Statistical model integrating interactions into genotype-phenotype association mapping: an application to reveal 3D-genetic basis underlying Autism

Qing Li[1,*], Chen Cao[1,*], Deshan Perera[1], Jingni He[1], Xingyu Chen[1], Feeha Azeem[1], Aaron Howe[2], Billie Au[4,6], Jun Yan[3,7], Quan Long[1,4,5,6,7,#]

1. Department of Biochemistry and Molecular Biology; 2. Heritage Youth Researcher Summer Program; 3. Department of Physiology and Pharmacology; 4. Department of Medical Genetics; 5. Department of Mathematics and Statistics; 6. Alberta Children's Hospital Research Institute; 7, Hotchkiss Brain Institute, University of Calgary, Alberta, Canada.

* = Joint first authors: Q.L. and C.C.

corresponding author: quan.long@ucalgary.ca

Abstract

Biological interactions are prevalent in the functioning organisms. Correspondingly, statistical geneticists developed various models to identify genetic interactions through genotype-phenotype association mapping. The current standard protocols in practice test single variants or single regions (that contain multiple local variants) sequentially along the genome, followed by functional annotations that involve various aspects including interactions. The testing of genetic interactions upfront is rare in practice due to the burden of testing a huge number of combinations, which lead to the multiple-test problem and the risk of overfitting. In this work, we developed interaction-integrated linear mixed model (ILMM), a novel model that integrates a *priori* knowledge into linear mixed models. ILMM enables statistical integration of genetic interactions upfront and overcomes the problems associated with combination searching.

Three dimensional (3D) genomic interactions assessed by Hi-C experiments have led to unprecedented biological discoveries. However, the contribution of 3D genomic interactions to the genetic basis of complex diseases has yet to be quantified. Using 3D interacting regions as *a priori* information, we conducted both simulations and real data analysis to test ILMM. By applying ILMM to whole genome sequencing data for Autism Spectrum Disorders, or ASD (MSSNG) and transcriptome sequencing data (GTEx), we revealed the 3D-genetic basis of ASD and 3D-eQTLs for a substantial proportion of gene expression in brain tissues. Moreover, we have revealed a potential mechanism involving distal regulation between FOXP2 and DNMT3A conferring the risk of ASD.

Software is freely available in our GitHub: https://github.com/theLongLab/Jawamix5

Keywords: Genotype-phenotype association mapping; Genetic interaction; Linear mixed model; 3D genomic interactions; Autism spectrum disorder.

Introduction

Genotype-phenotype association mapping has revealed thousands of loci associated with complex traits. As genes function by various forms of interactions and no gene operates in a vacuum (Costanzo et al. 2016; Phillips 2008), it is possible that the discovered single-gene associations may represent only the tip of an iceberg of the genetic-basis of complex disorders. In Statistical Genetics, genetic interaction is defined as the non-linear effects between multiple loci (Baryshnikova et al. 2013). Although researchers have developed many statistical models aiming to discover the role of genetic interactions underlying complex disorders (Fang et al. 2019; Greene et al. 2015; Jansen et al. 2019a; Miguel-Escalada et al. 2019; Watson et al. 2019; Wen et al. 2016), single locus analyses such as single variant models (Kang et al. 2010) or approaches jointly analyze multiple local variants in a single regions (Wen et al. 2016; Wu et al. 2010) are still dominant in the practice of association mapping (Jansen et al. 2019b; Watson et al. 2019). This may be partly due to the difficulties of interpreting the large number of outcomes from interaction analyses as well as the problem of multiple-test and overfitting (Cordell 2009). In contrast, researchers frequently utilize information regarding interactions, such as chromatin status (Tak and Farnham 2015), transcriptional regulations (Gallagher and Chen-Plotkin 2018), and protein bindings (Mao et al. 2016) and others (Schork et al. 2013), from various databases (Gallagher and Chen-Plotkin 2018; Ward and Kellis 2012) as sources for downstream annotations of peaks from single-region association mappings. On the other front, many methods quantitatively integrate single-locus functional information in association studies were also developed (Kichaev et al. 2019; Lu et al. 2016; Pickrell 2014; Sveinbjornsson et al. 2016; Yang et al. 2017). Moreover, methods integrating known biological network information in association studies were also recently proposed (Carlin et al. 2019), although genetic principle (e.g., heritability) was omitted in building such models. To our best knowledge, there is no tool of association mapping that integrates *a priori*, however yet incomplete, knowledge of interactions of multiple genomic regions into the statistical genetic models.

In this work, we developed ILMM, Interaction-Integrated Linear Mixed Model, a novel tool integrating *a priori* knowledge of genetic interactions with a statistical test to map associations between interacting genetic regions and the phenotypic variations. By leveraging a linear mixed-model, ILMM only requires the knowledge of the potential genetic regions; ILMM can test the existence of joint effects when the actual mechanism of interaction is unknown (**Methods & Materials**). By integrating biological *a priori* knowledge into the statistical models, ILMM has two major advantages over state-of-the-art models. First, as mentioned above, models searching for potential combinations of genetic loci *de novo* may lead to an astronomical number of candidates (Hoh and Ott 2003). However, ILMM tests a controllable number of combinations based on prespecified interacting regions, therefore, significantly relieving the risk of overfitting and the burden of multiple-test corrections. Second, instead of conducting statistical tests and the interpretation of interactions sequentially, ILMM allows simultaneous modeling and quantitative assessment of *a priori* partial knowledge using genetic and phenotypic variations in the disease cohort. ILMM

moves the efforts of collecting biological knowledge of interactions from downstream annotations to the upstream statistical tests. This is a meaningful innovation because the test P-values specifically quantify the strength of associations in terms of interactions as well as genetics rather than simply labelling marginal significance of a single gene as "found" or "not found" in the interactions databases.

The three-dimension (3D) chromatin structure is an important mechanism altering gene transcriptions that has been extensively studied for several years (Eres et al. 2019; Mah and Won 2019; Melo et al. 2020; Miguel-Escalada et al. 2019; Rao et al. 2014; Won et al. 2016). In addition to many biological insights revealed by the 3D structure of genomes, a fundamentally new view for statistical geneticists is that genes that are far apart when placed in the one-dimensional view could actually be spatially close to each other in 3D space when they function. This insight may bring a paradigm shift to statistical genetics, although more rigorous statistical analyses are required. A recent study suggested that the topologically associating domain (TAD) generally does not overlap with linkage disequilibrium (LD) block in large scale (Whalen and Pollard 2019). However, it is still unclear whether genetic interactions between the regions that are spatially close in a 3D domain exist and whether such interactions are functionally relevant to complex traits. Researchers have recently used 3D information in annotating peaks in association mappings (Fu et al. 2018; Giusti-Rodriguez and Sullivan 2019; Yu et al. 2019), however not integrated with association mapping models.

To illustrate the use of ILMM and further our understanding of the contribution of 3D genome structure to complex traits, we applied ILMM to multiple datasets of autism spectrum disorder (ASD) (Neale et al. 2012; Yuen et al. 2017) and gene expression data of 9 brain tissues and whole blood in the GTEx dataset (Aguet et al. 2017; Ardlie et al. 2015). In the analysis, we considered the interacting regions in 3D structure in brain tissues, assessed by Hi-C experiments (Rajarajan et al. 2018), as *a priori* information, and used them in the association mapping. We call this process 3D-Genome-wide association study, or 3D-GWAS. Indeed, we discovered interesting associations between 3D structure and ASD and expressions, which we called 3D-genetic basis of complex traits and 3D eQTLs, respectively. Additionally, we have identified substantial overlap between ASD and expressions in terms of 3D genetics, indicating pleiotropy effects shared by ASD and gene expressions. Through in-depth analysis of transcription factor binding motifs and protein docking, we have also revealed a mechanism underlying ASD that involves distal regulation between FOXP2 and DNMT3A.

This paper will explain the design intuition of ILMM, followed by its mathematical formulations. The simulations under various interaction mechanisms will be presented to demonstrate the universal power of ILMM which is robust to unknown mechanisms, contrasting to state-of-the-art alternatives. After that, analyses of real data and outcomes will be presented. Finally, limitations and future extensions will be discussed.

Materials and Methods

Design principle of ILMM.

Multiple genetic regions can jointly alter the phenotype through various mechanisms, including epistasis (Bateson 1903; Mackay 2014), compensatory (Brown et al. 2010), heterogeneity (Madsen et al. 2011), or sometimes just additive (Madsen et al. 2011). Designing a general model to test interactions without knowing a specific mechanism could be tricky. A test that exhaustively verifies all potential known mechanisms would lead to substantial risk of overfitting and multiple-test burden; not to mention the lack of a complete list on all potential mechanisms. We took advantage of a specific angle underlying the linear mixed models; linear mixed models (LMM), although being called "linear", can capture interactions implicitly. Although the pattern of genetic interactions could be complicated and largely unknown, the probability for two individuals carrying the same combinations of alleles at multiple genetic loci is proportional to the overall genetic similarity in these loci (e.g. identity by descendant). This similarity can be naturally captured by the genomic relationship matrix (GRM), which serves as the variance-covariance matrix in a multivariate normal distribution (MVN) of a random term in LMM.

