On the Prediction of Risk for Autism From Common Variants
– ADA Report

Lingxue Zhu [*]
Advisors: Kathryn Roeder [*] and Bernie Devlin [†]

September 23, 2015

1 Introduction

Autism spectrum disorder (ASD) is a set of heterogeneous neuro-developmental disorders causing significant social, communication, and behavioral deficits and challenges. ASD is known to be highly heritable, and it is important to understand its genetic architecture in order to develop more efficient and targeted treatments. However, although many studies have been conducted, our understanding of ASD is still very limited.

Genome-wide association studies have revealed a handful of risk genes and rare mutations that are associated with ASD. However, these mutations explain only a small fraction of the variability of ASD. Therefore, their predictive power is very limited. This phenomenon is referred to as "missing heritability", which also occurs in many other complex traits.

On the other hand, Klei *et al.* [9] revealed that a substantial amount of variability of ASD is contributed by additive effects of common genetic polymorphism (i.e., common SNPs), although not a single risk SNP has been identified. This suggests that ASD is affected by a myriad of common variants, each with very small effect. More recently, Gaugler *et al.* [15] show that around 50% of the heritability of ASD can be explained by common variants. These results raise a natural question of whether more predictive power for ASD can be extracted from common variants.

Meanwhile, multiple studies have also confirmed that damaging *de novo* mutations (i.e., mutations that are not inherited from parents), including loss-of-function (LoF) mutations and copy-number-variants (CNVs), are associated with ASD. Especially, in a recent largest Autism study, De Rubeis *et al.* [4] find that about 14% of probands carry damaging *de novo* variants, which is two times higher than the average in the population. Although these mutations are too rare to be useful for prediction, it is still of great interest to analyze their influence on ASD risks. Scientists have long hypothesized that Autistic children with damaging *de novo* mutations have lower risks inherited via the common variants. This raises another question that whether we can confirm this effect in the predicted ASD risk.

In this project, we aim to answer the above two questions. We build a prediction model for ASD based on common SNPs using a linear mixed-effect model, combined with a SNP selection procedure that incorporates our prior knowledge as well as a weighted Lasso selection. We analyze the model predictive power, and measure the effects of damaging *de novo* mutations in the predicted risk. We reveal that accurate prediction of ASD under current sample size is difficult. Nevertheless, we repeatedly observe significant effects of damaging *de novo* mutations in different data sets. This project provides a potentially valuable exploration of ASD risk prediction, as well as new insights of the genetic mechanism of ASD.

[*] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
[†] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260

2 Data

2.1 Data Sets

Our analysis is based on the three SNP-genotyped data sets: Simons simplex collection (SSC) [6], Autism genome project (AGP) [13], and Health ABC study (HABC) [1] (Table 1). All of these subjects are genotyped on over $810,000$ common SNPs, which form a huge pool of candidate covariates for prediction. Both SSC and AGP include ASD families with at least one proband (i.e., affected child), where SSC data contains only trio families with only one proband and parents are known to be unaffected, while AGP data contains families with multiple probands, sometimes even affected parents. Previous study has shown that the heritability of ASD is higher in AGP families [9], and this pattern will be confirmed in our model in later sections. On the other hand, HABC contains unrelated control subjects, and will serve as the control group in our analyses.

Data set	Total individuals	Family				
		Total	Trio	Quad	3-proband	2 fa.+2 mo.+2 prob.
SSC	3555	1185	1185	0	0	0
AGP	7880	2611	2571	36	1	3
HABC	1663	1663	–	–	–	–

Table 1: Three data sets: SSC, AGP, and HABC, where SSC and AGP contain ASD families, and HABC contains unrelated subjects. Trio families have only one proband (i.e., affected child), and quad families have two probands.

2.2 Damaging Rare Mutations

One of the goals of our project is to analyze the effects of damaging rare mutations. For this purpose, we divide the probands (i.e., affected children) into two groups, one with damaging rare mutations mutations, and the other one without. In the following sections, we will compare the predicted ASD risks between these two groups, and measure the difference. The damaging rare mutations are defined as follows:

SSC data In SSC data, the detailed *de novo* information is available, where *de novo* mutations are the ones that are not inherited from parents. A mutation is classified to be damaging if it is either (i) a *de novo* loss-of-function (LoF) mutation on at least one of the 50 risk ASD genes identified in [16]; or (ii) a *de novo* CNV that is recurrent or involves at least 14 genes. 69 probands in the SSC families carry at least one of these damaging *de novo* mutations.

AGP data Since AGP data has less detailed *de novo* information, we focus on the damaging CNV mutations that are either (i) considered to be pathogeni in previous AGP study; or (ii) *de novo* and with sive greater than 1 megabase. 66 probands in the AGP families carry at least one of these damaging CNV mutations.

2.3 Head Circumference Deviance

Head circumference deviaion (HC.DEV), after controlling for gender, age, height, weight, genetic ancestry, etc, is found to be significantly associated with ASD status [12]. Therefore, HC.DEV can be viewed as as a continuous trait that is associated with ASD, and will be used the evaluate the prediction power of our model. Among all of the 3555 SSC individuals, 3236 have measured HC.DEV (Table 2).

	Fathers	Mothers	Probands	Total individuals	Markers
Original SSC data	1185	1185	1185	3555	812,621
With HC.DEV	1050	1053	1133	3236	812,621

Table 2: SSC data used for HC.DEV analysis

3 G-BLUP Prediction of ASD Risk

3.1 Linear Random Effects Model

The G-BLUP prediction that is widely used in many genetic studies (for example, [7, 5]) is based on a linear random effects model. For every individual i, we model its phenotype y_i as

$$y_i = g_i + \epsilon_i\,,$$

where g_i is the "genomic value", which represents the expected value of phenotype given the genomic effects, and ϵ_i is the residual effects that accounts for any non-genomic effects. We assume that g_i and ϵ_i are independent. Stacking all the equations for individuals 1 to n, we have

$$\mathbf{y} = \mathbf{g} + \epsilon\,.$$

Without loss of generality, we assume that $\mathbf{y}$ is centered with mean 0. We model the genetic effects as random effects, so that $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2), \epsilon \sim N(0, \mathbf{I}\sigma_\epsilon^2)$, and $\mathbf{g} \perp \epsilon$. Therefore,

$$\mathbf{y} \sim N(0, \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2)\,. \tag{1}$$

