). Several examples of cancer associated lncRNAs exist. One previously mentioned example is HOTAIR, which has been shown to be overexpressed in and contribute to breast cancer (Gupta *et al*., 2010). Another interesting cancer associated lncRNA is metastasis-associated lung adenocarcinoma transcript 1 (MALAT1). It has been associated with 16 different cancers, and its overexpression correlates with metastasis and thus poor prognosis (Wei and Niu, 2015). These lncRNAs play critical roles in the progression of cancers and offer potential treatment targets, however, lncRNAs are particularly sensitive to alterations in cell state and therefore offer excellent means of diagnosis and prognosis (Xiong *et al*., 2016). The previously mentioned PCA3 is one such example. The role of lncRNAs in autism spectrum disorders (ASD) is still being determined, however given that lncRNAs impact development and show elevated expression in the brain (see above), they are likely contributors. This is further supported by the Ziats and Rennert (2013) study which detected 222 differentially expressed lncRNAs between ASD and control brain samples.

1.3 Co-expression network analysis

Network analysis is based upon graph theory, one of the disciplines in discrete mathematics. The first instance of the application of graph theory was in 1736 by the mathematician Euler. It is commonly used to map networks. In networks, nodes or vertices are entities such as genes, which are connected by edges. This connection implies a relationship. The nature of this relationship determines whether the network is directed or undirected. Directed networks imply causality. An example would be in a gene regulatory network. If gene A upregulates gene B then the directionality would be from A to B, however this does not imply that when gene B is downregulated, gene A is upregulated. In undirected networks, there is no causality or direction. For the previous example in an undirected network, gene A and gene B would simply share a connection. Mapping of biological networks is a common practice. Graph theory offers a semantic view of data and has a wide variety of applications. It has been used to map and understand brain function (Mears and Pollard, 2016), describe evolution (Shakarian *et al.*, 2012) and design and discover drugs (Takigawa and Mamitsuka, 2013). The idea of the co-expression networks evolved from the work of Butte and Kohane (1999), who initially proposed the employment of graph theory to biological networks. Bulle and Kohane were the first to point to the idea that nodes within biological networks can be connected via correlations. This eventually led to the idea of co-expression networks. Their idea was that if a correlation measured via any correlation measurement was above a threshold, then an edge or connection was established.

Co-expression networks are constructed through the measure of correlation between expression profiles for two genes. Examples of correlation measures include Pearson product moment correlation and Spearman correlation. Once this value has been determined then a connection is established either through a hard threshold such as ranking the interactions and taking a certain high percentile or by applying a weighted or soft threshold where the connectivity is a continuous measure (Langfelder and Horvath, 2008). Genes can then be clustered into what is referred to as modules based on their similarities in expression profiles. These modules can be measured for enrichment of terms, gene biotypes (i.e. lncRNAs, protein-coding genes, etc.) or any tag that can be assigned to the individual genes in comparison to a background. Therefore, genes of known function, such as many human protein-coding genes can be used to characterize genes of unknown function such as lncRNAs. Network analysis allows us to look for topology to determine the nature of interactions. Network topology refers to statistical measures which describe the distribution of nodes and edges. One commonly employed topology measure is the degree distribution. This measures the distribution of connections per node. For example in scale-free networks, plotting of the degree distribution should follow a linear pattern. These topologies can be overlaid to determine the strength of the network and note any changes between overlays. Network analysis also allows for a measure of centrality, which is the degree of the node or otherwise stated the number of connections or the summation of the connection weights for weighted networks. Genes showing high connectivity or centrality are assumed to be critical to the process being studied (Serin *et al*., 2016).

Co-expression network analysis is a "guilt by association" approach that has been widely employed in research (Serin *et al.*, 2016). The first gene co-expression network analysis was performed by Carter *et al.* (2004) to identify critical genes in the determination of cell state. Giuletti *et al*. (2016) built a co-expression network to identify critical genes in the development of pancreatic ductal adenocarcinoma. Lv *et al*. (2016) utilized a co-expression network to identify nitrogen responsive long intergenic non-coding RNAs (lincRNAs) in maize. Oliver *et al*. (2014) applied a co-expression network to disease gene prioritization for epileptic encephalopathy. These are just a few examples of the wide applications of this approach. The key assumption of co-expression network analysis is that over a range of samples, expression patterns will be shared by genes in the same pathways, between interacting partners, and genes that share functions.

1.4 Machine learning

Machine learning is a field comprised of methods for the identification of patterns from complex data to complete a given task. The idea of machine learning was first proposed by Alan Turing (1950) in his famous Turing test, in which he proposed that it may be possible for a machine learn to the point that it would be impossible for an interviewer to discern whether they were talking to a machine or a person. The first implementation of machine learning was the perceptron machine (Rosenblatt, 1958), which later became a binary classification algorithm that adjusts its decision boundary dependent on input data. This was shortly followed by the KNN algorithm which classify unknown instances by employing a distance metric and determining the classification of

the closest instances from the training set (Cover and Hart, 1967). The adoption of machine learning into the field biological science was heralded by Ted Shortliffe, who was the first to employ it to solve biological problems (Shortliffe, 1973).

Machine learning is becoming ubiquitous as it is employed in nearly every facet of business and technology (Jordan and Mitchell, 2015). In biology, its ability to make classifications based on large complex datasets make a valuable asset. The identification of a suitable machine learning problem can be the most time-consuming in the field of bioinformatics and requires domain-specific knowledge. Machine learning requires a task, a means of scoring the performance of the algorithm to perform this task, and experience upon which to learn. The experience itself is data, and the nature of this data determines the type of machine learning problem. If the data is unlabeled, then the task is unsupervised. Common tasks include clustering such as in a co-expression network, rule association, and dimensionality reduction. Rule association is the search for reliable association between fields in data. It is commonly used for sales transactions, in that if someone buys product A then they are likely to buy product B as well. A more pertinent example within biology would be the mining of a transcriptome to find that if genes are upregulated in tissue A, then they are down regulated in tissue B. Dimensionality reduction, as the name implies, are the methods for reducing the number of dimensions within data. This reduction allows for easier training of models and can decrease the probability of overfitting of models (high performance in training but poor real world performance). For example, Kim *et al*. (2016) recently applied self-organizing maps,

which are a form of artificial neural networks, to ultrasonography images. The group found patterns which led to the discernment of appendicitis.

However, if the data is labeled, for example disease versus non disease gene, then the machine learning algorithm is said to be supervised learning. The most common task for supervised machine learning is classification. These tasks seek to devise a model based upon examples with known classes to determine the class of unknown entities. Support vector machine (SVM) is a popular algorithm for machine learning. It is well suited to biological data as it deals well with high dimensionality and numerical versus categorical data (Cortes and Vapnik, 1995; Kourou *et al.*, 2014). It has been applied to diagnosing attention deficit hyperactive disorder using neuropsychological data (Bledsoe *et al*., 2016), diagnosing gastric cancer using serum biomarkers (Tong *et al*., 2016), and determining the onset of Alzheimer's disease using magnetic resonance imaging (Wei *et al*., 2016). The algorithm can be used for binary classifier that seeks to find a decision boundary between two groups plotted with their features acting as dimensions. The decision boundary that is found has the widest possible margin, which is the greatest distance from the closes training instances on either side of the decision boundary.

1.5 Candidate gene prioritization

Gene prioritization servers are particularly useful. Most disorders that are studied are complex in that they are multigenic. For example, there are over 400 genes associated with autism spectrum disorder (ASD) (Abrahams *et al*., 2013). When attempting to find genes associated with a disorder, association studies can lead to large gene lists. This

becomes a difficulty because each of those genes within the list is challenging to validate and test. Another difficulty in dealing with complex diseases is the extent of interaction. Two hypotheses are considered for complex diseases, one is the common disease common variant hypothesis, which argues that there exist many alleles with low penetrance which impact the expression of the disease. A counter to this hypothesis is the common disease rare variant hypothesis which argues that rare alleles with high penetrance lead to expression of the disease (Schork *et al*., 2009). While both hypotheses can be considered equally valid depending on the disease in question, disease prioritization servers operate on the common disease rare allele hypothesis. These systems are designed to prioritize gene lists based upon their likelihood to be associated with a given disease. Often these systems will utilize data from several sources including expression data, literature, physical interactions, and annotations. One guiding principle of prioritization systems is that disease genes are convergent on pathways which has been verified for ASD (Parikshak *et al*., 2013), and disruption of these pathways lead to disease. Therefore if a prioritization can identify genes within a shared pathway of disease genes, then these genes can be implicated in the disorder. GeneMANIA is an interesting example of a prioritization server. The user supplies their own gene list and genes with the highest association through analysis of co-expression, interactions, pathways, and co-localization, are returned to the user (Warde-Farley *et al*., 2010). The ENDEAVOUR system employs a similar methodology. Users submit a training set which is then passed through multiple models to allow for the prioritization of a given candidate gene list (Tranchevent *et al*., 2008).

1.6 My research

In chapter 2, we demonstrate the application of co-expression networks to determining the role of lncRNAs in cancer. In chapter 3 we employed co-expression network analysis to characterize the role of lncRNAs in neural development and autism spectrum disorders. In chapter 4, we apply the SVM algorithm to predict and prioritize ASD risk genes. In chapter 5, we show the construction of PGAR, a system incorporating co-expression analysis and machine learning results from previous studies (chapters 3 and 4) to prioritized genes for their association with ASD.

References

Abrahams, B. et al. (2013) SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism*, 4, 36.

Altschul, S. et al. (1990) Basic local alignment search tool. *J. Mol. Biol*, 215, 403-410.

Ashburner et al. (2000) Gene ontology: tool for the unification of biology *Nat Genet*, 25**,** 25-29.

Bartlett, A. et al. (2016) Generations of interdisciplinarity in bioinformatics. *New Genet Soc*, 35, 186-209.

Bateman, A. et al. (2015) UniProt: A hub for protein information. *Nucleic Acids Res*, 43, D204-212.

Benson, D. et al. (2013) GenBank. *Nucleic Acids Res*, 41, D36-D42.

Bernstein, F. et al. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *J Mol Biol*, 112, 535-542.

Bledsoe, J. et al. (2016) Diagnostic classification of ADHD versus control: Support vector machine classification using brief neuropsychological assessment. *J Atten Disord*, [Epub ahead of print].

Brannan, C. et al. (1990) The product of the H19 gene may function as an RNA. *Mol Cell Biol*, 20, 28–36.

Brenner, S. (1999) Errors in genome annotation. *Trends Genet*, 15, 132-133.

Brockdorff, N. et al. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71, 515-526.

Brown C. et al. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71, 527–542.

Bussemakers, M. et al. (1999) DD3: A new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res*, 59, 5975–5979.

Butte, A and Kohane, I. (1999) Unsupervised knowledge discovery in medical databases using relevance networks. *Proc AMIA Symp*, 711-715.

Carter, S. et al. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20, 2242-2250.

Chen, G. et al. (2013) LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*, 41, D983-D986.

Clerget, G. et al. (2015) Small non-coding RNAs: A quick look in the rearview mirror. *Methods Mol Biol*, 1296, 3-9.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13, 21-27.

Dai, L. et al. (2012) Bioinformatics clouds for big data manipulation. *Biol Direct*, 7, 43.

Derrien, T. et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775–1789.

Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet*, 17, 429-431.

Djebali, S. et al. (2012) Landscape of transcription in human cells. *Nature*, 489, 101-108.

Fernandez-Suarez, X. et al. (2014) The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Res*, 42, D1–D6.

Fickett, J. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*, 10, 5303-5318.

Finn, R. et al. (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res*, 44, D279-D285.

Gabory, A. et al. (2006) The H19 gene: regulation and function of a non-coding RNA. *Cytogenet Genome Res*, 113, 188-93.

Giulietti, M. et al. (2016) Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development. *Cell Oncol (Dordr)*, [Epub ahead of print].

Gupta, R. et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071-1076.

Guttman, M. et al. (2013) Ribosome profiling provides evidence that large non-coding RNAs do not encode proteins. *Cell*, 154, 240-51.

Hall, J. et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250, 1684-1689.

Harrow, J. et al. (2012) GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*, 22, 1760-1774.

Hawrylycz, M. et al. (2012) An Anatomically Comprehensive Atlas of the Adult Human Brain Transcriptome. Nature, 489, 391-399.

Hessels, D. et al. (2003) DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur Urol*, 44, 8–16.

Huang da, W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.

Huang, L. et al. (2014) Overexpression of long noncoding RNA HOTAIR predicts a poor prognosis in patients with cervical cancer. *Arch Gynecol Obstet*, 290, 717-723.

Iyer, M. et al. (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*, 47, 199-208.

Jacob F. and Monod J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3, 318–356.

Jinga, S. et al. (2015) Computational Prediction of RNA-Binding Proteins and Binding Sites. *Int J Mol Sci*, 16, 26303–26317.

Johnsson, P. et al. (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta*, 1840, 1063-1071.

Jordan, M. and Mitchell, T. (2015) Machine learning: Trends, perspectives, and prospects. *Science*, 349, 255-60.

Kapranov, P. et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316, 1484-1488.

Karapetyan, A. et al. (2013) Regulatory roles for long ncRNA and mRNA. *Cancers (Basel)*, 5, 462-490.

Kent, W. et al. (2002) The human genome browser at UCSC. *Genome Res*, 12, 996-1006.

Kim, K. et al. (2016) Automatic extraction of appendix from ultrasonography with self-organizing map and shape-brightness pattern learning. *Biomed Res Int*, [Epub ahead of print].

Kino, T. et al. (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal*, 3, ra8.

Koonin, E. and Galperin, M. (2003) Sequence - Evolution - Function: Computational approaches in comparative genomics. Boston: Kluwer Academic. Chapter 5, Genome Annotation and Analysis.

Kourou, K. et al. (2014) Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, 13, 8-17.

Kung, J. et al. (2013) Long noncoding RNAs: past, present, and future. *Genetics*, 193, 651-669.

Lander, E. et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559.

Laurent, G. et al. (2015) The landscape of long non-coding RNA classification. *Trends Genet*. 31, 239-251.

Lewitter, F. et al. (2009) The need for centralization of computational biology resources. *PLoS Comput Biol*, 5, e1000372.

Li, K. et al. (2010) A noncoding antisense RNA in tie-1 locus regulates tie-1 function in vivo. *Blood*, 115, 133-139.

Lin, M. et al. (2011) PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27, i275-i282.

Luscombe, N. et al. (2001) What is bioinformatics? A proposed definition and overview of the field. Methods *Inf Med*, 40, 346-358.

Lv, Y. et al. (2016) Genome-wide identification and functional prediction of nitrogen-responsive intergenic and intronic long non-coding RNAs in maize (Zea mays L.). *BMC Genomics*, 17, 350.

Lysenko, A. et al. (2016) Representing and querying disease networks using graph databases. *BioData Min*, 9, 23.

Ma, L. et al. (2013) On the classification of long non-coding RNAs. *RNA Biol*, 10, 925-933.

Mears, D. and Pollard, H. (2016) Network science and the human brain: Using graph theory to understand the brain and one of its hubs, the amygdala, in health and disease. *J Neurosci Res*, 94, 590-605.

Mewes, H. et al. (2010) MIPS: Curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res*, 39, D220-D224.

Nawrocki, E. et al. (2015) Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res*, 43, D130-D137.

Necsulea, A. et al. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505, 635-640.

Obeidat, M. et al. (2015) DRIP – Data rich, information poor: A concise synopsis of data mining. *Universal Journal of Management*, 3, 29-35.

Oliver, K. et al. (2014) Harnessing gene expression networks to prioritize candidate epileptic encephalopathy genes. *PloS One*, 9, e102079.

Parikshak, N. et al. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, 155, 1008-1021.

Percudani, R. et al. (2013) Ureidoglycolate hydrolase, amidohydrolase, lyase: how errors in biological databases are incorporated in scientific papers and vice versa. *Database (Oxford)*, 2013, bat071.

PubMed Health [Internet]. Bethesda (MD): National Library of Medicine (US); [updated 2011 Jan 1; cited 2011 Jan 6]. Available from: http://www.ncbi.nlm.nih.gov/pubmedhealth/

Quek, X. et al. (2015) LncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*, 43, D168-D173.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65, 386.

Sanger, F. and Tuppy, H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. I. the identification of lower peptides from partial hydrolysates. *Biochem J*. 49, 463-481.

Scalia-Wilbur, J. et al. (2016) Breast cancer risk assessment: Moving beyond BRCA 1 and 2. *Semin Radiat Oncol*, 26, 3-8.

Schatz, M. et al. (2010) Cloud computing and the DNA data race. *Nat Biotechnol*, 28, 691–693.

Schork, N. et al. (2009) Common vs. Rare Allele Hypotheses for Complex Diseases. *Curr Opin Genet Dev*, 19, 212–219.

Schulze, A. and Downward, J. (2000) Analysis of gene expression by microarrays: Cell biologist's gold mine or minefield? *J Cell Sci*, 113, 4151-4156.

Schnoes, A. et al. (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, 5, e1000605.

Serin, E. et al. (2016) Learning from co-expression networks: Possibilities and challenges. *Front Plant Sci*, 7, 444.

Sewell, J., & Thede, L. (2012). Informatics and nursing: Opportunities and challenges (4th ed.). Philadelphia: Lippincott,Williams &Wilkins.

Shakarian, P. et al. (2012) A review of evolutionary graph theory with applications to game theory. *Biosystems*, 107, 66-80.

Shortliffe, E. et al. (1973) An artificial intelligence program to advise physicians regarding antimicrobial therapy. *Comput Biomed Res*, 6, 544-560.

Taft, R. et al. (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29, 288-299.

Takigawa, I. and Mamitsuka, H. (2013) Graph mining: procedure, application to drug discovery and recent advances. *Drug Discov Today*, 18, 50-57.

Thompson, D. and Dinger, M. (2016) Endogenous microRNA sponges: evidence and controversy. *Nat Rev Genet*, 17, 272-283.

Tong, W. et al. (2016) Serum biomarker panels for diagnosis of gastric cancer. *Onco Targets Ther*, 9, 2455-2463.

Tranchevent, L. et al. (2008) Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res*, 36, W377-W384.

Turing, A. (1950) Computing machinery and intelligence. *Mind*, 59, 433-460.

Ulitsky, I. and Bartel, D. (2013) LincRNAs: genomics, evolution, and mechanisms. *Cell*, 154, 26-46.

Wan, Y. et al. (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505, 706-709.

Wang, J. et al. (2010) CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res*, 38, 5366-5383.

Wang, K. and Chang, H. (2011) Molecular mechanisms of long noncoding RNAs. *Mol Cell*, 43,904-914.

Wang, L. et al. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*, 41, e74.

Warde-Farley, D. et al. (2010) The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, 38, W214-W220.

Wei, R. et al. (2016) Prediction of conversion from mild cognitive impairment to Alzheimer's disease using MRI and structural network features. *Front Aging Neurosci*, 8, 76.

Wei, Y. and Niu, B. (2015) Role of MALAT1 as a Prognostic Factor for Survival in Various Cancers: A Systematic Review of the Literature with Meta-Analysis. *Dis Markers*, [Epub].

Welch, L. et al. (2014) Bioinformatics curriculum guidelines: Toward a definition of core competencies. *PLoS Comput Biol*. 10, e1003496.

Williamson, E (1987) Management information: Avoiding the data rich, information poor syndrome. *Mich Hosp*, 6, 19-23.

Xiong, X. et al. (2016) Long non-coding RNAs: An emerging powerhouse in the battle between life and death of tumor cells. *Drug Resist Updat*, 26, 28-42.

Xu, J. et al. (2016) A comprehensive overview of lncRNA annotation resources. *Brief Bioinform*, [Epub ahead of print].

Yang, J. et al. (2016) Non-coding RNAs: An introduction. *Adv Exp Med Biol*, 886, 13-32.

Zhao, Z. et al. (2016) Fast steerable principal component analysis. *IEEE Trans Comput Imaging*, 2, 1-12.

Ziats, M. and Rennert, O. (2013) Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci*, 49, 589-593.

Zou, D. et al. (2015) Biological databases for human research. *Genomics Proteomics Bioinformatics*, 13, 55-63.

CHAPTER II - CO-EXPRESSION NETWORK ANALYSIS OF HUMAN LNCRNAS AND CANCER GENES

Steven Cogill and Liangjiang Wang

Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

Published: *Cancer Inform*, 13, 49–59 (2014)

Abstract

We used gene co-expression network analysis to functionally annotate long noncoding RNAs (lncRNAs) and identify their potential cancer associations. The integrated microarray dataset from our previous study was used to extract the expression profiles of 1,865 lncRNAs. Known cancer genes were compiled from the Catalogue of Somatic Mutations in Cancer and UniProt databases. Co-expression analysis identified a list of previously uncharacterized lncRNAs that showed significant correlation in expression with core cancer genes. To further annotate the lncRNAs, we performed a weighted gene co-expression network analysis, which resulted in 37 co-expression modules. Three biologically interesting modules were analyzed in depth. Two of the modules showed relatively high expression in blood and brain tissues, whereas the third module was found to be downregulated in blood cells. Hub lncRNA genes and enriched functional annotation terms were identified within the modules. The results suggest the utility of this approach as well as potential roles of uncharacterized lncRNAs in leukemia and neuroblastoma.

2.1 Introduction

Long noncoding RNAs (lncRNAs) are a major class of noncoding RNAs and exceed 200 nucleotides in length. Originally suspected of being the result of transcriptional noise, lncRNAs have been shown to have a broad range of functions including transcriptional regulation, mediating protein interactions, and influencing mRNA splicing (Cech and Steitz, 2014). The ENCODE project has demonstrated that 74.7% of the human genome is transcribed, and more than 9,000 lncRNAs have been annotated (Djebali *et al*, 2012; Derrien *et al*., 2012). A large number of lncRNAs have also been identified in many other organisms. For instance, the FANTOM3 annotation project has discovered 34,030 lncRNA transcripts in the mouse genome (Maeda *et al*., 2006). These studies have led to the projection that there may be more lncRNAs than protein-coding genes. The roles in biological processes and mechanism of action for the majority of lncRNAs have not yet been determined (Calibi *et al*., 2011; Wang and Tran, 2013). For functional annotation, a weighted gene co-expression network analysis (WGCNA) of lncRNAs with well-annotated protein-coding genes offers an approach for insight into the biological roles of lncRNAs (Langfelder and Horvath, 2008).

A definitive link between cancer and lncRNAs has been established through disease state studies and their functions in development and cellular differentiation (Cheetham *et al*., 2013; Iyengar *et al*., 2014; Zhu *et al*., 2014). Examples of well-studied lncRNAs associated with cancer include HOX antisense intergenic RNA (HOTAIR), prostate cancer antigen 3 (PCA3) and metastasis-associated lung adenocarcinoma

transcript 1 (MALAT1). HOTAIR interacts with Polycomb Recessive Complex 2 (PRC2) and the LSD1/CoREST/ REST complex to modify histones, which results in silencing at multiple sites (Rinn *et al*., 2007; Tsai *et al*., 2010). PCA3 in contrast has no known function but acts as an effective noninvasive diagnostic marker for prostate cancer (Bussenmakers *et al*., 1999; Hessels *et al*., 2003). MALAT1, which was first discovered in a differential expression study of non–small-cell lung cancer tumors, has been linked to 16 different cancer types including cervical cancer and hepatocellular carcinoma (Ji *et al*., 2003; Chen *et al*., 2013; Guo *et al*., 2010; Luo *et al*., 2006). These three lncRNAs share the common feature found in most cancer associated lncRNAs, which are overexpressed in cancerous tissues (Bussemakers *et al*., 1999; Luo *et al*., 2006; Huang *et al*., 2014). The significant changes in expression levels aid in determining the function of these cancer-associated lncRNAs, which have become important for diagnosis and prognosis of cancers.