Based on the above rationale, we designed ILMM by embedding the focal genetic regions into an LMM (**Fig. 1**). In this LMM, we have two random terms: one term employs the whole-genome GRM as the variance-covariance matrix in its MVN, while the other term employs a "interacting regional" GRM as its variance-covariance matrix. The interacting GRM is calculated using genetic variants in the regions that are suspected to have interactions, e.g., the two regions that are interacting in 3D space (revealed by a Hi-C experiment) (**Fig. 1**). Using such an LMM model, we aggregate the genetic variants in regions that are potentially interacting into the "regional" random term; and the rest contributions (i.e., other genes or population structure) are captured by the "global" random term using the whole-genome GRM.

Mathematical formulations.

We use Y to denote the phenotype, U_i to denote the interaction term capturing the interacting regions, and U_g to denote the random term capturing the rest contribution. Our model then becomes:

$$Y = U_i + U_g + \varepsilon \quad (1)$$

Where $U_i \sim \mathrm{MV}N\left(0, \sigma_i^2 K_i\right)$, $U_g \sim \mathrm{MV}N\left(0, \sigma_g^2 K_g\right)$, and $\varepsilon \sim \mathrm{MV}N(0, \sigma_e^2 I)$.
where I is the identity matrix, K_g is the GRM calculated using the whole genome variants:

$K_g = \frac{1}{\mathrm{n}} X^T X$, where X is the centralized and standardized genotype matrix and n is the number of variants in whole genome. Denoting the corresponding centralized and

standardized genotype matrix in the *m*-th focal region (*m* = 1,2, …, *M*) as X_m, and the combined genotype matrix $X_{int} = (X_1, \ldots, X_m)$ then $K_i = \frac{1}{n_{int}} X_{int}{}^T X_{int}.$, where n_{int} is the total number of variants in all the *m* regions. The model is solved using an integration of de-correlation of the global GRM and the low-rank trick proposed by FaST-LMM (Lippert et al. 2011). The details of the mathematical derivations are in **Supplementary Materials**.

Software implementation.

ILMM is implemented as a function in our existing software Jawamix5 (Long et al. 2013; Xiong et al. 2019) that employs memory virtualization techniques (or "out-of-core" in computer science) based on HDF5 libraries. The software is scalable to very large genomic dataset that cannot be loaded into memory. The data file is stored in the disk with highly effective indexing so that the calculation is as fast as though the data were resided in the main memory (RAM).

Procedure of simulations and type-I error adjustment.

We thoroughly tested ILMM via simulations, contrasting to state-of-the-art alternatives. The control dataset of Wellcome Trust Case-Control Consortium (WTCCC) (Burton et al. 2007) were used as the template genotype (N=2,938).

To prepare *a priori* knowledge, we first collected a list of potential 3D-interacting pairs of regions, which has been reported by other researchers by conducting Hi-C experiments in brain tissues (Rajarajan et al. 2018; Watson et al. 2019; Won et al. 2016). Here, we utilized a Hi-C assessment in the developing brain which has 27,982 brain-specific paired 3D-interacting regions, measured from neurons derived from human induced pluripotent stem cells (hiPSCs) (Rajarajan et al. 2018). This dataset is available in the Synapse database (https://www.synapse.org/) with Synapse ID: syn12979149.

During each round of simulation, a pair of interacting regions was randomly selected. We then randomly selected 5 genetic variants in each region as causal to form the genetic contribution from the region. These 5 variants were modeled using a combination of "additive model" and "heterogeneity model": the aggregated contribution from a region will be 0, 1, or 2 if there are 0, 1 or 2 alternate alleles in the 5 variants, reflecting an additive pattern. However, if there are more than 3 alternate alleles, the total contribution is still 2, reflecting heterogeneity pattern in which different mutations can cause phenotypic changes independently. Based on the regional contributions, the total genetic contribution to phenotypes were simulated using four mechanisms of joint contributions from multiple genetic regions; the models used are the additive model, epistasis model, heterogeneity model and compensatory model, defined in **Table 1**. Finally, the phenotype was simulated by adding genetic contribution with a random noise. We rescaled the genetic contribution to ensure that the phenotype indeed has the prespecified heritability.

ILMM was compared to EMMAX (Kang et al. 2010), the most frequently used tools for single-marker analysis, and SKAT (Wu et al. 2010; Wu et al. 2011), a popular method testing aggregated effects of genetic variants in a single region. We also tested ILMM against our own implementation of single-region-based mixed model, which is a specific case of ILMM when M=1 (i.e., the number of interacting regions is just one). We call this method "LOCAL". The motivation of comparing ILMM to LOCAL is to quantify how much power gain ILMM will achieve when comparing with the single-region-based method with the same implementation.

To conduct a fair comparison, we simulated random phenotype (i.e., no genetic component) to adjust the Type-I-Errors (TIEs) to be the same. In particular, we ranked all the P-values calculated under the null distribution (i.e., using the random phenotype) and took the top 5% cut-off. If this cut-off is surrounding 0.05, we can conclude that the corresponding method's TIE is under control. Also, this cut-off, after multiple-test correction, will be the cut-off to claim significance when calculating powers under alternative hypothesis. How the other tools were executed are detailed in **Supplementary Materials.**

Genotype and phenotype data for 3D-GWAS

The *a priori* information of genetic regions suspected undergoing interactions, as stated in *Simulations,* is the 3D interaction data generated by other researchers using Hi-C experiments in brain tissues (Rajarajan et al. 2018).

The genotype and phenotype data are acquired from dbGaP and other genomic consortia. There are two ASD datasets: the first dataset is the influential MSSNG data (Yuen et al. 2017), containing 7,065 whole genome sequencing data; the second dataset contains 9,428 subjects (Neale et al. 2012) assessed by whole exome sequencing (phs000298.v4.p3.c1 & phs000298.v4.p3.c2). A more detailed description of the datasets is in **Supplementary Table S1**. We have conducted general quality control by removing variants with low MAF and derivation from HWE for association mapping but use the full dataset for downstream annotations.

Additionally, we applied ILMM to the gene expressions in brain tissues generated by the Genotype-Tissue Expression (GTEx) project (Aguet et al. 2017; Ardlie et al. 2015). The list of tissues and sample sizes are listed in **Supplementary Table S2**.

Annotating 3D-GWAS outcomes by distal regulation

In this work, relating to a pair of interacting regions assessed by Hi-C experiments, distal regulation refers to the situation where a transcription factor (TF) binding to one region of the pair interacts with a gene located in the other region of the pair. The input of this analysis is a list of identified candidate genes (that are located in regions significantly associated with ASD assessed by ILMM P-values, e.g. DNMT3A). The output is a list of TFs (together with their binding motifs) regulating the candidate genes. To detect such regulations, we took two

steps. First, we used the TF2DNA (Pujato et al. 2014) database to identify TFs which have been reported to interact with at least one of the candidate genes. These TFs, however, may or may not bind to the iterating regions. So, in the second step, we used the JASPAR (Khan et al. 2018) database to search for the binding sites (i.e., motifs) to filter out the TFs that do not bind to our candidate interacting regions. The TFs that have their binding sites located to the pairing regions will then be the output.

Next, we looked for genetic SNPs located at the binding sites (motifs) of these candidate TFs and further predicted the interactions between DNA and proteins (motif-TF complexes). This was achieved by utilizing HDock, a tool specifically designed to quantify the binding affinities between TF and motifs (Yan et al. 2020).

Calculating LD between interacting regions associated with phenotype and expressions

The paired regions that are in genetic interactions may be susceptible to have high linkage disequilibrium (LD), even though they may be far away from each other. For selected pairs of regions, we computed their LD in terms of both D' and r^2 using genotype data from the 1000 Genomes Project (Altshuler et al. 2015). Since the standard LD is defined between two genetic variants, we calculated the pairwise LD between all variants in one region and all variants in the other region and considered their average as the LD between two regions. To contrast the LDs between interacting regions with the background, i.e., non-interacting regions, we formed a null distribution by calculating LDs between randomly selected pairs of regions with the same sizes and between-region distance for 1000 times. These 1000 average LD values formed a null distribution. Then the standings of actual LDs of interacting regions will be ranked in the null distribution as an assessment of whether they are significantly high.

Results

Simulations

The type-I-errors of all the four competing methods were generally under control, although ILMM and LOCAL are slightly different to the supposed value of 0.05 (**Supplementary Table S3**). In particular, ILMM was more conservative (with a top 5% cut-off being 0.0713). Using the adjusted critical values ensuring type-I-errors to be 0.05, the fairness of power comparison was guaranteed. QQ-plot for the ILMM P-values under null hypothesis showed that the expected P-values are generally equal to observed P-values, although they were slightly under the diagonal (**Supplementary Fig. S1**), consistent to the slightly conservative type-I error.

As described in **Methods and Materials**, we compared the power of ILMM with other three methods: SKAT, EMMAX and LOCAL. The power is defined as the number of rounds that the corresponding method significantly identified the simulated pairs of regions. In the present setting of pairs of regions, there are two regions to be detected. For ILMM, naturally it will detect both regions in a single test; however, for the other tests, the criteria of defining

"success" could be detecting at least one region or detecting both regions. Here we used "(1)" to indicate the criteria of detecting at least one region, and "(2)" for the criteria of detecting both interacting regions. SKAT and LOCAL are naturally region-based, and we set up the window size of 5kb, which is the average of the length of one side of the pairs of interacting regions. For EMMAX, which is based on a single marker test, we claimed success of a region as long as there is at least one genetic variant significantly associated with the phenotype. The statistical significance of a particular test for a focal method is defined by its P-value smaller than the corresponding cut-off observed in the previous simulations to adjust type-I-error (**Supplementary Table S3)** divided by the number of tests (i.e., Bonferroni correction (Noble 2009)). Note that different methods had different numbers of tests. For SKAT and LOCAL, the number of tests was 1200K (=3 x 10^9/2500), which was the total number of tiling windows across the genome. For ILMM, it was 27,982, the total number of Hi-C assessed spatial-interacting regions. For EMMAX, it was the number of SNPs of the WTCCC array, which was 386,469.