The matrix $\mathbf{G}$ has diagonal values of 1, and is interpreted as the genetic relationship matrix (GRM), where $G_{i,j}$ measures the relatedness of individual i and j in the sense that

$$\text{cov}(y_i, y_j) = \begin{cases} \sigma_g^2 G_{i,j} & \text{if } i \neq j \\ \sigma_g^2 + \sigma_\epsilon^2 & \text{if } i = j \end{cases}\,.$$

One way to obtain $\mathbf{G}$ is from the pedigree information in family data. However, in more general situations where individuals have no known relatedness, the GRM is often estimated using all genotyped markers by

$$\mathbf{G} = \frac{1}{p}\mathbf{W}\mathbf{W}^T\,,$$

where$\mathbf{W}_{n\times p}$ is the standardized genotype matrix with $W_{ij} = \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}$, and $x_{ij} \in \{0, 1, 2\}$ is the number of copies of the reference allele for the ith SNP of the jth individual, and p_i is the frequency of the reference allele [7].

Finally, the prediction for model (1) can be obtained as follows. Suppose we have two groups of individuals $\mathbf{y}_1, \mathbf{y}_2$, where only $\mathbf{y}_1$ is observed and we want to predict $\mathbf{y}_2$. Note that

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{G}_{11}\sigma_g^2 + \mathbf{I}_1\sigma_\epsilon^2 & \mathbf{G}_{12}\sigma_g^2 \\ \mathbf{G}_{21}\sigma_g^2 & \mathbf{G}_{22}\sigma_g^2 + \mathbf{I}_2\sigma_\epsilon^2 \end{pmatrix}\right)\,,$$

where $\mathbf{G} = \frac{1}{p}WW^T = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}$ is the corresponding partition of GRM. If $\mathbf{G}$ is known, then it is well known that the best linear unbiased predictor (BLUP) of $\mathbf{y}_2$ is

$$\hat{\mathbf{y}}_2 = \mathbb{E}(\mathbf{y}_2|\mathbf{y}_1) = \sigma_g^2\mathbf{G}_{21}[\sigma_g^2\mathbf{G}_{11} + \sigma_\epsilon^2\mathbf{I}_1]^{-1}\mathbf{y}_1 = \mathbf{G}_{21}[\mathbf{G}_{11} + \lambda\mathbf{I}_1]^{-1}\mathbf{y}_1\,, \tag{2}$$

where $\lambda = \sigma_\epsilon^2/\sigma_g^2$. Note that $\mathbb{E}(\mathbf{y}_2|\mathbf{y}_1) = \mathbb{E}(\mathbf{g}_2|\mathbf{y}_1)$, so essentially we are estimating the genomic values. When $\mathbf{G}$ is estimated by $\frac{1}{p}\mathbf{W}\mathbf{W}^T$, the plug-in estimator of $\hat{\mathbf{y}}_2$ is referred to as Genomic-BLUP (G-BLUP) [5].

3.2 Heritability

Under the framework of model (1), the narrow sense heritability (h^2) is defined as the proportion of total phenotypic variation that is due to additive genetic factors:

$$h^2 = \frac{\text{var}(g)}{\text{var}(y)} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2} .$$

This is an important population parameter, since it indicates the proportion of information of y that can be obtained from additive effects of SNPs. The estimated $\hat{h}^2$ can be obtained by plugging in the REML estimation of $\hat{\sigma}_g^2, \hat{\sigma}_\epsilon^2$.

However, the above model is only well-defined for continuous traits. When coming to binary responses $\mathbf{y}$ in our case-control study, the following two issues need to be accounted for:

1. Ancestry control: variance estimation in case-control studies is especially sensitive to population structures. To avoid the confounding factors of ancestry structures, it is common to first obtain an estimation of ancestry among all samples, and restrict the analysis to a homogeneous group of subjects. In many cases, we will restrict our analysis to the European samples in the data.

   The genetically-estimated ancestry can be obtained using GemTools [8], where a random set of 20,000 SNPs are chosen for ancestry determination, and a random sample of 500 individuals are chosen to be bases. The identified European individuals in the three data sets are summarized in Table 3.

Data set	All individuals	European individuals	Ancestry estimated from
SSC	1185 probands	751 probands	SSC probands+HABC controls
AGP	2565 probands	2193 probands	AGP probands
HABC	1663 controls	1317 controls	SSC probands+HABC controls

Table 3: European ancestry in three data sets.

2. Liability scale: in a case-control study where $y \in \{0, 1\}$, the above estimated variance corresponds to the variation on the observed 0-1 scale (h_o^2). Under the threshold-liability model, the heritability in the unobserved continuous liability scale (h_l^2) can be obtained by a linear transformation

$$h_l^2 = h_o^2 \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)} ,$$

   where K is the population prevalence of the disease, z is the value of $N(0, 1)$ density at threshold point, and P is the proportion of cases in the collected data [14].

3.3 Predicting Head Circumference Deviance

Before diving into the binary case-control world, we first conduct some explorations of the model using a continuous trait that is closely related to ASD: head circumference deviance (HC.DEV). Head circumferences (HC) can be affected by many factors such as gender, age, height, weight, genetic ancestry, etc. After controlling for all these covariates, the deviation of HC from its expectation is found to be significantly associated with ASD status [12]. Therefore, HC deviation (HC.DEV) is considered as a continuous trait that is associated with ASD. Naturally, it is distributed as a normal distribution with mean 0, which fits the assumption of model (1).

Recall that 3236 SSC individuals have measured HC.DEV, which are used through out the analysis in this section (Table 2). All of the 812,621 markers have already been QCed with genotype missing rate < 0.05 on AGP+SSC data. We have also checked the missing rate within the SSC data, and find only 47 SNPs with missing rate between $0.05 \sim 0.07$, so we keep all these markers in the HC.DEV analysis. The heritability estimated from all of the 3236 individuals is 55.07%(se=3.42%). In the following subsections, we apply the G-BLUP model under two scenarios, to further evaluate its predictive power.