This study is unique in the application of co-expression analysis to normal (noncancerous) tissues to determine lncRNA and cancer gene associations. Previous studies have focused on differential expression between normal and cancerous tissues. An example is the genome-wide differential and co-expression analysis of hepatoblastoma tissues (Dong *et al*., 2014). Bipartite network analysis has also been performed to predict lncRNA–disease associations (Yang *et al*., 2014). In this study, we use a previously compiled dataset consisting of 2,968 microarray expression profiles across a wide spectrum of tissues (Wang *et al*., 2010). All expression profiles in this dataset were obtained using publicly available data from the Affymetrix HG-U133 Plus 2.0 Array

platform, which provides suitable genome coverage for known protein-coding genes with 98.6% of our cancer gene list being represented in the array probes. This microarray platform also contains probes for 1,970 lncRNAs (Zhang *et al*., 2012). By utilizing the available data for co-expression analysis, we have examined the previously uncharacterized lncRNAs for their potential role in cancer and functional annotation.

2.2 Methods

Gene lists

A core and an extended gene list of known cancer genes were compiled for this study (Additional file A-1). The core list comprised the known causal cancer genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census List from the Wellcome Trust Sanger Institute (Futreal *et al*., 2004). Redundant genes and genes that do not have protein-level expression were removed from the list. The UniProt knowledgebase was used to determine if there was evidence of protein-level expression (UniProt Consortium, 2014). The core list consisting of 472 protein-coding cancer genes was used to select microarray probes for the co-expression analysis of cancer genes and lncRNAs. To expand the core list for all plausible cancer genes, additional cancer genes not present within the core list were added to create the extended gene list (Additional file A-1). A custom search query was used to search the UniProt knowledgebase for additional cancer genes. Among the search criteria was a requirement for evidence of protein-level expression. The extended list consisting of 951 protein-coding cancer genes was used to select microarray probes for the WGCNA. The lncRNAs used in this study

(Additional file A-1) have at least one corresponding probe on the Affymetrix HG-U133 Plus 2.0 Array.

Microarray Expression Data

The microarray gene expression dataset was compiled in our previous study (Wang *et al*., 2010). The dataset had 2,968 microarray gene expression profiles generated using the Affymetrix HG-U133 Plus 2.0 Array with 54,675 probe sets. A data integration method was developed to combine the expression profiles from 131 different microarray studies into a single dataset (Wang *et al*., 2010). Most human tissue types were represented in the integrated microarray dataset, and the high quality of the dataset was demonstrated by examining tissue-specific gene expression patterns as well as for identifying co-expressed genes.

Co-expression analysis of cancer genes and lncRNAs

For each cancer gene probe in the core list, co-expression was calculated against all lncRNA probes individually using the microarray expression data. Co-expression was measured by Pearson product–moment correlation with Microsoft Excel (2013). The top 10 absolute correlation values were kept. *P*-values were calculated using R 3.0.2 (R Core Team, 2013). Due to the high degrees of freedom, the *P*-value after Bonferroni correction for multiple testing in each correlation measurement returned a significance of <6.53E-13. Cancer gene and lncRNA function were retrieved from the NCBI Gene database (Maglott *et al*., 2011). Cancer gene disease associations were provided in the COSMIC Cancer Gene Census List (Futreal, 2004).

Weighted gene co-expression network analysis

The co-expression network was constructed using the WGCNA package (Langfelder and Horvath, 2008). The normalized expression data for probes from the extended cancer gene list and lncRNAs were used as input. Given the relatively large dataset and our interest in finding all the co-expression modules, we opted for a smaller minimum module size at 10 probes. The merge cut height, defined as the threshold of dissimilarity, 1-Topological Overlap Matrix (TOM), below which separate modules would be merged, was set to 0.2. Visual inspection of the initial hierarchical clustering revealed no outliers, and soft thresholding was set to 4. An unsigned network with connections based upon absolute correlations was constructed. Module assignment of cancer genes and lncRNAs was performed using democratic vote method. A gene was assigned to the module that had the highest number of probes for the gene. Genes with equal numbers of probes in different modules were assigned using the highest mean module membership for the probes.

Functional term enrichment analysis

Each module was analyzed for gene ontology term enrichment using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang da *et al*., 2009). The analysis was performed using the Affymetrix probe identifiers in each module with the Affymetrix HG-U133 Plus 2.0 Array as the background. Where significant, functional annotation terms were selected for biological process, molecular function, and Online Mendelian Inheritance in Man (OMIM) disease association (Amberger *et al*., 2009). The *P*-value provided by DAVID was used as the measure of significance. The significance threshold was set to 0.1 for the reported functional terms.

Network visualization

Network visualization was performed using the VisANT software (Hu *et al*., 2013). All edges were based on TOM values with a threshold set for a minimum of one connection for each node. The 100 probes with the highest intramodal connectivity were analyzed. Node size was determined by connectivity.

2.3 Results

Normal cross-tissue expression profiles show high co-expression between lncRNAs and cancer genes

Cancer genes and lncRNAs appear to be involved in some common biological functions. Examples include the involvement in development and transcriptional regulation. Both cancer genes and lncRNAs have been shown to have tissue-specific expression patterns (Cabili *et al*., 2011). We thus hypothesize that associations between known cancer genes and lncRNAs could be demonstrated through correlations in expression across various tissue samples. Previous studies have also shown that different isoforms of a cancer gene or lncRNA may have specific activity, function, and impact on cancer progression (Li *et al*., 2013; Bochenek *et al*., 2013). Because of this possibility and our concern about the poor quality of some probe sets, we studied the microarray data at the probe level instead of combining multiple probe sets for a gene. We examined the highest co-expression correlations between the lncRNA and cancer gene probes. The degree of co-expression is shown here as a measure of the Pearson product–moment correlation. Since lncRNAs may have a silencing effect, the absolute correlation was used

in the ranking to account for a negative Pearson correlation (Negishi *et al*., 2014; Pandey *et al*., 2008).

The core cancer genes are a curated list of genes from the COSMIC cancer gene census database, and they are causal in that for 90% of the gene list, mutations in somatic cells causes some form of cancer and for 20% of the gene list mutations in germ line cells causes a predisposition to cancer (Futreal *et al*., 2004). From the integrated microarray expression dataset, which contains 2,968 profiles of various normal tissue samples (Wang *et al*. 2010), we extracted the expression profiles for the corresponding probes of all available lncRNAs and core cancer genes. The 10 highest correlations were compiled (Additional file A-2). Interestingly, a large majority of the co-expressed lncRNA probes show positive correlation with cancer genes and the minority show negative correlation. The well-known lncRNA HOTAIR showed a positive correlation (0.38) with homeobox C13 (HOXC13) and the lower level of positive correlation (0.28) with the transcription factor paired box 1 (PAX1). The lowly expressed lncRNA, PCA3, only showed a low level of positive correlation (0.23) with Rho guanine nucleotide exchange factor 12 (ARHGEF12).

To highlight the extent of co-expression between lncRNAs and cancer genes, the pairs with the highest correlations were compiled and annotated. Ten cancer genes with the highest absolute correlations with lncRNAs are shown in Table 2.1. The disease associations are from the COSMIC list (Futreal *et al*., 2004). All the correlation values are greater than 0.8 and well below the significance threshold of 0.05 (P-value < 6.53E-13). The majority of the lncRNAs analyzed in this study lack any functional annotation,

and this is reflected in the highly co-expressed lncRNAs. MEG3 is the only lncRNA to have functional annotation (Maglott *et al.*, 2011). Notably, two lncRNAs, LOC100505812 and ITGB2 antisense RNA 1 (ITGB2-AS1), demonstrate high co-expression with multiple cancer genes. For the cancer genes highly co-expressed with LOC100505812, three (PTPRC, FLI1, and IKZF1) are associated with acute lymphoid leukemia and two (IKZF1 and MYD88) are associated with diffuse large B-cell lymphoma. ITGB2-AS1 has high co-expression with IKZF1 and LCK, both of which are associated with acute lymphoid leukemia. The third cancer gene co-expressed with ITGB2-AS1, WAS, is associated with lymphoma. The proteins encoded by the cancer genes have various functions. Two of the 10 proteins function as transcription factors, and 5 have DNA- or RNA-binding capacity. While the majority of the proteins appear to have functions related to transcription, the other proteins include receptors, phosphatases, and kinases. Four of the 10 cancer genes are involved in the immune response.

Weighted gene co-expression network analysis shows close associations of lncRNAs and cancer genes

WGCNA with the extended gene list resulted in 37 distinct modules (Figure 2.1A and 2.1B). With the exceptions of Module 3 and Module 5, the six largest modules showed a greater number of cancer gene probes within the module than lncRNA probes (Figure 2.1B). Module 3 had twofold more lncRNAs than cancer genes. All of the modules contained at least one lncRNA probe, and Module 34 was the only module that contained only lncRNA probes. Nevertheless, the majority of the modules showed a relatively equal distribution of lncRNAs and cancer genes. There were 1,493 out of the

5,079 probes analyzed (29.4%) which were not assigned to any modules (shown in grey in Figure 2.1A). Out of the 2,632 cancer gene probes, 489 (18.6%) were not assigned, whereas 1,004 out of the 2,447 lncRNA probes (41.0%) were left out.

Table 2.1 Identification of lncRNAs highly co-expressed with known cancer genes

Cancer Gene	Function	Disease Associations	lncRNA	Function	Correlation Coefficient
PTPRC	Protein tyrosine phosphatase receptor involved in T-cell activation	Acute lymphoid leukemia (ALL)	LOC100505812	Uncharacterized	0.86215
FLI1	Transcription factor and proto-oncogene	Ewing sarcoma, ALL	LOC100505812	Uncharacterized	0.85744
IKZF1	Zinc finger transcription factor involved in lymphocyte differentiation	ALL, diffuse large B-cell lymphoma (DLBCL)	LOC100505812	Uncharacterized	0.84260
			ITGB2-AS1	Uncharacterized	0.83235
			C21orf96 (RUNX1-IT1)	Uncharacterized	0.82207
RBM15	RNA-binding motif protein	Acute megakaryocytic leukemia	LOC144438	Uncharacterized	0.83453
HNRNP A2B1	Ribonucleoprotein involved in pre-mRNA processing	Prostate cancer	FLJ31306	Uncharacterized	0.81977
CNBP	Zinc finger SSDNA and SSRNA-binding protein	Aneurysmal bone cyst	LOC388789	Uncharacterized	0.81457
MYD88	Adapter protein for Toll-like receptor and interleukin-1 (IL-1) signaling	DLBCL	LOC100505812	Uncharacterized	0.81449
LCK	Protein tyrosine kinase involved in T-cell development	ALL	ITGB2-AS1	Uncharacterized	0.81244
CHN1	GTPase-activating protein involved in neuronal signal-transduction	Extraskeletal myxoid chondrosarcoma	MEG3	Potential tumor suppressor that interacts with p53	0.81079
WAS	Signal transduction protein possibly involved in actin filament reorganization	Lymphoma	ITGB2-AS1	Uncharacterized	0.81049

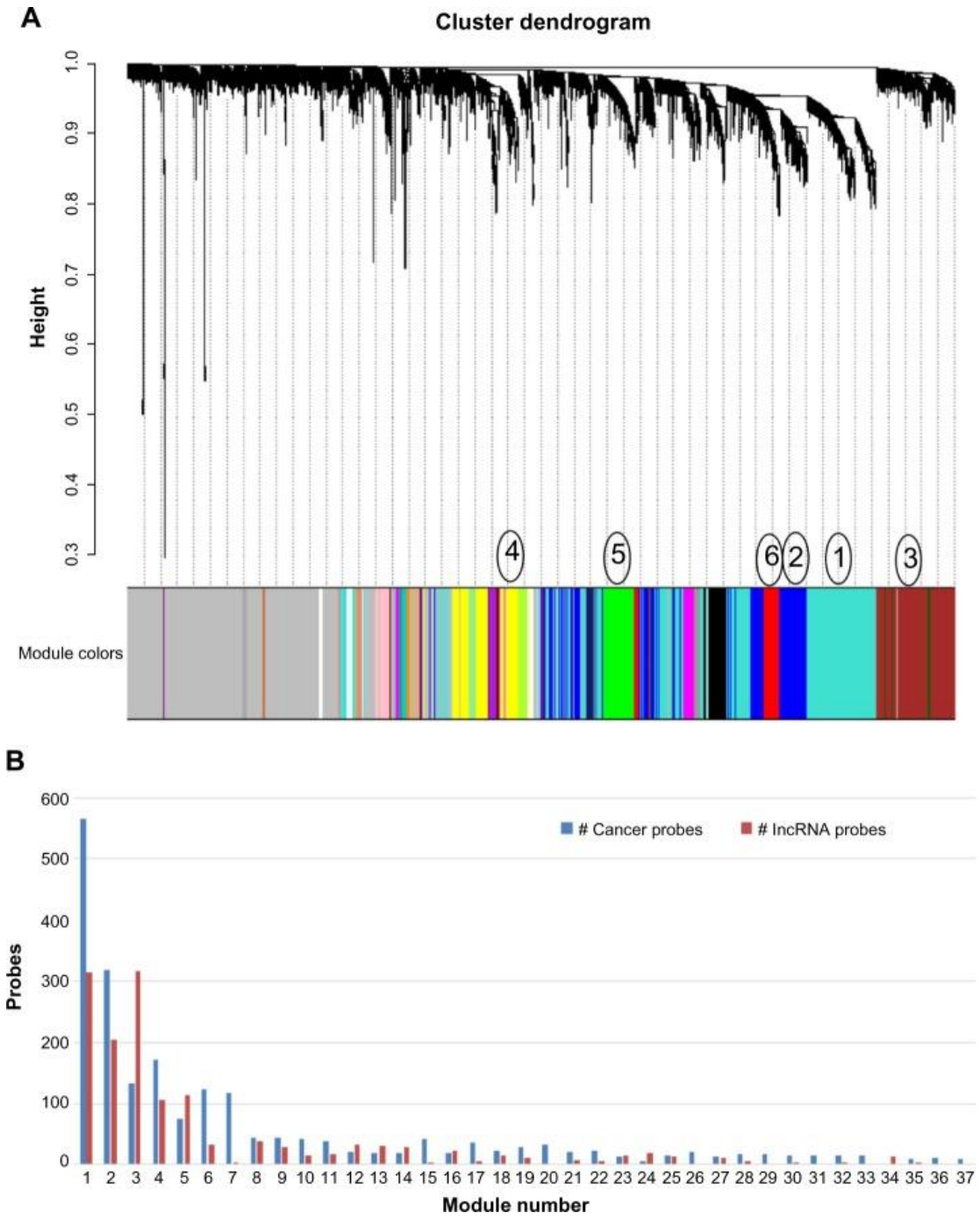


Figure 2.1 WGCNA of cancer genes and lncRNAs. (A) Cluster dendrogram of the co-expression modules. The modules were designated numerically based on size, and the six largest modules with Module 1 as the largest module are labeled adjacent to their respective color band. The grey band contains probes not assigned to any module. (B) Chart of the probe counts for cancer genes and lncRNAs respectively for each module.

Modules 1, 4, and 5 were chosen for further analysis. These selected modules were larger in size and showed high connectivity and module membership (data not shown) as well as divergence in expression patterns from one another.

Module 1 shows functional enrichment of transcriptional activity and blood-specific expression patterns

To examine the expression pattern of the module, samples were grouped by tissue types, and the mean expression level in each tissue type was calculated. As shown in Figure 2.2A, the average expression level of Module 1 genes is significantly higher in blood cells than the other tissues. Within the blood tissue type, neutrophils have the highest expression. The blood-specific expression pattern is also evident in the Module 1 expression heat map (Figure B-1A). The other tissue types have median to low expression for both lncRNAs and cancer genes. Interestingly, the cancer genes have a more uniform high expression pattern in blood cells in comparison to lncRNAs, which show moderate expression in blood cells.

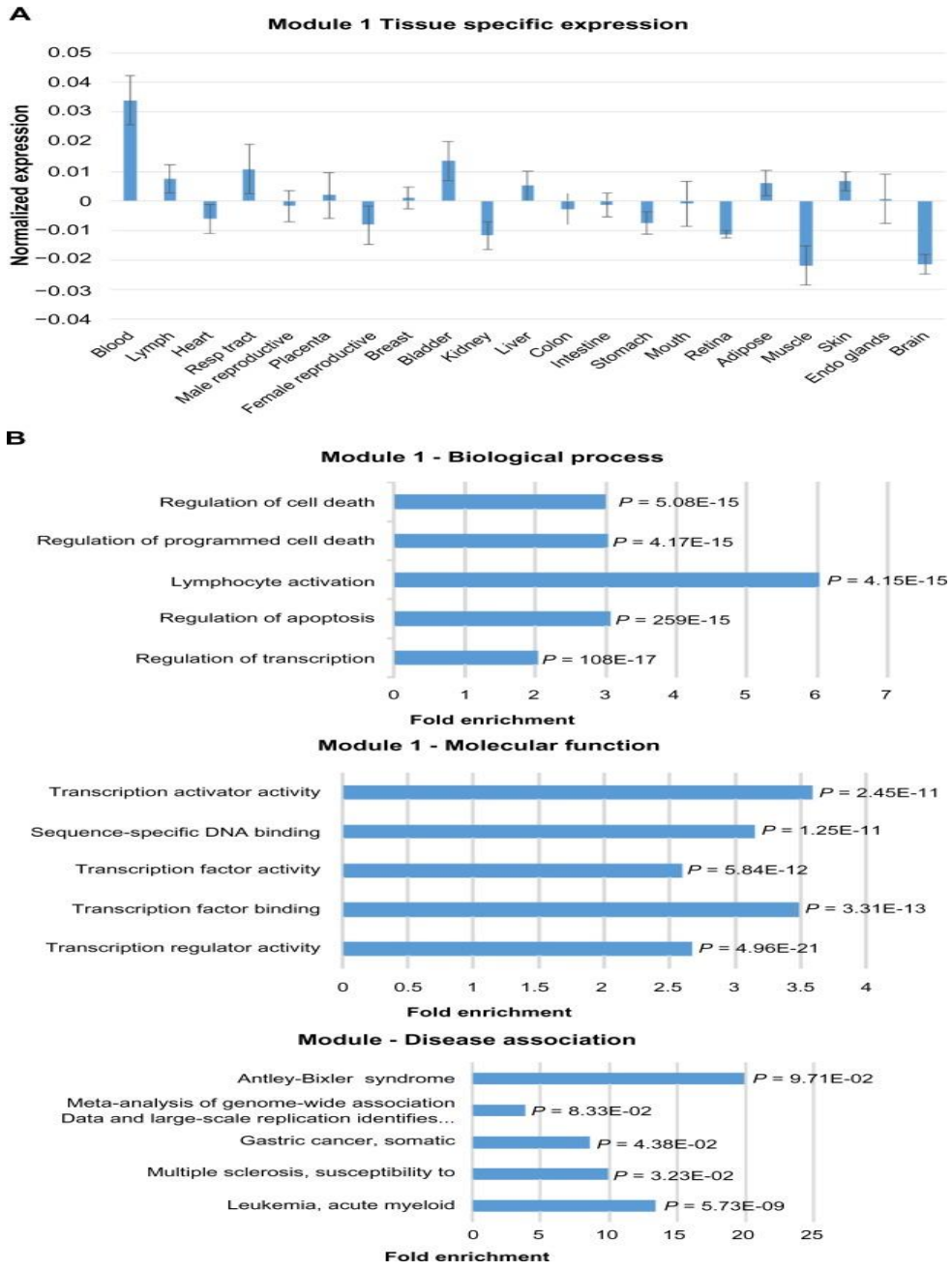


Figure 2.2 Expression and functional term enrichment of the largest Module 1 with high level of expression in blood. (A) Chart of the average expression levels of Module 1 genes in broad tissue types. Error bars represent standard deviations. (B) DAVID functional analysis of Module 1 genes. The enriched terms for biological process, molecular function, and OMIM disease association are plotted against fold enrichment with the corresponding *P*-value.

Figure 2.3 Network visualization of the largest Module 1 with high level of expression in blood. VisANT network visualization of the top 100 probes with the highest intramodal connectivity within Module 1. Node size is proportional to intramodal connectivity and edges are based upon TOM values with the minimum threshold set to 0.06.

To determine the biological significance of the module, functional term enrichment using the DAVID web server was performed (Figure 2.2B). While the highest fold enrichment has been found in a component of the innate immune response, there is a significant enrichment for lymphocyte activation for Module 1. Other terms show functional enrichment for processes involved in cell death. The tissue specificity and gene ontology term enrichment reinforce the OMIM disease association with acute myeloid leukemia (AML).

To visualize the co-expression network and identify hub genes, the 100 probes with the highest intramodal connectivity were analyzed using the VisANT software (Figure 2.3). The network visualization shows dense connectivity within the module. The lncRNA LOC100505812 is a hub gene for Module 1, providing further evidence of the

module's role in lymphocyte activation. Two other uncharacterized lncRNAs are present in Module 1, ITGB2-AS1and C17orf44. Module 1 is the largest module with 879 co-expressed probes. Thus, although the other lncRNAs do not represent hub genes in the network of the selected probes, they may have high connectivity degrees and possibly play a central role in the biological function of the module.

Module 4 expression is low in blood and enriched for genes associated with intracellular signaling pathways

In contrast to Module 1, Module 4 shows significantly lower expression in blood samples than the other tissue types (Figure 2.4A). Module 4 genes do not show obvious tissue-specific expression patterns. The lncRNAs show relatively low expression across tissues when compared to the cancer genes (Figure B-1B). Functional terms for Module 4 are enriched for intracellular signaling pathways involved in cell proliferation at the process level and phosphatase and kinase activity at the molecular level (Figure 2.4B). Interestingly, Module 4 shows an OMIM disease association for AML similar to Module 1. Module 4 also has less disparity between the proportion of lncRNAs to cancer genes and a larger number of higher intramodal connectivity for lncRNAs than Module 1 (Figure 2.5C). The network visualization reveals a tendency of the lncRNAs to not have connections with each other but many connections with the cancer genes. For the nodes with the highest connectivity in Module 4, only 1.3% of the potential lncRNA–lncRNA connections were above the TOM connection threshold of 0.06, and of the potential connections between lncRNAs and cancer genes, 13.5% were above the TOM connection threshold. The uncharacterized lncRNA, LOC100130776, is identified as a potential hub

gene within Module 4, and the lncRNA, AC009133.2 (GenBank accession) is of interest as well due to its high connectivity within the module.

Module 5 exhibits high expression in brain tissues, OMIM disease association with neuroblastoma, and functional enrichment for neural development

Module 5 genes show significantly higher levels of expression in the brain and retina tissues than the other tissue types (Figure 2.4A). Moreover, the expression level in the brain is higher than in the retina. The heat map of Module 5 expression shows that cancer genes generally have higher expression than the lncRNAs in the brain samples (Figure B-1C). Within the brain tissue group, dorsolateral prefrontal cortex has the highest mean expression level.

