A Bayesian framework to integrate genomic annotations for identification of Autism risk genes

By

Ying Ji

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE
in
Biostatistics

December 12, 2020

Nashville, Tennessee

Approved:

Ran Tao, Ph.D.

Bingshan Li, Ph.D.

DEDICATION

To my family, for always loving and supporting me.

To my friends, for being there for me.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor Dr. Ran Tao. I would like to thank my committee member who is also my advisor in Human Genetics graduate program, Dr. Bingshan Li. Without their assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. I feel incredibly lucky to be their students, they showed me the way to tackle scientific questions and enabled me to appreciate the beauty of statistical genetics.

I would like to thank the Director of Graduate Studies Dr. Robert Greevy and the Vice-Chair for Education Dr. Jeffery Blume for giving me the opportunity to be part of the Biostatistics graduate program and supporting me through my graduate study.

I would like to thank all the faculty members whom I had the pleasure to interact with, for being always responsive, patient and able to clearly explain complex topics and nuances.

I would like to thank the fellow graduate students in the Biostatistics program. I gained help and learned from our discussions and group projects.

Finally, I must express my very profound gratitude to my parents and to my friends for providing me with unfailing support and continuous encouragement throughout my years of study.

TABLE OF CONTENTS

Page

DEDICATION ii

ACKNOWLEDGMENTS iii

LIST OF TABLES vii

LIST OF FIGURES viii

1 Introduction 1

1.1 A primer on the genetics basis of Autism 1

1.2 Recent approaches to risk gene prioritization 2

1.2.1 Obtain seed genes 2

1.2.2 Gather relevant data and compute gene level score 3

1.2.2.1 Genetic sequence properties 3

1.2.2.2 Functional annotation 3

1.2.2.3 Network information 4

1.3 Proposed method overview 5

2 Materials and methods 7

2.1 Framework overview 7

2.1.1 Prior odds 8

2.1.2 Bayes factor 8

2.2 Seed genes 8

2.3 Distance to seed genes in a functional network 9

2.3.1 Construct an adjacency matrix from Gene Ontology (GO) annotations . 9

2.3.2 Construct a transition matrix from adjacency matrix 10

2.3.3 Apply random walk with restart algorithm to transition matrix to compute distance 10

2.4 Functional annotations 11
2.4.1 Empirical Bayes estimators 12
2.4.2 Binary annotations 13
2.4.2.1 Prior distribution of θ_{ij} 14
2.4.2.2 posterior distribution of θ_{ij} 15
2.4.2.3 Bayesian predictive distribution 16
2.4.3 Continuous annotations 17
2.4.3.1 Prior distribution of parameters 18
2.4.3.2 Posterior distribution of parameters 18
2.4.3.3 Bayesian predictive distribution 23
2.4.4 Selection of annotations 27
2.5 Model evaluation 28
2.5.1 Time-lapsed test 28
2.5.2 Enrichment in functional gene lists 29
2.5.3 Enrichment of ASD common SNP heritability 29
2.6 Expression of autism-associated genes in the brain 30
2.6.1 Measure of expression specificity 30
2.6.2 Tissue specific expression analysis 31
2.6.3 Cell-type specific expression analysis 32
2.6.4 Region- and time-specific expression analysis 32
3 Results 33
3.1 Selection of biological annotations 33
3.2 Validation of top candidate genes 35
3.2.1 Time-lapsed test 35
3.2.2 Enrichment in functional gene lists 37
3.2.3 Enrichment of common SNP heritability of ASD around top ranked genes 38

3.3 ASD-associated genetic changes in the spatiotemporal development of the brain. 40

3.3.1 Tissue Specific Expression Analysis 40

3.3.2 Cell-type specific expression analysis 42

3.3.3 Brain regions and developmental window Specific Expression Analysis 44

4 Conclusion 45

4.1 Remarks and observations 45

4.2 Drawbacks and limitations 46

4.3 Future work 47

References 48

LIST OF TABLES

Table Page

3.1 Annotations included in the framework 34

3.2 Enrichment of candidate genes in SFARI ASD genes 37

LIST OF FIGURES

Figure Page

3.1 Enrichment of top ranked genes in ASD2020 genes 36

3.2 Enrichment of ASD SNP heritability within or near prioritized genes by 2016NN and BN (error bars indicate the standard errors) 39

3.3 Candidate genes tend to be specifically expressed in brain tissue 41

3.4 Candidate genes tend to be specifically expressed in striatal medium spiny neurons and retina cells 43

3.5 Candidate genes show strong enrichment signals in early and mid fetal stages across different brain regions 44

Chapter 1

Introduction

1.1 A primer on the genetics basis of Autism

Autism, or autism spectrum disorder (ASD), refers to a broad range of conditions characterized by impairments in social communication and restrictive and repetitive behaviors. Globally, autism is estimated to affect 24.8 million people as of 2015 [1]. In the U.S., about 1.5% of children are diagnosed with ASD as of 2017 [2].

Genetics plays an important role in the etiology of ASD. Rare genetic variations, especially *de novo* variants (i.e. genetic variations that occur for the first time in an individual, not from parents) have led to the discovery of most known ASD risk genes [3–5]. Because *de novo* mutations generally have not been eradicated from the population by purifying selection, they are often more considered to be associated with sporadic genetic diseases like ASD. When such rare variation disrupts a gene in individuals with ASD more than expected by chance, it implicates that gene in risk [6]. Next generation-sequencing technologies have enabled the detection of de novo mutations and the investigation of their role in human disease, leading to the identification of dozens of high-confidence autism genes.

Despite great progress in risk gene identification by examining rare genetic variations, our understanding of the genetic etiology of the disorder is still limited. Dozens of ASD genes have been identified from rare genetic variations (e.g. *de novo* mutations). However, over 1000 genes are estimated to contribute to ASD risk. Also, rare variations alone are unlikely to uncover all genes associated with ASD risk as both rare variations of large effects and common variations of small to medium effects all contribute to the genetic etiology of ASD. Adding to the complexity of risk gene identification, ASD is a heterogeneous disease with different clinical sub-types. Thus, to close the gap between the number of anticipated and known autism genes, we will need to analyze diverse sources of biological evidence to supplement and amplify the signals observed through rare genetic variations. Recently, new computational approaches have been developed for ASD gene discovery recently. In this paper, we discuss the advantages and limitations of these approaches to risk gene prioritization and propose a framework to address some of the limitations of these approaches.

1.2 Recent approaches to risk gene prioritization

The process for ASD risk gene prioritization requires the scoring of the relevance of each gene within a set of candidate genes (can be the whole genome). The genes are then ranked according a decreasing order of relevance. The explosive growth of available data has stimulated the development of computational and statistical approaches for risk gene prioritization. Most current gene prioritization methods are based on the rationale that genes associated with or similar to known risk genes (”seed genes”) are more likely to also be risk genes. Using this rationale, we need to 1) obtain a set of seed genes: representative ASD associated genes to guide the prioritization; 2) gather relevant data about candidate genes and 3) compute gene-level score through summarizing data for each candidate gene.

1.2.1 Obtain seed genes

To prioritize risk genes, it is critical to use an appropriate and representative set of seed genes. For ASD, researchers usually obtain seed genes from the following sources:

1. Authors curate risk genes manually from published literature or authors set up automatic text-mining of ASD-gene co-occurrence in published abstracts.

2. Authors gather genes from publicly available databases including the Simons Foundation Autism Research Initiative Gene database (SFARI [7]) and other databases. SFARI is one of the most popular source for ASD-related genes: it is a continuously updated database that quantifies the available strength of ASD association. Currently, SFARI includes genes classified into several categories according to confidence levels of ASD association.

1.2.2 Gather relevant data and compute gene level score

To provide evidence for risk gene prioritization, researchers usually collect data from diverse gene-annotation sources. Evidence used for prioritization mainly includes the following classes: genetic sequence properties, functional annotation and network information [8, 9]. Seed genes or domain knowledge are sometimes utilized to compile data related to disease of interest.

1.2.2.1 Genetic sequence properties

Genetic mutations based approaches prioritize genes based on findings from sequencing studies through evidence like genes hit by multiple mutations in unrelated patients. Transmission And *de novo* Association (TADA) [6] is an important method to summarize genetic mutation data from sequencing studies into gene level *p*-values for ASD-risk. TADA includes a broad class of rare variations: *de novo* mutations, inherited variants, and variants identified within cases and controls in a gene-based likelihood model. For each gene, TADA tests the hypothesis: ”not being a risk gene” vs ”being a risk gene”. It scores genes based on the rarity of damaging events from *de novo* and transmitted sources from estimated distributions under null and alternative models. While TADA score provides confidence of a gene being an ASD risk gene based on genetic mutation evidence, it is restricted to sequence properties alone.

1.2.2.2 Functional annotation

Function annotation based methods use the idea of “guilt-by-association”: decide “guilt” of a gene if several of its partners (e.g. genes with the same functional annotation: biological process or pathway) share a corresponding status (association with disease). Scoring and ranking of genes are derived by the similarity of each candidate gene’s annotation profiles to seed gene profile. It is informative in finding some unknown risk genes, but an important limitation is that this method is biased to prioritize well-annotated genes. It cannot effectively give genome level prediction since only a fraction of the genome is annotated with pathways and phenotypes [10].

1.2.2.3 Network information

Network-based approaches are popular as they rely on network proximity of candidate genes to seed genes and are less biased to well-annotated genes. Networks are usually built on physical protein-protein interaction or gene co-expression information. Correlation between genes or topology based methods are commonly used to quantify the relative distance of each candidate gene to seed genes. DAWN [11] is an example of network-based approach to prioritize ASD genes. DAWN requires both genetic evidence from TADA and gene expression evidence for a risk gene to be identified, but the network used is highly specific (e.g. only use co-expression) and does not reflect the gene-gene relationships from other sources, limiting the potential to identify ASD genes.

Recently, there has been an increasing number of machine-learning based methods developed that could exploit more types of network information. In machine-learning based methods, gene prioritization is formed as a classification problem (classify genes into ASD risk genes and non ASD risk genes) to be solved by machine-learning algorithms. In a recent study [12], Krishnan et al. constructed a brain-specific functional network that integrated signals from gene expression, protein-protein interaction and regulatory-sequence data sets. Then they used a machine learning classifier: evidence-weighted linear support vector machine (SVM) to learn the connectivity pattern of seed genes (from TADA [6] and SFARI [7]) in the brain network and predict the level of potential ASD association for every gene in the genome. We recognize the power of machine learning approaches, but we have chosen not to use because of the inherent “black box” nature of machine learning limits interpretability of the final score. Instead, we present a readily interpretable output scores from Bayesian model selection approach for each gene in the genome.

1.3 Proposed method overview

Despite the demonstrated effectiveness of previous efforts in prioritizing autism risk genes, our preliminary investigations suggested there is still room for appreciable improvement in expanding relevant data evidence and computing gene-level scores.

We present a Bayesian model selection based approach, which integrates a more comprehensive set of evidence for ASD risk gene prioritization. In this study, we frame the problem of finding risk genes as a "Bayesian model selection" problem: for each gene, we select between models ”the gene is not associated with ASD risk" and ”the gene is associated with ASD risk." The Bayesian set-up provides readily interpretable output scores (i.e. Bayesian posterior odds of being a risk gene versus not being a risk gene) of each gene in the genome. We expect to discover more ASD risk genes and gain more functional insights of the genetic basis of ASD through this approach. We introduce the seed genes, relevant data, Bayesian model selection framework and model evaluation below.