3.3.1 Predicting Probands from Parents

First, we use the 1050 fathers and 1053 mothers as training data, apply the G-BLUP model, and obtain the predicted HC.DEV on the 1133 probands. The correlation between observed HC.DEV (y) and predicted HC.DEV ($\hat{y}$) of these 1133 probands is 0.3856 (Fig.1, left panel).

In addition, among these 1133 probands, 27 of them have damaging *de novo* CNV mutations, and 64 have damaging *de novo* mutations (CNV or LoF), as defined in section 2.2 (see Table 4).

	Damaging *de novo* LoF	Damaging *de novo* CNV	All damaging *de novo*
# of probands	37	27	64

Table 4: Probands with damaging *de novo* mutations in SSC data, with available HC.DEV.

We then compare the predicted HC.DEV of these 27 and 64 probands against others (Fig.1, middle and right panel). As indicated by the one-sided Mann-Whitney-Wilcoxon (MWW) tests, in both comparisons, the predicted HC.DEV for probands with mutations are significantly lower than other probands (p-value being 0.02 and 0.04, respectively). Recall that HC.DEV can be treated as a continuous proxy to ASD risk, this result supports the hypothesis that probands with damaging *de novo* mutations have lower risk inherited from common variants.

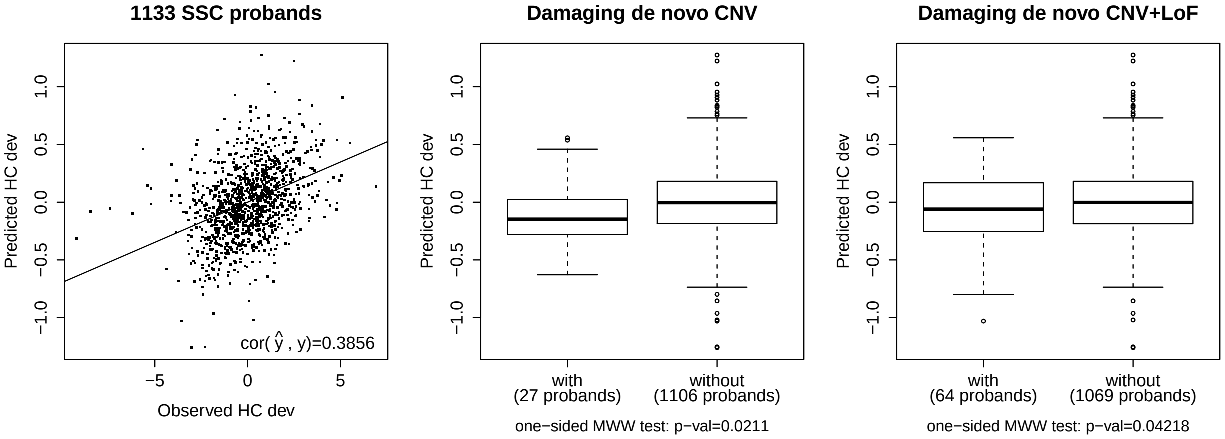

Figure 1: G-BLUP prediction of HC.DEV for 1133 SSC probands, using all parents as training data. ***Left***: Scatter plot of observed HC.DEV and predicted HC.DEV for 1133 probands; the straight line is a fitted linear regression line. ***Middle***: Boxplot of 27 probands with risk *de novo* CNVs, and the other 1106 probands; the *one-sided* Mann-Whitney-Wilcoxon (MWW) test shows p-value 0.0211. ***Right***: Boxplot of 64 probands with risk *de novo* CNV or LoF, and the other 1069 probands; the *one-sided* MWW test shows p-value 0.04218.

3.3.2 Predicting Probands from Probands

In order to check whether the G-BLUP model has predictive power among un-related individuals, we also try to randomly select 1000 probands as training data, and obtain the predicted HC.DEV on the remaining 133 probands. This is repeated 500 times, and in each time, we check the correlation between observed and predicted HC.DEV (cor$(y, \hat{y})$) for the 133 testing probands. The resulting average correlation is 0.0572(se=0.0855). In addition, among the 500 repetitions, only 129 of them showed negative correlations, and a sign test on these correlations gives p-value < 2.2e-16. In other words, the G-BLUP model still has predictive power among un-related individuals, although very limited. This also indicates that the high predictive power in section 3.3.1 is mostly obtained from the pedigree relatedness.

3.4 Predicting ASD Risk

After applying G-BLUP on HC.DEV, we continue to apply the model on the binary outcomes (case/control), and try to predict ASD risk. We first get the prediction of G-BLUP on SSC data, under 2 scenarios. The prediction on AGP data will be explained in section 4.

3.4.1 ASD Risk of SSC Probands: HABC Controls

Data pre-processing To avoid the confounding factor of ancestry, we restrict our samples to the European ancestry group identified in Table 3, including 751 SSC probands and 1317 HABC controls. The 833,050 SNPs in the SSC+HABC data set has been QCed for genotyping missing rate < 0.05. We also remove 55593 SNPs with MAF> 0.05, where the allele frequencies are calculated using SSC parents and HABC controls. Finally, there are 777,457 SNPs used in the analysis.

To further remove the confounding factors from ancestry, we not only restricting our samples to European ancestry, but also include the ancestry eigenvectors in our G-BLUP model to further remove any remaining ancestry effects. More precisely, in the G-BLUP model, we also add the 6 significant eigen vectors ($EV_j, j = 1, ..., 6$) identified by GemTools [8] as fixed effects:

$$y = \mu + \sum_{j=1}^{6} \beta_j EV_j + u + e\,,$$

where μ adjusts for the mean (intercept), u is the genetic random effects, and e is the Gaussian noise. Therefore, u is the genetic score for ASD risk after adjusting for ancestry effects, and when we refer to predicted ASD risk, we always mean the predicted random genetic effects $\hat{u}$.

Results The heritability estimated from all of the 2068 individuals, after accounting for the 6 eigen vectors, is 59.10%(se = 15.54%) in observational scale, and 35.26%(se = 9.25%) in liability scale, where the prevalence is taken to be 0.01.