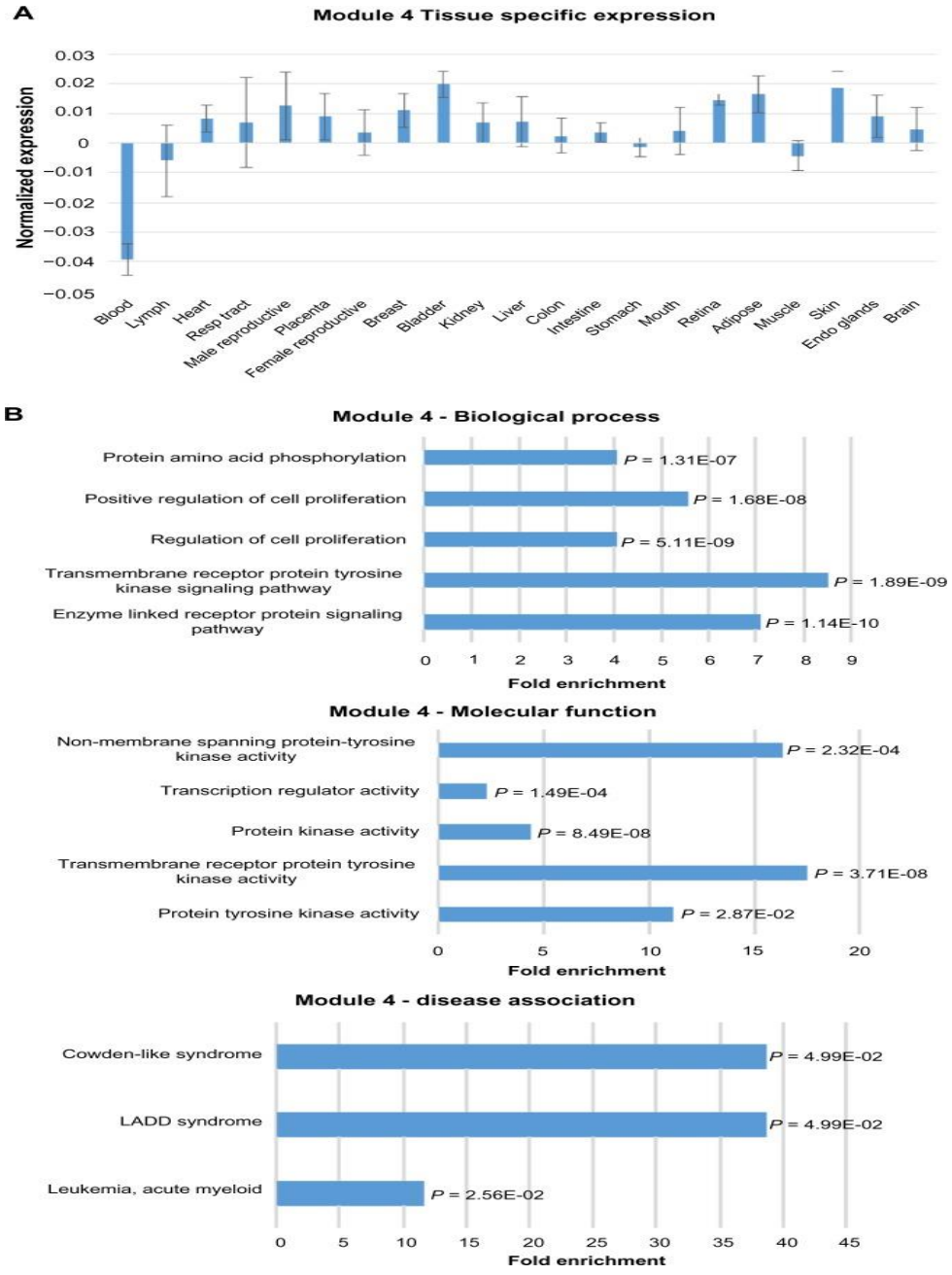


Figure 2.4 Expression and functional term enrichment of Module 4 genes with low level of expression in blood. (A) The average expression levels of Module 4 genes in broad tissue types with standard deviation bars. (B) The functional enrichment of biological process, molecular function, and OMIM disease association terms for Module 4 plotted against fold enrichment with the corresponding *P*-value.

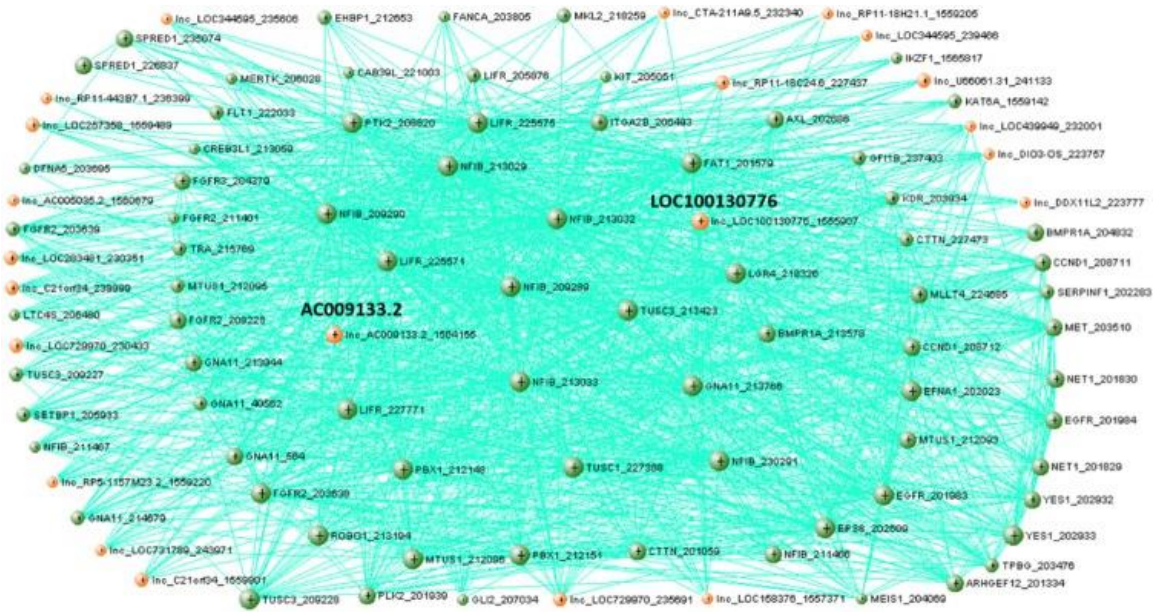


Figure 2.5 Network visualization of Module 4 genes with low level of expression in blood. VisANT network visualization of the top 99 probes with the highest intramodal connectivity within Module 4 (RNASEH2B excluded from visualization due to no connectivity above the edge threshold). The minimum TOM value threshold for edges is set to 0.06.

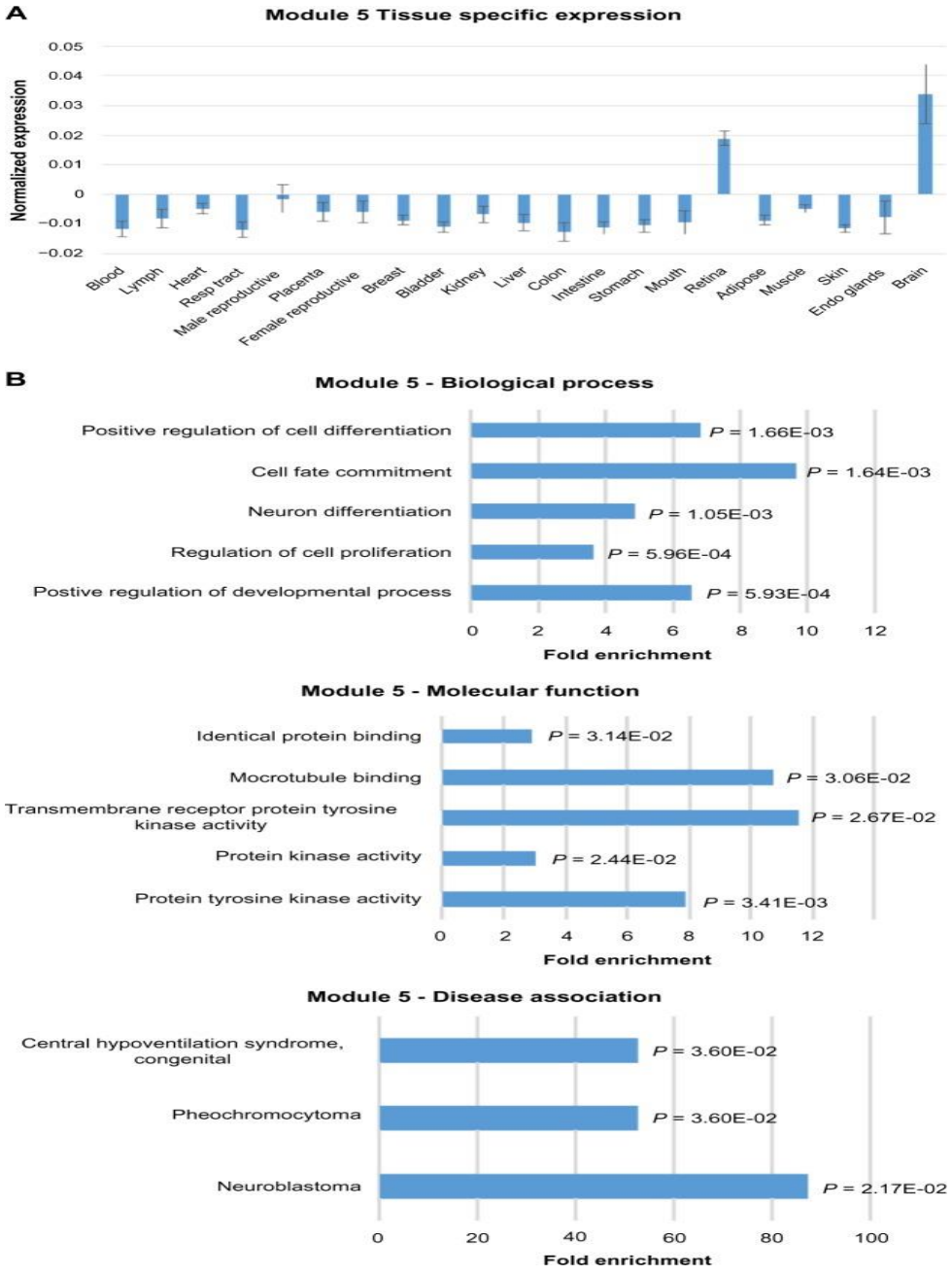


Figure 2.6 Expression and functional term enrichment of Module 5 genes with high proportion of lncRNAs and high level of expression in brain tissues. (A) Average expression levels of Module 5 genes in broad tissue types with standard deviation bars. (B) Functional enrichment analysis of Module 5 genes. The enriched terms for biological process, molecular function, and OMIM disease association are plotted against fold enrichment with the corresponding *P*-value.

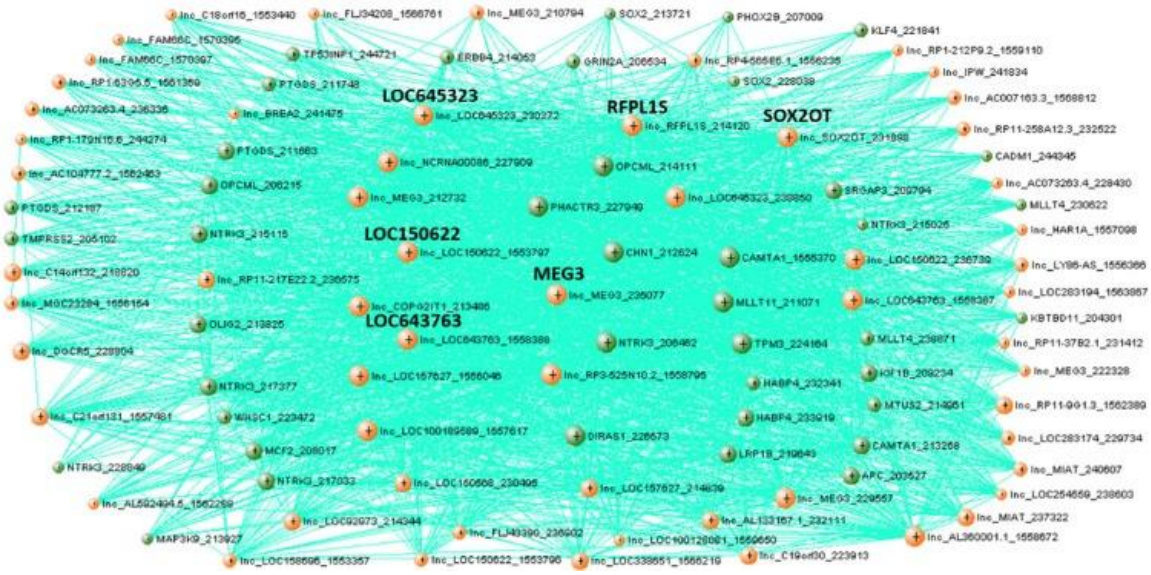


Figure 2.7 Network visualization of Module 5 genes with high proportion of lncRNAs and high level of expression in brain tissues. VisANT network visualization of the top 100 probes with the highest intramodal connectivity within Module 5. The minimum TOM value threshold for edges is set to 0.06.

Functional term enrichment indicates that Module 5 genes may play a role in neural development (Figure 2.4B). Module 5 has the significant enrichment of the biological process terms, positive regulation of developmental process and neuron differentiation. Not surprisingly given its brain-specific expression pattern, Module 5 is the only module to show an OMIM disease association with neuroblastoma. Although Module 5 does not show any significant molecular function term enrichment for transcriptional regulation as the previous two modules analyzed, it is similar to Module 4 in its term enrichment for intracellular signaling functions. Module 5 is also enriched for microtubule binding.

Network visualization of Module 5 shows less connectivity than Modules 1 and 4 (Figure 2.4C). The pattern of high numbers of connections between lncRNAs and cancer genes is also observed in this module. However, there is not a greater propensity of connections between lncRNAs and cancer genes as was observed in Module 4. For the nodes with the highest connectivity in Module 5, 59.3% of the potential lncRNA–lncRNA connections were above the TOM connection threshold of 0.06, and of the potential connections between lncRNAs and cancer genes, 57.1% were above the TOM connection threshold. Six lncRNAs are identified as hub genes within Module 5. Four of the hub genes, LOC645323, LOC643763, LOC150622, and RFPLS are uncharacterized, whereas the other two hub genes, MEG3 and SOX2OT, have been studied. SOX2OT has been shown to be expressed specifically in the brain and linked to neurogenesis in mice (Amaral *et al*., 2009). MEG3 is implicated in a variety of cancers, and MEG3 knockouts cause developmental disorders in mice (Benetatos *et al*., 2011).

2.4 Disucussion

In this study, we have demonstrated high degrees of co-expression between certain lncRNAs and cancer genes in noncancerous tissues. We have cataloged the lncRNAs that are highly co-expressed with the cancer genes in the core list. This catalog can serve as a prioritizing resource for research focused on the causal cancer genes and their potential interactions with lncRNAs. We have highlighted the biological significance of these interactions through the analysis of the highest correlations between lncRNAs and cancer genes. Interestingly, cancer genes that have high correlation with the

same lncRNA also tend to share a common disease association. The co-expression analysis has also provided new insights into the association of lncRNAs and cancer genes. The mainly positive correlations in expression between lncRNAs and cancer genes imply function beyond transcriptional inhibition. Tissue-specific cancer genes, especially those expressed in blood or brain tissues, tend to have higher degrees of co-expression with lncRNAs. The cancer gene with the highest lncRNA co-expression, CHN1, is predominantly expressed in the brain, consistent with the relatively high level of expression for lncRNAs in this tissue type (Wu *et al*., 2013). In addition, our results suggest a potential role of lncRNAs in the immune response. LOC100505812 is located on chromosome 19 adjacent to the caspase recruitment domain family member 8 (CARD8) gene, which is involved in the inflammation response. It is possible that LOC100505812 may have a functional role in the immune response as well as the leukemia and lymphoma disease states. However, the expression of LOC100505812 may also be the result of transcriptional noise due to its close proximity to the CARD8 gene since lncRNAs and protein-coding genes are equally likely to be transcribed with adjacent genes (Djebali *et al*., 2012). Both LOC100505812 and ITGB2-AS1 present interesting possibilities as leukemia or lymphoma biomarkers.

We have also performed gene co-expression network analysis to identify modules containing both lncRNAs and cancer genes. The expression patterns of the modules and their enrichment for biological process, molecular function, and disease association terms have provided the initial characterization for the previously uncharacterized lncRNAs. We have identified candidate lncRNAs that are hub genes within the biologically

significant modules and thus warrant further studies. For instance, LOC100505812 is a hub gene in Module 1, which shows functional term enrichment for transcriptional regulation and disease association for AML. Moreover, the analysis of three co-expression modules has provided new insights into the potential roles of lncRNAs in cancer. While Modules 1 and 4 share AML disease association, there is a stark contrast in the expression patterns between the two modules. Recent studies have suggested the involvement of several lncRNAs such as HOTAIRM1 and RUNX1 in AML, but still little is known about their roles in the disease (Zhang *et al*., 2014; Wang *et al*., 2014). Given the elevated expression pattern, the lncRNAs within Module 1 may have more functional potential related to the disease when compared with Module 4. However, overexpressed lncRNAs have previously served as diagnostic biomarkers. Thus, Module 4 with low level of expression in normal blood cells may provide some interesting diagnostic lncRNAs for cancer. Module 5 is of particular interest due to its brain-specific expression pattern, greater proportion of lncRNAs than cancer genes, and disease association with neuroblastoma. While lncRNAs have been shown to be involved in neural development, little is known about their role in neuroblastoma (Wu *et al*., 2013). Further characterization of the lncRNAs within Module 5 could provide insights into this disease.

We have shown the utility of our integrated microarray expression dataset for functional annotation of lncRNAs associated with cancer genes. The dataset contains 2,968 high-quality expression profiles of various normal tissue samples, which have been selected, after manual curation, from the vast amount of microarray data in public

databases (Wang *et al*., 2010). We have used this high-quality dataset for the co-expression analysis of cancer genes and lncRNAs. Since highly co-expressed genes are often involved in similar biological processes, the findings provide useful information for lncRNA annotation as well as cancer research. Our approach is different from the differential expression analysis of cancerous and normal samples, which is commonly used to identify disease-associated lncRNAs. Since cancer is a highly heterogeneous disease and lncRNAs are normally expressed at low levels, the analysis of gene co-expression in a wide range of normal tissue types may allow for the broader identification of cancer-associated lncRNAs and functional characterization. This approach can also be used to determine lncRNA associations with other disease states. Nevertheless, one limitation in this study is that only 1,865 lncRNAs are represented in the microarray platform (Affymetrix HG-U133 Plus 2.0 Array). This limitation can be overcome by utilizing RNA-seq data. With the rapid accumulation of RNA-seq data in public databases, a high-quality expression dataset containing all lncRNAs will be compiled and used for the gene co-expression network analysis in the future.

References

Amaral, P. et al. (2009) Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. *RNA*, 15, 2013–2027.

Amberger J. et al. (2009) McKusick's online mendelian inheritance in man (OMIM) *Nucleic Acids Res*, 37, D793–D796.

Benetatos, L. et al. (2011) MEG3 imprinted gene contribution in tumorigenesis. *Int J Cancer*, 129, 773–779.

Bochenek, G. et al. (2013) The large non-coding RNA ANRIL, which is associated with atherosclerosis, periodontitis and several forms of cancer, regulates ADIPOR1, VAMP3 and C11ORF10. *Hum Mol Genet*, 22, 4516–4527.

Bussemakers, M. et al. (1999) DD3: A new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res*, 59, 5975–5979.

Cabili, M. et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915–1927.

Cech, T. and Steitz, J. (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157, 77–94.

Cheetham S. et al (2013) Long noncoding RNAs and the genetics of cancer. *Br J Cancer*, 108, 2419–2425.

Chen, G. et al. (2013) LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*, 41, D983–D986.

Derrien, T. et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775–1789.

Djebali, S. et al. (2012) Landscape of transcription in human cells. *Nature*, 489, 101–108.

Dong, R., et al. (2014) Genome-wide analysis of long noncoding RNA (lncRNA) expression in hepatoblastoma tissues. *PLoS One*, 9, e85599.

Futreal, P. et al. (2004) A census of human cancer genes. *Nat Rev Cancer*, 4, 177–183.

Guo, F. et al. (2010) Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in CaSki human cervical cancer cells suppresses cell proliferation and invasion. *Acta Biochim Biophys Sin (Shanghai)*, 42, 224–229.

Hessels, D. et al. (2003) DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur Urol*, 44, 8–16.

Hu, Z. et al. (2013) VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic Acids Res*, 41, W225–W231.

Huang da, W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44–57.

Huang, L. et al. (2014) Overexpression of long noncoding RNA HOTAIR predicts a poor prognosis in patients with cervical cancer. *Arch Gynecol Obstet*, 290, 717-723.

Iyengar, B. et al. (2014) Non-coding RNA interact to regulate neuronal development and function. *Front Cell Neurosci*, 8, 47.

Ji, P. et al. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 22, 8031–8041.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559.

Li, C. et al. (2013) Differential Tks5 isoform expression contributes to metastatic invasion of lung adenocarcinoma. *Genes Dev*, 27, 1557–1567.

Liu, Q. et al. (2016) LncRNA RSU1P2 contributes to tumorigenesis by acting as a ceRNA against let-7a in cervical cancer cells. *Oncotarget*, [Epub ahead of print].

Luo, J. et al. (2006) Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. *Hepatology*, 44, 1012–1024.

Maeda, N. et al. (2006) Transcript annotation in FANTOM3: Mouse gene catalog based on physical cDNAs. *PLoS Genet*, 2, e62

Maglott, D. et al. (2011) Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res*, 39, D52–D57.

Negishi, M. et al. (2014) A new lncRNA, APTR, associates with and represses the CDKN1 A/p21 promoter by recruiting polycomb proteins. *PLoS One*, 9, e95216.

Pandey, R. et al. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell*, 32, 232–246.

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available at http://www.R-project.org/.

Rinn, J. et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129, 1311–1323.

Tsai, M. et al. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329, 689–693.

UniProt Consortium (2014) Activities at the universal protein resource (UniProt). *Nucleic Acids Res*, 42, D191–D198.

Wang, H. et al. (2014) An intragenic long noncoding RNA interacts epigenetically with the RUNX1 promoter and enhancer chromatin DNA in hematopoietic malignancies. *Int J Cancer*, 135, 2783-2794

Wang, L. et al. (2010) Microarray data integration for genome-wide analysis of human tissue-selective gene expression. *BMC Genomics*, 11, S15.

Wang, S. and Tran, E. (2013) Unexpected functions of lncRNAs in gene regulation. *Commun Integr Biol*, 6, e27610.

Wu, P. et al. (2013) Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res Bull*, 97, 69–80.

Yang, X. et al. (2014) A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One*, 9, e87797.

Zhang, X. et al. (2012) Long non-coding RNA expression profiles predict clinical phenotypes in glioma. *Neurobiol Dis*, 48, 1–8.

Zhang, X. et al. (2014) Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. *RNA Biol*, 11, 777-787.