Our framework requires both positive and negative seed genes as standards to specify the distribution of data evidence under models ”the gene is not associated with ASD risk" and ”the gene is associated with ASD risk." respectively. We include 65 positive genes identified from a large exome sequencing study [3]. We include 500 negative genes that are randomly selected from genome since there are no ”gold standard" for non-ASD risk genes available and we reasoned a larger set is needed to capture the properties of negative genes.

For each gene, we select between models M_0 and M_1:

M_0: this gene is not associated with ASD

M_1: this gene is associated with ASD

The output of our framework is a Bayesian posterior odds: the ratio of the posterior probability that a gene is an ASD risk gene versus this gene is not a risk gene. There are two important components of this: the prior belief of a gene's association to ASD and the Bayes factor based on observed weight of relevant data evidence. Here we use network distance to positive seed genes to form prior belief and functional annotation evidence summarized in a Bayes factor.

The assessment of classification quality requires known genes associated with ASD. While it is hard to get ”true risk genes” without experimentation, there are sources of well-supported risk genes based on sequencing studies in ASD families that could be used to evaluate our model: like from recent sequencing studies [13] and also from the publicly available SFARI database [7]. We used Fisher’s Exact test to test for enrichment of top scored genes in these functional gene lists. We also accessed the enrichment of heritability from a recent ASD GWAS within or near the top candidate genes.

The remainder of this paper is organized as follows. In chapter 2, we propose a Bayesian framework to probabilistically infer ASD risk genes by integrating three layers of information: (1) seed genes derived from de novo mutations observed in ASD patients; (2) multiple lines of supporting evidence from gene-level functional annotations; and (3) distance to known ASD genes in a biological network. In chapter 3, we evaluate the performance of our method by examining top scoring genes using curated high-confidence genes (from external sources and independent sequencing results) and evidence from GWAS studies. We also leverage the top candidate genes we identified to study the relationship between spatiotemporal gene expression in the brain to ASD risk. In chapter 4, we discuss the conclusions, limitations, and future directions for the current method.

Chapter 2

Materials and methods

2.1 Framework overview

Technically, the problem of finding risk genes can be framed as a “model selection” problem for each gene in the genome.

For each gene, we select between models M_0 and M_1:

M_0: this gene is not associated with ASD

M_1: this gene is associated with ASD

θ_1: the parameters under model M_1,

θ_0: the parameters under model M_0,

$P(M_1)$: prior probabilities of gene is associated with ASD, (i.e. under model M_1),

$P(M_2)$: prior probabilities of gene not associated with ASD, (i.e. model M_0)

We assume the prior probability is determined by distance in a gene-gene functional network:

N_P: distance to ”seed” genes (positive training genes) in a functional network

$D = (d_1, d_2, ...d_n)$ observed n dimensional vector of observed data for a gene (functional annotations)

We decide whether the gene is a risk gene by computing the posterior odds:

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{P(M_1)P(D|M_1)}{P(M_0)P(D|M_0)} \tag{2.1}$$

$$= \frac{P(M_1)\int p(D|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{P(M_0)\int p(D|\theta_0, M_0)p(\theta_0|M_0)d\theta_0} \tag{2.2}$$

$$= [\frac{P(M_1)}{P(M_0)}][\frac{\int p(D|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(D|\theta_0, M_0)p(\theta_0|M_0)d\theta_0}] \tag{2.3}$$

There are two pieces of information in the equation above: prior odds and ratio of the evidence for each model (Bayes factor).

2.1.1 Prior odds

We assume the the prior probability of a gene being a risk gene can be determined by two components: 1) overall fraction of risk genes in the genome; 2) distance of this gene to seed genes (i.e. known ASD risk genes) in a functional network. We'll discuss the details about the computation of distance in functional network in section 2.3.

We assume the fraction of risk genes in the genome is known: it's estimated around 1000 genes in 18000 genes are ASD risk genes [6]. In comparison between genes, we drop the $\frac{1000}{(18000-1000)}$ part since it is the same for all genes in the genome.

$$\frac{P(M_1)}{P(M_0)} = \frac{1000}{(18000-1000)} P(N_p) \tag{2.4}$$

2.1.2 Bayes factor

Here we consider D to be functional annotation data and derive Bayes factor from these annotations:

$$BF = \frac{P(D|M_1)}{P(D|M_0)} \tag{2.5}$$

We discuss the computation of Bayes factor for functional annotations in section 2.4.

2.2 Seed genes

We get positive genes from a whole exome sequencing study which used the enrichment of de novo and inherited mutations in ASD-affected families to identify 65 ASD risk genes. This is a well-established set of genes used by multiple gene prioritization approaches [3].

We get negative genes from random selection of the genome (after removing positive genes). While we know there might be positive genes in the 'negative' set derived in this way and inference based on which will lead to a more conservative estimate, we decide to choose this since it's hard to exclude a gene of being a risk gene. As for the number of negative genes to include, since some binary features are not widespread in the genome, we choose to use 500 random selected genes as negative genes to make it representative of the genome background.

2.3 Distance to seed genes in a functional network

We use the average distance of each gene to the 65 positive seed genes in a functional network to reflect our "prior" belief of a gene's probability of being ASD risk gene. The functional network was built to connect all pairs of human genes based on Gene Ontology (GO) [14, 15] annotations.

The rationale behind this is: genes don't function individuality but through interactions with each other. The closer two genes are in network, the higher chance that have similar functions. We reason genes closer to known ASD genes are also more likely to be ASD risk genes.

To compute distance between genes, we use the following steps:

1. Construct a adjacency matrix from Gene Ontology (GO) [14, 15] annotations

2. Construct a transition matrix from adjacency matrix

3. Apply random walk with restart algorithm on transition matrix to compute distance between genes

2.3.1 Construct an adjacency matrix from Gene Ontology (GO) annotations

We created a weighted network from publicly available Gene Ontology (GO) annotations. Each GO annotation term includes a gene product and a molecular function, biological process, or cellular component. Every gene pair in this network was assigned a score proportional to the log of the ratio of the likelihood that the two genes participate in the same GO annotation to the likelihood that they do not. We use the number and strength of shared GO annotations to build a weighted network in the following steps.

First, we assign a weight for each GO annotation term T (e.g. a molecular function) based on how specific the term is: measured by number of genes associated with the term (N_T) divided by total number of genes (N_{total}) with annotations.

$$S_T = \frac{N_T}{N_{total}} \tag{2.6}$$

Then, for a gene pair between gene i (g_i) and gene j (g_j), we compute the weight by summing all the terms shared by these two genes [16].

$$W_{g_i,g_j} = \sum_{g_i,g_j \in T} -2log(S_T) \tag{2.7}$$

Then we column normalized matrix W to make them unit-length, and used the resulting matrix W as adjacency matrix.

2.3.2 Construct a transition matrix from adjacency matrix

From the symmetric adjacency matrix W, we derive a transition matrix P. The rows of the transition matrix are probability vectors, in other words, the rows of matrix P are numerical vectors whose entries are real numbers between 0 and 1 whose sum is 1.

P_{g_i,g_j} reflects the probability of stepping to node j from node i informed by the adjacency matrix W.

$$P_{g_i,g_j} = \frac{W_{g_i,g_j}}{\sum_i W_{g_i,g_j}} \tag{2.8}$$

2.3.3 Apply random walk with restart algorithm to transition matrix to compute distance

Random walk with restart (RWR) provides a score to measure how closely related are two nodes in a weighted graph and has been used in numerous settings. Here we use RWR to measure how closely two genes are related from the transition matrix P [17, 18].

For each step starting from any node i of the network, the walker has two options: either moving to a neighbor with a probability $1-r$ or stay at node i with a probability r. The probability of the walker walks to each neighbor is proportional to the weight of the edge connecting them. The parameter r is called the restart probability. In current study, we fix $r = 0.3$.

Let q_t denote a vector with the reaching probability to all nodes at step t start from node i.

$$q_{t+1} = (1-r)Pq_t + rS_i \tag{2.9}$$

S_i is a indicator vector with the ith element as 1 and 0 for others, which means the starting node is node i.

q_t can be updated step by step until $|q_{t+1} - q_t| < T$, where T is a predefined threshold. We set T to $1e^{-6}$ in this study.

For each gene i, suppose q_t is the vector with the reaching probabilities to all the genes (after stabilize), we denote the average reaching probabilities to all positive training genes as $P(N_N)$, which is proportional to the prior odds of gene i being a risk gene.

2.4 Functional annotations

To find evidence showing gene's relatedness to ASD, we reason ASD genes converge on related biological processes and pathways. Then we aim to identify these functional processes and then summarize these information in a Bayes factor to reflect each gene's strength of evidence for ASD association.

Here we considered two forms of functional annotations: 1) binary: presence/absence in biological processes previously implicated in ASD (e.g. 1 if is a member of a major ASD-associated pathways and 0 if not), for instance, previous study have shown strong involvement of genes in developmental and specific spatial specificity circuits in pathogenesis of ASD [19]; 2) continuous: some continuous metrics describing a gene's properties, for example, mutation intolerant genes are much more likely to cause neurodevelopmental disorders than tolerant genes [20] so we include metrics suggesting intolerance to functional mutations that reflects gene essentiality.

For each gene, the evidence from binary and continuous annotations are summarized in the form of Bayes factor:

$$\frac{P(D_a|M_1)}{P(D_a|M_0)} = \frac{\int p(D_a|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(D_a|\theta_0, M_0)p(\theta_0|M_0)d\theta_0} \tag{2.10}$$

We estimate the distribution of θ_0, θ_1 under null and alternative models separately using Empirical Bayes estimators.

2.4.1 Empirical Bayes estimators

To estimate $P(D_a|M_1)$ and $P(D_a|M_0)$, we need to estimate the distribution of parameters $p(\theta_1|M_1)$ and $p(\theta_0|M_0)$. This requires us to specify a prior distribution of θ.

There are many different types of priors to choose from: such as flat, weakly informative or specific informative. Practically, flat priors tend not to work well and it is recommended to use at least weakly informative prior. However, it's challenging to specify informative priors for all the features in an automated way. The pre-specified priors also suffers from being subjective sometimes. Therefore, we decide to use Empirical Bayes estimators: learn the prior distributions of parameters from the training data at hand.

Empirical Bayes methods relies on conjugate prior modeling and estimate the hyperparameters of θ from the observations. It have been shown to be powerful and have been applied to a many problems since the first major use by Robbins [21] in the 1950s. The first major work in parametric empirical Bayes analysis is from a series of papers by Efron and Morris [22].

Specifically, we are using parametric empirical Bayes approaches: we specify a parametric family of prior distributions. The prior distribution is determined by some hyperparameters. Suppose we observe some random variables X_i from this distribution, we assume all the information about the prior parameters and hyperparameters is contained in the marginal distribution of X_i. Then we can use the marginal distribution to recover the hyperparameters in the prior distribution from the observations via maximum likelihood estimation method or moment method. Then we can use this ”estimated prior” in the subsequent inference of new data. Here, we model binary features using Beta-Bernoulli prior distribution and continuous features using Normal-Inverse Gamma prior distribution [23]

2.4.2 Binary annotations

Consider the binary representation of a functional annotation (i.e.,“present” or “absent” in a list, like FMRP target genes).