To obtain the predicted risks, we randomly divide the 2068 individuals into 5 folds. Every time, we obtain the predicted ASD risk ($\hat{u}$) on one fold using the other 4 folds as training data, and repeat this for every fold. We report the following results based on the predicted ASD risk:

(i) ROC curve: imagine we set a threshold of ASD risk t, and classify all individuals with $\hat{u} > t$ as affected, and all individuals with $\hat{u} < t$ as unaffected. We check the true positive rate (TPR) and false positive rate (FPR) of this classification. As the threshold t moves from small to large, the full ROC curve can be obtained (Fig.2, left panel).

(ii) Effects of damaging *de novo* CNV mutations: we compare the $\hat{u}$ for the 20 probands with damaging *de novo* CNVs, against the other 731 probands (Fig.2, middle panel). As expected, we observe that the probands with such mutations have significantly lower predicted ASD risks (p-value=0.02).

(iii) Effects of all damaging *de novo* mutations: we compare the $\hat{u}$ for the 52 probands with damaging *de novo* CNV or LoF mutations, against the other 699 probands (Fig.2, right panel). Again, we observe that the probands with such mutations have lower predicted ASD risks, but the difference is not statistically significant.

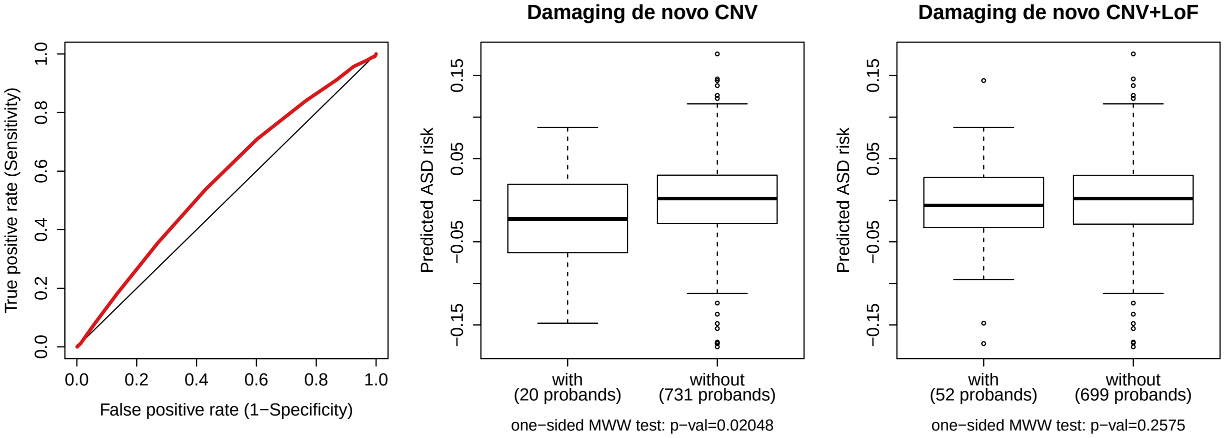

Figure 2: 5-fold G-BLUP prediction of ASD risk for 751 SSC probands and 1317 HABC controls. ***Left***: ROC curve of predicted risk. ***Middle***: Boxplot of 20 SSC probands with risk *de novo* CNVs, and the other 731 probands; the *one-sided* Mann-Whitney-Wilcoxon test shows p-value 0.02048. ***Right***: Boxplot of 52 probands with risk *de novo* CNV or LoF, and the other 699 probands; the *one-sided* MWW test shows p-value 0.2575.

3.4.2 ASD Risk of SSC Probands: Pseudo Controls

Instead of using HABC controls, another option is to use pseudo controls for prediction. A pseudo control contains all alleles that are not transmitted to the proband. This is known to be able to effectively control for false discoveries due to population structures, but it can also eliminate certain genetic signals.

Data pre-processing We use all of the 1185 SSC probands and their pseudo controls. We do not restrict to European ancestry since pseudo controls already account for the ancestry factor. We start from the 777,457 SNPs in section 3.4.1, and further remove those SNPs with genotype missing rate > 0.05 within the 1185 pairs of cases and pseudo controls. Finally, there are 754,060 SNPs remained for the following analysis. Unlike in section 3.4.1, we directly apply the G-BLUP model to the 1185 pairs of cases and pseudo controls, without regressing out any ancestry vectors.

Results The estimated heritability of the 1185 pairs of cases and pseudo controls hits the 0% boundary in observational scale (se=13.14%). Given this, we can expect the predictive power for ASD will be very limited when using pseudo controls. Again, we randomly divide the 1185 pairs into 5 folds, each with 237 pairs. Every time, we obtain the predicted ASD risk on one fold using the other 4 folds as training data, and repeat this for every fold. In the similar manner as in section 3.4.1, we obtain:

(i) ROC curve of the 1185 pairs of cases and pseudo controls (Fig.3, left panel). As expected, the predictive power is much less than using HABC controls. In fact, for every individual, the prediction is almost the same as random guessing.

(ii) Effects of damaging *de novo* CNV mutations: we compare the predicted risk for the 29 probands with damaging *de novo* CNVs, against the other 1156 probands (Fig.3, middle panel). We still observe that the probands with such mutations have somewhat significantly lower predicted ASD risks (p-value=0.07).

(iii) Effects of all damaging *de novo* mutations: we compare the predicted risk for the 69 probands with damaging *de novo* CNV or LoF, against the other 1116 probands (Fig.3, right panel). Again, we observe that the probands with such mutations have lower predicted ASD risks, but the difference is not statistically significant any more.

Note that the probands in (ii) and (iii) are just those in section 2.2. Again, the predicted risk is centered at 0. But compared to section 3.4.1, the predicted ASD risk has much smaller scales (i.e., absolute values much closer to 0). This is because the heritability under this scenario is much lower than that in section 3.4.1, so the genetic score accounts for a much smaller proportion of the ASD risk. Another way to think of this is, when using pseudo controls, the variance of the genetic score (u) becomes much smaller, so $\hat{u}$'s become closer to 0.