Zhu, S. et al. (2014) Differential expression profile of long non-coding RNAs during differentiation of cardiomyocytes. *Int J Med Sci*, 11, 500–507.

CHAPTER III - CO-EXPRESSION OF LONG NON-CODING RNA AND AUTISM RISK GENES IN THE DEVELOPING HUMAN BRAIN

Steven B. Cogill, Anand K. Srivastava, and Liangjiang Wang

Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

JC Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA

To be submitted: *PLoS One*

Support: This work was supported by a grant from the Self Regional Healthcare Foundation.

Abstract

Long non-coding RNAs (lncRNAs) have been implicated in autism spectrum disorder (ASD) and identified as potentially key regulators of neural development. Yet, many lncRNAs remain uncharacterized for their association with either. In this study, we performed co-expression network analysis on the developing brain transcriptome to identify potential lncRNAs associated with autism spectrum disorder and possible annotations for lncRNAs' functional role in brain development. We found co-enrichment of lncRNA genes and ASD risk genes in two distinct groups of modules showing elevated prenatal and postnatal expression patterns respectively. Further enrichment analysis of the module groups indicated that the early expression modules were

comprised mainly of transcriptional regulators while the later expression modules were associated with synapse formation. Finally, lncRNA genes were prioritized for their connectivity with the known ASD risk genes through analysis of an adjacency matrix. Collectively, the results imply early developmental repression of synaptic genes through lncRNAs and ASD transcriptional regulators.

3.1 Introduction

Long non-coding RNAs (lncRNAs) are defined as transcripts greater than 200 nucleotides in length, which do not code for protein. They serve a wide range of functions including, but not limited to, scaffolding for protein complexes, transcriptional regulation, and translational regulation (Shi *et al*., 2013; Tsai *et al*., 2010; Wang and Tran, 2013). Currently, the GENCODE consortium lists 15,941 lncRNA genes (Harrow *et al*., 2012). LncRNAs are potentially key regulators of brain development. Expression of lncRNAs has been shown to have increased temporospatial specificity in comparison to protein-coding genes, and lncRNAs are expressed in the brain at relatively high levels (Derrien *et al*., 2012; Cogill and Wang, 2014). Nescula *et al.* (2014) found that lncRNA genes of earlier evolutionary origin have been shown to contain homeobox transcription factor binding sites in their promoter regions at a frequency greater than two times that of protein coding genes. This indicates the potential role of lncRNAs in development. This group also found that younger lncRNAs, in terms of phylogenic split from a common ancestor, show lower interspecies conservation and a number of lncRNA families unique to primates offer potential insight into higher cognitive functions.

Autism spectrum disorder (ASD) is a heterogeneous group of neurodevelopmental disorders with a complex genetic etiology. The diagnosis is determined by significant deficit in reciprocal social interactions, impaired communication, and restricted, repetitive behaviors, and most documented cases are clinically diagnosed by the age of three (American Psychiatric Association, 2012). There is strong evidence to support a genetic causation model, including 88% pairwise concordance amongst monozygotic twins and 18.7% risk of ASD for siblings of affected individuals (Rosenberg *et al*., 2009; Ozonoff *et al*., 2011; Liu and Takumi, 2014). As with most complex genetic disorders, ASD could result from the accumulation of low risk common variants, high risk rare variants, or both. Approaches for ASD genetic studies have included copy number variation (CNV) studies, genome-wide association studies (GWAS) and rare de novo variant (RDNV) exome studies (Liu and Takumi, 2014). Ziats and Rennert (2013) found 222 differentially expressed lncRNAs in ASD. ASD risk genes are convergent on synaptic gene translation, transcription and chromatin remodeling (Liu and Takumi, 2014; Parikshak *et al*., 2013). These three processes can be controlled by lncRNAs (Wang and Chang, 2011).

This study used co-expression network analysis to identify lncRNAs potentially associated with ASD and provide possible functional annotations of lncRNAs for brain development. Since anatomical differences between ASD and control brain samples have been shown in several different structures, it is therefore beneficial in this study to examine all of the structures during the developmental period to place lncRNAs in a functional context within the developing brain (Lange *et al*., 2015). The BrainSpan

dataset offers a unique opportunity for identification of high-priority potential ASD associated lncRNA genes due to its comprehensive array of brain structures and developmental time points (Hawrylycz *et al*., 2012). We have compiled a comprehensive list of ASD risk genes from several sources to measure co-expression with lncRNA genes annotated in the GENCODE dataset (Harrow *et al*., 2012). Co-expression network analysis was performed on a curated set of genes from the BrainSpan dataset to cluster the genes into modules. Expression patterns and co-enrichment with lncRNA genes and ASD risk genes were used to identify modules of interest. Enrichment analysis and network topology analysis were carried out to associate biologically significant functions with the modules. Finally, to identify lncRNA genes of interest, lncRNAs were prioritized based upon their association with the known ASD risk genes within the network.

3.2 Methods

Datasets

The BrainSpan data set is a developmental transcriptome for the human brain (Hawrylycz *et al*., 2012). It is a RNAseq dataset in units of reads per kilobase million (RPKM), mapped to genes as annotated by the GENCODE consortium version 10. It consists of 524 samples covering a developmental time span of 8 weeks post conception to 40 years of age and 26 brain structures. Genes which did not show a minimum expression of 1 RPKM for one of the 524 samples and genes not present in the latest build of the GENCODE consortium (version 24) were removed from the dataset (Harrow

et al., 2012). Expression values were then $\log_2(RPKM+1)$ transformed. Next the sum of pairwise covariance was calculated for each gene. Using the KMeans class from the Scikit-learn Python library (Pedregosa *et al*., 2011), clustering with the total clusters set to 2 was performed on the sum of covariance values to filter out low information genes (Tritchler *et al*., 2009; Sakata *et al*., 2015). Then ASD risk genes and lncRNA genes within the dataset were identified (Additional file A-3). ASD risk genes were compiled from three different sources. We selected 290 genes from the Gene Scoring Module from the Simons Foundation Autism Research Initiative (SFRARI) on the criteria of a score of 1-4 with 1 being high confidence and 4 being minimal evidence (Basu *et al*., 2009). An additional 170 genes were from the core set of the Autism Knowledge Base from the Center for Bioinformatics in Peking University (Xu *et al*., 2012). The third source from which we selected 107 genes was from an exome sequencing study for de novo loss-of-function mutations in ASD cases (De Rubeis *et al*., 2014). Redundancy among the three datasets were removed resulting in an ASD risk gene set consisting of 433 genes. Genes of the lncRNA biotypes were indicated in the GENCODE build.

Co-expression network analysis

Genes were clustered into modules using the weighted gene co-expression network analysis (WGCNA) package in R (Langfelder *et al*., 2008). The package first generates a topological overlap matrix using neighbourhood analysis and the weighted pairwise correlation between genes $\left|corr\left(x_i x_j\right)\right|^P$ where P is a soft threshold for network scalability. For this study, we found that we reached a scale free topology with a soft threshold of 7 for this study. Then a dissimilarity dendrogram from the topological

overlap matrix is created, and the genes are grouped using a dynamic tree- cutting algorithm. The network in this study was an unsigned bi-weight network with a minimum module size of 30 and a merge cut-off height of 0.2. The heatmap of the expression patterns for the modules was generated using a modified version of the gplots package in R (Warnes *et al*., 2015). The expression patterns themselves are the eigengene (first principal component) for the respective modules. Enrichment of lncRNA and ASD risk genes within the modules was calculated by applying Fisher's exact test to gene type frequency within the module compared to gene type frequency for the entire dataset. The *P*-value was adjusted to a false discovery rate (FDR) to account for multiple testing using the p.adjust function in the stats package in R (R Core Team, 2014). For better visualization of enrichment, significance values were $-\log_{10}(FDR)$ transformed.

Enrichment analysis

Functional term enrichment for each module was implemented through the use of the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang da *et al*., 2009). This software receives a gene list and applies the EASE algorithm, which is a variation of the Fisher's exact test, using gene annotations present in the database and a designated background. In this study, enrichment was measured against a human genome background, and genes which could not be mapped were not considered in the enrichment calculations. The FDR values generated from DAVID were transformed as mentioned previously. Enrichment for significantly expressing genes within brain structures was calculated using the same methodology used to determine ASD and lncRNA gene enrichment in the previous section. Frequencies were grouped by

developmental time periods and gene types. The three developmental time periods used were prenatal (8pcw-37pcw), childhood (4mos-15yrs), and adulthood (18yrs-40yrs) (pcw=post conception weeks; mos=months; yrs=years). The gene types were lncRNA genes, ASD risk genes, and all genes within the module. The values were determined by the number of genes with expression greater than or equal to 1 RPKM for a given sample divided by the total possible number of genes compared to the appropriate background.

Network visualization and analysis

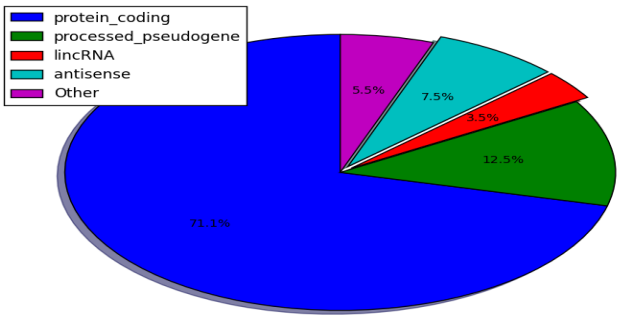
To construct a network visualization, we first sought to determine significant interactions between genes. We therefore constructed an adjacency matrix for the entire dataset using the absolute Pearson product moment correlation to a power of 7 as a measure of connectivity between genes. We then selected the top 5% of the correlations which became the edges in our network while the genes became the nodes. The network was then sub-divided based upon module assignments to determine changes in topology specifically in regard to lncRNA gene and ASD risk gene interactions. Representative modules were visualized using the Cytoscape software (Shannon *et al*., 2003). To prioritize lncRNA genes within our dataset for ASD association, we adapted a methodology used by Oliver et al. (2014), which used connectivity as a means of prioritization. Here we sum the pairwise connectivity from the adjacency matrix between the target lncRNA gene and all the known ASD risk genes in the dataset. The connectivity score is then normalized using $\frac{(x_i - \min(x))}{(\max(x) - \min(x))}$ for the range of all lncRNA genes analysed.

3.3 Results

Co-expression network analysis within the developing brain shows high co-enrichment of lncRNA genes and ASD risk genes in elevated pre- and postnatal expression modules

The BrainSpan dataset offers an opportunity to analyse in depth the gene expression patterns of the developing human brain (Hawrylycz et al., 2012). However, the dataset required further curation for efficient co-expression analysis to prevent noise from low expression or low variance genes. We first removed all genes which did not show sufficient expression (< 1 RPKM), and then selected for genes which showed high pairwise covariance with other genes within the dataset. The 20,456 genes which remained after the curation presented an interesting distribution of gene biotypes (Figure 3.1A). While protein-coding genes only account for 40.4% of the genes within the dataset, after curation they account for 71.1% while antisense lncRNA genes and long intervening RNA (lincRNA) genes originally comprised 19.1% of the dataset were reduced down to 11%.

A



B

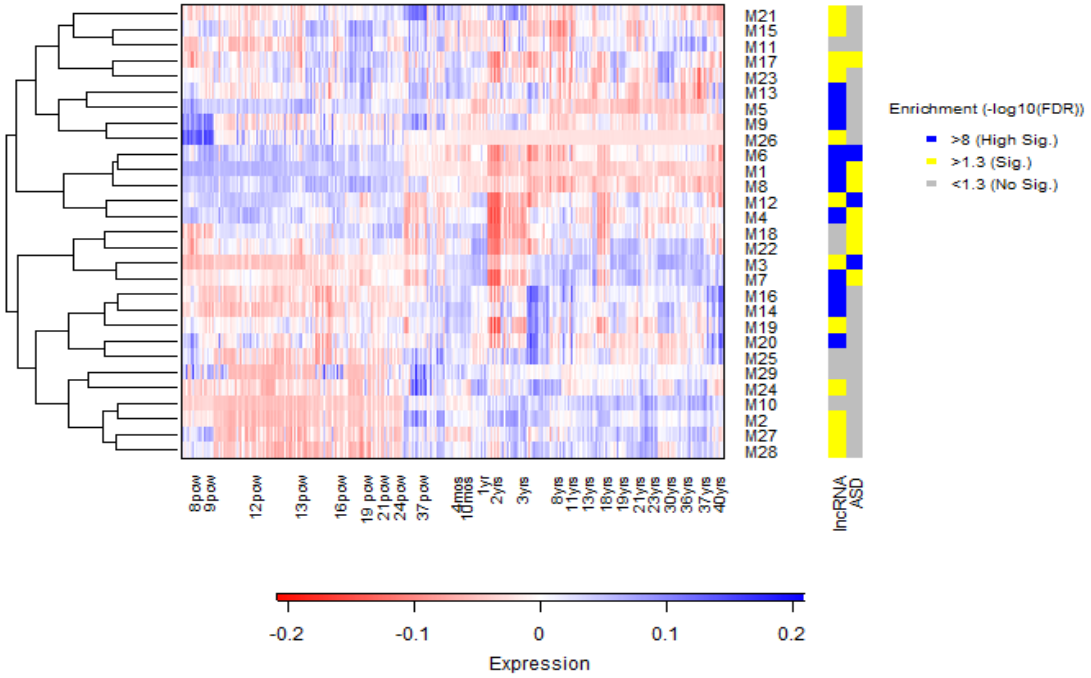


Figure 3.1 Co-expression analysis of BrainSpan dataset. (A) Pie chart of the distribution of genes by biotype for the curated gene set. (B) Heat map of module eigengenes from co-expression analysis for all samples chronologically. The row labels correspond to the module (M1=Module 1) and the column labels indicate the time point ranges (pcw=post conception weeks, mos=months, and yrs=years). To the right of the heatmap is a color sidebar mapping the enrichment of the modules for lncRNA genes and ASD risk genes respectively. The legend indicates the level of significance with the threshold set at 1.3 or FDR=0.05 (sig=significance).

After dataset curation, weighted gene co-expression network analysis (WGCNA) was performed (Langfelder *et al*., 2008). We found 29 modules. We mapped the expression pattern of the eigengene (first principal component) for each module onto a hierarchically clustered heatmap to better visualize shared expression patterns. Co-enrichment for lncRNA genes and ASD risk genes were also mapped to the modules (Figure 3.1B). We found two distinct clades from the module clustering, which show co-enrichment for lncRNA genes and ASD risk genes. One clade comprised of modules 1, 4, 6, 8, 12, which accounts for 9,355 genes within the dataset, shows elevated expression in prenatal samples and lower expression in postnatal samples. These sets are further referred to in this paper as early expression modules. Intriguingly, the other clade comprised of modules 3 and 7 shows an inverse pattern in that prenatal expression is low and postnatal expression is elevated. These sets are referred to as late expression modules. Only 5 of the 29 modules did not show significant enrichment (FDR<0.05) for lncRNA genes while 10 of the modules showed significant enrichment for ASD risk genes with module 6 being alone in showing high enrichment for both gene types.

Enrichment analysis of two module groups shows term enrichment for transcriptional regulation and synapse formation respectively and complementary structure enrichment for sensory cortical regions

To further characterize our module groups of interest, we performed term enrichment analysis. The Database for Annotation, Visualization and Integrated Discovery (DAVID) term enrichment analysis assigns Gene Ontology (GO) terms based upon their enrichment within the gene set (Huang da *et al*., 2009). It should be noted that

the gene sets are comprised of all of the genes within a module and not limited to ASD and lncRNA genes. While there are several categories for terms, we chose biological process, molecular function, and cellular component functional annotation terms to characterize our module groups. These categories offer the most relevant information for lncRNAs whereas the other categories are more relevant to protein coding genes or partially redundant to the given categories. For each module in either the early expression or late expression group, the most significant terms for each category are shown in Figure 3.2A and Figure 3.2B. The early expression modules (M1, M4, M6, M8, and M12) show overlap in biological process for the broad terms of transcription and modification-dependent macromolecule catabolic process, which corresponds to the breakdown of large macromolecules. There is also overlap in localization to the nuclear lumen, and the molecular function of DNA-binding as well as general nucleotide binding. Collectively this implies that the early expression modules are enriched for transcriptional regulators as well as partially involved in the breakdown of nucleotides. However, the late expression modules (M3 and M7) are enriched for a different aspect of brain development. While module 7 has enrichment for relatively ambiguous terms associated with protein transport, module 3 shows enrichment for genes involved in synaptic transmission and localized to the synapse.

Grouping together the samples based on structure and developmental period (prenatal, childhood, and adulthood), we analysed the enrichment of structures for expressed genes collectively, lncRNA genes, and ASD risk genes for the two module groups (Figure 3.3A and Figure 3.3B). Some structures had samples for the prenatal

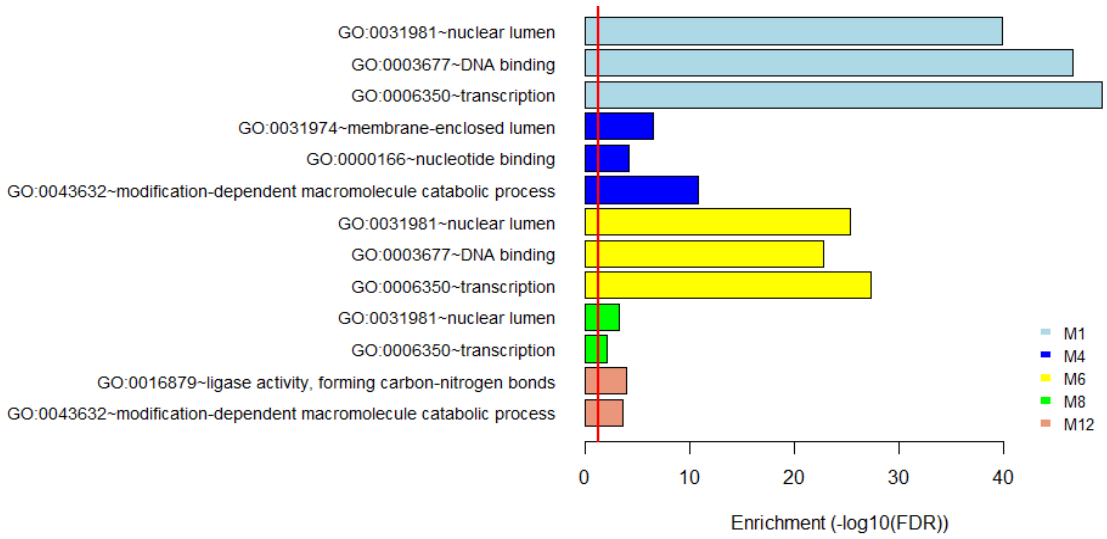
period but did not have samples for childhood and adulthood. Therefore for the early expression modules, we analysed all the structures for just the prenatal period as the later developmental periods (childhood and adulthood) showed little to no enrichment for structure-specific expression. For the late expression modules we analysed only structures present in all three developmental periods.

Enrichment of expressed genes in the early expression modules was significant for all of the structures. Interestingly, expressed lncRNA genes and ASD risk genes show similar patterns of enrichment in the different brain structures for the early expression modules. Not surprisingly, expressed lncRNA genes and ASD risk genes are highly enriched for the sensory cortical regions, striatum, and amygdaloid complex. These three structures have been implicated in ASD (Di Martino *et al*., 20011; Zalla and Sperduti, 2013; Marco *et al*., 2011).

Enrichment of expressed genes in the brain structures for the late expression modules shows distinct patterns based on gene type. With the exception of the mediodorsal nucleus of thalmus, all of the structures were significantly enriched (FDR<0.05) for expressed genes during the prenatal period. Thirteen of the structures were significantly enriched for expressed genes during childhood, and thirteen structures were significantly enriched for expressed genes during adulthood. The hippocampus became significantly enriched during the transition from childhood to adulthood, and the cerebellar cortex lost significant enrichment in the same transition period. Enrichment values for expressed lncRNA genes and ASD risk genes within structures do not show the same similarities as was observed for the early expression modules. Expressed

lncRNA genes are significantly enriched in the prenatal period for every structure, only significantly enriched for the cerebellar cortex in the childhood developmental period, and significantly enriched for six structures in the adult developmental period. Expressed ASD risk genes show no enrichment for any structure in the prenatal period, significant enrichment in eight structures during the childhood developmental period, and significant enrichment in eight structures during the adult developmental period. Intriguingly, there is no significant enrichment for expressed ASD risk genes in the striatum, which has been implicated in the physiopathology of ASD (Di Martino *et al*., 2011).

A



B

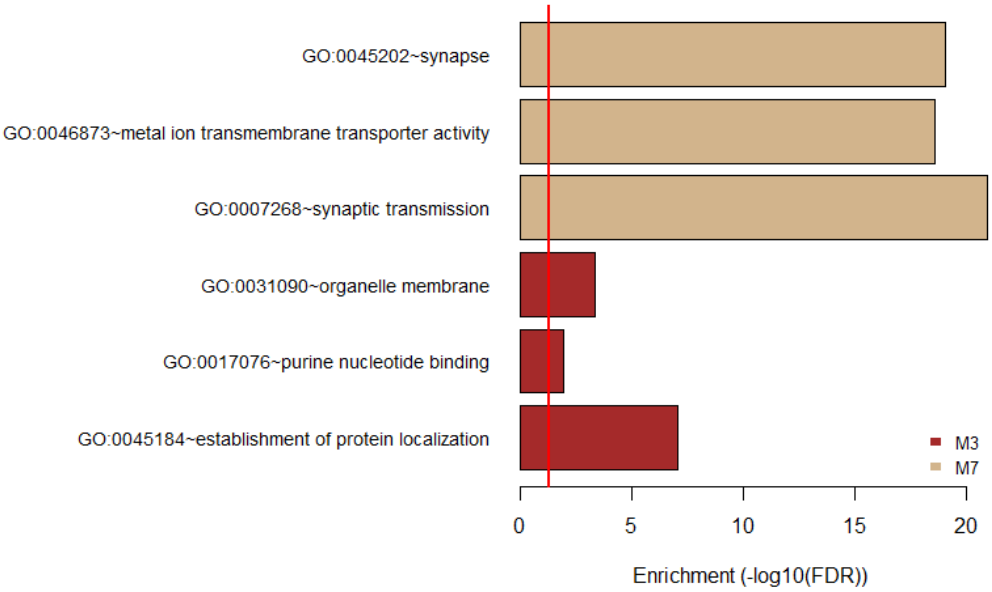


Figure 3.2 Term enrichment analysis of early and late expression module groups. (A) Term enrichment color coded by module for early expression modules. Term categories are from the top to bottom: cellular component, molecular function, and biological function for each module in the group. The red vertical line indicates the significance cutoff (FDR=0.05). Modules 8 and 12 did not have significant terms for molecular function and cellular compartment respectively. (B) Term enrichment for late expression modules.