The outcome is depicted in **Fig. 2**. Evidently, the other methods had very small powers when the criterion is to detect both interacting regions. This is consistent with the motivation that testing multiple interacting regions together will substantially improve power. Additionally, ILMM outperforms all the methods with the criteria of identifying at least one region. This implied that the current standard protocol, which discovers an associated region in single-region tests followed by database-search-based annotations, is suboptimal compared to the interaction-integrated test.

3D-genetic basis of Autism Spectrum Disorder

Using the potential 3D interacting regions revealed by Hi-C experiments in neurons induced by human iPSC brains (Rajarajan et al. 2018) as *a priori* knowledge, we applied ILMM to the MSSNG whole-genome sequencing data (Yuen et al. 2017) to identify 3D-genetic basis of ASD. As important follow-ups, we carried out annotations for the statistically significant genes from various aspects including enrichment analysis for pathway and gene ontology, as well as functional annotation towards distal regulations.

Statistically significant genes. By applying ILMM to MSSNG dataset, we identified 1,164 pairs of regions that are significantly associated with ASD (whole genome FDR =0.05). There are 1,445 genes located in these significant regions (**Supplementary Table S4**). As a comparison, we also applied SKAT (Wu et al. 2010; Wu et al. 2011) to this dataset with sliding windows of 5kb and revealed 32,682 significant regions (with the same criteria of whole-genome FDR =0.05), containing 6,322 genes, among which 663 genes (10.4%) also have been identified by ILMM method (**Supplementary Table S5**). This comparison indicates that majority genes still impact ASD by marginal effects, in which around 10.4% may be involved in 3D-interactions. On the other hand, by directly integrating interactions in a statistical model, ILMM identified around 2.1 (= 1,445 / 663) times genes than a single region-based model.

To carry out functional enrichment analyses for the 1,445 potential genes located in significant regions, we utilized an R package named clusterProfiler (Yu et al. 2012) which quantifies the level of enrichment of a gene set with regards to various functional annotations. Using clusterProfiler, we performed both pathway enrichment analysis based on Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2002) and gene ontology (GO) (Ashburner et al. 2000) enrichment. The top 20 KEGG significantly enriched pathways were reported (**Fig. 3a**). Among them, Huntington disease, which belongs to the nervous system category, was the pathway that covered the highest proportion of genes. Notch signaling pathway, which is central to a wide range of development processes in human organs (Lasky and Wu 2005) was also among the most significantly enriched pathways. Notably, the Glutamatergic synapse pathway, which is labelled by KEGG to be associated with ASD, was also present in our results. Several additional pathways, such as antigen processing and presentation, SNARE interactions in vesicular transport and base excision repair, were also reported to be related with ASD (Bennabi et al. 2018; Castermans et al. 2010; Markkanen et al. 2016).

The GO analysis was applied to three ontologies: biological process (BP), cellular component (CC), and molecular function (MF); and the top 10 significant GO terms are depicted (**Fig. 3b**). Among the top 10 highly enriched BP terms, four of them are leukotriene associated, i.e., leukotriene metabolic process, leukotriene D4 metabolic process, leukotriene D4 biosynthetic process, and leukotriene biosynthetic process. The elevated levels of leukotrienes have been reported in autistic patients in several studies (El-Ansary and Al-Ayadhi 2012; Theoharides et al. 2016). In addition, leukotriene can be used as a biomarker for the early diagnostic of autistic patients (El-Ansary and Al-Ayadhi 2012; Qasem et al. 2016). Among the significant CC terms, U1 snRNP (small nuclear ribonucleoprotein) is the most significant cellular component (P-value = 1.06×10^{-4}) and is one of the 9 snRNA blood signatures boosting diagnostic accuracy for ASD in clinical practice(Zhou et al. 2019). For MF, the most significant enrichment was pre-mRNA 5'-splice site binding (P-value = 2.23×10^{-5}), consistent with the report of the important role of alternative splicing of mRNA in ASD blood (Stamova et al. 2013). The above analyses for KEGG and GO enrichment showed that, at the gene-set level, our discoveries generally align to reported ASD pathology.

In addition to the enrichment analyses, we conducted further annotations of significant results to figure out a hypothetical novel mechanism underlying ASD. To narrow down candidates, we first searched the SFARI database, an established repository for existing ASD genes (Abrahams et al. 2013), and found that 49 ILMM-identified genes in are in SFARI. The proportion that is in SFARI, 49/1,445 = 3.4%, is relatively low. This might be due to that firmly verified ASD genes are limited. Indeed, for the established method for single region analysis, SKAT, the corresponding ratio is also 217/6,322 = 3.4%. These 49 genes include 12 genes scored as 1, which are known as high-confident ASD genes in SFARI (**Table 2**).

Using the TF2DNA + JASPAR pipeline (described in **Materials and Methods**), we generated the distal regulatory TFs for the above score-1 genes. As the MSSNG dataset is

whole-genome sequencing that provides all polymorphisms, we were enabled to characterize the consequence of a genetic mutation located in a motif to the binding complex by using HDock (Yan et al. 2020). In particular, reference and mutant binding motifs as well as the TFs (represented by their protein data bank ID, or PDB) were submitted to the HDock server. We found that (1) FOXP2 is a TF to the gene DNMT3A (a score-1 candidate gene in **Table 2**), and (2) a single nucleotide polymorphism, or SNP (NC_000002.11: g.133032477T>C) on the binding motif of FOXP2 (ATTGTT**T**TATT), will affect FOXP2 binding affinity. More specifically, on the structure of FOXP2, there is a positively charged interface (J:542R, J:543R, K:553R, K:554H, K:549K, K:583R, where J and K are protein chain index) in the minimum energy protein-ligand complex around the reference allele T (**Fig. 4a**). According to a previous study (Luscombe et al. 2001), because thymine has the highest acidities among the four nucleotides, thymine (reference allele) preferentially interacts with arginine (R), histidine (H), and lysine (K) than cytosine (mutant allele). Thus, thymine can reduce the binding energy by forming stable protein-DNA electric fields (Luscombe et al. 2001). Consistent to this interpretation, the docking score for wild type motif is -362.00, in contrast to the docking score for the mutated motif (ATTGTT**C**TATT) being -330.00, which has higher binding energy with FOXP2. Together, this suggested that the mutation (g.133032477T>C) reduces the binding affinity of FOXP2 to its motif. As a result, DNMT3A, which is regulated distally by FOXP2 through chromatin loop, may have diminished expression (**Fig. 4b**).

DNMT3A encodes an enzyme named DNA methyltransferase 3 alpha, which is involved in DNA methylation and plays a crucial role in epigenetic regulation in cells. In particular, DNMT3A binds preferentially to intergenic regions and across transcribed regions of genes, which primarily induces methylated CA sequences (mCA) in the early-life neuron (Stroud et al. 2017) (**Fig. 4c**). These mCA functions act as landmarks for MECP2, whose function is to restrain gene expression in the maturing brain (**Fig. 4d**). As such, DNMT3A, mCA, and MECP2 coordinate to precisely tune gene expressions that are crucial for normal brain development and function (Stroud et al. 2017). The hypomethylation in adult neurons caused by insufficient DNMT3A at their early life will lead to overexpression of many genes that lead to risk of ASD (Stroud et al. 2017) (**Fig. 4d**). Indeed, disruption of the cooperation either through DNMT3A or MECP2 has been reported to cause Rett syndrome, a severe neurological disorder with features of autism. (Chahrour and Zoghbi 2007; Gabel et al. 2015). Additionally, mutations on DNMT3A have been widely reported in ASD (Alex et al. 2019; Yokoi et al. 2020) as well as those with intellectual disabilities (Tatton-Brown et al. 2014). As such, our proposed mechanism is that mutations in the region on or surrounding DNMT3A in conjunction with mutations of FOXP2 binding sites jointly confer the risk of ASD.

In summary, our in-depth analysis including ILMM-based genetic mapping, functional annotation, motif search, and protein docking has jointly revealed a plausible mechanism for ASD: the SNV (g.133032477T>C), presenting in one of FOXP2 motifs, may lead to decreased gene expression of DNMT3A through distal regulation. The low level of DNMT3A may further cause the hypomethylation of CA, which reduces the recruitment of MECP2 and

results in the increased expression of some genes, causing higher risk to ASD. This regulation through 3D interactions may jointly confer the risk of ASD with local mutations surrounding DNMT3A.