Let D_i to denote the i-th binary annotation.

For a gene k, $D_{ik} = 1$ if gene is in annotation i, $D_{ik} = 0$ otherwise

For annotation i, we estimate two separate models (M_{i0} and M_{i1}) parameterized by θ_{i0} and θ_{i1}

As these annotations are binary, we assume data is generated from Bernoulli distribution:

θ_{i1}: fraction of disease-associated genes that also have annotation $D_i = 1$

θ_{i0}: fraction of non disease-associated genes that also have annotation $D_i = 1$

* Under model M_{i1} (i.e. the gene is an ASD risk gene)

$D_i \sim Bernoulli(\theta_{i1})$

Under M_{i1}, the probability of observing D_{ik}:

$$f(D_{ik};\theta_{i1}) = \theta_{i1}^{D_{ik}}(1-\theta_{i1})^{1-D_{ik}} \tag{2.11}$$

with $D_{ik} = 0/1$

* Under model M_{i0} (i.e. the gene is not an ASD risk gene)

$D_i \sim Bernoulli(\theta_{i0})$

Under M_{i0}, then the probability of observing D_{ik}:

$$f(D_{ik};\theta_{i0}) = \theta_{i0}^{D_{ik}}(1-\theta_{i0})^{1-D_{ik}} \tag{2.12}$$

with $D_{ik} = 0/1$

To estimate the parameter θ_{ij}, suppose we have a prior distribution of the parameter (determined by hyperparameters of beta distribution), and we have observed training set genes along with their annotations information and labels (disease-associated/non disease-associated), we used empirical Bayes approach to estimate the distribution of these parameters and hyperparameters.

2.4.2.1 Prior distribution of θ_{ij}

To obtain conjugate prior for θ, we use Beta distribution here.

Under M_{i1}:

$\theta_{i1}|\alpha_{i1},\beta_{i1} \sim Beta(\alpha_{i1},\beta_{i1})$

$$f(\theta_{i1}|\alpha_{i1},\beta_{i1}) = \frac{1}{B(\alpha_{i1},\beta_{i1})}\theta_{i1}^{\alpha_{i1}-1}(1-\theta_{i1})^{\beta_{i1}-1} \tag{2.13}$$

with $0 < \theta_{i1} < 1$

Under M_{i0}:

$\theta_{i0}|\alpha_{i0},\beta_{i0} \sim Beta(\alpha_{i0},\beta_{i0})$

$$f(\theta_{i0}|\alpha_{i0},\beta_{i0}) = \frac{1}{B(\alpha_{i0},\beta_{i0})}\theta_{i1}^{\alpha_{i0}-1}(1-\theta_{i0})^{\beta_{i0}-1} \tag{2.14}$$

with $0 < \theta_{i0} < 1$

2.4.2.2 posterior distribution of θ_{ij}

From training data, we can update the distribution of θ_{ij}

Under M_{i1},

For binary annotation i, suppose we observe n positive training genes (i.e the 65 ASD genes) on whether they are included in the annotation: $x_{1,2\ldots,n}$ $(x_i = 0/1)$.

Each annotation x_i is from bernoulli distribution with Beta prior:

$x_i|\theta_{i1} \sim Bernoulli(\theta_{i1})$

$\theta_{i1}|\alpha_{i1}, \beta_{i1} \sim Beta(\alpha_{i1}, \beta_{i1})$

Our goal is to estimate the hyperparameters α_{i1}, β_{i1} from the data using method of moments.

$p(x_{1:p}) = \prod_1^p p(x_i|\alpha_{i1}, \beta_{i1})$

For a single observation, the marginal likelihood is

$$
\begin{aligned}
p(x_i = k|\alpha_{i1}, \beta_{i1}) &= \int p(X_i = k|\theta_{i1})p(\theta_{i1}|\alpha_{i1}, \beta_{i1})d\theta_{i1} && (2.15)\\
&= \int Bernoulli(k|\theta_{i1})Beta(\theta_{i1}|\alpha_{i1}, \beta_{i1})d\theta_{i1} && (2.16)\\
&= \frac{B(\alpha_{i1} + k, n - k + \beta_{i1})}{B(\alpha_{i1}, \beta_{i1})} && (2.17)
\end{aligned}
$$

The moments of Beta-Bernoulli distributions are:

$$E(X_i) = \frac{\alpha_{i1}}{\alpha_{i1} + \beta_{i1}} \tag{2.18}$$

$$Var(X_i) = \frac{\alpha_{i1}\beta_{i1}}{(\alpha_{i1} + \beta_{i1})^2} \tag{2.19}$$

$$E(X_i^2) = E(X_i)^2 + Var(X_i) \tag{2.20}$$

$$= \frac{\alpha_{i1}^2 + \alpha_{i1}\beta_{i1}}{(\alpha_{i1}+\beta_{i1})^2} \tag{2.21}$$

$$= \frac{\alpha_{i1}}{\alpha_{i1}+\beta_{i1}} \tag{2.22}$$

We notice that $E(X_i) = E(X_i^2)$.

Let $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ (kth moment of X)

We equate $E(X_i)$ to A_1.

As we have one equation and two parameters to estimate, we assume $\tilde{\beta_{i1}} = 1$ so that we can get estimate $\tilde{\alpha_{i1}}$ by solving $E(x_i) = A_1$

The resulting estimates for hyperparameters [24]:

$$\tilde{\beta_{i1}} = 1 \tag{2.23}$$

$$\tilde{\alpha_{i1}} = \frac{A_1}{1-A_1} \tag{2.24}$$

With these in hand, the estimated mean of θ_{i1}:

$$E(\tilde{\theta_{i1}}) = \frac{\tilde{\alpha_{i1}}}{\tilde{\alpha_{i1}}+\tilde{\beta_{i1}}} \tag{2.25}$$

Similarly, under M_0, we can derive the posterior estimate of α_{i0} and β_{i0} from negative training genes.

2.4.2.3 Bayesian predictive distribution

Suppose we observe a ”new” gene (i.e. not in training positive or negative gene set) with binary feature i equal to D_i. We can compute the probability of observing this under

M_{i1} and M_{i0}: $p(D_i|\alpha_{i1}, \beta_{i1})$ and $p(D_i|\alpha_{i0}, \beta_{i0})$ for M_{i1} and M_{i0} respectively. Then we can compare which model gives the observed data a better fit.

Here we plug in the posterior estimates for α_{i1}, β_{i1} and α_{i0}, β_{i0} we estimated from last section:

$$\begin{align} p(D_i = k|M_{i1}) &= p(D_i = k|\tilde{\alpha}_{i1}, \tilde{\beta}_{i1}) \tag{2.26} \\ &= \frac{B(\tilde{\alpha}_{i1} + k, 1 - k + \tilde{\beta}_{i1})}{B(\tilde{\alpha}_{i1}, \tilde{\beta}_{i1})} \tag{2.27} \end{align}$$

Similarly under M_{i0}:

$$\begin{align} p(D_i = k|M_{i0}) &= p(D_i = k|\tilde{\alpha}_{i0}, \tilde{\beta}_{i0}) \tag{2.28} \\ &= \frac{B(\tilde{\alpha}_{i0} + k, 1 - k + \tilde{\beta}_{i0})}{B(\tilde{\alpha}_{i0}, \tilde{\beta}_{i0})} \tag{2.29} \end{align}$$

Therefore we can predict the Bayes factors for a new observation D_i using the ratio of Bayesian predictive distribution we derived:

$$\frac{p(D_i|M_{i1})}{p(D_i|M_{i0})} = \frac{p(D_i|\tilde{\alpha}_{i1}, \tilde{\beta}_{i1})}{p(D_i|\tilde{\alpha}_{i0}, \tilde{\beta}_{i0})} \tag{2.30}$$

2.4.3 Continuous annotations

We assume the distribution of continuous annotations (e.g. score of a gene's intolerance to mutations) follow normal distributions.

μ_1: mean of D_i among disease-associated genes

θ_1: variance of D_i among disease-associated genes

μ_0: mean of D_i among non disease-associated genes

θ_0: variance of D_i among non disease-associated genes

Then we have:

Under model M_0: $D_{i0} \sim N(\mu_{i0}, \theta_{i0})$

Under model M_1: $D_{i1} \sim N(\mu_{i1}, \theta_{i1})$

2.4.3.1 Prior distribution of parameters

Suppose observations $X_1,...,X_n$ come from normal distribution with Normal-Inverse Gamma prior, in other words, we assume prior for mean is under Normal distribution, prior of variance is under Inverse Gamma distribution. :

$X_i|\mu, \theta \sim N(\mu, \theta)$

$\mu|\theta \sim N(\mu_0, \frac{\theta}{\kappa_0})$

$\theta \sim IG(\upsilon_0/2, \upsilon_0\sigma_0^2/2)$

We have these hyperparameters: μ_0,κ_0,υ_0,σ_0

We assume this distribution form for both M_1 and M_0 but under different hyperparameters.

2.4.3.2 Posterior distribution of parameters

Suppose observations $X_1,...,X_n$ from normal distribution with Normal-Inverse Gamma prior:

$X_i|\mu, \theta \sim N(\mu, \theta)$

$\mu|\theta \sim N(\mu_0, \frac{\theta}{\kappa_0})$

$\theta \sim IG(\upsilon_0/2, \upsilon_0\sigma_0^2/2)$

Let

$\eta = (\mu_0, \kappa_0, \nu_0, \sigma_0^2)$ be the hyperparameters we got following previous section, we can compute the marginal distribution of x in model,following [25]

To lighten notations, the η will be dropped in the densities

$$p(x|\eta) = \frac{f(x|\mu, \theta)\pi(\mu, \theta|\eta)}{\pi(\mu, \theta|x, \eta)} \tag{2.31}$$

$$\pi(\mu, \theta|x, \eta) \propto f(x|\mu, \theta)\pi(\mu, \theta)$$

$$\pi(\mu,\theta) = \pi(\mu|\theta)\pi(\theta)$$

$$\pi(\mu,\theta|x) \propto f(x|\mu,\theta)\pi(\mu|\theta)\pi(\theta)$$

We have:

$$\pi(\theta) \propto (\frac{1}{\theta})^{-\frac{\upsilon}{2}+1} exp(-\frac{\upsilon_0\sigma_0^2}{2}\frac{1}{\theta})\ ,\ \theta > 0$$

$$\pi(\mu|\theta) \propto (\frac{1}{\theta})^{-\frac{1}{2}} exp(-\frac{\kappa_0}{2\theta}(\mu-\mu_0)^2)\ ,\ -\infty < \mu < \infty$$

$$f(x|\mu,\theta) = \frac{1}{\sqrt{2\pi\theta}} exp(-\frac{(x-\mu)^2}{2\theta}) \propto (\frac{1}{\theta})^{1/2} exp(-\frac{1}{2\theta}(x-\mu)^2$$

Thus,

$$\begin{aligned}\pi(\mu,\theta|x) \propto (\frac{1}{\theta})^{-\frac{\upsilon}{2}+1} exp(-\frac{\upsilon_0\sigma_0^2}{2}\frac{1}{\theta})(\frac{1}{\theta})^{-\frac{1}{2}} exp(-\frac{\kappa_0}{2\theta}(\mu-\mu_0)^2)(\frac{1}{\theta})^{1/2} exp(-\frac{1}{2\theta}(x-\mu)^2 \\ = \theta^{\frac{-1}{2}}\theta^{-\frac{\upsilon_0}{2}+1} exp\{-\frac{1}{2\theta}[(x-\mu)^2+\kappa_0(\mu-\mu_0)^2+\upsilon_0\sigma_0^2]\}\end{aligned}$$