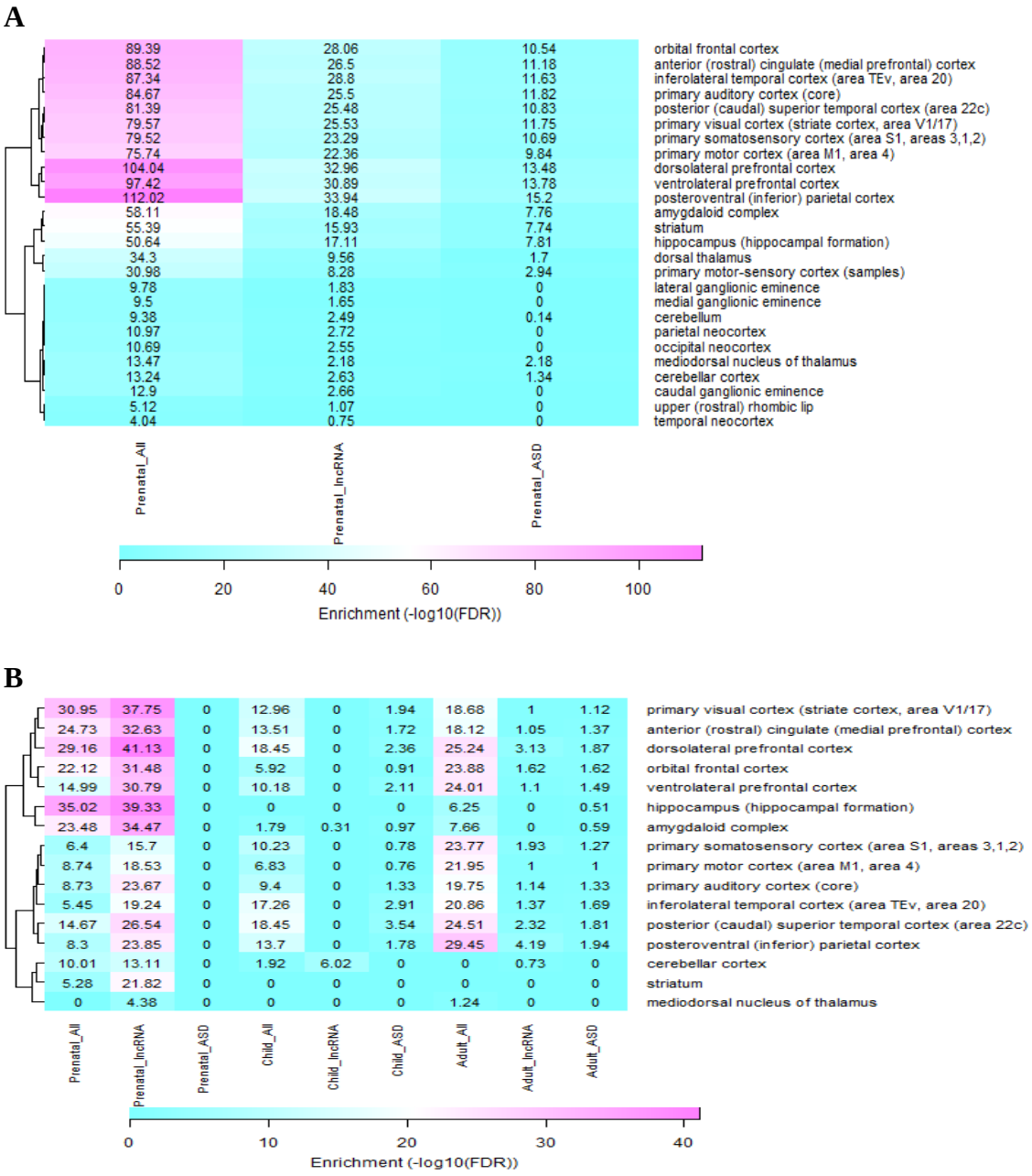


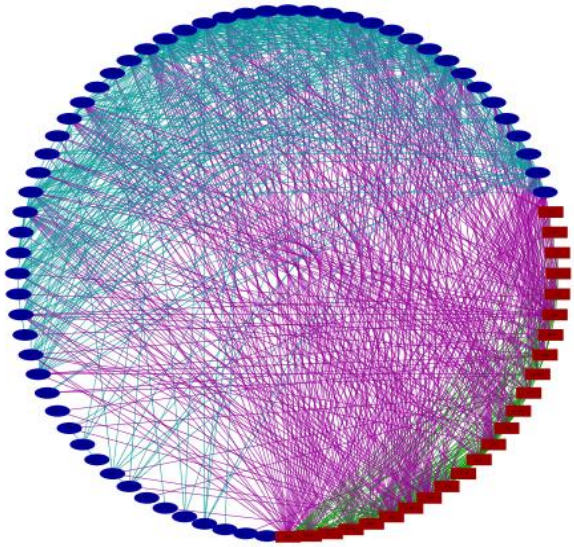
Figure 3.3 Gene expression enrichment analysis of early and late expression module groups. (A) Heatmap for structure specific enrichment for early expression modules. Row labels indicate the brain structure and the columns indicate the developmental period and the gene type with all corresponding to all of the genes within the module. The enrichment values are shown in each cell, and rows are clustered based on their enrichment values. (B) Heatmap for structure specific enrichment for late expression modules.

Visualization of the topology of module networks demonstrates high connectivity between lncRNA genes and ASD risk genes for early expression modules

While term and structure enrichment can give general information on the biological roles associated with modules and provide potential annotation for uncharacterized lncRNA genes, it does not indicate the interactivity between ASD risk genes and lncRNA genes. The enrichment analysis did demonstrate the possibility that lncRNA and ASD risk genes may be more closely associated in the early expression modules than in the later ones. This is confirmed by network analysis. To form the network we used an adjacency matrix to establish pairwise correlation for all of the genes. We then selected for the most significant (highly correlated) interactions for the network. For each of the modules of interest, we observed the significant connections between lncRNA genes and ASD risk genes and found that for the early expression modules there was greater connectivity between the two gene types. Figure 3.4A and Figure 3.4B show the networks for a representative early expression module (M12) and late expression module (M7) respectively. One module from each group was chosen to demonstrate the contrast in topologies between them. Module 12 shows dense connectivity for all of the nodes but does have a greater number of interactions between ASD risk genes and lncRNA genes than between lncRNA genes and ASD risk genes respectively. Notably there is a high degree of interaction between lncRNA genes. However, module 7 shows a less dense network even though the number of genes present is comparable to that of module 12. Interactions between ASD risk genes and lncRNA

genes are greater than other interactions in the network with few interactions between lncRNA genes.

A



B

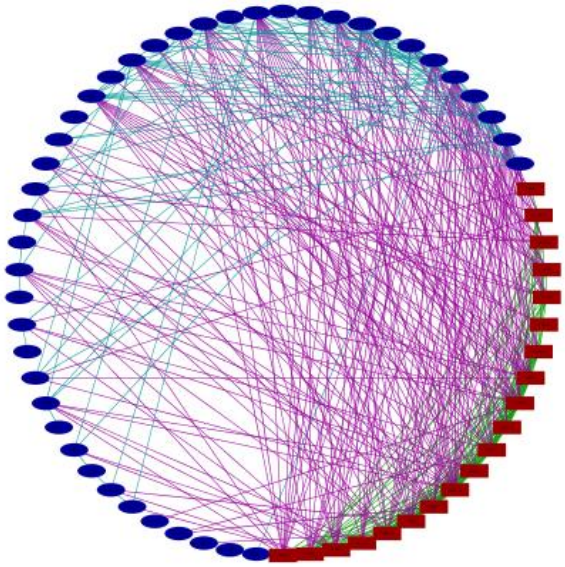


Figure 3.4 Network topology for modules of interest (A) Network topology for lncRNA genes and ASD risk genes for module 12. ASD genes are red rectangles and lncRNA genes are blue ellipses. Interactions are color coded as follows: ASD to ASD=Green, ASD to lncRNA=Purple, lncRNA to lncRNA=turquoise. Modules are in an attribute based circular layout with the attribute being gene type. (B) Network topology for lncRNA genes and ASD risk genes for module 7.

Prioritization of lncRNA genes using connectivity with known ASD risk genes implicates biologically relevant targets

To identify high-priority targets for further study, we prioritized the lncRNA genes in our dataset based on their connectivity with known ASD risk genes. For each lncRNA gene, the pairwise correlation from our adjacency matrix was summed for all ASD risk genes. Genes were ranked relative to their sums of connectivity with higher values associated with greater potential association with ASD. The complete list of lncRNAs with their module assignment and normalized values for ASD gene connectivity and adjusted intramodular connectivity are provide in Additional file A-3. Adjusted intramodular connectivity is the sum of the pairwise connectivity of a gene for all other genes within the module with the sum of pairwise connectivity for the gene for all genes not in the module subtracted from it. The normalized value is based upon the range for all the genes in the dataset and calculated using the same method as for the normalized value for ASD gene connectivity. Table 3.1 shows highly prioritized lncRNAs which have been previously characterized and demonstrate tentative links to ASD. The gene prioritized the highest is RP11-281C10.5, an antisense lncRNA to CEP170, which is a component of the centromere and critical to cell division (Wellburn and Cheeseman, 2012). KDM4A-AS1 is antisense to KDM4A, a lysine de-methylase, which has been shown to increase copy number gains in CNVs associated with ASD (Black *et al*., 2013). LINC-PINT is a lincRNA, which is activated by P53 and like the well-characterized lncRNA, HOTAIR, has been shown to associate with polycomb

recessive complex 2 (Marin-Bejar *et al*., 2013). TUG1 is one of the few highly prioritized genes not grouped to module 1, and it has low intramodular connectivity. It has no direct link to ASD but has been implicated in neurodegenerative disorders (Wu *et al*., 2013). The role of lncRNAs in ASD is still being elucidated, so it is not surprising that many of the genes have no direct link to ASD. The list itself acts as a putative implication of the highly prioritized lncRNAs for their role in ASD.

Table 3.1 List of selected biologically significant and highly prioritized for ASD association lncRNA genes.

Name	Biotype	ASD Connectivity (Normalized)	Module	Intramodular Connectivity (Normalized)
RP11-261C10.5	Antisense	1	1	0.8482
KDM4A-AS1	Antisense	0.9015	1	0.8479
LINC-PINT	Antisense	0.7091	1	0.7510
TUG1	Antisense	0.6992	6	0.3157

3.4 Discussion

With the relatively recent expansion of the Autism Disorder to include Asperger's Syndrome, Rett Syndrome, uncharacterized pervasive developmental disorders, and Autistic Disorder under the common banner of Autism Spectrum Disorders, the complexity of finding its causality increases (American Psychiatric Association, 2012). While there have been significant advances in clinical diagnostic tools, the number of ASD-affected individuals has increased at a rate greater than what is estimated to be due to improved diagnostics with the CDC reporting a 10-fold increase over a 20-year period

(Okamura *et al*., 2004). There are competing theories on the underlying cause of the disorder which are not mutually exclusive (Liu and Takumi, 2014). However, the leap from genetic abnormalities to phenotypic causation has been difficult due to a multitude of factors. Among them include the difficulty of studying the brain physiology of affected individuals and the complexity of genetic interactions associated with the disorder. Within this study we utilized the most comprehensive expression dataset currently available for the developing human brain to further elucidate the complex interactions in an effort to show the role of lncRNAs in brain development and ASD.

LncRNAs have been shown to be in evolutionarily conserved gene families unique to primates and even further to humans alone (Necsulea, 2014), yet their functional roles in brain development warrant further definition. This study indicates a critical role for lncRNAs in transcriptional regulation and synaptic formation in the brain during development. Within this study, we have broadly characterized the role of lncRNAs in brain development and ASD. Clustering our curated gene list, we found that lncRNAs were enriched nearly ubiquitously across our modules but only co-enriched with ASD risk genes in two distinct module groups showing high prenatal and high postnatal expression respectively. This distinction in expression at that particular developmental point is interesting as it has been previously implicated as a critical time for ASD development (Parikshak *et al.* 2013). This data combined with term enrichment suggesting transcriptional regulation and the network topologies showing higher numbers of significant interactions between lncRNA genes and lncRNA genes strongly suggest that lncRNAs within the group of early expression modules regulate brain development

through repression of genes controlling synapse formation possibly in the late expression modules.

ASD is a neurodevelopmental disorder, and in identifying potentially key lncRNA regulators of brain development we have also begun to identify putative high priority targets for potential therapeutics and diagnostics. Due to their tight regulatory control (Derrien *et al*., 2012), lncRNAs are excellent biomarkers. One of the most notable examples was in 1995, when the lncRNA PCA3 was discovered and has since become a diagnostic for pancreatic cancer (Angata *et al*., 2000). It was recently found that 90% of disease associated SNPs from genome-wide association studies were found outside of protein coding regions (Wang and Chang, 2011), which indicates non-coding genes and regulatory regions within the genome could have a major role in disease. These regions may also provide insight into the etiology of complex disorders such as ASD. Our group has previously applied the approach of co-expression network analysis to define high priority disease-associated lncRNA genes based upon normal tissue expression patterns when we published work showing strong associations between cancer genes and lncRNAs (Cogill and Wang, 2014). Our approach allows for disease associations to be implied based solely on expression patterns. It is our hope that this study will highlight lncRNA genes that can act as diagnostic markers to the disorder as well as genes that can further elucidate the etiology of ASD. We also hope that this study further demonstrates the utility of co-expression network analysis on non-disease samples to implicate lncRNAs in disorders.

References

American Psychiatric Association. (2012). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). Washington, DC: Author.

Angata, K. et al. (2000) Differential biosynthesis of polysialic acid on neural cell adhesion molecule (NCAM) and oligosaccharide acceptors by three distinct alpha 2,8-sialyltransferases, ST8Sia IV (PST), ST8Sia II (STX), and ST8Sia III. *J Biol Chem*, 275, 18594-18601.

Basu, S. et al. (2009) AutDB: A gene reference resource for autism research. *Nucleic Acids Res*, 37, D832-D836.

Black, J. et al. (2013) KDM4A lysine demethylase induces site-specific copy gain and rereplication of regions amplified in tumors. *Cell* 154, 541-555.

Cogill, S. and Wang, L. (2014) Co-expression network analysis of human lncRNAs and cancer genes. *Cancer Inform*, 13, 49-59.

De Rubeis, S. et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515, 209-215.

Derrien, T. et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-1789.

Di Martino, A. et al. (2011) Aberrant striatal functional connectivity in children with autism. *Biological Psychiatry*, 69, 847-856.

Harrow, J. et al. (2012) GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*, 22, 1760-1774.

Hawrylycz, M. et al. (2012) An Anatomically Comprehensive Atlas of the Adult Human Brain Transcriptome. Nature, 489, 391-399.

Huang da, W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.

Lange, N. et al. (2015) Longitudinal volumetric brain changes in autism spectrum disorder ages 6-35 years. *Autism Res*, 8, 82-93.

Langfelder, P. et al. (2008) Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics*, 24, 719-720.

Liu, X. and Takumi, T. (2014) Genomic and genetic aspects of autism spectrum disorder. *Biochem Biophys Res Commun*, 452, 244–253.

Marco, E. et al. (2011) Sensory processing in autism: A review of neurophysiologic findings. *Pediatr Res*, 69, 48R-54R.

Marin-Bejar, O. et al. (2013) Pint lincRNA connects the p53 pathway with epigenetic silencing by the polycomb repressive complex 2. *Genome Biol*, 14, R104.

Necsulea, A. et al. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505, 635-640.

Okamura, A. et al. (2004) Involvement of casein kinase iepsilon in cytokine-induced granulocytic differentiation. *Blood*, 103, 2997-3004.

Oliver, K. et al. (2014) Harnessing gene expression networks to prioritize candidate epileptic encephalopathy genes. *PloS One*, 9, e102079.

Ozonoff, S. et al. (2011) Recurrence risk for autism spectrum disorders: A baby siblings research consortium study. *Pediatrics*, 128, e488-e495.

Parikshak, N. et al. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, 155, 1008-1021.

Pedregosa, F. et al. (2011) Scikit-Learn: Machine Learning in Python. *J Mach Learn Res*, 12, 2825-2830.

R Core Team. (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014). URL http://www.R-project.org/.

Rosenberg, R. et al. (2009) Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch Pediatr Adolesc Med*, 163, 907-914.

Sakata, K. et al. (2015) System-wide analysis of the transcriptional network of human myelomonocytic leukemia cells predicts attractor structure and phorbol-ester-induced differentiation and dedifferentiation transitions. *Sci Rep*, 5, 8283.

Shannon, P. et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2498-2504.

Shi, X. et al. (2013) Long non-coding RNAs: A new frontier in the study of human diseases. *Cancer Lett*, 339, 159-166

Tritchler, D. et al. (2009) Filtering genes for cluster and network analysis. *BMC Bioinformatics*, 10, 193.

Tsai, M. et al. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329, 689-93.

Wang, K. and Chang, H. (2011) Molecular mechanisms of long noncoding RNAs. *Mol Cell*, 43, 904-914.

Wang, S. and Tran, E. (2013) Unexpected functions of lncRNAs in gene regulation. *Commun Integr Biol*, 6, e27610.

Warnes, G. et al. (2015) gplots: Various R Programming Tools for Plotting Data. R package version 2.17.0. http://CRAN.R-project.org/package=gplots

Wellburn, J. and Cheeseman, I. (2012) The microtubule-binding protein Cep170 promotes the targeting of the kinesin-13 depolymerase Kif2b to the mitotic spindle. *Mol Biol Cell*, 23, 4756-4795.

Wu, P. et al. (2013) Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res Bull*, 97, 69-80.

Xu, L. et al. (2012) AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res*, 40, D1016-D1022.

Zalla, T. and Sperduti, M. (2013). The amygdala and the relevance detection theory of autism: An evolutionary perspective. *Front Hum Neurosci*, 7, 894.

Ziats, M. and Rennert, O. (2013) Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci*, 49, 589-593.

CHAPTER IV - SUPPORT VECTOR MACHINE MODEL OF DEVELOPMENTAL BRAIN GENE EXPRESSION DATA FOR PRIORITIZATION OF AUTISM RISK GENE CANDIDATES

Steven B. Cogill and Liangjiang Wang

Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

Published: *Bioinformatics* [Epub ahead of print] (2016)

Support: This work was supported by a grant from the Self Regional Healthcare Foundation.

Abstract

Motivation: Autism spectrum disorders (ASD) are a group of neurodevelopmental disorders with clinical heterogeneity and a substantial polygenic component. High-throughput methods for ASD risk gene identification produce numerous candidate genes that are time-consuming and expensive to validate. Prioritization methods can identify high-confidence candidate genes. Previous ASD gene prioritization methods have focused on *a priori* knowledge, which excludes genes with little functional annotation or no protein product such as long non-coding RNAs (lncRNAs).

Results: We have developed a support vector machine (SVM) model, trained using brain developmental gene expression data, for the classification and prioritization of ASD risk genes. The selected feature model had a mean accuracy of 76.7%, mean specificity of 77.2% and mean sensitivity of 74.4%. Gene lists comprised of an ASD risk gene and adjacent genes were ranked using the model's decision function output. The known ASD risk genes were ranked on average in the 77.4^{th}, 78.4^{th} and 80.7^{th} percentile for sets of 101, 201 and 401 genes respectively. Of 10,840 lncRNA genes, 63 were classified as ASD-associated candidates with a confidence greater than 0.95. Genes previously associated with brain development and neurodevelopmental disorders were also prioritized highly within the lncRNA gene list.

4.1 Introduction

Autism spectrum disorder (ASD) is the umbrella term for the neurodevelopmental disorders: autistic disorder, Asperger's syndrome, pervasive developmental disorder not otherwise specified, Rett syndrome and childhood disintegrative disorder. It is generally diagnosed at an age greater than four years old based predominantly on the behavioral phenotype described as delayed communication, difficulty acknowledging social cues, and engaging in repetitive behaviors (American Psychiatric Association, 2012). In 2010, the Centers for Disease Control and Prevention (CDC) estimated the prevalence of ASD at 1 in 68 children aged 8, and this was an increase from the 2007 estimate of 1 in 150 (CDC, 2014). The increase in prevalence may be attributable to more public awareness and implementation of prescreening technology (Chlebowski *et al.*, 2013). Twin and

sibling studies indicate that ASD etiology is influenced heavily by genetics and to a lesser extent environmental factors (Kim and Leventhal, 2015). The actual physiological cause of ASD is currently unknown, but leading theories include imbalance between excitatory and inhibitory synapses and substandard signaling between brain structures due to poor axonal growth (McFadden and Minshew, 2013; Fakhoury, 2015). It is possible that a myriad of genetic and environmental factors could lead to distinct physiological conditions convergent on the behavioral phenotype.

ASD is complex with hundreds of genes implicated in its etiology. The predominant focus of previous research has been on protein-coding genes, but there is the potential of more genes such as non-coding genes being implicated as well. Disease gene identification studies are usually large-scale and high-throughput. Examples include genome-wide association studies (GWAS), copy number variation studies (CNV) and whole exome sequencing (WES). These studies in themselves are time-consuming and expensive especially when considering the sample size required for an effective study. The output can contain numerous potential candidate genes, which are also expensive and time-consuming to validate, with minimal impact on risk of the disease itself. GWAS studies have been shown to be particularly susceptible to weak SNP associations for ASD (Anney *et al.*, 2012). Disease gene prioritization systems seek to determine high confidence for disease association targets amongst gene lists, and while disease gene prioritization methods have become somewhat ubiquitous, they generally do not have methodologies accounting for prioritization of non-coding genes.

Support vector machine (SVM) approaches have previously been applied to ASD research. Bruing *et al*. (2014) provided support for a phenotype-genotype relationship by using symptom profiles as features and genetic disorders as classes for a multi-class extension of SVM. Magnetic resonance image (MRI) offers one of the most potentially fruitful routes for ASD diagnostics, and SVM approaches have been applied to MRI data to classify and further characterize the morphology of the disorder. Retico *et al*. (2016) used SVM to determine differences in the morphologies between young male and female patients with ASD. A similar method was employed by Ecker *et al*. (2010) on whole-brain structural imaging to reveal a correlation for the distance from the decision boundary and the severity of the disorder.

In this study, we have developed a machine learning method to prioritize genes for ASD risk. Based on domain-specific knowledge of ASD, it is hypothesized that expression patterns offer a potential means of prioritization for all gene types for ASD risk. In particular, we design the novel approach of using the normal developmental brain expression patterns found in the BrainSpan dataset to leverage previous research focused primarily on protein-coding genes to classify ASD risk gene candidates. The validity of this approach is supported by weighted gene co-expression network analysis on the BrainSpan dataset, which has shown convergence of ASD risk genes on developmental pathways (Parikshak *et al*., 2013). It is further supported by evidence of the potential role in ASD of lncRNAs, whose function is closely linked with their expression patterns (Derrien *et al*., 2012; Necsulea *et al*., 2014; Ziats and Rennert, 2013). In this study, we first generated a gene list of high-confidence developmental ASD risk genes. We were

then able to create a support vector machine (SVM) model for ASD risk gene prediction with a 76.7% accuracy capable of prioritizing ASD candidate genes based solely on expression patterns in the developing brain. Utilizing a wrapper methodology and a best-first search method during feature selection, we were able to drastically reduce the dimensionality and identify biologically relevant and novel temporospatial features within the dataset. The performance of the feature subset showed improvement over the full feature set. To further test our model, we used the ASD risk gene list to generate hypothetical loci similar to what would be expected from an association study. The genes within the loci were prioritized to determine the relative rank within the list of the known risk gene. Finally, the model was applied to the prioritization of long non-coding RNA (lncRNA) genes. Overall, the study demonstrates the effective application of a machine learning approach to ASD risk gene identification using normal tissue expression patterns.