3D-cis eQTL in brain tissues

EQTLs are the genetic mutations associated with gene expression. Analog to the above 3D-GWAS that discovers 3D-genetic basis of complex traits, here we aimed to extend the concept of eQTL to spatially interacting regions using the list of interacting regions assessed by Hi-C experiments (that were used above). To keep this first attempt simple, we only looked at the eQTLs in *cis*, which means the gene body is located in or surrounding one of the interacting regions (20,000 base pair upstream or downstream of that gene). Such 3D-cis eQTLs may be deemed as *trans* in standard analysis (with 1D genome) but are considered as *cis* here as we hypothesized that the regulatory mutations are spatially nearby in the 3D domain. We applied ILMM on 10 tissues in GTEx datasets including 9 brain tissues and the whole blood to identify interacting regions functioning as 3D-cis eQTLs. First, we selected genes with high variance (variance >= 10 in RPKM value), which led to around five thousand genes for each tissue. Then we performed eQTL mapping between expressions of these selected genes and genotypes using ILMM to uncover paired regions that are significantly associated with gene expressions. Based on a cut-off of P-value lower than 0.05 after FDR correction (Noble 2009), we discovered hundreds of significant 3D-cis eQTL from these 10 tissues. All results are listed in **Supplementary Table S6-15.** To assess the proportion of genes that are were able to detect 3D-cis eQTLs, we calculated the number of genes that are located in (or around) a pair of interacting regions and the number of genes for which we indeed identified 3D-cis eQTLs. It is observed that 3D-cis eQTL accounts for a small proportion (3% - 8%) of genes (**Fig. 5a**). This indicates that 3D-cis eQTLs do exist, however, they are not dominant for gene regulations.

We checked whether the 3D interacting regions are in higher linkage disequilibrium (LD) compared to background (**Materials and Methods**). To narrow down the candidates, we selected only the pairs of regions that are significantly associated with both ASD and gene expressions. In total, there were 69 paired regions as the 3D co-localized loci for both ASD and eQTL (**Supplementary Table S16**). We then calculated the average r^2 and D' between the two regions and contrasted them to the background (**Materials & Methods**). The outcome is depicted in **Fig. 5b & Supplementary Fig. S2**, showing that some of the regions indeed experienced significantly higher LD than the background. However, many regions did not. This showed that the 3D interactions may be very weak to the extent of not being able to impact the LD between regions, an observation consistent with the recent reported LD study at a larger scale (Whalen and Pollard 2019).

Conclusion & Discussion

By applying ILMM to GTEx data, we identified 3D-cis eQTLs for only 3% - 8% genes in the brain tissues. This indicated that such 3D-genetic basis of expressions is not dominant, compared with standard cis eQTLs based on the 1D view of a genome. However, this may be due to our current limitation of accurately pinpointing the 3D interacting regions. Furthermore, the genetic interaction may not be limited to chromatin conformations, and other sources of interactions may also contribute as well. Therefore, there is potentially substantial room to utilize ILMM in practice.

As the GO and KEGG dataset may not be specifically designed for ASD, based on our hypothesis that ASD might be mediated by communication disorders, we specifically searched for annotations of genes association with hearing deficiencies using DisGeNET (Pinero et al. 2020). In total 41 genes that have been reported to be associated with Sensorineural Hearing Loss (disorder) or Nonsyndromic Deafness. Notably, 8 genes among these 41 genes were reported to be associated with ASD (**Supplementary Table S17**). Starting from these genes, further exploration to the functional mechanism of ASD medicated by communication disorders will be an interesting future work.

In addition to the whole genome data from MSSNG, we have also applied ILMM together with the same 3D interacting regions (Rajarajan et al. 2018) to a whole exome sequencing (WES) dataset containing 7,766 individuals (4944 control and 2822 ASD cases, dbGaP ID: phs000298.v4.p3.c1), which yielded no significant results. In contrast, applying a single-region tool with sliding windows of 5kb and 25kb led to the discovery of several significant regions (**Supplementary Table S18)**. This outcome indicates that the 3D genetic interactions are generally distributed in non-genic regions and may need sequencing data to analyze.

In summary, we developed a novel statistical method, ILMM, that integrates *a priori* knowledge of genetic interaction with linear mixed models to statistically identify genetic interactions. To demonstrate its use in practise, we used 3D chromatin conformation information assessed by Hi-C experiments as example *a priori* of potentially interacting regions. Using this list of paired regions, we applied ILMM to the whole-genome sequencing data generated by MSSNG and revealed the 3D genetic basis of ASD. Additionally, we also applied ILMM to transcriptome data generated by the GTEx consortium. The real data analysis revealed substantial insights into the 3D-genetic basis of both complex traits and expressions. Specifically, we formed a hypothetical mechanism including both local mutations and a motif impacting distal regulation that jointly confer the risk of ASD. This novel statistical method offers a complementary protocol to the standard practice that conducts association mappings for single regions followed by annotations and co-localization analyses. Additionally, compared with pure statistical methods searching for interactions *de novo*, ILMM does not suffer from the problem of choosing or optimizing the size of candidate genetic regions as the *a priori* knowledge will naturally offer such information. Therefore, ILMM further reduces the burden of multiple-test and risk of overfitting. Finally, the outcome of ILMM is naturally interpretable as the potential annotation, e.g., 3D-interactions, has been

built in. We expect ILMM will be broadly used in practise to discover novel genetic interactions.

Declaration

Funding

This research is supported by the Campbell McLaurin Chair for Hearing Deficiencies (J.Y.), New Frontiers in Research Fund (Q.L.), Canada Foundation for Innovation (Q.L.), and Alberta Children's Hospital Research Institute postdoctoral fellowship (C.C.).

Conflicts of interest/Competing interests

The authors declare that there is no conflict of interests.

Availability of data and material

All datasets used in this study are available in the following URL through application.

1. WTCCC data: https://www.wtccc.org.uk/info/access_to_data_samples.html
2. MSSNG data: https://www.mss.ng/
3. phs000298.v4.p3: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v4.p3
4. Genotype-Tissue Expression (GTEx) : https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2
5. Brain Hi-C interacting regions: https://www.synapse.org/#!Synapse:syn12979149

Code availability

Software is freely available in our GitHub: https://github.com/theLongLab/Jawamix5

Authors' contributions

[Q.L.1 = Qing Li and Q.L.2 = Quan Long]
Conceived the study: Q.L.2 and J.Y.; Developed the software: Q.L.2; Simulations: Q.L.1; Real data analysis: Q.L.1, C.C., and D.P.; Contributed to the analysis: J.H, X.C., F.A., and A.H.; Contributed to the interpretation of the data: B.A. and J.Y. Wrote the manuscript: Q.L.2, Q.L.1, and C.C. with the input of all authors.

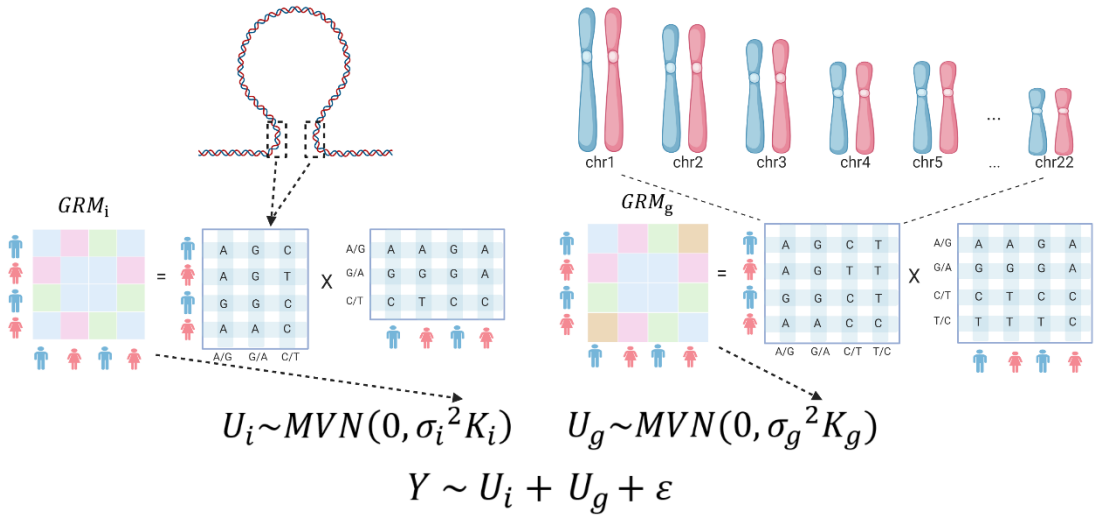


Figure 1. The protocol of ILMM. GRM_i refers to the genomic relationship matrix (GRM) calculated by the genotype present in the interacting regions. GRM_g refers to the GRM calculated by all genotype from the whole genome.