Denote

$\mu_n = \frac{x+\mu_0}{1+\kappa_0}$

$\kappa_n = \kappa_0 + 1$

$\upsilon_n = \upsilon_0 + 1$

$\upsilon_n\sigma_n^2 = \upsilon_0\sigma_0^2 + \frac{\kappa_0}{1+\kappa_0}(x-\mu_0)^2$

Thus,

$$(x-\mu)^2 + \kappa_0(\mu-\mu_0)^2 + \upsilon_0\sigma_0^2 = \kappa_n(\mu-\mu_n)^2 + \upsilon_n\sigma_n^2$$

Then we can show

$$
\begin{aligned}
& f(x|\mu,\theta)\pi(\mu,\theta) \\
& = f(x|\mu,\theta)\pi(\mu|\theta)\pi(\theta) \\
& = (2\pi)^{-1/2}\theta^{-1/2}exp(-\frac{1}{2\theta}(x-\mu)^2) \\
& \times \frac{1}{\sqrt{2\pi\theta/\kappa_0}}exp(-\frac{\kappa_0}{2\theta}(\mu-\mu_0)^2) \\
& \times \frac{(\frac{\upsilon_0\sigma_0^2}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})}(\frac{1}{\theta})^{\upsilon_0/2+1}exp(-\frac{\upsilon_0\sigma_0^2}{2\theta}) \\
& = (2\pi)^{-1}\sqrt{\kappa_0}\frac{(\frac{\upsilon_0\sigma_0^2}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})} \\
& \times \theta^{-\frac{1}{2}}\theta^{-(\frac{\upsilon_n}{2}+1)}exp\{-\frac{1}{2\theta}[\kappa_n(\mu-\mu_n)^2+\upsilon_n\sigma_n^2]\}
\end{aligned}
$$

Denote

$$C_1 = (2\pi)^{-1}\sqrt{\kappa_0}\frac{(\frac{\upsilon_0\sigma_0^2}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})}$$

It is shown that $\pi(\mu,\theta|x)$ is a normal-inverse-gamma distribution:

$\mu|\theta,x \sim N(\mu_n,\theta/2)$

$\theta|x \sim IG(\alpha^*,\beta^*)$

With $\alpha^* = \frac{\upsilon_n}{2}\beta^* = \frac{\upsilon_n\sigma_n^2}{2}$

The joint posterior distribution:

$$
\begin{aligned}
\pi(\mu,\sigma|x) &= \pi(\mu|\theta,x)\pi(\theta|x) \\
&= \frac{1}{\sqrt{2\pi\theta/\kappa_n}}exp[-\frac{\kappa_n}{2\theta}(\mu-\mu_n)^2] \\
&\times \frac{(\frac{\upsilon_0\tilde{\sigma_0}^2}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})}\theta^{-\frac{\upsilon_n}{2}+1}exp(-\frac{1}{\theta}\frac{\upsilon_n\sigma_n^2}{2}) \\
&= \frac{\sqrt{\kappa_n}}{\sqrt{2\pi}}\frac{(\frac{\upsilon_0\tilde{\sigma_0}^2}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})} \\
&\times \theta^{-\frac{1}{2}}\theta^{-(\frac{\upsilon_n}{2}+1)}exp\{-\frac{1}{2\theta}[\kappa_n(\mu-\mu_n)^2+\upsilon_n\sigma_n^2]\}
\end{aligned}
$$

Denote

$$C_2 = \frac{\sqrt{\kappa_n}}{\sqrt{2\pi}} \frac{(\frac{\upsilon_n \sigma_n^2}{2})^{\upsilon_n/2}}{\Gamma(\frac{\upsilon_n}{2})}$$

Consequently,

$$\begin{aligned}
p(x|\eta) &= \frac{f(x|\mu,\theta)\pi(\mu,\theta|\eta)}{\pi(\mu,\theta|x,\eta)} \\
&= \frac{C_1 \theta^{-1/2} \theta^{-\frac{\upsilon_n}{2}+1} exp\{-\frac{1}{2\theta}[\kappa_n(\mu-\mu_n)^2 + \upsilon_n\sigma_n^2]\}}{C_2 \theta^{-1/2} \theta^{-\frac{\upsilon_n}{2}+1} exp\{-\frac{1}{2\theta}[\kappa_n(\mu-\mu_n)^2 + \upsilon_n\sigma_n^2]\}} \\
&= \frac{C_1}{C_2}
\end{aligned}$$

Given a gene with value x for the continuous annotation, we have $n = 1$, and the marginal probability becomes:

$$\begin{aligned}
p(x|\eta) &= \frac{C_1}{C_2} \\
&\propto [\frac{\upsilon_0\sigma_0^2 + \frac{\kappa_0}{1+\kappa_0}(x-\mu_0)^2}{2}]^{-\frac{\upsilon_0+1}{2}} \\
&\propto [1 + \frac{1}{\upsilon_0} \frac{(x-\mu_0)^2}{\frac{(1+\kappa_0)}{\kappa_0^2}\sigma_0^2}]^{-\frac{\upsilon_0+1}{2}} \\
&\sim t_{\upsilon_0}(\mu_0, \frac{(1+\kappa_0)\sigma_0^2}{\kappa_0})
\end{aligned}$$

which is a non standardized student-t distribution where $\tilde{\mu_0}$ is a location parameter, $\sigma = \sqrt{\sigma_0^2(1+\kappa_0)/\kappa_0}$ is a scale parameter, υ_0 is degrees of freedom

The marginal distribution of X in model can be shown to be non standardized student t distribution, that is

$$X \sim t_{\upsilon_0}(\mu_0, \frac{\sigma_0^2(1+\kappa_0}{\kappa_0}) \tag{2.32}$$

with

μ_0 as location parameter

$\sigma = \sqrt{\frac{\sigma_0^2(1+\kappa_0}{\kappa_0}}$ as scale parameter

υ_0 as degree of freedom parameter

For $X \sim t_\upsilon(\mu, \sigma^2)$, it can be shown that the first six moments of X are:

$$EX = \mu$$

if $\upsilon > 2$:

$$EX^2 = \mu^2 + \frac{\upsilon\sigma^2}{\upsilon - 2}$$

$$EX^3 = \mu^3 + 3\mu\frac{\upsilon\sigma^2}{\upsilon - 2}$$

if $\upsilon > 4$:

$$EX^4 = \mu^4 + 6\mu^2\frac{\upsilon\sigma^2}{\upsilon - 2} + 3\frac{\upsilon^2\sigma^4}{(\upsilon-2)(\upsilon-4)}$$

$$EX^5 = \mu^5 + 10\mu^3\frac{\upsilon\sigma^2}{\upsilon - 2} + 15\mu\frac{\upsilon^2\sigma^4}{(\upsilon-2)(\upsilon-4)}$$

if $\upsilon > 6$:

$$EX^6 = \mu^6 + 15\mu^4\frac{\upsilon\sigma^2}{\upsilon - 2} + 45\mu^2\frac{\upsilon^2\sigma^4}{(\upsilon-2)(\upsilon-4)} + 15\frac{\upsilon^3\sigma^6}{(\upsilon-2)(\upsilon-4)(\upsilon-6)}$$

The hyperparameters in the model are $\mu_0, \kappa_0, \upsilon_0, \sigma_0$

Let

$$\mu = \frac{\sigma_0^2(1+\kappa_0)}{\kappa_0}$$

Here we cannot directly obtain estimators for σ_0 and κ_0, but we can obtain estimators for μ if we set $\tilde{\kappa_0} = 1$

We calculate the empirical Bayes estimators of mean and variance using method of moment [25]:

Let $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ (kth moment of X)

we have:

$$\tilde{\mu_0} = A_1 \tag{2.33}$$

$$\tilde{\kappa_0} = 1 \tag{2.34}$$

$$\tilde{\upsilon_0} = \frac{-\frac{14}{3}A_1^4 + 4A_1^2 A_2 + 2A_2^2 - \frac{4}{3}A_4}{-\frac{2}{3}A_1^4 + A_2^2 - \frac{1}{3}A_4} \tag{2.35}$$

$$\tilde{\sigma_0^2} = \frac{-\frac{1}{6}(A_2 - A_1^2)(5A_1^4 - 6A_1^2 A_2 + A_4)}{-\frac{7}{3}A_1^4 + 2A_1^2 A_2 + A_2^2 - \frac{2}{3}A_4} \tag{2.36}$$

We use these above equations to estimate prior parameters under M_1 using positive training genes and parameters under M_0 under negative training genes.

2.4.3.3 Bayesian predictive distribution

Plugging in the empirical Bayes estimators for hyperparameters, we have:

$X_i|\mu, \theta \sim N(\mu, \theta)$

$\mu|\theta \sim N(\tilde{\mu_0}, \theta)$

$\theta \sim IG(\tilde{\upsilon_0}/2, \tilde{\upsilon_0}\tilde{\sigma_0^2}/2)$

Let

$\eta = (\tilde{\mu_0}, \tilde{\kappa_0}, \tilde{v_0}, \tilde{\sigma_0}^2)$ be the hyperparameters we got following previous section, we can compute the marginal distribution of x in model,following [25]

$$p(x|\eta) = \frac{f(x|\mu,\theta)\pi(\mu,\theta|\eta)}{\pi(\mu,\theta|x,\eta)} \tag{2.37}$$

$$\pi(\mu,\theta|x,\eta) \propto f(x|\mu,\theta)\pi(\mu,\theta)$$

To lighten notations, the η will be dropped in the densities

$$\pi(\mu,\theta) = \pi(\mu|\theta)\pi(\theta)$$

$$\pi(\mu,\theta|x) \propto f(x|\mu,\theta)\pi(\mu|\theta)\pi(\theta)$$

We have:

$$\pi(\theta) \propto (\frac{1}{\theta})^{-\frac{\tilde{\upsilon}}{2}+1} exp(-\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2}\frac{1}{\theta}) \text{ , } \theta > 0$$

$$\pi(\mu|\theta) \propto (\frac{1}{\theta})^{-\frac{1}{2}} exp(-\frac{1}{2\theta}(\mu-\tilde{\mu_0})^2) \text{ , } -\infty < \mu < \infty$$

$$f(x|\mu,\theta) = \frac{1}{\sqrt{2\pi\theta}} exp(-\frac{(x-\mu)^2}{2\theta}) \propto (\frac{1}{\theta})^{1/2} exp(-\frac{1}{2\theta}(x-\mu)^2$$

Thus,

$$\begin{aligned}\pi(\mu,\theta|x) &\propto (\frac{1}{\theta})^{-\frac{\tilde{\upsilon}}{2}+1} exp(-\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2}\frac{1}{\theta})(\frac{1}{\theta})^{-\frac{1}{2}} exp(-\frac{1}{2\theta}(\mu-\tilde{\mu_0})^2)(\frac{1}{\theta})^{1/2} exp(-\frac{1}{2\theta}(x-\mu)^2 \\ &= \theta^{\frac{-1}{2}} \theta^{-\frac{\tilde{\upsilon_0}}{2}+1} exp\{-\frac{1}{2\theta}[(x-\mu)^2+(\mu-\mu_0)^2+\tilde{\upsilon_0}\tilde{\sigma_0}^2]\}\end{aligned}$$

