bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Clustering by phenotype and genome-wide association study in autism

Akira Narita[1,2], Masato Nagai[1,2], Satoshi Mizuno[1,2], Soichi Ogishima[1,2], Gen Tamiya[1,2,3], Masao Ueki[1,2,3], Rieko Sakurai[1,2,3], Satoshi Makino[1,2,3], Taku Obara[1,2,4], Mami Ishikuro[1,2], Chizuru Yamanaka[1,2], Hiroko Matsubara[1,2], Yasutaka Kuniyoshi[2], Keiko Murakami[1,2], Fumihiko Ueno[1,2], Aoi Noda[1,2], Tomoko Kobayashi[1,2,4], Mika Kobayashi[1], Takuma Usuzaki[1], Hisashi Ohseto[1], Atsushi Hozawa[1,2], Masahiro Kikuya[1,2,5], Hirohito Metoki[1,2,6], Shigeo Kure[1,2,4], Shinichi Kuriyama[1,2,7,*]

[1]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan

[2]Graduate School of Medicine, Tohoku University, Sendai, Japan

[3]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

[4]Tohoku University Hospital, Tohoku University, Sendai, Japan

[5]School of Medicine, Teikyo University, Tokyo, Japan

[6]School of Medicine, Tohoku Medical and Pharmaceutical University, Sendai, Japan

[7]International Research Institute of Disaster Science, Tohoku University, Sendai, Miyagi, Japan

These authors contributed equally: Akira Narita, Masato Nagai, and Satoshi Mizuno.

[*]Corresponding author

Shinichi Kuriyama, MD, PhD

Department of Molecular Epidemiology, Environment and Genome Research Center, Tohoku University Graduate School of Medicine.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2-1, Seiryo-machi, Aoba-ku, Sendai, Miyagi, 980-8575, Japan.

Phone: +81-22-717-8102, Fax: +81-22-717-8106

E-mail: kuriyama@med.tohoku.ac.jp

Short title

Clustering and genome-wide association study

Keywords

Clustering, phenotype, genome-wide association study, autism, machine learning, k-means

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Background: Autism spectrum disorder (ASD) has phenotypically and genetically heterogeneous characteristics. A simulation study demonstrated that attempts to categorize patients with a complex disease into more homogeneous subgroups could have more power to elucidate hidden heritability.

Methods: We conducted cluster analyses using the k-means algorithm with a cluster number of 15 based on phenotypic variables from the Simons Simplex Collection (SSC). As a preliminary study, we conducted a conventional genome-wide association study (GWAS) with a dataset of 597 ASD cases and 370 controls. In the second step, we divided cases based on the clustering results and conducted GWAS in each of the subgroups vs controls (cluster-based GWAS). We also conducted cluster-based GWAS on another SSC dataset of 712 probands and 354 controls in the replication stage.

Results: In the preliminary study, we observed no significant associations. In the second step of cluster-based GWASs, we identified 65 chromosomal loci, which included 30 intragenic loci located in 21 genes and 35 intergenic loci that satisfied the threshold of $P<5.0\times10^{-8}$. Some of these loci were located within or near previously reported candidate genes for ASD: *CDH5*, *CNTN5, CNTNAP5, DNAH17, DPP10, DSCAM*, *FOXK1*, *GABBR2, GRIN2A*5, *ITPR1, NTM, SDK1, SNCA* and *SRRM4.* Of these 65 significant chromosomal loci, rs11064685 located within the *SRRM4* gene had a significantly different distribution in the cases vs. controls in the replication cohort.

Conclusions: These findings suggest that clustering may successfully identify subgroups with relatively homogeneous disease etiologies. Further cluster validation and replication studies are warranted in larger cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Autism spectrum disorder (ASD) has heterogeneous characteristics in terms of both phenotypic features and genetics. ASD is mainly characterized by difficulties in communication and repetitive behaviors (1), but ASD also shows many other symptoms (2). Regarding genetics, previous studies have not consistently identified genetic variants that are associated with an increased risk of ASD (3), although several lines of evidence suggest that genetic factors strongly contribute to the increased risk of ASD. Monozygotic twins have higher concordance rates of ASD (92%) than dizygotic twins (10%) (4). The recurrence risk ratio is 22 for ASD among siblings (5). The Human Gene module of the Simons Foundation Autism Research Initiative (SFARI) Gene provides a comprehensive reference for suggested human ASD-related genes in an up-to-date manner (6) and currently demonstrates ~1,000 genes that may have links to ASD, potentially indicating the heterogeneity of ASD. In addition to phenotype and genotype heterogeneities, ASD shows heterogeneous responses to interventions. Several kinds of pharmacological treatments are suggested, but the effects of these treatments are controversial (7).