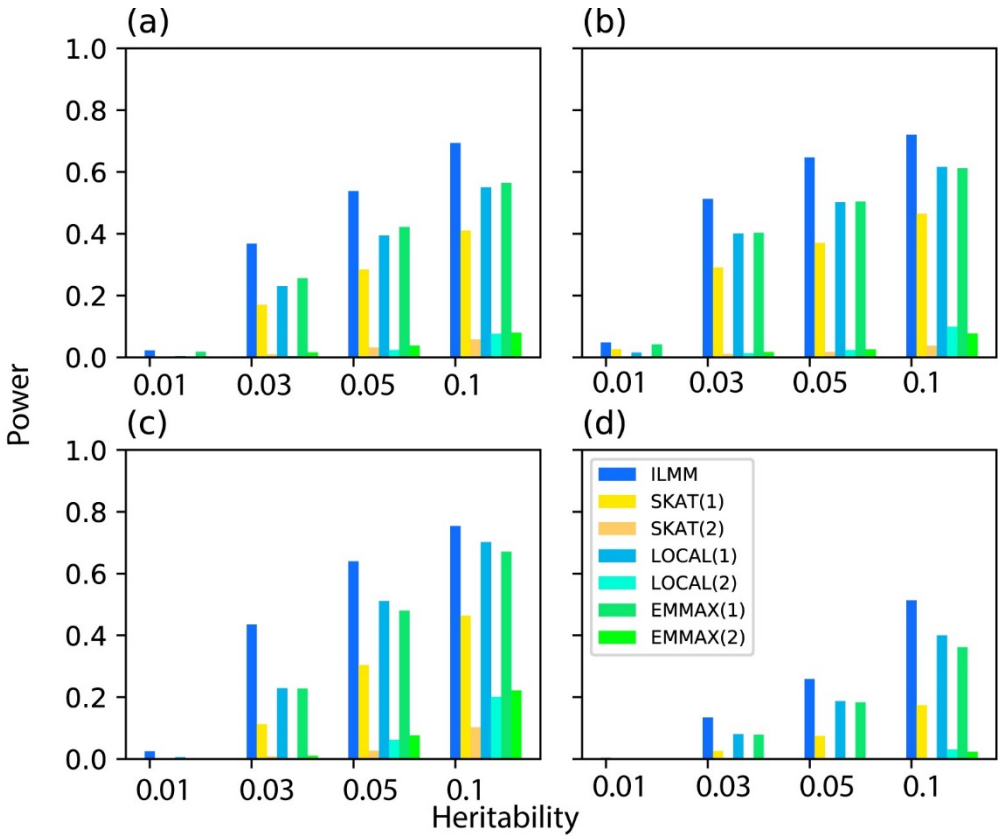


Figure 2. Statistical power (y-axis) of four methods under four interacting models with various regional trait heritability (x-axis). (a): Additive model, (b): Epistasis model, (c): Heterogeneity model, (d): Compensatory model

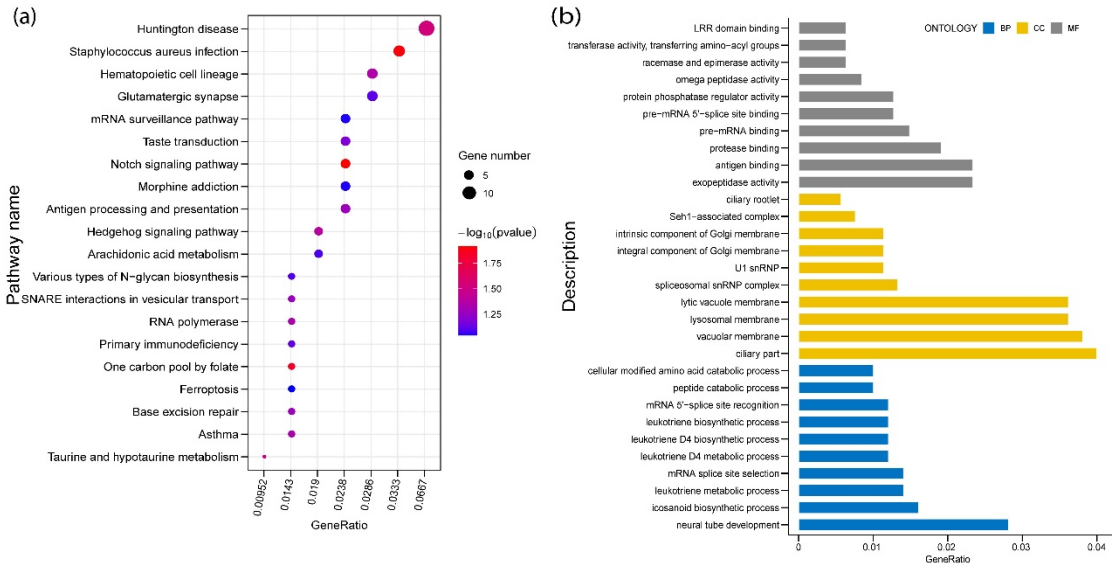


Figure 3. KEGG and GO enrichment analysis of significant genes. (a): Top 20 KEGG pathways (ranked by p-value). Gene ratio (x-axis) is the percentage of significant genes over the total genes in a given pathway. (b): Top 10 (ranked by p-value) GO terms of three categories (BP: biology process, CC: cell component, MF: molecular function). Gene ratio (x-axis) is the percentage of the number of genes present in this GO term over the total number of genes in this category.

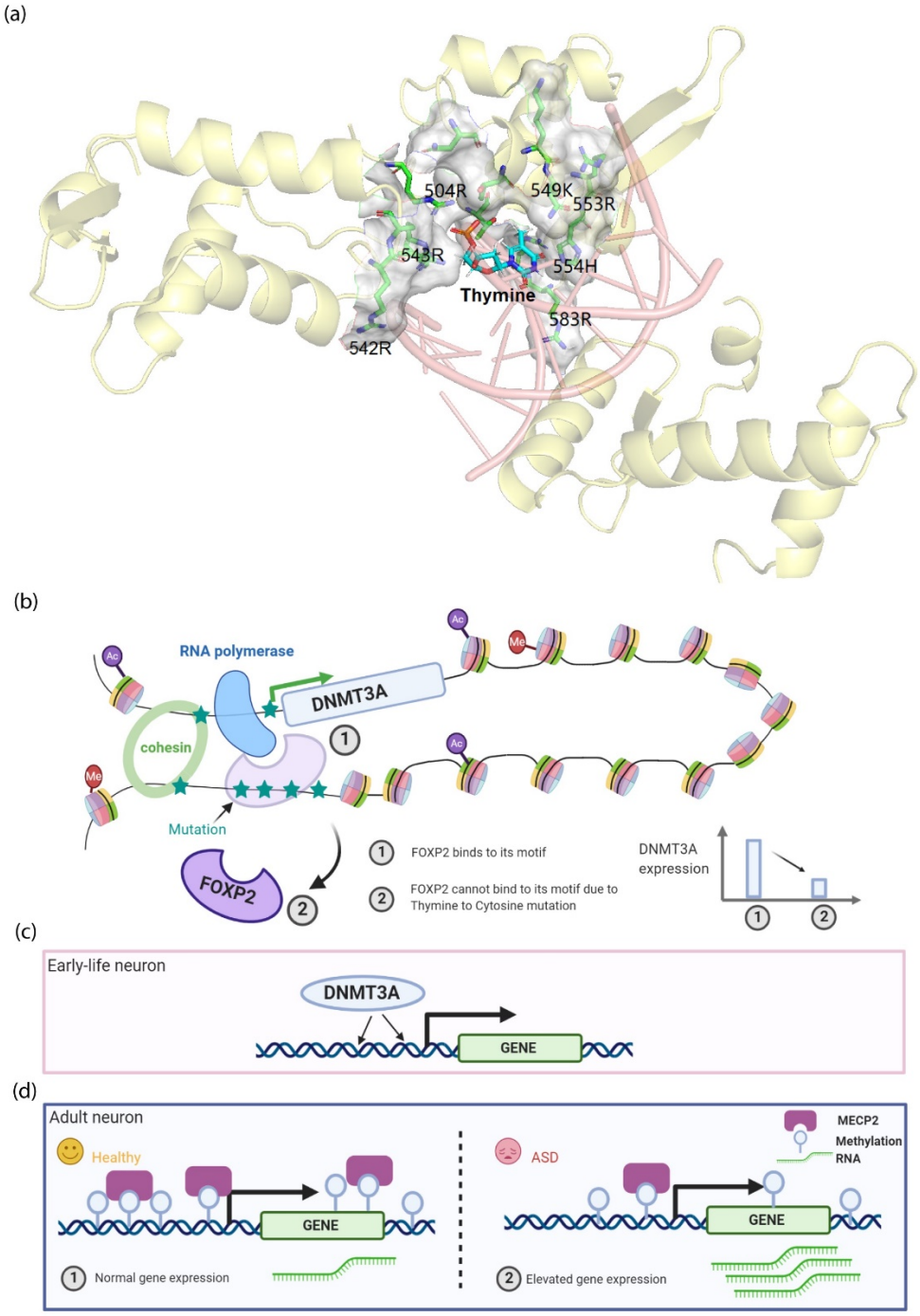
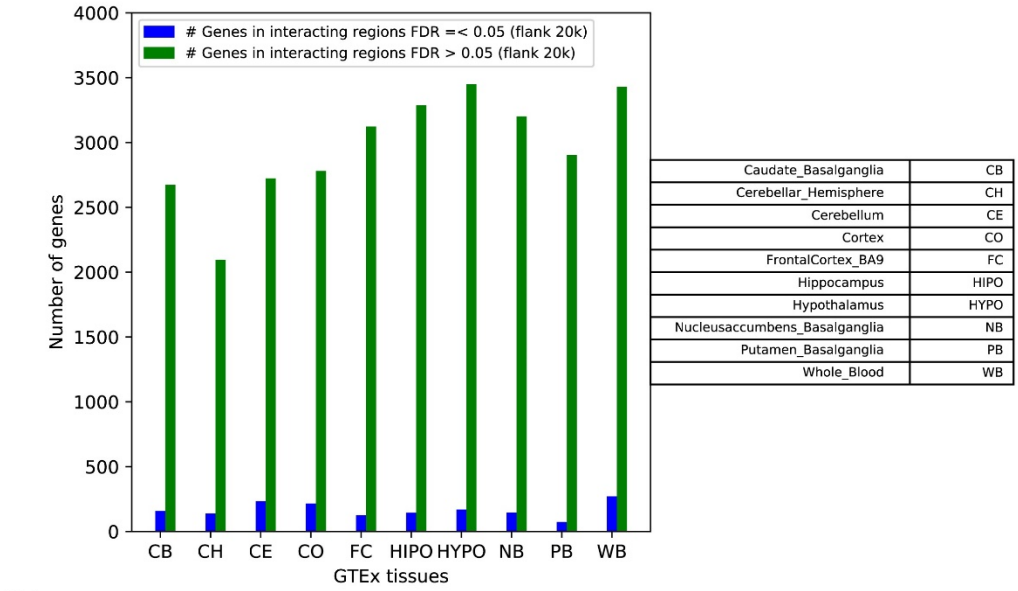