The expression in the square brackets:

$$
\begin{aligned}
&(x-\mu)^2+(\mu-\mu_0)^2+\tilde{\upsilon_0}\tilde{\sigma_0}^2\\
&=2\mu^2-2\mu(x+\tilde{\mu_0})+x^2+\tilde{\mu_0}^2\\
&=2(\mu-\frac{x+\tilde{\mu_0}}{2})^2+\frac{(x-\tilde{\mu_0})^2}{2}
\end{aligned}
$$

Denote

$\mu_n=\frac{x+\tilde{\mu_0}}{2}$

$\upsilon_n=\tilde{\upsilon_0}+1$

$\upsilon_n\sigma_n^2=\tilde{\upsilon_0}\tilde{\sigma_0}^2+\frac{1}{2}(x-\tilde{\mu_0})^2$

Thus,

$$(x-\mu)^2+(\mu-\tilde{\mu_0})^2+\tilde{\upsilon_0}\tilde{\sigma_0}^2=(\mu-\mu_n)^2+\upsilon_n\sigma_n^2$$

Then we can show

$$
\begin{aligned}
&f(x|\mu,\theta)\pi(\mu,\theta)\\
&=f(x|\mu,\theta)\pi(\mu|\theta)\pi(\theta)\\
&=(2\pi)^{-1/2}\theta^{-1/2}exp(-\frac{1}{2\theta}(x-\mu)^2)\\
&\times\frac{1}{\sqrt{2\pi\theta}}exp(-\frac{1}{2\theta}(\mu-\tilde{\mu_0})^2)\\
&\times\frac{(\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2})^{\tilde{\upsilon_0}/2}}{\Gamma(\frac{\tilde{\upsilon_0}}{2})}(\frac{1}{\theta})^{\tilde{\upsilon_0}+1}exp(-\frac{\tilde{\upsilon_0}\tilde{\sigma_0^2}}{2\theta})\\
&=(2\pi)^{-1}\frac{(\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2})^{\tilde{\upsilon_0}/2}}{\Gamma(\frac{\tilde{\upsilon_0}}{2})}\\
&\times\theta^{-\frac{1}{2}}\theta^{-(\frac{\upsilon_n}{2}+1)}exp\{-\frac{1}{2\theta}[(\mu-\mu_n)^2+\upsilon_n\sigma_n^2]\}
\end{aligned}
$$

Denote

$$C_1=(2\pi)^{-1}\frac{(\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2})^{\tilde{\upsilon_0}/2}}{\Gamma(\frac{\tilde{\upsilon_0}}{2})}$$

It is shown that $\pi(\mu,\theta|x)$ is a normal-inverse-gamma distribution:

$\mu|\theta,x \sim N(\mu_n,\theta/2)$

$\theta|x \sim IG(\alpha^*,\beta^*)$

With $\alpha^* = \frac{\upsilon_n}{2}$

$\beta^* = \frac{\upsilon_n\sigma_n^2}{2}$

The joint posterior distribution:

$$\begin{aligned}
\pi(\mu,\sigma|x) &= \pi(\mu|\theta,x)\pi(\theta|x) \\
&= \frac{1}{\sqrt{2\pi\theta}} exp[-\frac{2}{2\theta}(\mu-\mu_n)^2] \\
&\times \frac{(\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2})^{\tilde{\upsilon_0}/2}}{\Gamma(\frac{\tilde{\upsilon_0}}{2})} \theta^{-\frac{\upsilon_n}{2}+1} exp(-\frac{1}{\theta}\frac{\upsilon_n\sigma_n^2}{2}) \\
&= \frac{\sqrt{2}}{\sqrt{2\pi}} \frac{(\frac{\tilde{\upsilon_0}\tilde{\sigma_0}^2}{2})^{\tilde{\upsilon_0}/2}}{\Gamma(\frac{\tilde{\upsilon_0}}{2})} \\
&\times \theta^{-\frac{1}{2}}\theta^{-(\frac{\upsilon_n}{2}+1)} exp\{-\frac{1}{2\theta}[(\mu-\mu_n)^2+\upsilon_n\sigma_n^2]\}
\end{aligned}$$

Denote

$$C_2 = \frac{\sqrt{2}}{\sqrt{2\pi}} \frac{(\frac{\tilde{\upsilon_n}\tilde{\sigma_n}^2}{2})^{\tilde{\upsilon_n}/2}}{\Gamma(\frac{\tilde{\upsilon_n}}{2})}$$

Consequently,

$$\begin{aligned}
p(x|\eta) &= \frac{f(x|\mu,\theta)\pi(\mu,\theta|\eta)}{\pi(\mu,\theta|x,\eta)} \\
&= \frac{C_1\theta^{-1/2}\theta^{-\frac{\upsilon_n}{2}+1} exp\{-\frac{1}{2\theta}[2(\mu-\mu_n)^2+\upsilon_n\sigma_n^2]\}}{C_2\theta^{-1/2}\theta^{-\frac{\upsilon_n}{2}+1} exp\{-\frac{1}{2\theta}[2(\mu-\mu_n)^2+\upsilon_n\sigma_n^2]\}} \\
&= \frac{C_1}{C_2}
\end{aligned}$$

Given a gene with value x for the continuous annotation, we have $n = 1$, and the marginal probability becomes:

$$
\begin{aligned}
p(x|\eta) &= \frac{C_1}{C_2} \\
&\propto [\frac{\upsilon_0\sigma_0^2 + \frac{1}{2}(x-\mu_0)^2}{2}]^{-\frac{\upsilon_0+1}{2}} \\
&\propto [1 + \frac{1}{\upsilon_0}\frac{(x-\mu_0)^2}{2\sigma_0^2}]^{-\frac{\upsilon_0+1}{2}} \\
&\sim t_{\tilde{\upsilon_0}}(\tilde{\mu_0}, 2\tilde{\sigma_0}^2)
\end{aligned}
$$

which is a non standardized student-t distribution where $\tilde{\mu_0}$ is a location parameter, $\sigma = \sqrt{2\sigma_0^2}$ is a scale parameter, υ_0 is degrees of freedom

Suppose for a "new" gene (i.e. not in positive or negative training set), we observe the continnuous annotation i value to be D_i We derive the empirical Bayes estimates of hyperparameters under both M_0 and M_1 following the equations above.

Then, for this gene, we have the probability of observing D_i in the non standardized student-t distribution where $\tilde{\mu_{i1}}$ is a location parameter, $\sqrt{2\tilde{\sigma_{i1}}^2}$ is a scale parameter, $\tilde{\upsilon_{i1}}$ is degrees of freedom (under M_1) divided by the probability of observing D_i in the non standardized student-t distribution where $\tilde{\mu_{i0}}$ is a location parameter, $\sqrt{2\tilde{\sigma_{i0}}^2}$ is a scale parameter, $\tilde{\upsilon_{i0}}$ is degrees of freedom (under M_0). We divide these two probabilities to get the Bayes factor:

$$
\begin{aligned}
\frac{p(D_i|M_1)}{p(D_i|M_0)} &= \frac{p(D_i|\eta_1)}{p(D_i|\eta_0)} \qquad (2.38) \\
&= \frac{p(D_i|\tilde{\upsilon_{i1}}, \tilde{\mu_{i1}}, \tilde{\sigma_{i1}}, \tilde{\kappa_{i1}})}{p(D_i|\tilde{\upsilon_{i0}}, \tilde{\mu_{i0}}, \tilde{\sigma_{i0}}, \tilde{\kappa_{i0}})} \qquad (2.39)
\end{aligned}
$$

2.4.4 Selection of annotations

We included two types of annotations in the model, binary and continuous, to derive the Bayes factor reflecting the strength of evidence in favor of a gene being an ASD gene based on biological functional processes. Initially, we collected 61 annotations from literature and used training genes to guide the selection of annotations to be included in the model. We use the set of 65 seed genes as positive genes and randomly selected genes as negative genes (see methods 2.2). We then used Fisher's Exact test to identify binary annotations enriched for positive genes and t-test to identify continuous annotations with significant differences between positive and negative genes. A complete list of the selected features are shown in Table 3.1.

2.5 Model evaluation

2.5.1 Time-lapsed test

Large exome sequencing studies of ASD families and case control have lead to well-supported candidate genes for ASD. A study published in 2015 [3] (refer to as ”2015 study”) identified the 65 ASD genes we used in the training process through excess of de novo and transmitted loss-of-function mutations (n = 10,220 total samples). In 2020, the largest exome sequencing study of ASD (n = 35,584 total samples, 11,986 with ASD) was published [13] (refer to as ”2020 study”) and 102 risk genes was identified with strong evidence.

The two studies are ”nested”: the 2020 study integrated all samples from 2015 study with other studies and newly sequenced samples. The structure of the two large exome sequencing studies enabled us to do ”time-lapse” data experiments [26]: we attempted to predict genes which have been associated with ASD in 2020 study using seed genes from 2015 study. Specifically, we rank genes by their scores using both our method and ”2016 NN” method and select the candidate genes based on selected ranking cutoffs. Then we use one-sided Fisher's exact test to test for enrichment of candidate genes in the risk genes identified in 2020 study compared to the rest of the genome.

2.5.2 Enrichment in functional gene lists

The assessment of classification quality requires multiple positive and negative instances, i.e. genes associated and not associated with a disease. However, it's extremely hard to find the 'true' risk genes. Thus, we used lists of autism-associated genes from databases like SFARI [7]. These genes were linked to ASD from a variety of evidence source: usually recurrent mutations in patients with autism or implicated by a genome-wide association study. A team of expert autism geneticists established a set of criteria to rank genes into several categories. To evaluate our genome-wide prediction, we deleted training genes (65 genes) from the SFARI set, and formed 3 tiers of "gold standard" genes according to SFARI classifications. Genes of high confidence from SFARI that have at least three de novo likely-gene-disrupting mutations being reported in the literature or meet the most rigorous threshold of genome-wide significance are designated as T1 (tier 1 evidence); genes classified as "strong candidate" from SFARI, which have two reported de novo likely-gene-disrupting mutations or implicated in by a genome-wide association study are designated as T2 (tier 2 evidence); genes of "suggestive evidence" from SFARI which have a single reported de novo likely-gene-disrupting mutation or evidence from a significant but unreplicated association study are designated as T3 (tier 3 evidence).

Using these ASD-associated gene lists, we conducted enrichment test for the top ranking genes from our method to the rest of the genome using Fisher's exact test (one-tailed).

2.5.3 Enrichment of ASD common SNP heritability

Partitioned Linkage disequilibrium score regression (LDSC) is a method to estimate the proportion of genome-wide SNP-heritability attributable to a SNP set, using information from all SNPs and explicitly modeling LD. In this study, we use partitioned LDSC to compute the enrichment of SNP heritability near or within predicted candidate genes [27, 28]. Template files and code are obtained from the LDSC Github repository [29].

LDSC require annotation files with a row per SNP and a column for each annotation (1= a SNP is part of this annotation). We create annotation files for each gene list by mapping SNPs within a 10kb distance to candidate genes' transcription start site both upstream and downstream.