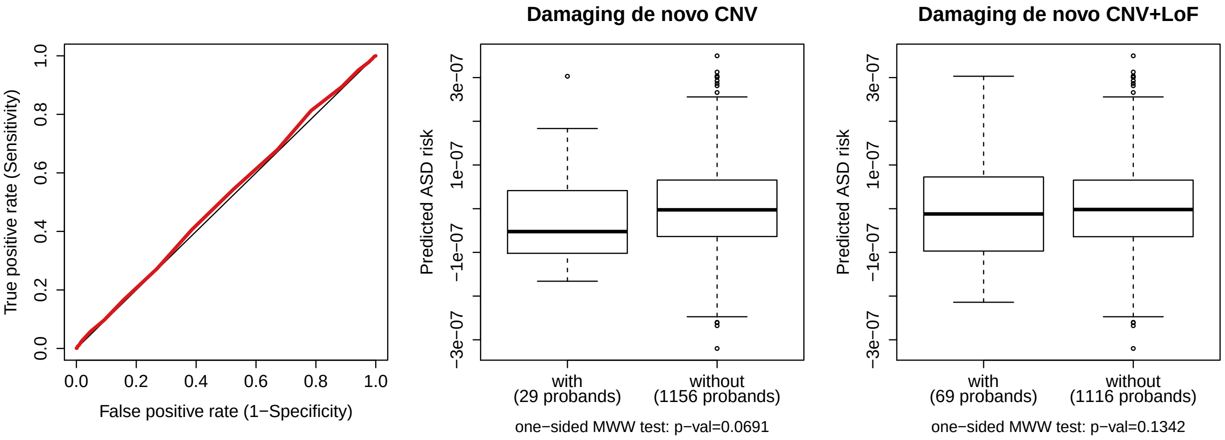

Figure 3: 5-fold G-BLUP prediction of ASD risk for 1185 pairs of SSC probands and pseudo controls. ***Left***: ROC curve of predicted risk. ***Middle***: Boxplot of 29 SSC probands with risk *de novo* CNVs, and the other 1156 probands; the *one-sided* Mann-Whitney-Wilcoxon test shows p-value 0.0691. ***Right***: Boxplot of 69 probands with risk *de novo* CNV or LoF, and the other 1116 probands; the *one-sided* MWW test shows p-value 0.1342.

4 Improving G-BLUP Prediction

4.1 A linear mixed effects model

The linear random effects model (1) assumes all SNPs to have small effects. However, Chatterjee *et al.* analyzed a fixed-effect model for risk prediction using common SNPs, where they include only a moderate amount of SNPs, each with moderately large effects [11]. Inspired by this, we propose a linear mixed effects model for risk prediction as follows:

$$\mathbf{y} = \mathbf{X}\gamma + \mathbf{W}\beta + \epsilon, \text{ where } \beta \sim N(0, \sigma_\beta^2 \mathbf{I}_{p-L}), \epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n), \beta \perp \epsilon \,. \tag{3}$$

In this model, we assume that there are L SNPs that are important and have larger effects than others, and we model them as fixed effects γ. Typically, $L \ll p$. For the other SNPs, we still model them as random effects. This can help to capture the relatedness among individuals. If there are indeed a subset SNPs with large effects, this model would perform better than model (1).

LMM prediction The prediction of model (3) is also straightforward. Suppose we have two sets of individuals $\mathbf{y}_1, \mathbf{y}_2$, where only $\mathbf{y}_1$ is observed and we want to predict $\mathbf{y}_2$. We will describe in details how to select the L SNPs in section 4.2, and for now, suppose we have chosen L SNPs to model as fixed effects. Note that we do not need to predict $\hat{\beta}$'s, so we write mode (3) in the following equivalent form:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \gamma + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix},$$

$$\text{where } \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \sim N\left(0, \sigma_g^2 \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}\right), \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n), \; \mathbf{u} \perp \epsilon \, .$$

Then the estimated best predictor of random effects can be easily obtained as

$$\hat{\mathbf{u}}_1 = \mathbf{G}_{11} \left[\mathbf{G}_{11} + \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_g^2} \mathbf{I}_1 \right]^{-1} (\mathbf{y}_1 - \mathbf{X}_1 \hat{\gamma}) \; ,$$

where the estimated fixed coefficients, $\hat{\gamma}$, is obtained from the SNP selection procedure, specifically, the Weighted-lasso estimation. We will elaborate this in the next section. Finally, the predicted $\hat{\mathbf{y}}_2$ is obtained by

$$\hat{\mathbf{y}}_2 = \mathbf{X}_2 \hat{\gamma} + \mathbf{G}_{21} \mathbf{G}_{11}^{-1} \hat{\mathbf{u}}_1 \, .$$

Note that if we replace $\hat{\gamma}$ by the MLE, this becomes the regular LMM prediction.

4.2 Selecting Large-effect SNPs

The remaining problem for model (3) is how to select the L risk SNPs for the fixed effects. Here, we carefully incorporate the following knowledge of risk genes and SNPs into the selection procedure:

(i) eQTLs of ASD risk genes: [16] have developed Transmission And De novo Association (TADA), which provides a powerful tool to discover promising risk genes for Autism using whole-exome sequencing (WES) data. More recently, [10] developed the DAWN algorithm that utilizes TADA score and gene co-expression network. This further increases the power of detecting Autism genes. Combining the results from both methods, we are able to get a list of risk genes for Autism with false discovery rate (FDR) lower than some threshold. After that, we can obtain the eQTLs of these risk genes by Braineac database [3]. These eQTLs form our first pool of candidate SNP predictors.

(ii) Schizophrenia risk SNPs: lately, 128 risk SNPs are identified for Schizophrenia in [18]. Since Schizophrenia is also a neurodevelopment disorders that often share risk factors with Autism, these 128 identified SNPs are also considered as promising candidate predictors.

Due to the issue of different genotyping platform, many of the eQTLs and Schizophrenia risk SNPs are not overlapped with our data. We will then include their proxy SNPs. The proxy SNPs are obtained using SNAP [2], with cut-off point of the correlation being 0.5.