4.2 Materials and Methods

The machine learning problem in this study can be defined in the following way: genes serving as instances are to be classified for autism spectrum disorder (ASD) risk using their respective expression profiles which serve as the feature set. A model for this decision would allow the prioritization of gene lists based on the strength of predicted ASD associations. Using known ASD risk genes and non-ASD genes with their expression profiles, we seek to perform supervised training of a model.

Datasets

The BrainSpan Atlas of the Developing Human Brain is a developmental transcriptome dataset compiled by a consortium consisting of the Allen Institute for Brain Science and five collaborating universities (Hawrylycz *et al.*, 2012). The dataset consists of 524 samples with a developmental time point range from 8 weeks post-conception to 40 years of age from 26 brain structures. While the dataset demonstrates a lack of availability for multiple samples at each temporospatial time point in development, the BrainSpan dataset is currently the most comprehensive transcriptome of the human developing brain. Expression values were RNA-sequencing reads that were assembled and aligned using the GENCODE consortium's annotation release v10 (Harrow *et al.*, 2012). They were in the units of Reads Per Kilobase of transcript per Million mapped reads (RPKM). A $\log_2(RPKM+1)$ transformation was applied to the data. Genes in the dataset were instances, and their expression values for the temporospatial time points acted as features for the training dataset.

To build a model for ASD risk gene classification, negative and positive gene instances were required. While many genes can be considered non-ASD, two considerations were made to enhance potential model performance. Genes associated with diseases unrelated to the disease being studied have previously been used as negative controls in prioritization studies in an effort to reduce potential systematic bias (Thienpont *et al.*, 2010; Erlich *et al.*, 2011; Moreau and Tranchevent, 2012), and here we employ that same methodology by using non-ASD disease-associated genes as our negative instances. Since many individuals afflicted with ASD are also diagnosed with

some form of intellectual disability (ID) (Bakken *et al.*, 2010; Hoekstra *et al.*, 2010), and there is considerable overlap between ID and ASD-associated genes (Pinto *et al.*, 2010), ID genes were not among the negative instances. The positive instances were ASD risk genes compiled from the Simons Foundation Autism Research Initiative Gene database (Abrahams *et al.*, 2013), AutismKB (Xu *et al.*, 2012), and De Rubeis *et al.*'s (2014) large exome sequencing study for *de novo* mutations in individuals with ASD. To curate for developmental ASD risk genes, the top 85% of the genes based upon expression variance within the BrainSpan dataset were used (Additional file A-4).

Support vector machines

Support vector machine (SVM) is a supervised machine learning algorithm that is effective for high-dimensional datasets comprised of real numerical as opposed to categorical values. It is commonly used for biological classification problems (Cortes and Vapnik, 1995; Kourou *et al.*, 2014). Genes within our training dataset are vectors defined as $x_i \in \mathbb{R} \mid 0 \le x_i \le 1$ (after normalization) for $i = 1,\ldots,l$, where i is a temporospatial feature in the BrainSpan dataset and l is the size of the feature vector. The classification of each gene, non-ASD or ASD, is defined here as $y_i \in \{-1,+1\}$. When the model is trained, the algorithm seeks to maximize the distance between margins for a decision boundary separating the positive and negative instances in hyper-dimensional space. The margins are determined from a subset of the total instances nearest in Euclidian distance to the decision boundary referred to as support vectors. The distance between margins is defined as $2/\|\omega\|$ where ω is a vector orthogonal to the decision boundary such that its dot product with a support vector is zero. The sign of the decision function (positive or

negative) is used for binary classification. Classification problems generally require richer feature space than what is defined originally within the dataset to separate the variables. Plotting vectors in higher dimensions is computationally expensive, but application of a kernel function allows for model optimization in higher dimensional space without having to plot the points. Popular kernel methods for SVM models include radial basis, linear, polynomial and sigmoidal, and for our initial feature selection and final model after optimization, we used the radial basis function (RBF) kernel:

$$K(x \bullet x') = \exp\left(-\gamma \left\| x - x' \right\|^2\right) \quad (1)$$

The parameter γ determines the 'smoothness' of the decision boundary. For this study, we use the SVM SVC class from the Scikit-learn Python library (Pedregose *et al.*, 2011) for all SVM implementations with the exception of the feature selection process where the libSVM package from the WEKA data mining software was used (Hall *et al.*, 2009).

The training dataset was imbalanced, consisting of 366 ASD risk genes as positive instances and 1762 non-ASD disease genes as negative instances, a ratio of 1:4.8. For model construction, there are methods of balancing the dataset such as oversampling using synthetic minority over-sampling technique (SMOTE) (Chawla *et al*., 2002) and randomized under-sampling of the majority class (Kubat and Matwin, 1997). However, adjusting class weights within the parameters of the learning algorithm may offer an optimal solution. The class weights balance the misclassification cost in establishing soft margins (see *Model parameter optimization and performance evaluation*) without altering the underlying data space. In this study, we have empirically demonstrated that there is no performance loss in using the class weight parameter versus

randomized under-sampling of the majority class for optimized SVM models with the full feature set (Table C-1) based on 50 repetitions of tenfold cross-validations.

The SVM algorithm was chosen over other machine learning algorithms due to its effectiveness in producing a more generalized model through its maximization of the decision boundary and not the minimization of training errors (Yang, 2004). It has been shown to outperform many other learning algorithms on various biological problems, including prediction of proteolytic cleavage (duVerle and Mamitsuka, 2012), DNA-binding residues in proteins (Si *et al*., 2015; Wang and Brown, 2006) and linear B-cell epitopes (Wang and Pai, 2014). It is favorable due to its low computational cost for training a model and easily optimized parameters. The SVM algorithm also has the benefit of a function output, which allows for gene prioritization (see *Candidate Gene Prioritization*). To verify its suitability for ASD risk gene prediction, we compared the performance of the weighted SVM model to other commonly used machine learning algorithms (Table C-1). The dataset used for this analysis was balanced through randomized under-sampling of the majority class, and the performance evaluation was the same as outlined in *Model parameter optimization and performance evaluation* for 50 repetitions of tenfold cross-validations.

Feature selection

Gene expression datasets such as the BrainSpan dataset can have high dimensionality. The feature selection process removes redundant and irrelevant features to improve model performance, reduce computational load, and decrease the ratio of features to samples, which reduces the probability of overfitting. Wrapper methods

evaluate feature subsets in the context of the learning algorithm. Given that these methods take into account the learning algorithm that is to be optimized, wrapper methods generally perform better than filtering or embedded techniques especially for gene expression datasets, but they are computationally expensive (Hira and Gillies, 2015) and can potentially lead to overfitting (Saeys *et al*., 2007). To address potential overfitting in this study, the approach of candidate gene prioritization (see *Candidate Gene Prioritization*) has been employed as external validation in lieu of an independent test set. In addition to model performance improvement, the methodological approach of the wrapper method allows us further knowledge discovery in evaluating the performance of the model with the addition of each new feature. This is of particular interest in this study given that the features represent temporospatial points in brain development and the molecular etiology of ASD is still unclear.

The number of potential feature subsets for a brute force search is equivalent to a power set or 2^N, where $N = 524$ for our dataset. This is infeasible. Deterministic feature selection methods such as the sequential forward selection (SFS) method, which incrementally adds features in a greedy hill-climbing search, have been used successfully in cancer machine learning studies with expression datasets (Kourou *et al.*, 2014). In a greedy hill-climbing search, for a machine learning algorithm such as SVM, a feature set of size n, and a feature subset F,

$$f_{SVM}(F) \text{ where } F \subseteq \{f_1, f_2, \ldots, f_n\} \tag{2}$$

$$f_i \in \{f_1, f_2, \ldots, f_n\}$$

the performance of the classifier is measured by a scoring metric. For this study, overall accuracy was used (TP=true positive, TN=true negative, FP=false positive, and FN=false negative):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

After testing multiple performance measures, overall accuracy was determined to be optimal for searching subsets. It showed steady increases and converged on a subset with minimal processing time. The subset is then built in the following manner where t is representative of iterations:

$$if\ f_{SVM}\left(F^{t-1} \cup \{f_i\}\right) > f_{SVM}\left(F^{t-1}\right)\ then\ Set\colon F^t = F^{t-1} \cup \{f_i\} \tag{4}$$

$$if\ f_{SVM}\left(F^{t-1} \cup \{f_i\}\right) \leq f_{SVM}\left(F^{t-1}\right)\ then\ Feature\ Set = F$$

While SFS is an effective wrapper search method, it does present the possibility of local maxima. One way to partially alleviate this while maintaining a heuristic search is to allow for hill traversal. The best-first search algorithm implements a greedy hill climbing algorithm but allows for backtracking and expansion of previously evaluated nodes. It has also been shown to outperform the greedy algorithm (Kohavi and John, 1997). In this study using the WEKA data mining software, we searched the feature subset using the best-first search algorithm with the overall accuracy from a fivefold cross-validation using libSVM at default settings as the performance measure (Hall *et al.*, 2009). The best-first search was forward starting from an empty feature set and allowed for the expansion of five non-improving nodes.

Model parameter optimization and performance evaluation

To improve SVM classification performance with the full and selected feature sets respectively, we optimized the three parameters, cost (C), γ, and kernel using a grid search approach, which evaluates all combinations based on a performance measure. Parameter optimization for high-dimensional datasets can be sensitive to class imbalance if overall accuracy is used as the performance measure. Optimized parameters can favor a model with low sensitivity where positive instances are in the minority such as ASD genes. Therefore, we used G-mean (geometric mean) as the performance metric, which measures the classifier's ability to balance specificity and sensitivity (Lin and Chen, 2012):

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Gmean = \sqrt{Sensitivity \times Specificity} \tag{7}$$

C is the penalty assigned for misclassifications. In the maximization of $2/\|\omega\|$, which alternatively is the minimization of $\|\omega\|^2/2$ during training of the SVM model, when misclassifications are allowed, the problem takes on the form,

$$min_{\omega,\varepsilon}\{\frac{1}{2}\|\omega\|^2 + C\sum_i \varepsilon_i\} \tag{8}$$

where ε is the quantification of the misclassification. The γ parameter can be found in Equation (1). Previous work has shown that the most efficient method of optimization for C and γ, is to exponentially increase the values across a range (Hsu *et al*. 2003). In this study, we used $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$. Using the radial basis function (RBF) kernel, we measured the performance for each parameter pairing of C and γ. Using the linear based kernel, we measured the performance for each C. The optimal kernel, C and γ parameter combinations were selected based on the highest G-mean returned from tenfold cross validations. Model performance was evaluated using sensitivity (Equation 5), specificity (Equation 6), overall accuracy (Equation 3), and the Matthews correlation coefficient (MCC), which measures the correlation between the predicted and actual classifications on a scale of $MCC = \mathbb{R} \mid -1 \leq MCC \leq 1$ (Matthews, 1975):

$$MCC = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (9)$$

Candidate Gene Prioritization

To rank candidate genes, we used the output from the SVM model with greater output corresponding to higher rank. To test the ability of the model to prioritize ASD risk genes, we generated gene lists containing at least one ASD risk gene. This methodology was adapted from the work by Prio *et al.* (2010). For each ASD risk gene in the training dataset, we identified N flanking genes on the same chromosome using the GENCODE release v10 annotation. We then constructed a hypothetical locus with $2N+1$ genes centered on the ASD risk gene. If an ASD risk gene was close to a chromosome terminal, the number of genes in the opposing flank were extended to ensure that each

gene list was of equal length. Three gene list lengths were tested: 101, 201, and 401. Genes present in the training dataset were removed from each gene list. The SVM model was trained for each hypothetical locus, and the candidate gene list was prioritized. Model performance was evaluated by the percentile rank assigned to the known ASD risk gene in its respective candidate gene list:

$$Percentile\ Rank = \frac{L}{N} \times 100\% \quad (10)$$

In the above equation, L is the number of SVM scores less than the target, and N is the total number of candidate genes.

Long Non-Coding RNA Gene Candidate Prioritization

The 10,840 lncRNA genes in the GENCODE release v10 were prioritized using the SVM model to further demonstrate its capabilities and performance. The model was built with all instances from the training dataset, the feature subset from feature selection, and the optimized parameters. Confidence measures of the classification (ASD or non-ASD associated) were assigned for each gene. If the instance was classified as positive, the confidence was (1 - false positive rate), and if the instance was classified as negative, the confidence was (1 - false negative rate) (Wang and Brown, 2006).

4.3 Results

Support vector machine classification of ASD risk genes

Table 4.1 shows the performance of the support vector machine (SVM) classifier for 50 repetitions of tenfold cross-validations using all 524 features available in the dataset. The model was optimized on the G-means performance measure and used the

radial basis function (RBF) kernel with a cost $(C) = 8$ and $\gamma = 0.0078125$. The SVM model with the full feature set achieved a mean accuracy of 0.739 with 0.748 sensitivity, 0.737 specificity and 0.385 Matthews Correlation Coefficient (MCC). The receiver operator characteristic (ROC) curve for ASD risk gene prediction using the full feature set is shown in Figure 4.1. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1 - specificity) for varying output thresholds of the SVM classifier (Hajian-Tilaki, 2013), and in this study it was generated using the ROC class from the Scikit-learn Python library (Pedregosa *et al.*, 2011) The ROC curve and the area under the curve (ROC-AUC) are considered to be the most robust measures of model performance. The ROC-AUC for the full feature set is 0.8045. This value is significantly greater than the random guess value of 0.5. Given the novelty of the study, there are no real means of comparison to other model performance, and therefore the SVM model performance can serve as a benchmark for future models. The heterogeneous nature of ASD warrants against over-optimization as the probability of overfitting may increase, and the current performance appears to be indicative of an effective generalized model.

Table 4.1 The mean sensitivity, specificity, overall accuracy and Matthews Correlation Coefficient (MCC) of each model for 50 repetitions of tenfold cross-validations.

	Full Feature Set	**Selected Feature Set**
Sensitivity	0.748 ± 0.006	0.744 ± 0.005
Specificity	0.737 ± 0.003	0.772 ± 0.002
Accuracy	0.739 ± 0.003	0.767 ± 0.002
MCC	0.385 ± 0.003	0.419 ± 0.005

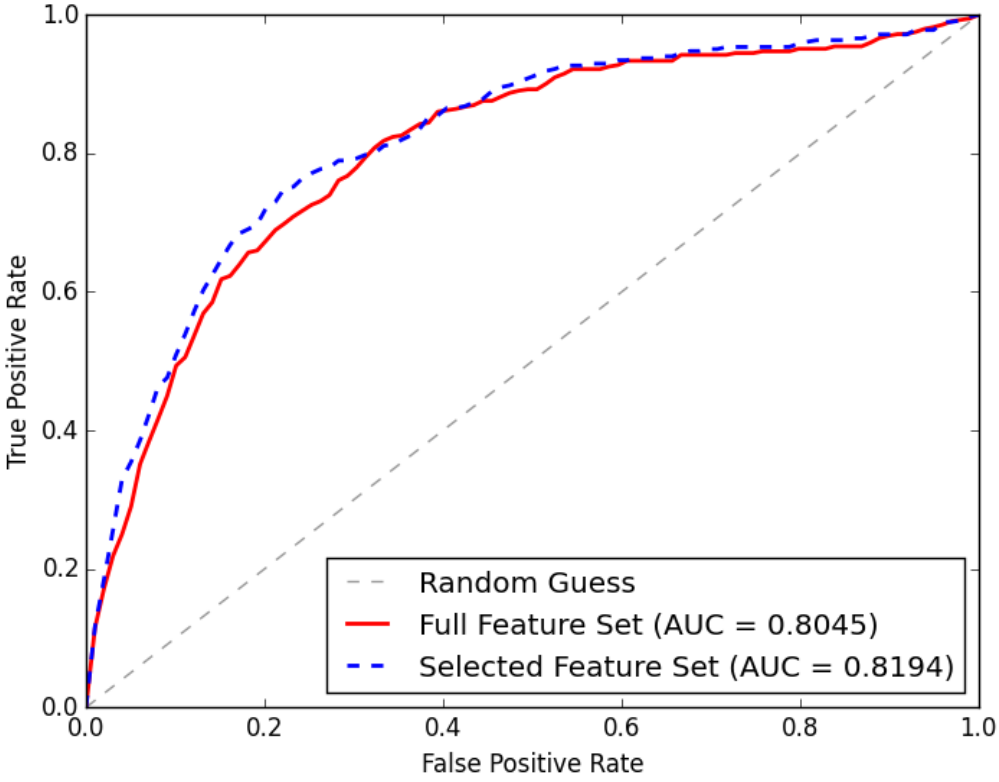


Figure 4.1 ROC curves of the selected and full feature set SVM models. The AUCs for the ROC curves are given in the legend.

Wrapper method with best-first heuristic search for feature selection

Reduction of the dimensionality for a dataset decreases computational load and the probability of overfitting. Here we demonstrate SVM model performance gain with a selected feature subset over the full feature set. Our methodology applied a forward heuristic best-first feature selection search using a wrapper method. Forward searches begin with empty sets and build them incrementally with the addition of one feature at a time, and this approach allows for evaluation of each feature added to what will become

the final subset. The best-first search is a modification of the greedy stepwise method and allows for backtracking if no improvement is seen with further feature additions. Table 4.2 shows the incremental building of the feature subset. The features are in the order of their additions to the subset, and the overall accuracy is listed for each resulting subset. For example, the feature subset of the primary somatosensory cortex (area S1, areas 3,1,2) (S1C) at 13 post-conception weeks (pcw) had an accuracy of 0.662, and the subset of S1C at 13 pcw and the dorsolateral prefrontal cortex (DFC) at 8 pcw had an accuracy of 0.694. The selected feature model was evaluated with the same parameters as the full feature set model. The specificity of 0.772, accuracy of 0.767 and MCC of 0.419 are all significantly improved over the model with the full feature set, and there was no significant difference in the specificity (Table 4.1). The selected feature model used the RBF kernel with $C = 32$ and $\gamma = 0.03125$. Its ROC-AUC of 0.8194 shows a performance improvement over the full feature model (Figure 4.1).

Table 4.2 The selected features from the best-first search algorithm. The features are described by the time point when the sample was collected and the brain structure where the sample was collected. They are listed in order of their addition to the cumulative set. The overall accuracy, sensitivity, and specificity of each subset is listed. pcw=post conception weeks, yr(s)=years of age.

Developmental Time Point	**Structure**	**Accuracy**	**Sensitivity**	**Specificity**
13 pcw	primary somatosensory cortex (area S1, areas 3,1,2)	0.661071	0.73918	0.644847
8 pcw	dorsolateral prefrontal cortex	0.694323	0.703607	0.692395
9 pcw	parietal neocortex	0.709699	0.721585	0.70723
37 pcw	mediodorsal nucleus of thalamus	0.719803	0.70153	0.723598
1 yr	dorsolateral prefrontal cortex	0.728092	0.745902	0.724393
4 yrs	dorsolateral prefrontal cortex	0.746203	0.722787	0.751067
1 yr	primary visual cortex (striate cortex, area V1/17)	0.747989	0.740929	0.749455
16 pcw	orbital frontal cortex	0.750761	0.755628	0.74975
8 pcw	orbital frontal cortex	0.754389	0.751694	0.754949
30yrs	primary auditory cortex (core)	0.754474	0.754754	0.754415
8 pcw	occipital neocortex	0.756053	0.752787	0.756731
40 yrs	primary motor cortex (area M1, area 4)	0.756147	0.75153	0.757106
21 yrs	inferolateral temporal cortex (area TEv, area 20)	0.758778	0.748798	0.760851
8 pcw	hippocampus (hippocampal formation)	0.759596	0.751093	0.761362
8 yrs	primary somatosensory cortex (area S1, areas 3,1,2)	0.759746	0.747978	0.762191

Prioritization of ASD risk gene candidates using SVM model output

Given the relatively low number of high-confidence ASD risk genes and the heterogeneity of the disorder, all of the positive instances were used in the training of the model to maximize the available data space. This however precludes the use of an independent test dataset as a means of external validation. As an alternative to the use of

an independent test dataset, both models were evaluated on their ability to prioritize hypothetical loci. We generated gene lists for each ASD risk gene and its surrounding genes, and model performance was measured by the ability to highly prioritize the known ASD risk genes. While the lists may contain previously unknown ASD risk genes, a liberal estimate of ASD risk gene frequency in the genome at 2% (De Rubeis *et al*., 2014) and varied location of CNV's (Liu and Takumi, 2014) would allow us to assume that a high performing model would prioritize a known ASD risk gene within the 90th percentile. Since our interests are in the relative ranks of the genes for the hypothetical loci, we perform a form of ordinal regression in ordering by the decision function output. Here we compare the prioritization capability for the selected and full feature set SVM models.

Table 4.3 shows the mean percentile rank for known ASD risk genes in their respective hypothetical loci of varying sizes. Again, the selected feature model outperformed the full feature model. The selected feature model showed 2-3% greater mean percentile rank of ASD risk genes for each locus length than the full feature model. The mean percentile ranks increase with the size of the loci, which is to be expected given the assumption that ASD risk genes highly prioritized would remain so in an expanding list. Figure 4.2 shows the distributions of ASD risk genes grouped by percentile rank for the two models. There were little to no genes ranked in or near the 50th percentile for the two models. ASD risk genes were either ranked in the low or high percentiles, and the amounts in each are consistent with the overall accuracy of the models. Risk genes classified as positive instances were predominantly in the 95th

percentile or above. This is consistent with expectations and indicative of strong performance for both models. Genes in the lower percentiles were principally classified as negative instances. Given the heterogeneity of the disorders, it is also possible that the misclassified genes are false positives or that their etiology for ASD is independent of brain development.

Table 4.3 The mean percentile rank of known ASD risk genes for three prioritized gene set sizes for the selected and full feature set SVM models.

Number of Genes	Selected Features Model Mean Percentile Rank	Full Feature Set Model Mean Percentile Rank
101	77.4	75.6
201	78.4	75.6
401	80.7	77.2

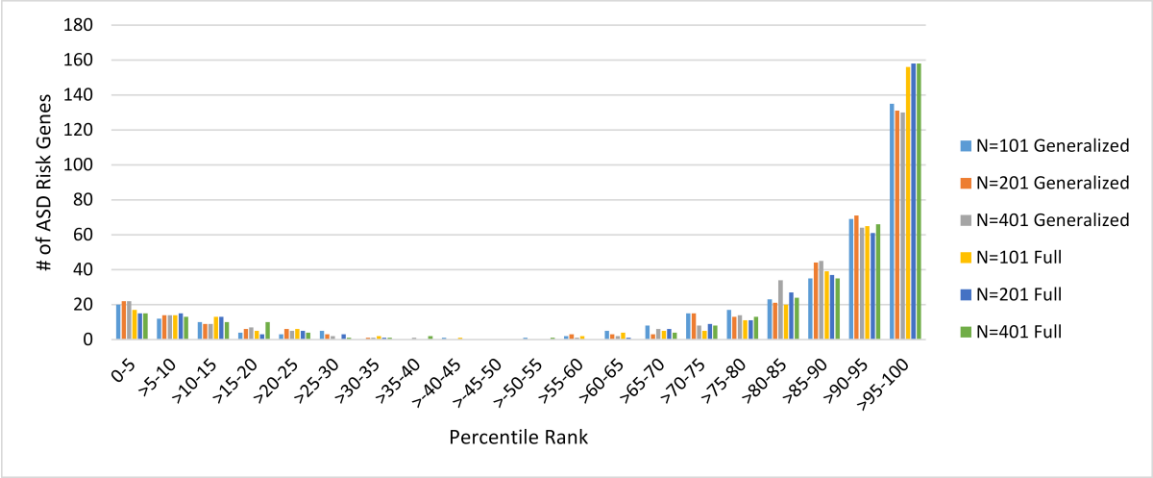


Figure 4.2 Histogram of ASD risk gene count grouped by percentile rank for the selected and full feature set SVM models for the three gene set sizes. For each hypothetical locus, the percentile rank of the known ASD risk gene within the gene list was calculated, and here those ranks are grouped in 5 percentile point increments.