We restricted analysis to Hapmap3 SNPs according to the recommendations.A recent ASD GWAS was used to access the heritability [30]. Partitioned LDSC was used to compute the proportion of SNP heritability associated with our constructed annotation while taking into account all other annotations. Using the proportion of SNP heritability and proportion of SNPs in each annotation, we can get an enrichment value associated with each annotation and an associated enrichment p-value (one-tailed test).

2.6 Expression of autism-associated genes in the brain

2.6.1 Measure of expression specificity

We use the specificity index (*SI*) defined in literature [31] to measure the gene level expression specificity in a specific tissue, cell type, brain region or developmental stage. This metric was used to define the genes specifically expressed in a tissue/cell type/brain region/developmental stage under a cutoff.

Here, the tissue/cell type/brain region/developmental stage are the conditions used in the experiment. As an example, we discuss the calculation of *SI* of genes in a ”group”, group can be tissue/cell type/brain region/developmental stage here.

Suppose we want to calculate the *SI* for a given gene (n), in a given group (1), compared to other groups, $k = 2...m$.

Let $E_{1,n}$ denote the expression of gene n in group 1.

We first compute a ”fold-change” value to measure relative expression in group 1 to other groups for all genes and use the rank of a gene relative to all other genes in group 1.

For example, we compute fold change for group 1 to group k: $E_{1,n}/E_{k,n}$, and rank all the genes in group 1 in descending order according this fold change value, then we can get the rank of gene n compared to all other genes in group 1, denote that as: $R_{1/k,n}$

SI for gene n is defined as the average rank of $R_{1/k,n}$ for $k = 2, ..., m$:

$$SI_{n,1} = \frac{\sum_{k=2}^{m} R_{1/k,n}}{m-1} \tag{2.40}$$

SI is only calculated for those genes in groups with an absolute expression above 50, and with $log2(E_k/E_{total})$ values above a threshold. E_{total} is the expression of total RNA from the source the group was from (e.g. if group is a cell, total is the expression from the tissue the cell is extracted from). Threshold is based on a set of negative control genes known not to be expressed in this cell type.

Raw *SI* scores are not directly comparable across groups due to the number of genes in comparison depends on the number of genes expressed and level of filtering used in each group. To ease comparison, a permutation testing was used to calculate a *p*-value for each SI: the expression values are randomly shuffled and SIs are calculated many times to determine the frequency of a particular SI value occur. This *p*-value is referred to as *pSI*. Code can be found at [32].

2.6.2 Tissue specific expression analysis

To evaluate tissue specific expression of genes, The Genotype-Tissue Expression (GTEx) data was used [33]. We used Reads Per Kilobase of transcript per Million mapped reads (RPKM), which is a normalized unit of transcript expression, for the analysis. There are tissues with multiple replicates, RPKM values for these tissues were averaged resulting in 25 unique tissues. For each tissue, *pSI* is calculated for all genes, and genes with *pSI* less than a cutoff value was defined as specific to that tissue. The smaller the cutoff, the more stringent the criteria.

Using *pSI*, a gene list of specifically expressed genes can be defined for each tissue.

Then candidate genes lists (prioritized risk genes here) are tested for enrichment in each tissue's specific gene list by Fisher's exact test followed with Benjamini-Hochberg correction.

2.6.3 Cell-type specific expression analysis

Data from several cell-type specific gene expression studies from the mouse brain was used to identify sets of transcripts specifically expressed in particular mouse cells [34]. The methods of cell type specific gene list determination and enrichment test for candidate gene lists is similar to the above section.

2.6.4 Region- and time-specific expression analysis

Data from Brainspan [35] was condensed into 6 major regional divisions across 10 developmental times to enable the determination of region and time specific genes in human brain. The enrichment test for candidate gene lists is similar to the above section.

Chapter 3

Results

3.1 Selection of biological annotations

We included two types of annotations in the model, binary and continuous, to derive the Bayes factor reflecting the strength of evidence in favor of a gene being an ASD gene based on biological functional processes (methods 2.4.4). After selecting of annotations we collected, 17 remained and were included in the model, as shown in Table 3.1.

The binary annotations are from literature reported evidence such as a gene belongs to a particular biological function, process or pathway involved in ASD. Specifically, these genes are either experimentally identified targets of major ASD-associated regulators (e.g. FMRP [36], CHD8 [37]) or belong to the major pathways implicated in ASD (e.g. presynaptic density genes [38]). The continuous features mainly include the evolutionary constraint measurement of genes from different parts of the genome (e.g. coding and noncoding) using different measurement strategies. The intuition is that the the absence of genetic variation (i.e. more ”constraint”) within or near a gene from large human cohorts implies strong purifying selection due to essential function or disease pathology. We include these measurements since there are multiple reports supporting that ASD associated genes tend to more evolutionarily constrained [3, 12].

Table 3.1: Annotations included in the framework

Gene.set	short.description
EG[39]	essential genes (human orthologs of genes with an essential role)
HIS[40]	Haploinsufficiency Score (probability of being haploinsufficient)
GHIS[41]	Genome-Wide Haploinsufficiency Score
FMRP_target[36]	Fragile X mental retardation (FMRP) protein targets
pLI[42]	the probability of being loss-of-function intolerant
RVIS_maf_0.05.[20]	Genic Intolerance to Functional Variation (variants of MAF $>$ 0.05)
OMIM[20, 43]	OMIM disease genes
ncRVIS[44]	noncoding region genic Intolerance to Functional Variation
embryonic[45, 46]	genes expressed preferentially in embryos
chromatin_modi[47]	genes encoding chromatin modifiers
rbfox[48]	Rbfox Splicing-Regulatory Network target genes
ca[49]	Calcium channel and signaling genes
miR-137[50]	miR-137 target genes
presynap[38]	presynaptic genes
CHD8[37]	chromatin remodeler CHD8 regulated genes
PSD-95[51]	post synaptic genes
his_mod_enz[47]	histone modifying enzymes

3.2 Validation of top candidate genes

We constructed a Bayesian selection model based on ASD-related functional annotations and obtained a systematic genome-wide ranked list of autism candidate genes. As previous literature suggests, around 1000 genes contribute to ASD risk [6]. We took the top 1000 genes in our ranked list as candidate genes and validated them using recent sequencing study results [13], public available SFARI database [7] and enrichment in ASD SNP heritability from GWAS [30].

3.2.1 Time-lapsed test

The 65 ASD seed gene we used in the modeling process was identified from an excess of de novo and transmitted loss-of-function mutations in a large exome sequencing study (n = 10,220 total samples) published in 2015 [3]. In 2020, the largest exome sequencing study of ASD (n = 35,584 total samples) [13] was published and 102 risk genes were identified using an identical sequencing process as the 2015 study. This 2020 study integrated all samples from 2015 study with newly sequenced samples.

The ”nested” structure of the two large exome sequencing studies enabled us to do ”time-lapse” data experiments. We attempted to predict genes which have been associated with ASD in 2020 study using seed genes from the 2015 study. Although they are limited in the size of the test set (only 102 newly identified risk genes), the experiments provide a rather realistic evaluation of our model’s performance as they mimic more closely the gene discovery process in real-life.

We found our top ranked genes to be enriched in ASD2020 set compared to the rest of the genome using Fisher’s Exact test. Since our method includes Bayes factor and network information, we refer to this method as ”BN” in the figures. For comparison, we included the genome-wide prediction from another study (refer to as ”2016NN”) that also included the same set of 65 genes in the training process [12]. In Figure 3.1, we observe our top ranked genes to be more significantly enriched for ASD2020 genes when we compare genes ranked top 100, 101 to 500 or 501 to 1000. We also included the results from genes ranked from 1001 to 2000, 2001 to 3000, 3001 to 4000 for comparison and found much lower enrichment of ASD102 genes in these sets, indicating most risk genes are concentrated in top ranked 1000 genes by our method. For enrichment of top 1000 genes, our method has an odds ratio of 19.83 (*p*-value = 1.80E-24) while method 2016NN ranked top 1000 genes has an odds ratio of 5.58 (*p*-value = 9.60E-7). Therefore, we focused on top 1000 genes ranked by our method as major set of candidate genes.

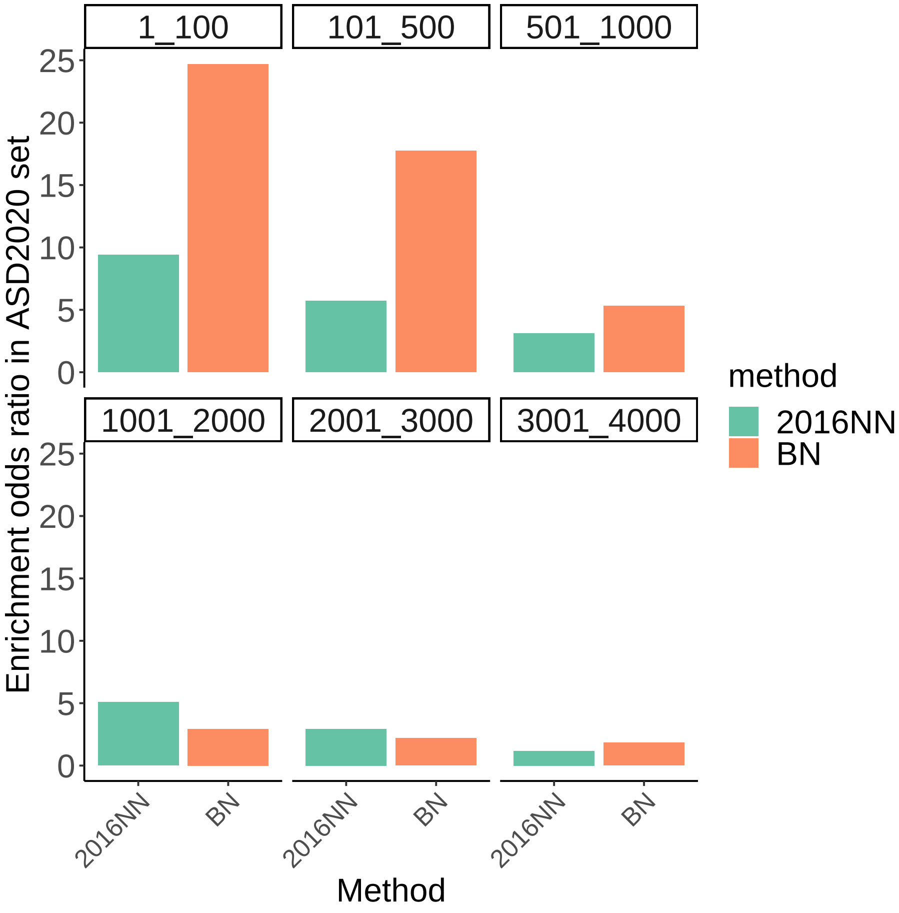

Figure 3.1: Enrichment of top ranked genes in ASD2020 genes

3.2.2 Enrichment in functional gene lists

There are a few databases contained lists of autism-associated genes like SFARI (cite: https://gene.sfari.org/about-gene-scoring/). These genes were linked to ASD from a variety of evidence source: usually recurrent mutations in patients with ASD or implicated by a genome-wide association study. We used the 3 tiers of expert curated genes according to SFARI classifications: genes of ”high confidence” are designated as T1 (tier 1 evidence); genes classified as ”strong candidate” are designated as T2 (tier 2 evidence); genes of ”suggestive evidence” are designated as T3 (tier 3 evidence), details in section 2.