Figure 4. Interaction between FOXP2 and its motif for distal regulation on DNMT3A. (a) Thymine (reference allele) on the motif and its nearby positively charged interface of FOXP2 visualized by PyMOL (DeLano 2002). (b): The single nucleotide mutation on FOXP2 motif is likely to reduce the binding affinity of FOXP2 which may decrease the gene expression of DNMT3A through distal regulation. (c): A gene present in early-life neuron and DNMT3A deposits methylation on its CA sequences. (d): The low level of DNMT3A across the neuron development process may cause hypomethylation of many genes resulting in elevated expression of these genes in adult neurons.

(a)



(b)

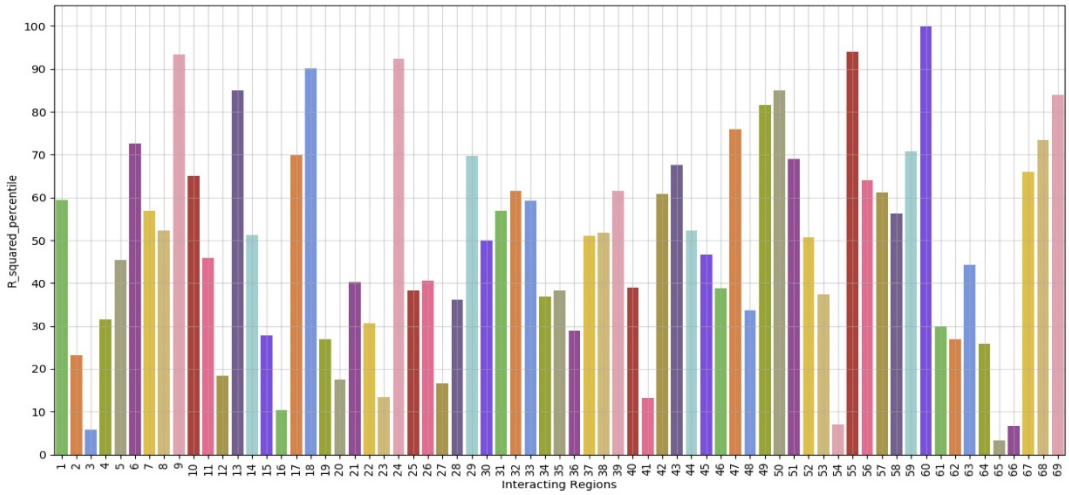


Figure 5. 3D-cis eQTL analysis. (a) Number of genes for which we can and cannot identify 3D-cis eQTLs. (b) LD (r^2) value for 69 interacting regions associated with both ASD and gene expressions. X-axis: coordinate of one interacting region (details in Supplementary Table S15). Y-axis: percentile of LD value among background (higher percentile indicates stronger LD in interacting regions).

Table 1. Definition of four different models used to simulate genetic component to phenotype (before adding random noise).

Additive		Region A contribution		
		0	1	2
Region B contribution	0	0	1	2
	1	1	2	3
	2	2	3	4
Epistasis		Region A contribution		
		0	1	2
Region B contribution	0	0	0	0
	1	0	1	1
	2	0	1	1
Heterogeneity		Region A contribution		
		0	1	2
Region B contribution	0	0	1	1
	1	1	1	1
	2	1	1	1
Compensatory		Region A contribution		
		0	1	2
Region B contribution	0	0	1	2
	1	1	0	1
	2	2	1	0

Table 2. Significant genes identified by ILMM which are also reported as high confident ASD genes in the SFARI database. An FDR of 0.05 is used to adjust P-values.

GENE	*SFARI SCORE*	*CHR*	*START*	*END*	*ILMM ADJUSTED P-VALUE*
CACNA1C	1	chr12	2079952	2802108	1.98×10^{-15}
DHCR7	1	chr11	71139239	71163914	3.86×10^{-02}
DNMT3A	1	chr2	25455845	25565459	1.29×10^{-10}
DPYSL2	1	chr8	26371791	26515694	1.49×10^{-03}
GABRB3	1	chr15	26788693	27184686	4.31×10^{-04}
KANSL1	1	chr17	44107282	44302733	1.19×10^{-05}
KMT2C	1	chr7	151832010	152133090	5.39×10^{-05}
MYT1L	1	chr2	1792885	2335032	1.47×10^{-10}
NF1	1	chr17	29421945	29709134	2.77×10^{-06}
PHF21A	1	chr11	45950871	46142985	9.63×10^{-07}
SHANK2	1	chr11	70313961	70963623	3.86×10^{-02}
ZNF292	1	chr6	87862551	87973914	3.92×10^{-09}

References

Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, Packer A (2013) SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism 4: 36. doi: 10.1186/2040-2392-4-36

Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, Mohammadi P, Park Y, Parsana P, Segre AV, Strober BJ, Zappala Z, Cummings BB, Gelfand ET, Hadley K, Huang KH, Lek M, Li X, Nedzel JL, Nguyen DY, Noble MS, Sullivan TJ, Tukiainen T, MacArthur DG, Getz G, Management NP, Addington A, Guan P, Koester S, Little AR, Lockhart NC, Moore HM, Rao A, Struewing JP, Volpi S, Collection B, Brigham LE, Hasz R, Hunter M, Johns C, Johnson M, Kopen G, Leinweber WF, Lonsdale JT, McDonald A, Mestichelli B, Myer K, Roe B, Salvatore M, Shad S, Thomas JA, Walters G, Washington M, Wheeler J, Bridge J, Foster BA, Gillard BM, Karasik E, Kumar R, Miklos M, Moser MT, Jewell SD, Montroy RG, Rohrer DC, Valley D, Mash DC, Davis DA, Sobin L, Barcus ME, Branton PA, Grp EMW, Abell NS, Balliu B, Delaneau O, Fresard L, Gamazon ER, Garrido-Martin D, Gewirtz ADH, Gliner G, Gloudemans MJ, Han B, He AZ, Hormozdiari F, Li X, Liu B, Kang EY, McDowell IC, Ongen H, Palowitch JJ, Peterson CB, Quon G, Ripke S, Saha A, Shabalin AA, Shimko TC, Sul JH, Teran NA, Tsang EK, Zhang H, Zhou YH, et al. (2017) Genetic effects on gene expression across human tissues. Nature 550: 204. doi: 10.1038/nature24277

Alex AM, Saradalekshmi KR, Shilen N, Suresh PA, Banerjee M (2019) Genetic association of DNMT variants can play a critical role in defining the methylation patterns in autism. Iubmb Life 71: 901-907. doi: 10.1002/iub.2021

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu YM, Wang J, Chang YQ, Feng Q, Fang XD, Guo XS, Jian M, Jiang H, Jin X, Lan TM, Li GQ, Li JX, Li YR, Liu SM, Liu X, Lu Y, Ma XD, Tang MF, Wang B, Wang GB, Wu HL, Wu RH, Xu X, Yin Y, Zhang DD, Zhang WW, Zhao J, Zhao MR, Zheng XL, Lander ES, Altshuler DM, Gabriel SB, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Flicek P, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Bentley DR, Grocock R, Humphray S, James T, Kingsbury Z, Lehrach H, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, et al. (2015) A global reference for human genetic variation. Nature 526: 68. doi: 10.1038/nature15393

Ardlie KG, DeLuca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou YH, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M, Reverter F, Goldmann JM, Koller D, Guigo

R, McCarthy MI, Dermitzakis ET, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen XQ, Stephens M, Pritchard JK, Tu ZD, Zhang B, Huang T, Long Q, Lin L, Yang JL, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S, Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD, Branton PA, Sobin LH, Barcus M, Qi LQ, McLean J, Hariharan P, Um KS, Wu SP, Tabor D, Shive C, Smith AM, Buia SA, et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348: 648-660. doi: 10.1126/science.1262110

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Consortium GO (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29. doi: 10.1038/75556

Baryshnikova A, Costanzo M, Myers CL, Andrews B, Boone C (2013) Genetic Interaction Networks: Toward an Understanding of Heritability. Annual Review of Genomics and Human Genetics, Vol 14 14: 111-133. doi: 10.1146/annurev-genom-082509-141730

Bateson W (1903) Mendel's principles of heredity in mice. Nature 68: 33-34. doi: 10.1038/068033c0