Application of the SVM model to lncRNAs

Long non-coding RNA (lncRNA) genes code for transcripts greater than 200 nucleotides in length, which are not translated to peptide sequences. They are ideal genes for testing the performance of an expression-based prioritization model since they lack protein product, generally lack functional annotation, may be more numerous than protein-coding genes in the genome, and are highly expressed in the brain (Derrien *et al*., 2012). They also have high temporospatial expression specificity, and conservation studies have identified them as potentially key developmental regulators (Necsulea *et al*., 2014). LncRNAs have been found to be differentially expressed in individuals with ASD (Ziats and Rennert, 2013), but the role of lncRNAs in ASD is still an emerging field of research.

To further evaluate the model performance, we prioritized a gene list comprised of the available lncRNA genes within the dataset using the selected feature model. For each gene, confidence values for the prediction were assigned based upon the SVM output for the gene (Additional file A-5). Of the 10,840 lncRNA genes, 962 (8.87%) were classified as potential ASD risk genes, but only 63 had a confidence measure greater than 0.95.

While an exhaustive investigation of the high-priority candidate lncRNAs and their potential ASD association is outside the scope of this study, we highlight the most interesting genes based on existing annotations as a means of further demonstrating the validity of the approach. Table 4.4 shows four lncRNA genes that are highly ranked for ASD association. CHL1-AS1 is antisense to the ASD risk gene CHL1 (Salyakina *et al*.,

2011). MALAT1 has been shown to be a regulator of synapse formation (Bernard *et al*., 2010). MIAT has been shown to influence cell fate during neurogenesis (Aprea *et al*., 2013). TUG1 has been linked previously to neurodegenerative disorders (Wu *et al*., 2013). Given the limited knowledge of the role of lncRNAs in ASD, the high prioritization of genes with roles in brain development or adjacency to known ASD risk genes demonstrates the high performance of the SVM model.

Table 4.4 Genes of interest from the prioritization of lncRNA genes with their biotype, SVM output, and confidence score. lincRNA= long intervening non-coding RNA.

Gene	Type	SVM Output	Confidence
CHL1-AS1	lncRNA-antisense	3.150	0.996
MIAT	lincRNA	2.933	0.974
MALAT1	lincRNA	2.266	0.919
TUG1	lincRNA-antisense	1.915	0.846

4.4 Discussion

Autism spectrum disorder (ASD) has a complex physiologic etiology. Combinations of environmental and genetic factors causing aberrant development of brain regions have been linked to the disorder (Fakhoury, 2015). Studies utilizing non-invasive brain imaging procedures such as magnetic resonance imaging (MRI) and positronic emission tomography (PET) have also implicated brain structures and biological processes contributing to ASD (Ecker *et al*., 2015; Zürcher *et al*., 2015). While the underlying mechanisms remain poorly understood, there is a large body of knowledge in the field to be utilized for further study. Although treatment of ASD has been shown to be effective, it is dependent on early detection. Imaging can offer diagnostic input, but the process is expensive and not always practical. Biomarkers may offer the best means

for early detection of the disorder, but the complexity of the disorder and the increase in sequencing capacity have led to a multitude of potential targets too numerous to test. The need for an ASD risk gene prioritization system encompassing all gene types is evident. In this study, we employ a novel approach to bridging this gap.

By reframing the task of ASD risk gene identification as a supervised machine learning problem, we were able to construct an accurate classification model. Leveraging the existing extensive research on protein-coding genes in ASD and the comprehensive view of the transcriptome for the developing human brain offered by the BrainSpan dataset, we were able to discern a pattern from the expression profile, which distinguishes ASD risk genes. Applying a heuristic search through the possible feature space, we were also able to refine the model through feature selection, which improved performance and implicated potentially critical temporospatial features in the onset of ASD. To test the performance of our model in the absence of an independent test set, we used both standard cross-validations and the prioritization of candidate genes within hypothetical loci. The performance measures confirmed that the model achieved high accuracy and had the capability to highly prioritize ASD risk genes within gene lists of varying lengths. In light of the gathering evidence that lncRNAs are associated with ASD, we further demonstrated the utility of our model through the prioritization of lncRNA candidates. Biologically significant candidates were highly prioritized, providing further validation to our approach.

Collectively, the feature subset has interesting aspects (Table 4.2). It is intriguing that while ASD as early developmental disorders can be diagnosed by the age of two

years old with standard methods (American Psychiatric Association, 2012), the developmental time points of the selected features span the entirety of the transcriptome studied (from 8 pcw to 40 years of age). Most notably, the time points are enriched for early development, particularly from 8 pcw to one year of age, and these points were predominantly added to the subset before later developmental time points such as those greater than four years of age. Therefore, the early developmental period appears to have a larger influence on ASD risk gene identification. Co-expression modules enriched with ASD risk genes have been shown to have either dramatic upward or downward trends in expression patterns between 8 pcw to one year of age which indicates a critical timespan in relation to ASD etiology (Parikshak *et al*., 2013). Not surprisingly, the structures that were selected are mainly cortical regions, which are associated with sensory input processing and behavior. Generally, cortical regions have been found to be enlarged in children with ASD around three years of age (Schumann *et al*., 2010). Two non-cortical regions selected are the hippocampus and the mediodorsal nucleus of thalamus (MD). Both the hippocampus and the thalamus have been found to be proportionately smaller for individuals between the ages of four and eighteen years old with ASD (Sussman, *et al*., 2015). However, there is little to no evidence of a role for the MD in ASD. The selected feature set also contained later development cortical structures. These unanticipated features may present new avenues of research for ASD.

4.5 Conclusions

In this study, a novel approach is proposed for knowledge transfer from known ASD risk protein-coding genes to all gene types. We have demonstrated that a model built using only expression patterns within normal brain development as features can accurately classify and prioritize ASD risk genes. This provides a distinct advantage over previous models. It does not require *a priori* knowledge and allows for the prioritization of non-protein coding genes. It is our hope that this will lay the groundwork for an accurate prioritization tool utilizing this model.

4.6 Acknowledgements

We would like to thank Dr. Anand Srivastava for his discussion, and Jose Guevara for his review and comments on the manuscript.

References

Abrahams, B. et al. (2013) SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism*, 4, 36.

American Psychiatric Association. (2012) Diagnostic and statistical manual of mental disorders (5th ed., text rev.).

Anney, R., et al. (2012) Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet*, 21, 4781-4792.

Aprea, J. et al. (2013) Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment. *EMBO J*, 32, 3145-3160.

Bakken, T. et al. (2010) Psychiatric disorders in adolescents and adults with autism and intellectual disability: A representative study in one county in norway. *Res Dev Disabil*, 31, 1669-1677.

Bernard, D. et al. (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J*, 29, 3082-3093.

Bruing, H. et al. (2014) Behavioral signatures related to genetic disorders in autism. *Mol Autism*, 5, 11.

Chawla, N. et al. (2002) SMOTE: Synthetic Minority Oversampling Technique. *J Artif Intell Res*, 16, 321-357.

Chlebowski, C. et al. (2013) Large-scale use of the modified checklist for autism in low-risk toddlers. *Pediatrics*, 131, e1121-e1127.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273-297.

De Rubeis, S. et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209-215.

Derrien, T. et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-1789.

Developmental disabilities monitoring network surveillance year 2010 Principal Investigators and Centers for Disease Control and Prevention (CDC). (2014) Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2010. *MMWR Surveil Summ*, 63, 1-21.

duVerle, D. and Mamitsuka, H. (2012) A review of statistical methods for prediction of proteolytic cleavage. *Brief Bioinform*, 13, 337-349.

Ecker, C. et al. (2010) Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *Neuroimage*, 49, 44-56.

Ecker, C. et al. (2015) Neuroimaging in autism spectrum disorder: Brain structure and function across the lifespan. *Lancet.Neuro*, 14, 1121-1134

Erlich, Y. et al. (2011) Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res*, 21, 658-664.

Fakhoury, M. (2015) Autistic spectrum disorders: A review of clinical features, theories and diagnosis. *Int J Dev Neurosci*, 43, 70-77.

Hajian-Tilaki, K. (2013) Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*, 4, 627-635.

Hall, M. et al. (2009) The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11, 10-18.

Harrow, J. et al. (2012) GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*, 22, 1760-1774.

Hawrylycz, M. et al. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489, 391-399.

Hira, Z. and Gillies, D. (2015) A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinf*, 2015, 198363.

Hoekstra, R. et al. (2009) Association between extreme autistic traits and intellectual disability: Insights from a general population twin study. *Br J Psychiatry*, 195, 531-536.

Hsu, C. et al. (2003) A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University.

Kim, Y. and Leventhal, B. (2015) Genetic epidemiology and insights into interactive genetic and environmental effects in autism spectrum disorders. *Biol Psychiatry*, 77, 66-74.

Kohavi, R. and John, G. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, 97, 273-324.

Kourou, K. et al. (2014) Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, 13, 8-17.

Kubat, M. and Matwin, S. (1997) Addressing the curse of imbalanced training sets: One sided selection. In: *Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, Tennessee. Morgan Kaufmann, pp. 179–186

Lin, W. and Chen, J. (2012) Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform*, 12, 13-26.

Liu, X. and Takumi, T. (2014) Genomic and genetic aspects of autism spectrum disorder. *Biochem Biophys Res Commun*, 452, 244-253.

Matthews, B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405, 442-451.

McFadden, K. and Minshew, N. (2013) Evidence for dysregulation of axonal growth and guidance in the etiology of ASD. *Front Hum Neurosci*, 7, 671.

Moreau, Y. and Tranchevent, L. (2012) Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat Rev.Genet*, 13, 523-536.

Necsulea, A. et al. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505, 635-640.

Parikshak, N. et al. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155, 1008-1021.

Pedregosa, F. et al. (2011). Scikit-Learn: Machine learning in Python. *J Mach Learn Res*, 12, 2825-2830.

Pinto, D. et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466, 368-372.

Piro, R. et al. (2010) Candidate gene prioritization based on spatially mapped gene expression: An application to XLMR. *Bioinformatics*, 26, i618-i624.

Retico, A. et al. (2016) The effect of gender on the neuroanatomy of children with autism spectrum disorders: A support vector machine case-control study. *Mol Autism*, 7, 5.

Saeys, Y. et al. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23, 2507-2517.

Salyakina, D. et al. (2011) Copy number variants in extended autism spectrum disorder families reveal candidates potentially involved in autism risk. *PloS One*, 6, e26049.

Schumann, C. et al. (2010) Longitudinal magnetic resonance imaging study of cortical development through early childhood in autism. *J Neurosci*, 30, 4419-4427.

Si, J. et al. (2015) An Overview of the Prediction of Protein DNA-binding Sites. *Int J Mol Sci*, 16, 5194-5215.

Sussman, D. et al. (2015) The autism puzzle: Diffuse but not pervasive neuroanatomical abnormalities in children with ASD. *NeuroImage Clin*, 8, 170-179.

Thienpont, B. et al. (2010) Haploinsufficiency of TAB2 causes congenital heart defects in humans. *Am J Hum Genet*, 86, 839-849.

Wang, H. and Pai, T. (2014) Machine learning-based methods for prediction of linear B-cell epitopes. *Methods Mol Biol*, 1184, 217-236.

Wang, L. and Brown, S. (2006) BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res*, 34, W243-W248.

Wüu, P. et al. (2013) Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res Bull*, 97, 69-80.

Xu, L. et al. (2012) AutismKB: An evidence-based knowledgebase of autism genetics. *Nucleic Acids Res*, 40, D1016-D1022.

Yang, Z. (2004) Biological applications of support vector machines. *Brief Bioinform*, 5, 328-38.

Ziats, M. and Rennert, O. (2013). Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci*, 49, 589-593.

Zürcher, N. et al. (2015) A systematic review of molecular imaging (PET and SPECT) in autism spectrum disorder: Current state and future research opportunities. *Neurosci Biobehav Rev*, 52, 56-73.

CHAPTER V – PRIORITIZATION SYSTEM OF GENES FOR AUTISM RISK (PGAR): AUTISM CANDIDATE GENE PRIORITIZATION SYSTEM USING EXPRESSION PATTERNS

Steven B. Cogill and Liangjiang Wang

Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA.

Published: *Proceedings of the 17th International Conference on Bioinformatics and Computational Biology* 38-43 (2016)

Support: This work was supported by a grant from the Self Regional Healthcare Foundation.

Abstract

The Prioritization system of Genes for Autism Risk (PGAR) is a web-based tool for autism spectrum disorder (ASD) candidate gene prioritization. It is built on a database which stores information from machine learning and co-expression analysis for a majority of known human genes. Users submit a gene list, and for each gene, a classification score from the machine learning model is retrieved. A prioritized gene list is returned based on the classification score with links to gene profiles with co-expression analysis. The system is novel in its use of expression patterns, which allows for prioritization of non-coding RNA genes and genes lacking functional annotation. The

user-friendly design, high accuracy classification model, and depth of information for all genes make PGAR useable for researchers studying ASD.

5.1 Introduction

Autism spectrum disorder (ASD) is a heterogeneous group of disorders convergent on a behavioral phenotype with a strong genetic component (American Psychiatric Association, 2012). Hundreds of genes are currently associated with ASD, and given an etiology that is still not definitively known, many more genes are hypothesized to be associated (Banerjee-Basu and Packer, 2013). Current high-throughput screening such as genome-wide association studies produce multiple targets, which are not feasible to research on an individual level. Therefore, prioritization is needed. However, many candidate gene prioritization systems have focused on annotated protein-coding genes through the use of protein-protein interaction networks and literature mining (Erten *et al*., 2011; Hristovski *et al*., 2005). This neglects the other gene types. This is very prominent in ASD for one gene type in particular, long non-coding RNA (lncRNA) genes. They are defined as genes, which have transcripts greater than 200 nucleotides that do not code for protein. These genes have been shown to be developmental regulators, highly expressed in the brain, and differentially expressed in ASD cases (Derrien *et al*., 2012; Ziats and Rennert, 2013). Here we present the Prioritization system of Genes for Autism Risk (PGAR), a prioritization system which employs support vector machines (SVM) and co-expression network analysis to gene expression profiles to prioritize and annotate gene lists. To our knowledge, the PGAR

system is unique in its use of expression data from developmental brain tissue for ASD gene prioritization. PGAR is a web tool with a database backend which facilitates ASD research through the identification of high priority targets within gene lists.

5.2 Database

Our database uses a star schema with a gene profile table at the center. Each gene was assigned a unique ID for the PGAR system, and the location and type of the gene were documented in this center table. For this current version of the system, these annotations were from GENCODE (Harrow *et al*., 2012). This schema allows for an extensible database in that there are three distinct table groups linked to the main table. One section is comprised of lookup tables for potential identifiers such as the ENSEMBL ID. The second group comprises machine learning results, and the third group is made up of co-expression data. This design allows for multiple analyses and identifications of the genes in our system. Currently there is one machine learning analysis and co-expression analysis respectively in our system. The two analyses were run using the same expression dataset and list of known ASD risk genes. These analyses were from previous studies using the BrainSpan dataset, which consists of 524 samples from 8 weeks post conception to 40 years of age and 26 brain structures (Hawrylycz *et al*., 2012). The known ASD risk gene list used in the studies was compiled from three different resources (Banerjee-Basu and Packer, 2013; Basu *et al*., 2009; Xu *et al*., 2012). The dataset has values for all the genes in the GENCODE v10 build, and genes not in the most recent build (v24) were removed leaving 46,782 genes currently in PGAR. Therefore we have

prioritization information for the majority of known genes of all types in the human genome based on their expression patterns in the developing human brain.

5.3 Supervised machine learning model

The values used for prioritization are from a previous study (Cogill and Wang, 2016) performed by this group. We developed a supervised machine learning model. Briefly, the BrainSpan dataset was used, and the 524 features in the dataset were reduced down to fifteen features using a wrapper method with the SVM algorithm and a best-first search method. Once the model was generated, all of the available genes in the dataset were analyzed using the model to generate an SVM output value. This value is what the genes are prioritized on. The output itself has a sign value associated with it, and this determines the classification of the gene as either ASD risk or non-ASD risk, which is what is shown in the output. To assign a meaningful numeric value, a confidence score is given for the classification. This is on a scale of 0-1 for both negative and positive classifications. For example, a gene with high confidence for ASD risk could have a value of 0.9 and a gene with high confidence for non-ASD risk could also have a value of 0.9. These values are based on the range of outputs for the genes used to train the model.

5.4 Gene Co-expression network

In another study we sought to provide functional annotation of lncRNAs through weighted gene co-expression analysis (WGCNA) (Cogill *et al*., unpublished). In that study, we curated the BrainSpan dataset down to 20,456 genes with the highest

covariance sums, and then clustered those genes based on co-expression into modules. Further enrichment analyses were performed. To incorporate our study results into the PGAR system, we used the adjacency matrix generated in the study and summed the weighted connectivity between each gene in the curated dataset and all the known ASD risk genes in the dataset. This provided a means of measuring co-expression with known ASD risk genes and offered further insight into the role of the gene in ASD. Next we analyzed each module as a means of providing partial functional annotation for the genes within our system through their module assignment. We calculated the enrichment of ASD risk genes within the module using a Fisher's exact test for the frequencies of ASD genes within the module and those for the total set. Then we performed term enrichment analysis using the DAVID bioinformatics tool (Huang da *et al*., 2009).

5.5 PGAR input and output

To begin the analysis users submit a gene list and have the option of pasting an existing list into the field under the "Paste Gene List:" heading or they can submit a text file (Figure 5.1). Currently PGAR supports gene identifiers in the formats of ENSEMBL IDs (Flicek *et al*., 2014) and gene symbols as dictated by the HUGO Gene Nomenclature Committee (Gray *et al*., 2015). Alternatively, users can provide loci. For instance, if a region from a copy number variant study is identified, users can simply indicate that region in the input field, and PGAR will return a prioritized list of all the available genes in that region.

The format is as follows:

"chromosome:start position to end position"

For example, 1:10,000-50,000 would refer to all genes on chromosome 1 which overlap or are between positions 10,000 and 50,000. When the loci list is submitted, the user is given the option of either prioritizing the loci as a group or as individual lists.

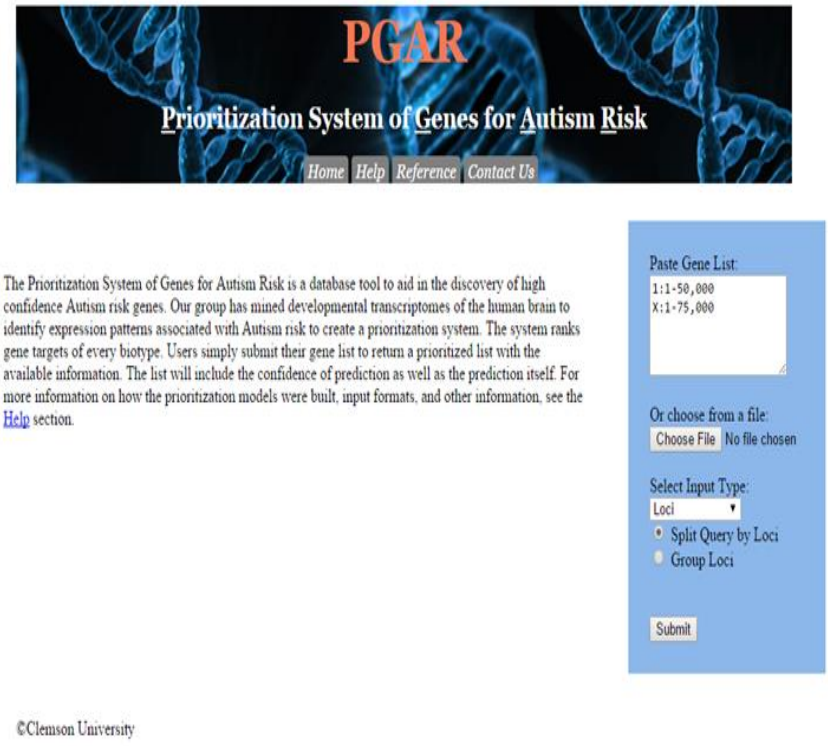


Figure 5.1 Screenshot of the PGAR home page. A brief introduction of the system is provided. The panel on the right provides options for the uploading of gene lists for prioritization.

After submission of a gene list, the user is directed to a page containing a result table of the prioritized gene list (Figure 5.2). The far left column is the PGAR ID which is the systems' unique identifier. The next column is the gene name, which is the identifier in the submitted list, and the third column is the list rank. This number is the

prioritized rank within the submitted gene list, and it should be noted that this is not the rank within the entire system. The next column is the classification as ASD risk gene or non-ASD risk gene, and it should be noted that this is the candidate classification prediction by the PGAR system. Finally, the confidence of this classification is given in the last column. For loci submissions, the locus searched is given at the top of the table and for loci prioritized separately, a table is generated for each locus. For all tables, there is a button to export the table to a ".csv" file located at the bottom right of the table.

Each PGAR ID has a link to the profile for that gene. In that profile, general information for the gene is provided, which includes the gene symbol, location, and type (Figure 5.3). Next the classification and confidence from the machine learning model-based prioritization is shown, which for the current prototype is from the SVM model outlined above. Following that, the co-expression information is displayed. This consists of the gene's weighted connectivity with known ASD risk genes as well as the gene's module assignment. It should be noted that not all genes have associated co-expression information because as stated previously, to form the co-expression network, the original dataset was curated. However given the nature of our machine learning approach, this does not necessarily preclude them from being high priority candidates or there utility as negative instances. Therefore, genes without co-expression data were not removed from the system.

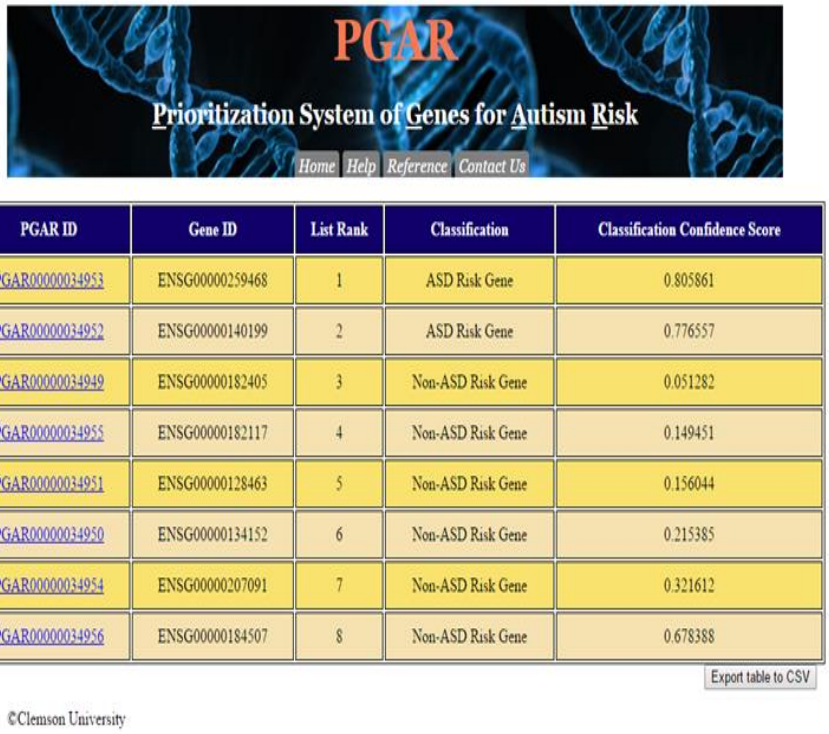


Figure 5.2 Screenshot of the PGAR results page. Below the banner, a table is generated for the prioritized gene list with links to the system profiles for each gene.