Finally, the SNP specific procedures as follows:

1. Obtain 15,604 eQTLs from Braineac data set [3], including all brain tissues.

2. Among the 15,604 eQTLs, find available proxy SNPs in our data (around 8,000 SNPs).

3. Calculate the marginal logistic p-values of the proxy SNPs using the training AGP probands and controls. Keep those with p-value < 0.1.
4. Calculate the marginal logistic p-values of the proxy SNPs using the 1185 pairs of SSC probands and pseudo controls. Keep those with p-value < 0.1.
5. Also keep the SNPs with both p-values in step 3,4 < 0.3.
6. Check the p-values of these proxy SNPs in the Schizophrenia study, and keep those with p-value < 0.001.
7. Check the TADA and DAWN score of the genes of the eQTLs. Keep those with at least one of the scores < 0.3.
8. Keep the 77 proxy SNPs of the 128 Schizophrenia risk SNPs.
9. Combine all SNPs selected in step 3–7 and remove the highly correlated SNP pairs. Specifically, we check the LD using `--clump` option in PLINK, with 1000kb windows and cut-off correlation 0.7. Remove those that are clumped with other SNPs.
10. Apply weighted-Lasso on the pre-screened SNPs after step 9. Choose the smoothing parameter such that approximately 50, 250 or 1100 SNPs are selected. The weighted version of Lasso accounts for the fact that the individuals in our data set are correlated [17]. Specifically, we first transform $\mathbf{y}, \mathbf{X}$ to make the observations independent of each other, and then perform the standard Lasso:

$$\mathbf{y}^* = \Sigma^{-\frac{1}{2}}\mathbf{y} = \Sigma^{-\frac{1}{2}}\mathbf{X}\gamma + \Sigma^{-\frac{1}{2}}(\mathbf{g} + \epsilon) = \mathbf{X}^*\gamma + \epsilon^*, \quad \epsilon^* \sim N(0, I)$$
$$\hat{\gamma} = \arg\min_{\gamma} ||\mathbf{y}^* - \mathbf{X}^*\gamma||_2^2 + \lambda\,||\gamma||_1\ ,$$

where Σ is the variance of $\mathbf{y}$, which is estimated by REML using a linear random-effects model:

$$\hat{\Sigma} = \hat{\sigma}_g^2\mathbf{G} + \hat{\sigma}_e^2\mathbf{I}\,.$$

4.3 ASD risk of AGP Probands: HABC Controls

In this section, we use HABC controls to contrast with AGP probands for the risk prediction. Recall from section 3.4 that using HABC controls can provide more power in predictions than using pseudo controls.

Data pre-processing Similar as before, to control for ancestry confounders, we restrict our analysis to the European ancestry identified in Table 3, which includes 2193 AGP probands and 1317 HABC controls. Starting from 810,067 markers in the AGP+HABC data set, we further remove the SNPs with MAF>0.05 and genotype missing rate > 0.05, where the allele frequencies are calculated using AGP parents and HABC controls. Finally, there are 752,893 SNPs remained for the following analysis.

Results The heritability estimated from all 2193 probands and 1317 controls hits the boundary of 100% in observational scale, with se=7.62%. Similar to section 3.4, we randomly divide the 3510 individuals into 5 folds. Every time, we obtain the predicted ASD risk on one fold using the other 4 folds as training data, using the G-BLUP model and LMM model with selected 50, 250 and 1100 SNPs. The SNP selection procedure is illustrated in Fig.4. This is repeated this for every fold. Note that the numbers in Fig.4 are based on fold 1 in one realization, and will change slightly when we switch to other folds.

Finally, we calculate:

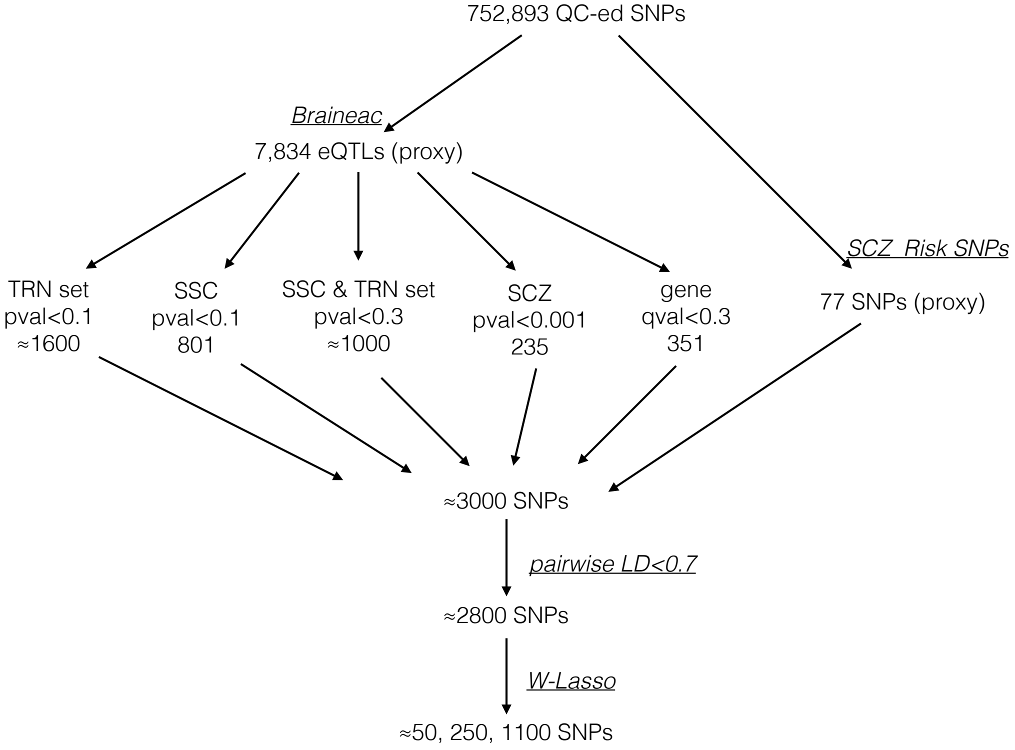

Figure 4: Procedure to select SNPs to be modeled as fixed effects in the linear mixed-effects model.

(i) ROC curve of the 2193 AGP probands and 1317 HABC controls, for 4 different models: G-BLUP, LMM model with selected 50, 250 or 1100 SNPs as fixed effects (Fig.5, left panel). Note that the predictive power is much better than in SSC data, which is consistent with previous research where AGP data was shown to have stronger genetic signals. In addition, G-BLUP model gives the best prediction, and fixed-effects model has very limited predictive powers.