Bennabi M, Gaman A, Delorme R, Boukouaci W, Manier C, Scheid I, Mohammed NS, Bengoufa D, Charron D, Krishnamoorthy R, Leboyer M, Tamouza R (2018) HLA-class II haplotypes and Autism Spectrum Disorders. Scientific Reports 8. doi: 10.1038/S41598-018-25974-9

Brown KM, Costanzo MS, Xu WX, Roy S, Lozovsky ER, Hartl DL (2010) Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase. Molecular Biology and Evolution 27: 2682-2690. doi: 10.1093/molbev/msq160

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans D, Leung HT, Marchini JL, Morris AP, Spencer CCA, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop DT, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Mathew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678. doi: 10.1038/nature05911

Carlin DE, Fong SH, Qin Y, Jia TQ, Huang JK, Bao BK, Zhang C, Ideker T (2019) A Fast and Flexible

Framework for Network-Assisted Genomic Association. Iscience 16: 155. doi: 10.1016/j.isci.2019.05.025

Castermans D, Volders K, Crepel A, Backx L, De Vos R, Freson K, Meulemans S, Vermeesch JR, Schrander-Stumpel CTRM, De Rijk P, Del-Favero J, Van Geet C, Van De Ven WJM, Steyaert JG, Devriendt K, Creemers JWM (2010) SCAMP5, NBEA and AMISYN: three candidate genes for autism involved in secretion of large dense-core vesicles. Human Molecular Genetics 19: 1368-1378. doi: 10.1093/hmg/ddq013

Chahrour M, Zoghbi HY (2007) The story of Rett syndrome: From clinic to neurobiology. Neuron 56: 422-437. doi: 10.1016/j.neuron.2007.10.001

Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 10: 392-404. doi: 10.1038/nrg2579

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan GH, Wang W, Usaj M, Hanchard J, Lee SD, Pelechano V, Styles EB, Billmann M, van Leeuwen J, van Dyk N, Lin ZY, Kuzmin E, Nelson J, Piotrowski JS, Srikumar T, Bahr S, Chen YQ, Deshpande R, Kurat CF, Li SC, Li ZJ, Usaj MM, Okada H, Pascoe N, San Luis BJ, Sharifpoor S, Shuteriqi E, Simpkins SW, Snider J, Suresh HG, Tan YZ, Zhu HW, Malod-Dognin N, Janjic V, Przulj N, Troyanskaya OG, Stagljar I, Xia T, Ohya Y, Gingras AC, Raught B, Boutros M, Steinmetz LM, Moore CL, Rosebrock AP, Caudy AA, Myers CL, Andrews B, Boone C (2016) A global genetic interaction network maps a wiring diagram of cellular function. Science 353. doi: 10.1126/science.aaf1420

DeLano WL (2002) Pymol: An open-source molecular graphics tool. CCP4 Newsletter on protein crystallography 40: 82-92.

El-Ansary A, Al-Ayadhi L (2012) Lipid mediators in plasma of autism spectrum disorders. Lipids in Health and Disease 11: 1-9. doi: 10.1186/1476-511x-11-160

Eres IE, Luo KX, Hsiao CJ, Blake LE, Gilad Y (2019) Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. Plos Genetics 15. doi: 10.1371/journal.pgen.1008278

Fang G, Wang W, Paunic V, Heydari H, Costanzo M, Liu XY, Liu XT, VanderSluis B, Oately B, Steinbach M, Van Ness B, Schadt EE, Pankratz ND, Boone C, Kumar V, Myers CL (2019) Discovering genetic interactions bridging pathways in genome-wide association studies. Nature Communications 10: 1-18. doi: 10.1038/S41467-019-12131-7

Fu Y, Tessneer KL, Li C, Gaffney PM (2018) From association to mechanism in complex disease genetics: the role of the 3D genome. Arthritis Research & Therapy 20: 216. doi: 10.1186/S13075-018-1721-X

Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, Hemberg M, Ebert DH, Greenberg ME (2015) Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature 522: 89-93. doi: 10.1038/nature14319

Gallagher MD, Chen-Plotkin AS (2018) The Post-GWAS Era: From Association to Function. American Journal of Human Genetics 102: 717-730. doi: 10.1016/j.ajhg.2018.04.002

Giusti-Rodriguez P, Sullivan P (2019) Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. bioRxiv. 406330.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T, Troyanskaya OG (2015) Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics 47: 569-576. doi: 10.1038/ng.3259

Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nature Reviews Genetics 4: 701-709. doi: 10.1038/nrg1155

Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hagg S, Athanasiu L, Voyle N, Proitsi P, Witoelar A, Stringer S, Aarsland D, Almdahl IS, Andersen F, Bergh S, Bettella F, Bjornsson S, Braekhus A, Brathen G, de Leeuw C, Desikan RS, Djurovic S, Dumitrescu L, Fladby T, Hohman TJ, Jonsson PV, Kiddle SJ, Rongve A, Saltvedt I, Sando SB, Selbaek G, Shoai M, Skene NG, Snaedal J, Stordal E, Ulstein ID, Wang YP, White LR, Hardy J, Hjerling-Leffler J, Sullivan PF, van der Flier WM, Dobson R, Davis LK, Stefansson H, Stefansson K, Pedersen NL, Ripke S, Andreassen OA, Posthuma D (2019a) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genetics 51: 404-413. doi: 10.1038/s41588-018-0311-9

Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hagg S, Athanasiu L, Voyle N, Proitsi P, Witoelar A, Stringer S, Aarsland D, Almdahl IS, Andersen F, Bergh S, Bettella F, Bjornsson S, Braekhus A, Brathen G, de Leeuw C, Desikan RS, Djurovic S, Dumitrescu L, Fladby T, Hohman TJ, Jonsson PV, Kiddle SJ, Rongve A, Saltvedt I, Sando SB, Selbaek G, Shoai M, Skene NG, Snaedal J, Stordal E, Ulstein ID, Wang YP, White LR, Hardy J, Hjerling-Leffler J, Sullivan PF, van der Flier WM, Dobson R, Davis LK, Stefansson H, Stefansson K, Pedersen NL, Ripke S, Andreassen OA, Posthuma D (2019b) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genetics 51: 404. doi: 10.1038/s41588-018-0311-9

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Research 30: 42-46. doi: 10.1093/Nar/30.1.42

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348-54. doi: 10.1038/ng.548

Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G, Baranasic D, Arenillas DJ, Sandelin A, Vandepoele K, Lenhard B, Ballester B, Wasserman WW, Parcy F, Mathelier A (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research 46: D260-D266. doi: 10.1093/nar/gkx1126

Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL (2019) Leveraging Polygenic Functional Enrichment to Improve GWAS Power. American Journal of Human Genetics 104: 65-75. doi: 10.1016/j.ajhg.2018.11.008

Lasky JL, Wu H (2005) Notch signaling, brain development, and human disease. Pediatric Research 57: 104-109. doi: 10.1203/01.Pdr.0000159632.70510.3d

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed

models for genome-wide association studies. Nature Methods 8: 833-835. doi: 10.1038/Nmeth.1681

Long Q, Zhang Q, Vilhjalmsson BJ, Forai P, Seren Ü, Nordborg M (2013) JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models. Bioinformatics 29: 1220-1222.

Lu QS, Yao XW, Hu YM, Zhao HY (2016) GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics 32: 542-548. doi: 10.1093/bioinformatics/btv610

Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Research 29: 2860-2874. doi: 10.1093/nar/29.13.2860

Mackay TFC (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nature Reviews Genetics 15: 22-33. doi: 10.1038/nrg3627

Madsen AM, Ottman R, Hodge SE (2011) Causal Models for Investigating Complex Genetic Disease: II. What Causal Models Can Tell Us about Penetrance for Additive, Heterogeneity, and Multiplicative Two-Locus Models. Human Heredity 72: 63-72. doi: 10.1159/000330780

Mah W, Won H (2019) The three-dimensional landscape of the genome in human brain tissue unveils regulatory mechanisms leading to schizophrenia risk. Schizophr Res. doi: 10.1016/j.schres.2019.03.007

Mao FB, Xiao LY, Li XF, Liang JL, Teng HJ, Cai WS, Sun ZS (2016) RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins. Nucleic Acids Research 44: D154-D163. doi: 10.1093/nar/gkv1308

Markkanen E, Meyer U, Dianov GL (2016) DNA Damage and Repair in Schizophrenia and Autism: Implications for Cancer Comorbidity and Beyond. International Journal of Molecular Sciences 17. doi: 10.3390/Ijms17060856

Melo US, Schopflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, Klever MK, Turkmen S, Heinrich V, Pluym ID, Matoso E, de Sousa SB, Louro P, Hulsemann W, Cohen M, Dufke A, Latos-Bielenska A, Vingron M, Kalscheuer V, Quintero-Rivera F, Spielmann M, Mundlos S (2020) Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. American Journal of Human Genetics 106: 872-884. doi: 10.1016/j.ajhg.2020.04.016

Miguel-Escalada I, Bonas-Guarch S, Cebola I, Ponsa-Cobas J, Mendieta-Esteban J, Atla G, Javierre BM, Rolando DMY, Farabella I, Morgan CC, Garcia-Hurtado J, Beucher A, Moran I, Pasquali L, Ramos-Rodriguez M, Appel EVR, Linneberg A, Gjesing AP, Witte DR, Pedersen O, Grarup N, Ravassard P, Torrents D, Mercader JM, Piemonti L, Berney T, de Koning EJP, Kerr-Conte J, Pattou F, Fedko IO, Groop L, Prokopenko I, Hansen T, Marti-Renom MA, Fraser P, Ferrer J (2019) Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nature Genetics 51: 1137. doi: 10.1038/s41588-019-0457-0