We found our ASD candidate genes (i.e. top 1000 ranked genes) are significantly enriched in those ASD-associated gene lists using Fisher’s exact tests. The results are shown in Table 3.2. The candidate genes show a strong enrichment in T1 genes (OR=26.89, *p*-value = 1.36E-65). We also observe a trend of decreasing enrichment odds ratio from T1 to T3. We found after top 2000 genes, the signal for enrichment is not significant, this is consistent with our expectation of around 1000 genes to be ASD risk genes [6]. The other method (”2016NN”) included most SFARI genes as training set, therefore we were not able to include that for comparison in this analysis.

Table 3.2: Enrichment of candidate genes in SFARI ASD genes

	lb	rb	T1_*p*-value	T1_OR	T2_*p*-value	T2_OR	T3_*p*-value	T3_OR
1	1	1000	1.36E-65	26.89	2.83E-27	8.26	1.35E-20	4.04
2	1001	2000	4.02E-4	2.69	3.42E-3	2.12	4.52E-7	2.35
3	2001	3000	1.2E-1	1.54	1.56E-2	1.87	1.25E-5	2.13
4	3001	4000	1.2E-1	1.54	3.07E-2	1.75	1.59E-1	1.24

3.2.3 Enrichment of common SNP heritability of ASD around top ranked genes

The positive seed genes that guide our prioritization were mostly identified from disruptive de novo variations in patients with ASD [3]. Despite the strong contribution of de novo genetic variations to ASD, common variant-related polygenic risk also play important roles in conferring ASD risk. We reason that: if our predicted candidate gene list is enriched for ”true” ASD risk genes, then more of the genome-wide association signal of ASD would be concentrated around these genes. Here we sought to evaluate the strength of evidence from common genetic variation to our candidate genes using stratified Linkage disequilibrium score regression (LDSC) [28]. We assessed enrichment of the common SNP heritability of ASD within or near the candidate genes. We map genes to SNPs by including the SNPs located within a distance of 10 kb up or downstream of the transcription start sites.

We evaluated the enrichment of SNP-level heritability from the largest to date ASD GWAS results [30] within or near the candidate genes predicted by 2016NN and our BN method using stratified LDSC. As shown in Fig 3.2, both our method BN and 2016NN method are strongly enriched for ASD heritability and exceeded significance level under several cutoffs. We found our method to be more significantly enriched and this is not limited to top 1000 genes but continues to top 4000 genes. We stopped at 4000 since the enrichment is not significant afterwards.

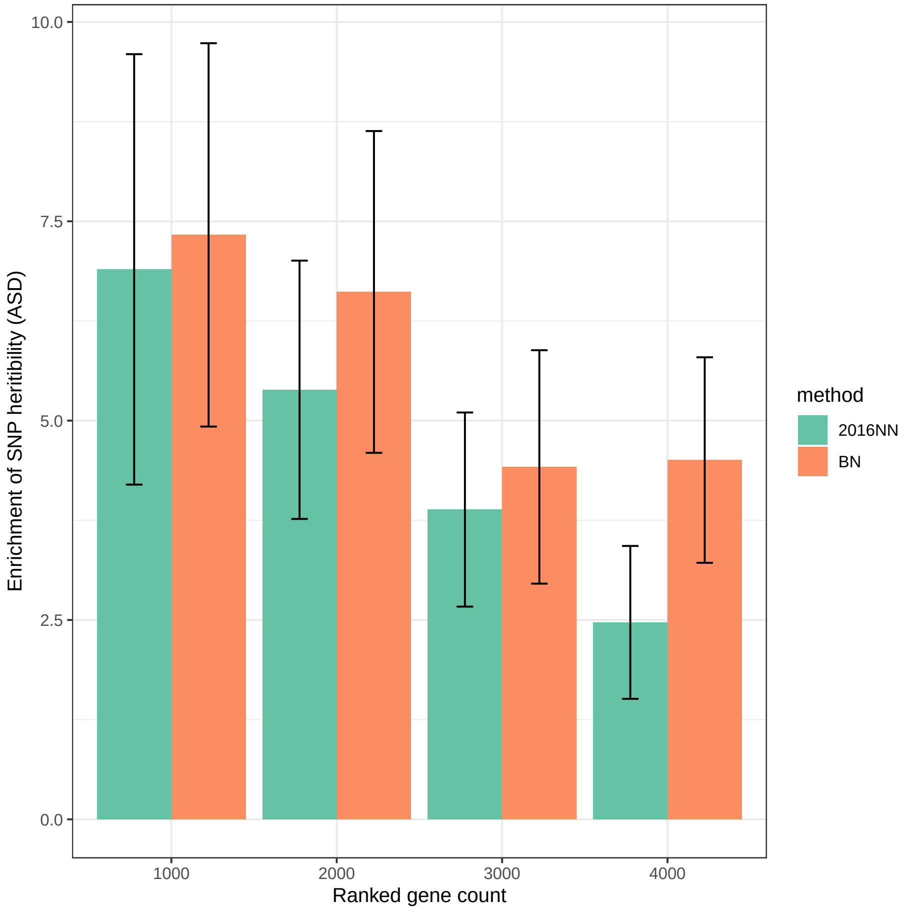

Figure 3.2: Enrichment of ASD SNP heritability within or near prioritized genes by 2016NN and BN (error bars indicate the standard errors)

3.3 ASD-associated genetic changes in the spatiotemporal development of the brain.

3.3.1 Tissue Specific Expression Analysis

While ASD is a disease mainly manifested in the brain, studying the relationship between all available tissues to ASD gave us the opportunities to explore the complex genetic foundations that may contribute to this disorder and identify potential peripheral tissues that are relevant to ASD research. The advantages of finding other tissues includes the large sample sizes and multiple replication opportunities, as target-tissue brain studies are often limited to post-mortem sampling and relatively small sample sizes.

Therefore, using the candidate genes (top 1000 genes) from our prediction, we explore the tissue specific expression pattern of these genes using data from different human tissues (from GTEx project [33]). For each tissue, a list of genes that are specifically expressed in this tissue was defined using *pSI* scores (see methods 2.6.1). Fisher's Exact tests were used to test for enrichment of candidate genes in each tissue specific gene lists, the results are shown in Fig 3.3. We found brain to be most enriched for candidate genes (p-value=5.58E-12, BH corrected p-value=1.42E-10, under *pSI* threshold = 0.5); we didn't observe other tissues to be significantly enriched for candidate genes.

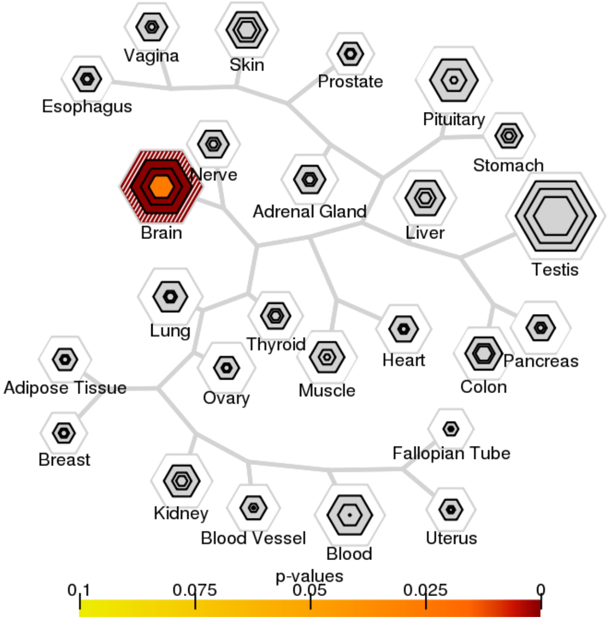

Figure 3.3: Candidate genes tend to be specifically expressed in brain tissue

3.3.2 Cell-type specific expression analysis

The cellular disruptions that lead to ASD is still largely unclear. There is a theory that disease-causing genes tend to be specifically expressed in the cell types disrupted by the disorder [34]. While this might not be true for all situations, it is possible for us to find enrichment of disease-relevant genes in the vulnerable cell types. Therefore, we study the cell-type expression specificity of our candidate genes to find potential cell types relevant to ASD [31, 34, 35].

Fig 3.4 shows over-representation of candidate genes in striatal medium spiny neurons and retina specific genes dervied from cell-type level gene expression data from mice (details in methods 2.6.3). Interestingly, we found literature support that defects in the striatum seem to specifically contribute to the motor, social and communication impairments seen in ASD patients. Previous reports disrupting striatum in mice leads to repetitive behaviors and social problems [52]. Brain imaging studies also showed that some parts of the striatum are enlarged in ASD patients [53]. Striatum was also suggested to shrink as children matures, but it keeps growing in autistic patients [54]. On the other hand, researchers have been using retina as an accessible window to understand brain wiring and functions. Retina is part of the central nervous system (CNS). It uses mainly glutamate and GABA to transmit and modulate visual signals and produces most neurotransmitters found in the brain. Therefore it is possible that the developmental disturbances that lead to ASD also affect retina functions. Indeed, there is report that adolescent male mice of VPA-induced ASD model have alterations in retinal function and protein expression compatible with those found in brain regions of other ASD models [55]. Also, it is reported that ASD patients perform better than usual in some visual tasks, and worse in others [56].

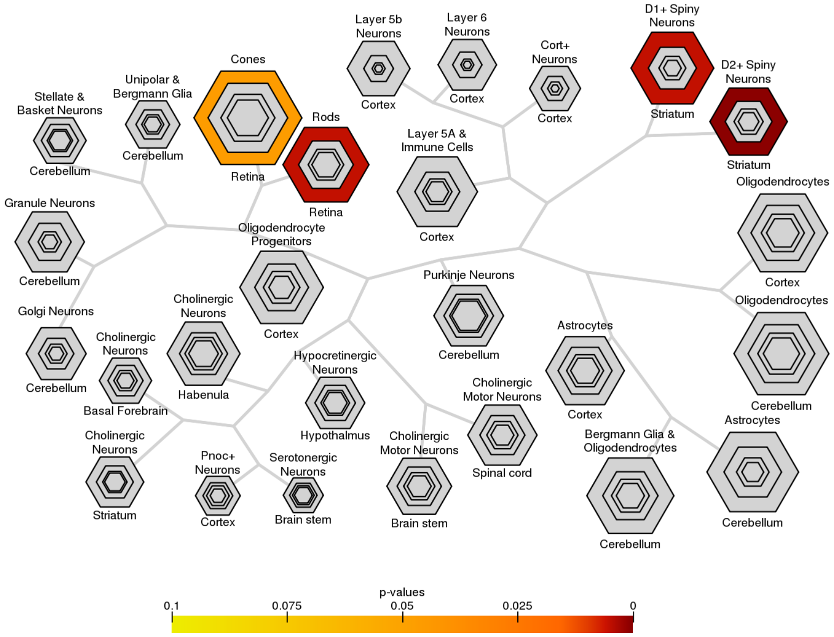

Figure 3.4: Candidate genes tend to be specifically expressed in striatal medium spiny neurons and retina cells

3.3.3 Brain regions and developmental window Specific Expression Analysis

The precise brain region and developmental stage vital to ASD is not clear yet. Therefore, we study candidate genes for enrichment in different brain region and developmental stages using transcriptome data from Brainspan [35].