(ii) Effects of damaging CNV mutations: we compare the predicted risk for the 61 AGP probands with damaging CNVs, against the other 2132 probands. Since as indicated by the ROC curve, G-BLUP gives the best prediction, so we use its predicted risk for this comparison (Fig.5, right panel). As expected, we find the group with these mutations have significantly lower predicted risks (p-value=0.02).

4.4 ASD risk of AGP Probands: Pseudo Controls

We also apply our model to pseudo controls versus AGP probands. As illustrated in section 3.4, we expect the predictive power to be weaker in this scenario than using HABC controls.

Data pre-processing We use all of the 2565 pairs of AGP probands and pseudo controls, since pseudo controls can already account for the ancestry effects. Starting from the 812,621 markers in the SSC+AGP data set, we further remove the SNPs with MAF>0.05, where the allele frequencies are calculated using AGP parents and SSC parents. We also remove the SNPs with genotype missing rate > 0.1 in the 2565 pairs of cases and pseudo controls. Finally, there are 739,129 SNPs remained for the following analysis.

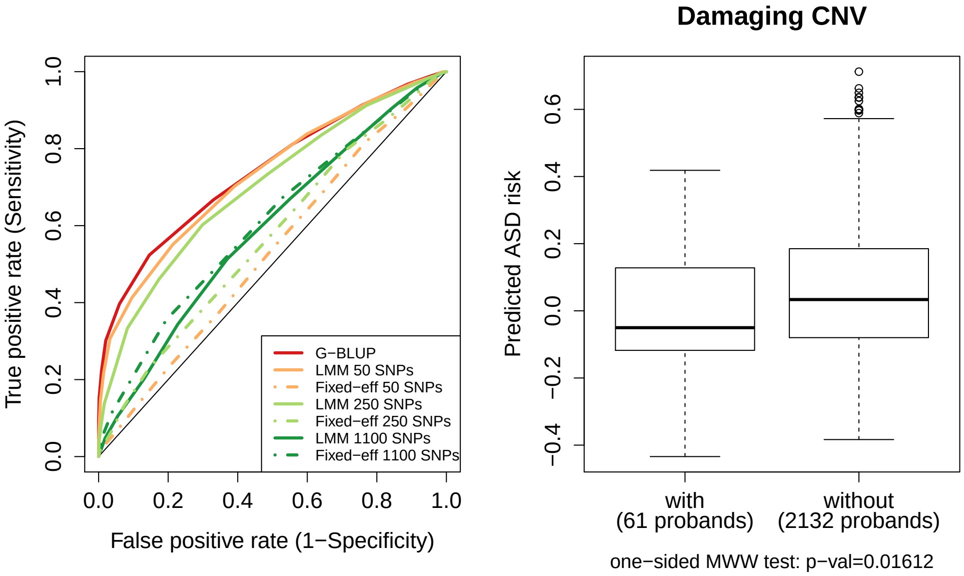

Figure 5: 5-fold G-BLUP prediction of ASD risk for 2193 AGP probands and 1317 HABC controls. ***Left***: ROC curve of predicted risk. ***Right***: Boxplot of 61 SSC probands with damaging CNVs, and the other 2132 probands, where the predicted risk is obtained by G-BLUP; the *one-sided* Mann-Whitney-Wilcoxon test shows p-value 0.01612.

Results The heritability estimated from all of the 2565 pairs of cases and pseudo controls is 54.52%(se = 6.60%) in observational scale, and 30.09%(se = 3.64%) in liability scale, where the prevalence is taken to be 0.01. Again, we randomly divide the 2565 pairs into 5 folds, each with 513 pairs. Every time, we obtain the predicted ASD risk on one fold using the other 4 folds as training data, using the G-BLUP model and LMM model with selected 50, 250 and 1100 SNPs. The SNP selection procedure is the same as in section 4.3, with the number of SNPs changed slightly (as illustrated in Fig.6). This is repeated this for every fold.

We calculate:

(i) ROC curve of the 2565 pairs of AGP probands pseudo controls, for 4 different models: G-BLUP, LMM model with selected 50, 250 or 1100 SNPs as fixed effects (Fig.7, left panel). As expected, the predictive power is less than using HABC controls, but is still better than in the SSC data. Again, note that G-BLUP model has the best prediction.

(ii) Effects of damaging CNV mutations : compare the predicted risk for the 66 AGP probands with damaging CNVs, against the other 2499 probands. Since as indicated by the ROC curve, G-BLUP gives the best prediction, so we use its predicted risk for this comparison (Fig.7, right panel). This time, the predicted risks for the group with these mutations are higher, but the large p-value in the two-sided MWW test indicates that this is not statistically significant.

5 Discussion and Future Work

In this project, we build a prediction model for ASD risk prediction from common variants using a linear mixed-effect model combined with feature selection. Our results illustrate that the prediction under current sample size is difficult, even after carefully utilizing prior information from risk genes, eQTLs and Schizophrenia studies. Our results also indicate that G-BLUP model can outperform