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer

C, Liu H, Zhao T, Cai GQ, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu YQ, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242-245. doi: 10.1038/nature11011

Noble WS (2009) How does multiple testing correction work? Nature Biotechnology 27: 1135-1137. doi: 10.1038/nbt1209-1135

Phillips PC (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9: 855-867. doi: 10.1038/nrg2452

Pickrell JK (2014) Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. American Journal of Human Genetics 94: 559-573. doi: 10.1016/j.ajhg.2014.03.004

Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research 48: D845-D855. doi: 10.1093/nar/gkz1021

Pujato M, Kieken F, Skiles AA, Tapinos N, Fiser A (2014) Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. Nucleic Acids Research 42: 13500-13512. doi: 10.1093/nar/gku1228

Qasem H, Al-Ayadhi L, El-Ansary A (2016) Cysteinyl leukotriene correlated with 8-isoprostane levels as predictive biomarkers for sensory dysfunction in autism. Lipids in Health and Disease 15: 1-10. doi: 10.1186/S12944-016-0298-0

Rajarajan P, Borrman T, Liao W, Schrode N, Flaherty E, Casino C, Powell S, Yashaswini C, LaMarca EA, Kassim B, Javidfar B, Espeso-Gil S, Li A, Won H, Geschwind DH, Ho SM, MacDonald M, Hoffman GE, Roussos P, Zhang B, Hahn CG, Weng Z, Brennand KJ, Akbarian S (2018) Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science 362: 1269. doi: 10.1126/science.aat4311

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159: 1665-1680. doi: 10.1016/j.cell.2014.11.021

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Tobacco, Genetics C, Bipolar Disorder Psychiatric Genomics C, Schizophrenia Psychiatric Genomics C, Schork NJ, Andreassen OA, Dale AM (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet 9: e1003449. doi: 10.1371/journal.pgen.1003449

Stamova BS, Tian YF, Nordahl CW, Shen MD, Rogers S, Amaral DG, Sharp FR (2013) Evidence for differential alternative splicing in blood of young boys with autism spectrum disorders. Molecular Autism 4. doi: 10.1186/2040-2392-4-30

Stroud H, Su SC, Hrvatin S, Greben AW, Renthal W, Boxer LD, Nagy MA, Hochbaum DR, Kinde B,

Gabel HW, Greenberg ME (2017) Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. Cell 171: 1151. doi: 10.1016/j.cell.2017.09.047

Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G, Holm H, Kong A, Thorsteinsdottir U, Sulem P, Gudbjartsson DF, Stefansson K (2016) Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nature Genetics 48: 314-317. doi: 10.1038/ng.3507

Tak YG, Farnham PJ (2015) Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics & Chromatin 8. doi: 10.1186/s13072-015-0050-4

Tatton-Brown K, Seal S, Ruark E, Harmer J, Ramsay E, Duarte SD, Zachariou A, Hanks S, O'Brien E, Aksglaede L, Baralle D, Dabir T, Gener B, Goudie D, Homfray T, Kumar A, Pilz DT, Selicorni A, Temple IK, Van Maldergem L, Yachelevich N, van Montfort R, Rahman N, Consortium CO (2014) Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nature Genetics 46: 385. doi: 10.1038/ng.2917

Theoharides TC, Tsilioni I, Patel AB, Doyle R (2016) Atopic diseases and inflammation of the brain in the pathogenesis of autism spectrum disorders. Translational Psychiatry 6. doi: 10.1038/tp.2016.77

Ward LD, Kellis M (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research 40: D930-D934. doi: 10.1093/nar/gkr917

Watson HJ, Yilmaz Z, Thorntont LM, Hubel C, Coleman JRI, Gaspar HA, Bryois J, Hinney A, Leppa VM, Mattheisen M, Medland SE, Ripke S, Yao SY, Giusti-Rodriguez P, Hanscombe KB, Purves KL, Adan RAH, Alfredsson L, Ando T, Andreassen OA, Baker JH, Berrettini WH, Boehm I, Boni C, Perica VB, Buehren K, Burghardt R, Cassina M, Cichon S, Clementi M, Cone RD, Courtet P, Crow S, Crowley JJ, Danner UN, Davis OSP, de Zwaan M, Dedoussis G, Degortes D, DeSocio JE, Dick DM, Dikeos D, Dina C, Dmitrzak-Weglarz M, Docampo E, Duncan LE, Egberts K, Ehrlich S, Escaramis G, Eskos T, Estivill X, Farmer A, Favaro A, Fernandez-Aranda F, Fichter MM, Fischer K, Focker M, Foretova L, Forstner AJ, Forzan M, Franklin CS, Gallinger S, Giegling I, Giuranna J, Gonidakis F, Gorwood P, Mayora MG, Guillaume S, Guo YR, Hakonarson H, Hatzikotoulas K, Hauser J, Hebebrand J, Helder SG, Herms S, Herpertz-Dahlmann B, Herzog W, Huckins LM, Hudson JI, Imgart H, Inoko H, Janout V, Jimenez-Murcia S, Julia A, Kalsi G, Kaminska D, Kaprio J, Karhunen L, Karwautz A, Kas MJH, Kennedy JL, Keski-Rahkonen A, Kiezebrink K, Kim YR, Klareskog L, Klump KL, Knudsen GPS, La Via MC, Le Hellard S, Levitan RD, et al. (2019) Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nature Genetics 51: 1207. doi: 10.1038/s41588-019-0439-2

Wen XQ, Lee Y, Luca F, Pique-Regi R (2016) Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. American Journal of Human Genetics 98: 1114-1129. doi: 10.1016/j.ajhg.2016.03.029

Whalen S, Pollard KS (2019) Most chromatin interactions are not in linkage disequilibrium. Genome research 29: 334-343.

Won HJ, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu DN, Lee C, Eskin E, Voineagu I, Ernst J, Geschwind DH (2016) Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538: 523. doi: 10.1038/nature19847

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X (2010) Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86: 929-942. doi: 10.1016/j.ajhg.2010.05.002

Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH (2011) Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American Journal of Human Genetics 89: 82-93. doi: 10.1016/j.ajhg.2011.05.029

Xiong Z, Zhang QR, Platt A, Liao WY, Shi XH, de los Campos G, Long Q (2019) OCMA: Fast, Memory-Efficient Factorization of Prohibitively Large Relationship Matrices. G3-Genes Genomes Genetics 9: 13-19. doi: 10.1534/g3.118.200908

Yan YM, Tao HY, He JH, Huang SY (2020) The HDOCK server for integrated protein-protein docking. Nature Protocols 15: 1829-1852. doi: 10.1038/s41596-020-0312-x

Yang JJ, Fritsche LG, Zhou X, Abecasis G, Degene IA-RM (2017) A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. American Journal of Human Genetics 101: 404-416. doi: 10.1016/j.ajhg.2017.08.002

Yokoi T, Enomoto Y, Naruto T, Kurosawa K, Higurashi N (2020) Tatton-Brown-Rahman syndrome with a novel DNMT3A mutation presented severe intellectual disability and autism spectrum disorder. Human Genome Variation 7: 1-3. doi: 10.1038/s41439-020-0102-6

Yu GC, Wang LG, Han YY, He QY (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omics-a Journal of Integrative Biology 16: 284-287. doi: 10.1089/omi.2011.0118

Yu JT, Hu M, Li C (2019) Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. Bmc Genetics 20. doi: 10.1186/s12863-019-0744-x

Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang ZZ, Pellecchia G, Buchanan JA, Walker S, Marshall CR, Uddin M, Zarrei M, Deneault E, D'Abate L, Chan AJS, Koyanagi S, Paton T, Pereira SL, Hoang N, Engchuan W, Higginbotham EJ, Ho K, Lamoureux S, Li WL, MacDonald JR, Nalpathamkalam T, Sung WWL, Tsoi FJ, Wei J, Xu LZ, Tasse AM, Kirby E, Van Etten W, Twigger S, Roberts W, Drmic I, Jilderda S, Modi BM, Kellam B, Szego M, Cytrynbaum C, Weksberg R, Zwaigenbaum L, Woodbury-Smith M, Brian J, Senman L, Iaboni A, Doyle-Thomas K, Thompson A, Chrysler C, Leef J, Savion-Lemieux T, Smith IM, Liu XD, Nicolson R, Seifer V, Fedele A, Cook EH, Dager S, Estes A, Gallagher L, Malow BA, Parr JR, Spence SJ, Vorstman J, Frey BJ, Robinson JT, Strug LJ, Fernandez BA, Elsabbagh M, Carter MT, Hallmayer J, Knoppers BM, Anagnostou E, Szatmari- P, Ring RH, Glazer D, Pletcher MT, Scherer SW (2017) Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nature Neuroscience 20: 602-611. doi: 10.1038/nn.4524

Zhou JX, Wang XJ, Cheng W, Pan CL, Xing XB (2019) Development and validation of a novel and

robust blood small nuclear RNA signature in diagnosing autism spectrum disorder. Medicine 98. doi: 10.1097/MD.0000000000017858