We defined brain region- and developmental stage- specific gene lists using approach similar to above sections (methods 2.6.1). As shown in Fig 3.5, we found strong enrichment signals from candidate genes in early and mid fetal stage genes, also this enrichment was spread across all brain regions. Later in development, we found significant enrichment signals in cerebellum during mid-late childhood and cortex during young adulthood. This is consistent with the reported heterogeneity of ASD as abruptions from many brain regions might all contribute to ASD [12, 57].

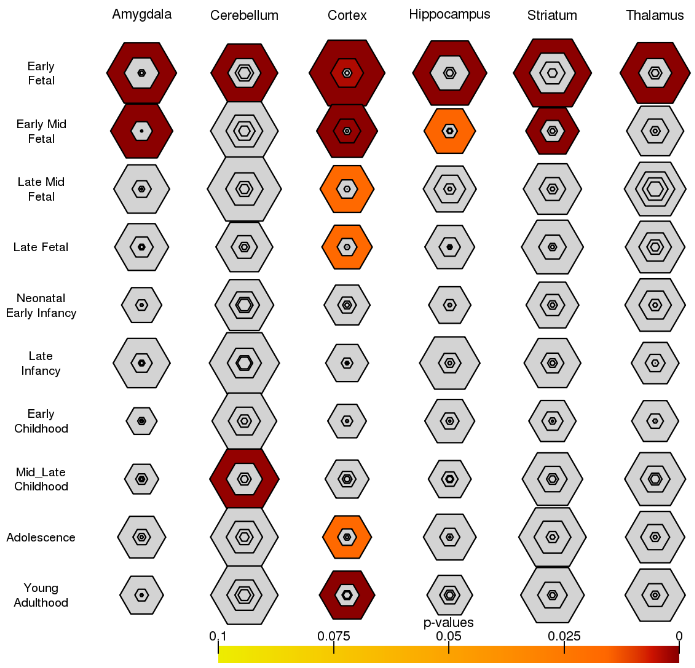

Figure 3.5: Candidate genes show strong enrichment signals in early and mid fetal stages across different brain regions

Chapter 4

Conclusion

4.1 Remarks and observations

In this study, we present a Bayesian model selection based framework to facilitate genome-wide ASD gene discovery. Our approach integrates diverse functional annotations and network information to obtain genome-wide prediction of ASD risk. The validity of this genome wide prediction is tested by 1) a recent exome sequencing study published not used in our initial analyses (Fig 3.1); 2) genes from a comprehensive database [7] that includes genes associated with ASD risk (not used in training process) (Table 3.2) and 3) enrichment of heritability from a recent GWAS study (Fig 3.2). Then we studied the gene expression patterns of top candidate genes and found strong enrichment in striatal medium spiny neurons and early developmental stages.

The major innovative points of our approach are: 1) integrate functional annotations and network information together in risk gene prediction, which is uncommon in published methods; 2) direct interpretability of final results as Bayesian posterior odds ratios. Our approach were able to achieve more accurate predictions and pinpoint 1000 top candidate genes of relatively high confidence compared to other methods (Fig 3.1, Fig 3.2).

A common difficulty in using Bayesian models is the need to specify priors. We explored full Bayesian treatment and found the specification of prior for diverse annotations hard to achieve and can suffer from subjective belief which makes it harder to justify. The use of flat/uninformative priors would be an option if there are large amounts of data, but given the limited size of data (e.g. 65 positive genes), we decide it is not an option for us. Instead, we chose to use empirical Bayes estimation in a parameteric model with analytic solutions. The usage of empirical Bayes estimation enabled us to learn the prior from seed genes we collected. This parametric formulation with conjugate prior also enables us to get estimates in computational efficient ways.

This Bayesian model has some advantages compared to Frequentist 'hypothesis-testing' approaches and 'black-box' machine learning models. Compared to Frequentist approaches that could only reject hypotheses and do not offer assessments of strength of evidence in favor of the hypotheses, Bayesian approach enables us to evaluate evidence about hypotheses. Also in contrast to "black-box" machine-learning predictors, Bayesian predictions are readily interpretable as they represent conditional probability relationships among information sources.

This framework also makes it flexible to adding new evidence to the existing model: adding a conditional independent evidence transfers to multiplying a Bayes factor summarizing the newly added feature. Thus, it is also useful for guiding an evolving model-building process.

4.2 Drawbacks and limitations

We used parametric model specification in our framework due to the computation efficiency and empirical data support. However, nonparametric model specifications tend to be more validity-robust. If our parametric model assumptions are "true", the parametric approach can yield less uncertainty in estimations. But in reality, we don't know the "true" model and some nonparametric approaches could successfully provide estimations without making strong assumptions about the underlying model. Contemporary computing resources and MCMC methods for integration approximate also have made nonparametric approaches more feasible.

Another limitation of our approach is its dependency on existing patterns in training genes. Our approach is more powerful to identify new disease candidate genes that similar to "known" disease genes. Thus, we interpret our results with caution: we can implicate candidate genes but we are not confident to exclude genes since they can lead to diseases through entirely unexpected mechanisms.

The functional annotations used in our current approach is limited to biological processes related to ASD and generic gene-level annotations. We didn't include epigenomics information, which might provide important link between genetic variations and diseases. Also, we didn't include tissue or cell type specific annotations in the prediction process. These could be explored in future models and might contribute to accuracy gains in predictions.

4.3 Future work

In the future, with the expansion of both genomic data and epigenomic data, the identification of risk genes could be greatly improved by expanding our framework to include more annotations. Instead of selecting the annotations individually as we have done in this study, the recent development of automatic feature selection approaches (e.g. autoML [58]) might provide opportunities to integrate more annotations with ease. Nonparametric model specifications can also be explored to provide more robust estimations.

It is our hope that this framework can contribute to advancing our understanding of the biology of ASD and the goal of guiding effective diagnosis and intervention of ASD.

References

1. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388,** 1545–1602 (Oct. 2016).

2. Lyall, K. *et al.* The Changing Epidemiology of Autism Spectrum Disorders. *Annu Rev Public Health* **38,** 81–102 (Mar. 2017).

3. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87,** 1215–1233 (Sept. 2015).

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

5. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (Nov. 2014).

6. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9,** e1003671 (2013).

7. *SFARI* http://gene.sfari.org (2020).

8. Doncheva, N. T., Kacprowski, T. & Albrecht, M. Recent approaches to the prioritization of candidate disease genes. *Wiley Interdiscip Rev Syst Biol Med* **4,** 429–442 (2012).

9. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol* **24,** 537–544 (May 2006).

10. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37,** W305–W311 (2009).

11. Liu, L. *et al.* DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol Autism* **5,** 22 (Mar. 2014).

12. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* **19,** 1454–1462 (Nov. 2016).

13. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180,** 568–584 (Feb. 2020).

14. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).

15. Consortium, G. O. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research* **47,** D330–D338 (2019).

16. Gilman, S. R. *et al.* Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70,** 898–907 (2011).

17. Tong, H., Faloutsos, C. & Pan, J.-Y. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems* **14,** 327–346 (2008).

18. Wang, Q. *et al.* A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nature neuroscience* **22,** 691–699 (2019).

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

20. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9,** e1003709 (2013).

21. Robbins, H. *An Empirical Bayes Approach to Statistics* in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (University of California Press, Berkeley, Calif., 1956), 157–163. https://projecteuclid.org/euclid.bsmsp/1200501653.

22. Efron, B. & Morris, C. Stein's paradox in statistics. *Scientific American* **236,** 119–127 (1977).

23. Casella, G. An introduction to empirical Bayes data analysis. *The American Statistician* **39,** 83–87 (1985).

24. Wolpert, R. L. *beta-binomial* https://www2.stat.duke.edu/courses/Spring16/sta532/lec/ebhb.pdf (2020).

25. Zhang, Y.-Y., Rong, T.-Z. & Li, M.-M. The empirical Bayes estimators of the mean and variance parameters of the normal distribution with a conjugate normal-inverse-gamma prior by the moment method and the MLE method. *Communications in Statistics-Theory and Methods* **48,** 2286–2304 (2019).

26. Cáceres, J. J. & Paccanaro, A. Disease gene prediction for molecularly uncharacterized diseases. *PLoS computational biology* **15,** e1007078 (2019).

27. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47,** 1236 (2015).

28. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47,** 1228 (2015).

29. *LDSC Git Repo* https://github.com/bulik/ldsc (2020).

30. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics* **51,** 431–444 (2019).

31. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic acids research* **38,** 4218–4230 (2010).

32. Dougherty, J. *SI code* www.bactrap.org (2020).

33. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45,** 580–585 (2013).

34. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *Journal of Neuroscience* **34,** 1420–1431 (2014).

35. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic acids research* **41,** D996–D1008 (2012).

36. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146,** 247–261 (2011).

37. Gompers, A. L. *et al.* Germline Chd8 haploinsufficiency alters brain development in mouse. *Nature neuroscience* **20,** 1062 (2017).

38. Pirooznia, M. *et al.* SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **28,** 897–899 (2012).

39. Ji, X., Kember, R. L., Brown, C. D. & Bućan, M. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proceedings of the National Academy of Sciences* **113,** 15054–15059 (2016).

40. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6,** e1001154 (2010).

41. Steinberg, J., Honti, F., Meader, S. & Webber, C. Haploinsufficiency predictions without study bias. *Nucleic acids research* **43,** e101–e101 (2015).

42. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

43. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33,** D514–D517 (2005).

44. Petrovski, S. *et al.* The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet* **11,** e1005492 (2015).

45. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).

46. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).

47. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).

48. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell reports* **6,** 1139–1152 (2014).

49. Of the Psychiatric Genomics Consortium, C.-D. G. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381,** 1371–1379 (2013).

50. Mahmoudi, E. & Cairns, M. MiR-137: an important player in neural development and neoplastic transformation. *Molecular psychiatry* **22,** 44–55 (2017).

51. Bayés, À. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience* **14,** 19–21 (2011).

52. Fuccillo, M. V. Striatal circuits as a common node for autism pathophysiology. *Frontiers in neuroscience* **10,** 27 (2016).

53. Nickl-Jockschat, T. *et al.* Brain structure anomalies in autism spectrum disorder—a meta-analysis of VBM studies using anatomic likelihood estimation. *Human brain mapping* **33,** 1470–1489 (2012).

54. Langen, M. *et al.* Changes in the developmental trajectories of striatum in autism. *Biological psychiatry* **66,** 327–333 (2009).

55. Guimarães-Souza, E. M., Joselevitch, C., Britto, L. R. G. & Chiavegatto, S. Retinal alterations in a pre-clinical model of an autism spectrum disorder. *Molecular autism* **10,** 19 (2019).

56. Little, J.-A. Vision in children with autism spectrum disorder: a critical review. *Clinical and Experimental Optometry* **101,** 504–513 (2018).

57. Dinstein, I., Heeger, D. J. & Behrmann, M. Neural variability: friend or foe? *Trends in cognitive sciences* **19,** 322–328 (2015).

58. Vitsios, D. & Petrovski, S. Mantis-ml: Disease-Agnostic Gene Prioritization from High-Throughput Genomic Screens by Stochastic Semi-supervised Learning. *The American Journal of Human Genetics* **106,** 659–678 (2020).