Each module assignment on the profile pages links to a profile for that module. That profile includes basic information about the module, which is the number of genes within the module and ASD gene enrichment *P*-value (Figure 5.4). The profile also includes a table of the enrichment terms for the module. At the top of the list is a link to the source for the enrichment terms which currently is the DAVID bioinformatics server. For each listing, the term itself, its broader category designation, fold enrichment, and *P*-value are given.

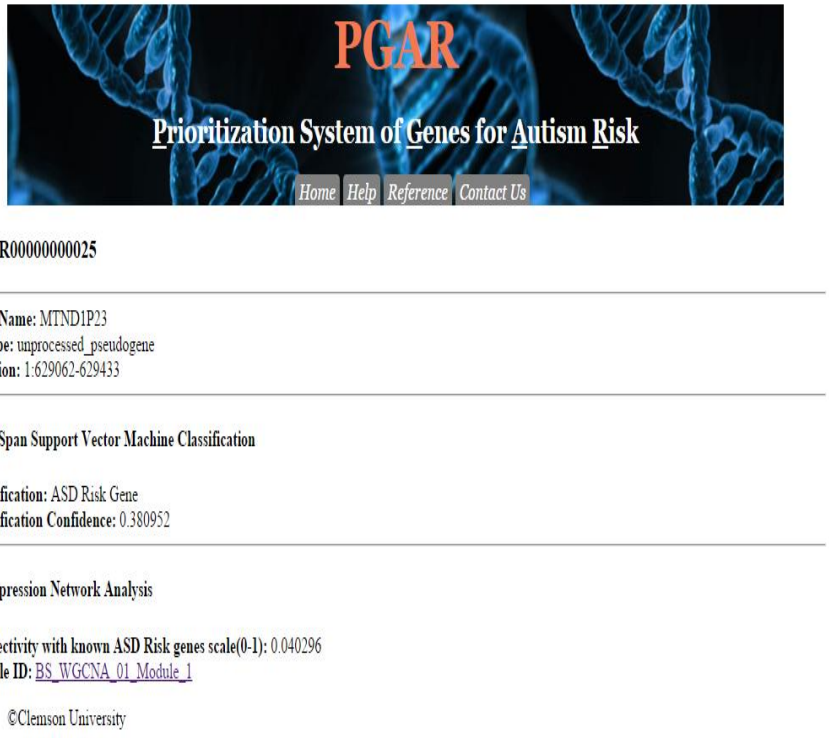


Figure 5.3 Screenshot for the PGAR gene profile page. This page shows the profile for an unprocessed pseudogene which is in module 1 for the WGCNA analysis.

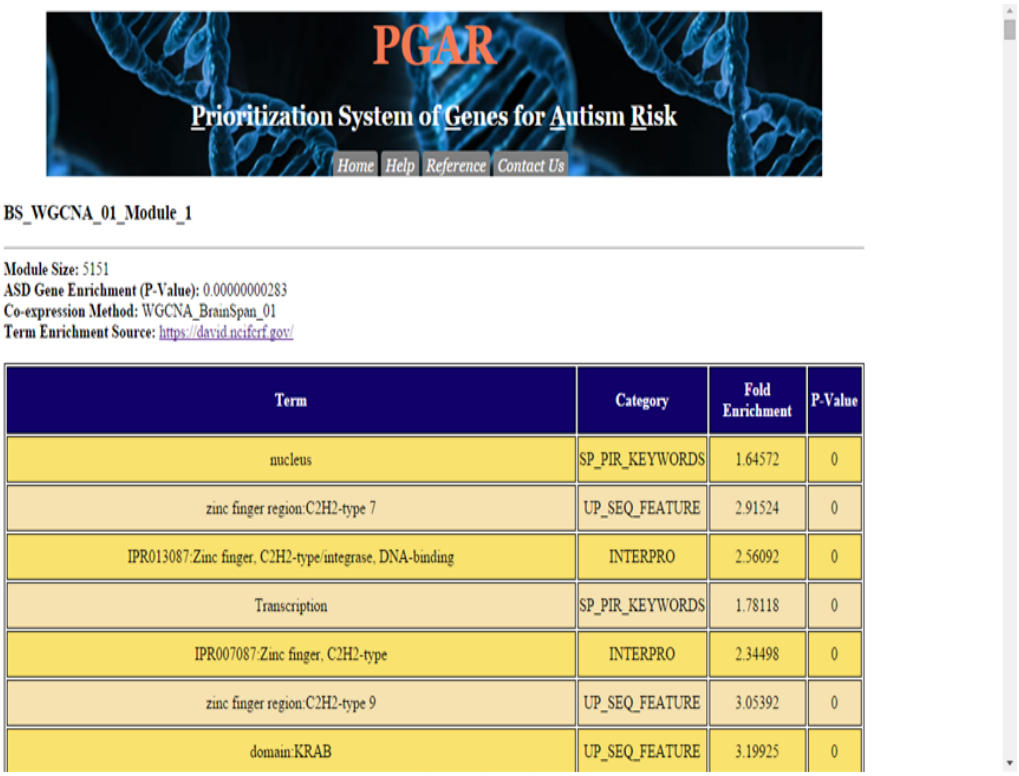


Figure 5.4 Screenshot of PGAR module summary page. The page gives a brief summary of the module attributes and the displays a table of the enrichment terms associated with the module.

5.6 Validation of the prioritization system

Future plans for the PGAR software include multiple machine learning methods as options or considered together in an ensemble approach, but currently the system uses the high performance model built using a supervised SVM approach. The model boasts a 77% classification accuracy for ASD vs non-ASD risk genes and has been demonstrated to prioritize known ASD genes highly (Cogill and Wang, 2016). Given that a majority of known ASD risk genes were used in our machine learning and co-expression studies, testing of the prioritization system requires new data. We are currently in collaboration

with another group at the Greenwood Genetic Center, which is independently testing the PGAR system using copy number variant studies where a larger region was associated with ASD and that region was subsequently parsed down through a process of elimination. Given that the novel risk gene(s) is known, the performance of the system can be determined by its ability to both classify the gene(s) as an ASD risk gene and to prioritize it highly in the system. The system is currently residing on a test server at http://scogill.people.clemson.edu/PGAR.php. We are in the process of testing all of the utilities as well as compatibility with various browsers.

5.7 Comparison with other systems

ASD is currently a high profile area of study. There are many data repositories of ASD risk genes including AutKB (Basu *et al*., 2009) and SFARI (Xu *et al*., 2012), which characterize genes based on empirical evidence from previous studies. While this is useful, it does not allow for the identification of novel candidate genes. Many of the existing popular prioritization systems such as DADA (Erten *et al*., 2011), ENDEAVOR (Tranchevent *et al*., 2008), and GeneMANIA (Warde-Farley *et al*., 2010) use broad expression networks, existing annotations, or rely upon protein-protein interactions. Our system is novel in that it uses brain developmental data in a targeted approach to ASD risk gene prioritization. ASD is a neurodevelopmental disorder, and the study of expression patterns in developing brain tissue is essential to identifying high priority candidate genes. This specialization for ASD gives our system a performance advantage

over existing systems, and we believe this will lead to the identification of many novel non-coding ASD candidate genes.

5.8 Conclusions

This system offers a novel approach to the identification of high-confidence ASD candidate genes. The system is easy to use with a simple interface for input in the form of gene lists. The site provides a 'Help' section found in the banner at the top. One of the future goals is the expansion of the allowable inputs to afford the user more options for data analysis. This may include transcript IDs as well as microarray probes. We also plan to expand the system in several areas. As more known ASD risk genes are discovered or genes within are current set are further curated due to new evidence, we will update and rerun our analyses to determine any significant performance benefits in the form of higher classification accuracy or identification of more relevant interactions within the co-expression network. We also plan to increase the number of expression datasets used. The BrainSpan dataset is unique in its comprehensive coverage of the developmental brain transcriptome, but we have begun looking for other suitable expression datasets to be incorporated into our prioritization analysis. The use of new expression data may allow us to increase the number of genes within our system. The database design allows for multiple machine learning models and co-expression networks. With additional expression datasets, we can add more analysis data to the system and potentially improve on the existing entries. We are currently in the process of testing the system, and in the future we will start updating PGAR.

References

American Psychiatric Association. (2012). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). Washington, DC: Author.

Banerjee-Basu, S. and Packer, A. (2013) SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism*, 4, 36.

Basu, S. et al. (2009) AutDB: A gene reference resource for autism research. *Nucleic Acids Res*, 37, D832-D836.

Cogill, S. and Wang, L. (2016) Support vector machine model of developmental brain gene expression data for prioritization of autism risk gene candidates. *Bioinformatics*, [Epub ahead of print].

Derrien, T. et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-1789.

Erten, S. et al. (2011) DADA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Min*, 4, 19.

Flicek, P. et al. (2014) Ensembl 2014. *Nucleic Acids Res*, 42, D749-D755.

Gray, K. et al. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res*, 43, D1079-D1085.

Harrow, J. et al. (2012) GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*, 22, 1760-1774.

Hawrylycz, M. et al. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489, 391-399.

Hristovski, D. et al. (2005) Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 74, 289-298.

Huang da, W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.

Tranchevent, L. et al. (2008) ENDEAVOUR update: A web resource for gene prioritization in multiple species. *Nucleic Acids Res*, 36, W377-W384.

Warde-Farley, D. et al. (2010) The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, 38, W214-W220.

Xu, L. et al. (2012) AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res*, 40, D1016-D1022.

Ziats M. and Rennert, O. (2013) Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci*. 49, 589-593.

CHAPTER VI - CONCLUSIONS

Within this study, data mining approaches are applied to the annotation of lncRNAs for function and disease association. We have leveraged the existing knowledge of protein-coding genes to elucidate the expression patterns associated with function and disease. We have then applied this knowledge to the identification of lncRNAs of interest. Co-expression network analysis is a particularly useful approach for functional annotation. In this study, we have constructed two separate co-expression networks. One was built using microarray data with expression profiles for known cancer genes and the lncRNA genes available in the microarray platform. The results demonstrate the efficacy of the approach of using co-expression between protein-coding genes and non-coding genes to identify disease associations. As more expression data becomes available, the construction of a similar tissue specific expression dataset with a more comprehensive lncRNA list becomes possible, and this approach could be applied again for co-expression between lncRNA genes and cancer genes to potentially identify more lncRNAs associated with cancer. Additionally, several high-priority targets were identified, which may act as treatment routes or potential diagnostic measures. For the second co-expression network analysis study, we sought to find autism spectrum disorder (ASD) associations for lncRNA genes and functionally annotate their role in brain development. The second network used an RNAseq dataset, which contained all known lncRNA genes at the time of its construction. We again showed the benefit of using expression patterns in healthy tissue samples to identify potential disease genes based

upon shared expression patterns. This study also further demonstrated significant roles of lncRNAs in human brain development. Beyond potential diagnostic measures, this co-expression network began to identify some lncRNAs affecting development with may further elucidate the complex etiology of ASD.

This study also demonstrated the novel application of machine learning to the prediction of ASD risk genes from their expression patterns in brain development. This expression profile approach was demonstrated to accurately predict ASD risk genes, and the lncRNAs prioritized within the list could potentially act as biomarkers or further elucidate the etiology of this complex disorder. We have also extracted knowledge from the dataset in finding the minimum combination of temporospatial features required for the accurate prediction of ASD risk genes. While this approach was applied specifically to the identification of ASD risk genes, future directions for the research might find that binary classification of genes via machine learning applied to expression patters is applicable to other disease gene types.

Finally, the study describes a prioritization system for ASD risk genes. Often is the case where large gene lists are returned with a tentative association with ASD, but a lack of means to effectively test all genes within the list. In this study, we outline the design and use of our inclusive prioritization system. The design of the system allows for the inclusion of multiple machine learning models and co-expression networks and it is our hope that the system will continue to improve in performance through updated information.

Predicting the functional role of lncRNAs is difficult. They have little annotation, little sequence conservation, are potentially greater in number than protein-coding genes, and their structures have yet to be elucidated. This is compounded with their emerging importance in development, tissue and cell differentiation, and multiple disorders. Currently, in the absence of an efficient high-throughput method to determine the structure of lncRNAs, expression patterns offer the best means to determining the functional roles on a global scale. Within this study, we have demonstrated the effective use of expression patterns to determine lncRNA disease associations and functional annotations.

APPENDICES

APPENDIX A – ADDITIONAL FILES

Additional file A-1 Lists of all the genes and their module assignments. Included are the core and extended lists of cancer genes compiled from the COSMIC and UniProt databases and the list of lncRNAs with probes available on the Affymetrix HG-U133 Plus 2.0 Array.

Additional file A-2 Highest co-expression lncRNAs for each cancer gene in the core list. The core cancer gene and corresponding probe are listed with the ten lncRNA probes with the highest absolute Pearson product moment correlation.

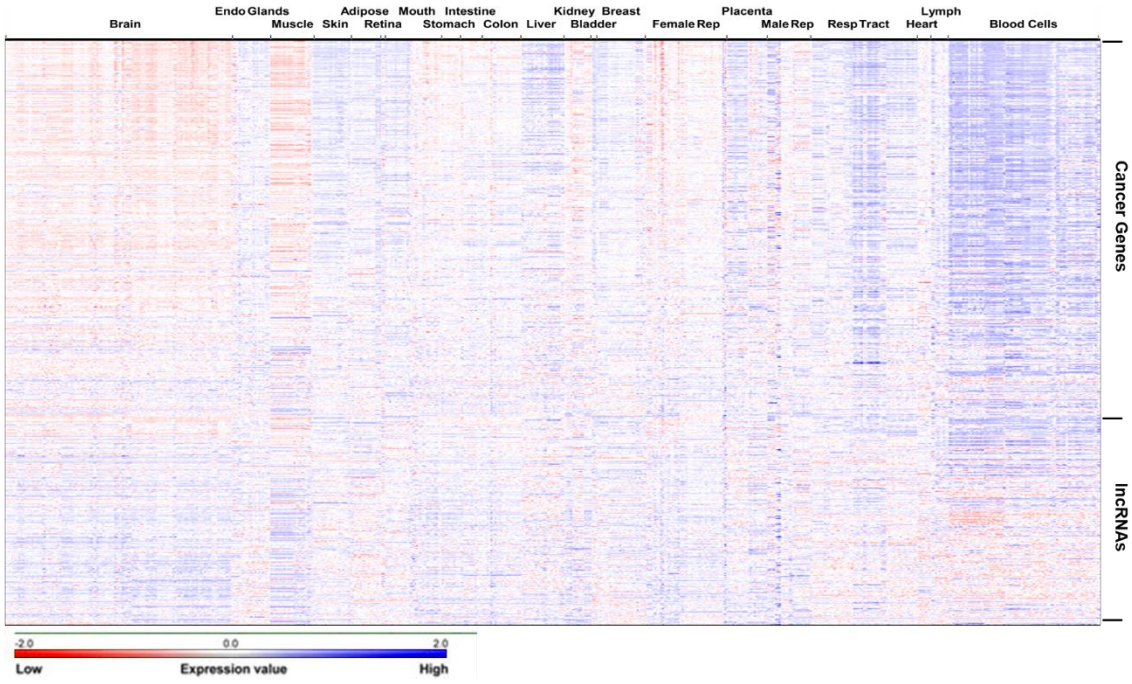
Additional file A-3 List of lncRNAs with their module assignment and normalized values for ASD gene connectivity and adjusted intramodular connectivity.

Additional file A-4 Disease and ASD genes after curation with expression values for the BrainSpan dataset.

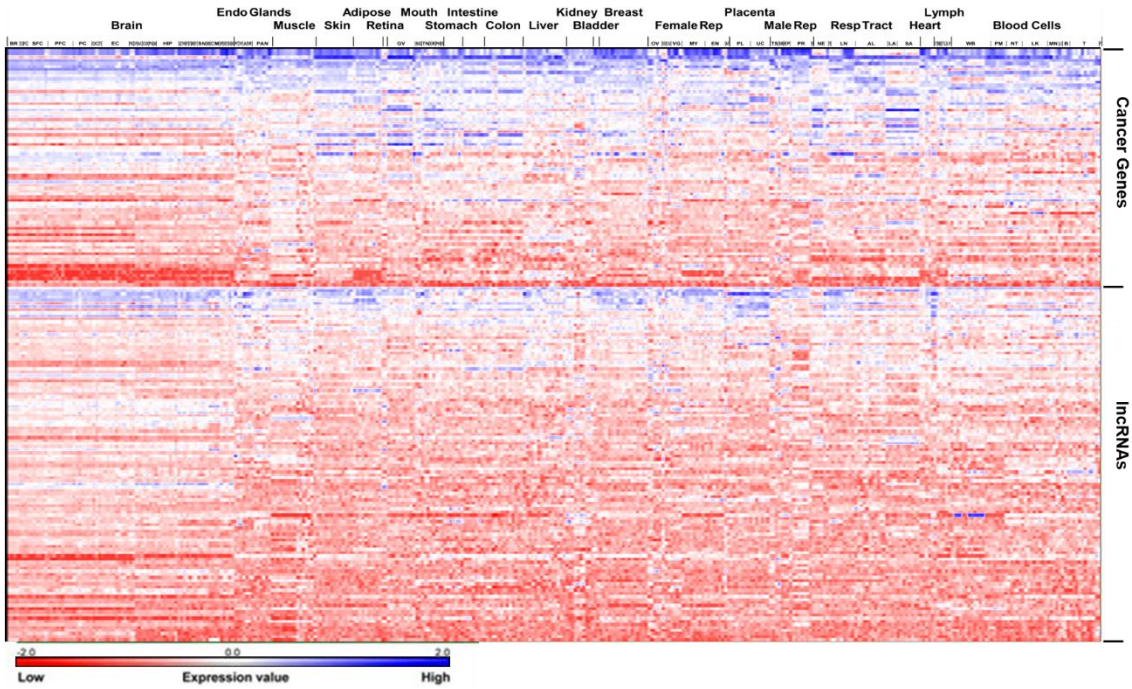
Additional file A-5 Prioritized lncRNA gene list with confidence values for the ASD non-ASD classification assigned based upon the SVM output for the gene.

APPENDIX B – SUPPLEMENTARY FIGURES

A



B


Brain
Endo
Glands
Muscle
Adipose
Skin
Mouth
Retina
Intestine
Stomach
Colon
Liver
Kidney
Bladder
Breast
Placenta
Female Rep
Male Rep
Resp Tract
Lymph
Heart
Blood Cells
Cancer Genes
lncRNAs
-2.0
0.0
2.0
Low
Expression value
High

C

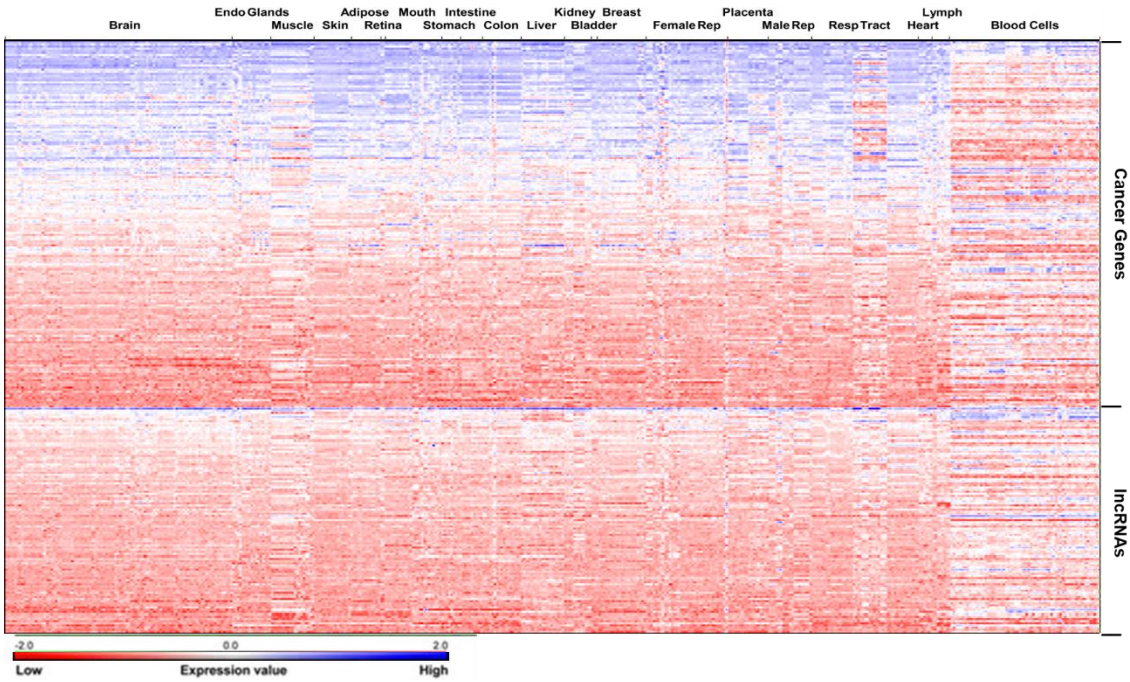


Figure B-1 Heat maps to show expression patterns across normal tissue samples in (A) Module1, (B) Module 4, and (C) Module 5. The probes are sorted by their average expression levels across the tissue types highest to lowest.

APPENDIX C – SUPPLEMENTARY TABLES

Table C-1 A comparison of the performance of different machine learning approaches for ASD risk gene prediction. The mean sensitivity, specificity, overall accuracy (acc.) and Matthews Correlation Coefficient (MCC) of each model for 50 repetitions of tenfold cross-validations is shown.

Methods	**Sensitivity**	**Specificity**	**MCC**	**Accuracy**
SVM (Balanced)	0.807 ± 0.009	0.694 ± 0.003	0.388 ± 0.008	0.714 ± 0.003
SVM (Weighted)	0.748 ± 0.006	0.737 ± 0.003	0.385 ± 0.006	0.739 ± 0.003
K-nearest Neighbor (Balanced)	0.712 ± 0.012	0.684 ± 0.005	0.307 ± 0.010	0.689 ± 0.004
Gaussian Naïve Bayes	0.684 ± 0.004	0.674 ± 0.001	0.278 ± 0.003	0.676 ± 0.001
Random Forest (Balanced)	0.789 ± 0.008	0.654 ± 0.003	0.338 ± 0.007	0.677 ± 0.003
Adaboost Decision Tree (Balanced)	0.758 ± 0.015	0.651 ± 0.008	0.312 ± 0.012	0.669 ± 0.007