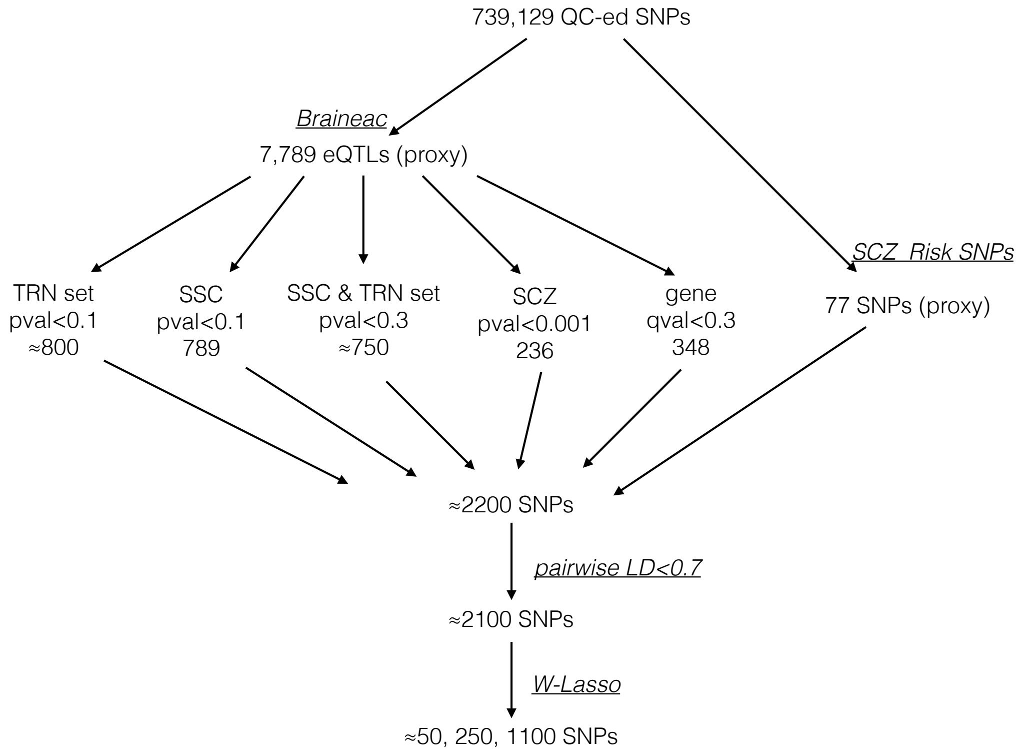

Figure 6: Procedure to select SNPs to be modeled as fixed effects in the linear mixed-effects model.

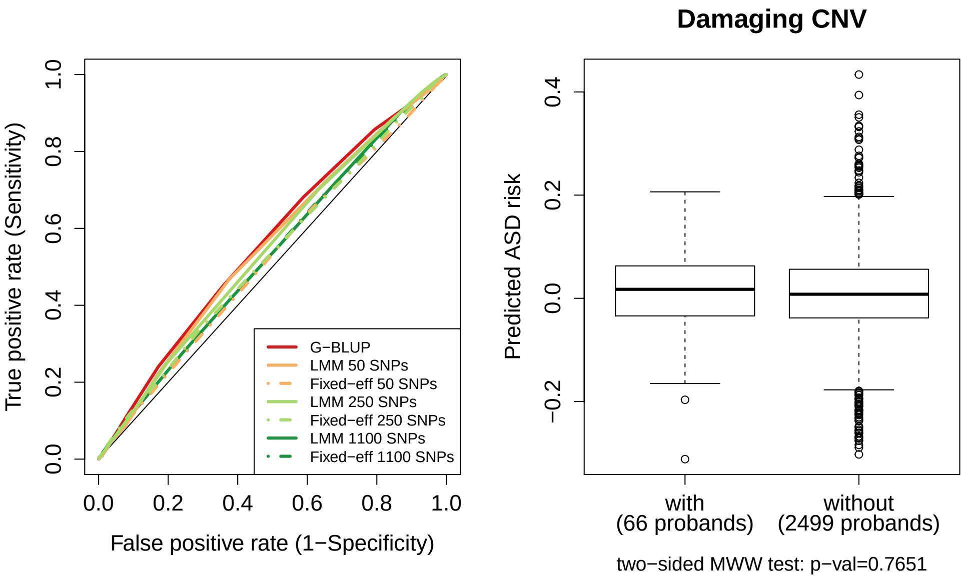

Figure 7: 5-fold G-BLUP prediction of ASD risk for 2565 pairs of AGP probands and pseudo controls. ***Left***: ROC curve of predicted risk. ***Right***: Boxplot of 66 SSC probands with damaging CNVs, and the other 2499 probands, where the predicted risk is obtained by G-BLUP; the *two-sided* Mann-Whitney-Wilcoxon test shows p-value 0.7651.

both the mixed-effects model and fixed-effects model, which indicates that almost no SNP has large effects, at least cannot be identified in current data. This provides support to the infinitesimal model, which suspects ASD being influenced by a large number of SNPs, each with very tiny effects.

Moreover, we confirm the effects of damaging *de novo* mutations on ASD risks in multiple scenarios, where we observe a significant difference between the probands with such mutations and those without. We also note that there are some scenarios where the effect is not significant, but we believe that it is due to the lack of predictive power of the current model.

One might wonder whether a generalized linear model should be used for ASD risk predictions. In our experiments, the generalized model gives very similar prediction results with the linear model. Considering the computation load, we choose to present the simple linear model.

Finally, we observe that using real controls (i.e., HABC controls) can gain much more predictive power than using pseudo controls. We expect that as more and more control data become available, the predictive power of the model can be improved. However, an accurate prediction of ASD risk for every individual will still be a challenging task.

References

[1] *http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000169.v1.p1.*

[2] http://www.broadinstitute.org/mpg/snap/ldsearch.php.

[3] A. Ramasamy et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neurosci.*, 2014.

[4] De Rubeis S. et al. Synaptic, transcriptional andchromatin genes disrupted in autism. *Nature*, 2014.

[5] G. Campos et al. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.*, 2013.

[6] GD. Fischbach et al. The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron*, 2010.

[7] J. Yang et al. Gcta: a tool for genome-wide complex trait analysis. *Am J Hum Genet.*, 2011.

[8] L. Klei et al. Gemtools: a fast and efficient approach to estimating genetic ancestry. 2011.

[9] L. Klei et al. Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism*, 2012.

[10] L. Liu et al. Dawn: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol Autism*, 2014.

[11] N. Chatterjee et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*, 2013.

[12] P. Chaste et al. Adjusting head circumference for covariates in autism: clinical correlates of a highly heritable continuous trait. *Biol Psychiatry*, 2013.

[13] R. Anney et al. A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, 2010.

[14] SH. Lee et al. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.*, 2011.

[15] T. Gaugler et al. Most genetic risk for autism resides with common variation. *Nature Genetics*, 2014.

[16] X. He et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.*, 2013.

[17] S. Gupta and F. Bunea. A study of the asymptotic properties of lasso estimates for correlated data. *Ph.D. Thesis, Florida State University*, 2009.

[18] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 2014.