If the heterogeneous phenotypes and responses to treatment in some way correspond to differences in genotype, grouping persons with ASD according to phenotype and responses to treatment variables may increase the chances of identifying genetic susceptibility factors. Traylor and colleagues (8) demonstrated that attempts to categorize patients with a complex disease into more homogeneous subgroups could have more power to elucidate the hidden heritability in a simulation study. Several studies on Alzheimer's disease, neuroticism, or asthma indicated that items or symptoms were to some degree more useful for identifying high-impact genetic factors than broadly defined diagnoses (9-11), although a study of ASD demonstrated modest effects of

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

two-way stratification by individual symptoms (12). Additionally, medical researchers have begun to use machine learning methods (13), which is an artificial intelligence technique that can reveal masked patterns of data sets. In view of the abovementioned circumstances, clustering algorithms of machine learning could be hypothesized to make novel and more genetically homogeneous clusters, but these algorithms using phenotypic variables have not, to the best of our knowledge, been applied to subgrouping ASD to date.

We therefore explored whether grouping persons with ASD using a clustering algorithm with phenotype and responses to treatment variables can be used to discriminate more genetically homogeneous persons with ASD. In the present study, we conducted cluster-based genome-wide association studies (named cluster-based GWASs) using real data based on the concept of a previous simulation study (8) adopting a machine learning k-means (14) algorithm for cluster analysis.

Methods and Materials

We conducted the present study in accordance with the guidelines of the Declaration of Helsinki (15) and all other applicable guidelines. The protocol was reviewed and approved by the institutional review board of Tohoku University Graduate School of Medicine, and written informed consent was obtained from all participants over the age of 18 by the Simons Foundation Autism Research Initiative (SFARI) (16). For participants under the age of 18, informed consent was obtained from a parent and/or legal guardian. Additionally, for participants 10 to 17 years of age, informed assent was obtained from the individuals.

Datasets

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We used phenotypic variables, history of treatment, and genotypic data from the Simons Simplex Collection (SSC) (16). The SSC establishes a repository of phenotypic data and genetic data/samples from mainly simplex families.

The SSC data were publicly released in October 2007 and are directly available from the SFARI. From the SSC dataset, we used data from 614 affected white male probands who had no missing information regarding Autism Diagnostic Interview-Revised (ADI-R) scores (17) and vitamin treatment (18, 19) and 391 unaffected brothers for whom genotype data, generated by the Illumina Human Omni2.5 (Omni2.5) array, were available for subsequent clustering and genetic analyses. We excluded participants whose ancestries were estimated to be different from the other participants using principal component analyses (PCAs) performed by EIGENSOFT version 7.2.1 (20, 21) for the genotype data. Based on the PCAs, we excluded data beyond 4 standard deviations of principal components 1 or 2 (Figure S1). Therefore, we used data from 597 probands and 370 unaffected brothers.

In the replication study, we used another SSC dataset genotyped using the Illumina 1Mv3 (1Mv3) array. In the dataset, data from 735 affected male probands with no missing information regarding ADI-R scores or vitamin treatment and 387 unaffected brothers were available. After conducting PCA, we excluded data beyond 4 standard deviations of principal components 1 or 2 as outliers. In this way, we used data from 712 probands and 354 unaffected brothers in the replication study.

Clustering

We conducted cluster analyses using phenotypic variables of ADI-R (17) scores and history of vitamin treatment (18, 19). We chose these variables because the ADI-R is one of the most

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reliable estimates of ASD and has the ability to evaluate substructure domains of ASD (17). Among the ADI-R scores, "the total score for the Verbal Communication Domain of the ADI-R minus the total score for the Nonverbal Communication Domain of the ADI-R", "the total score for the Nonverbal Communication Domain of the ADI-R", "the total score for the Restricted, Repetitive, and Stereotyped Patterns of Behavior Domain of the ADI-R", and "the total score for the Reciprocal Social Interaction Domain of the ADI-R" were included in the preprocessed dataset.

Among the treatments, we selected the variable of history of vitamin treatment because we recently found that a cluster of persons with ASD is associated with potential responsiveness to vitamin B6 treatment (18, 19). The history of treatment is not always compatible with responsiveness, but we considered that continuous treatment indicates responsiveness to some degree. The SSC dataset includes history of treatment but not variables of responsiveness.

We applied the machine learning k-means (14) algorithm to conduct a cluster analysis to divide the dataset including data from ASD persons into subgroups using phenotypic variables and history of treatment. The k-means algorithm requires cluster numbers determined by researchers. When using k-means algorithms, we set a priori the cluster numbers of 2, 3, 4, 5, 10, 15, and 20. We performed the analyses using the scikit-learn toolkit in Python 2.7 (Supplementary Information S1) (22).

Clustering is an exploratory data analysis technique, and the validity of the clustering results may be judged by external knowledge, such as the purpose of the segmentation (23). Several methods have proposed to prespecify a cluster number (k), such as visual examination of the data, and likelihood and error-based approaches; however, these methods do not necessarily provide results that are consistent with each other (24). Although there are measures for

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

evaluating the quality of the clusters (25), the number of clusters should also be determined according to the research purposes. We regarded the inflation factor (λ) of quantile-quantile (Q-Q) plots of the logarithm of the P-value to base 10 ($-\log_{10}P$) as one of the indicators of successful clustering in the present study. We calculated λ for each cluster number.

When conducting clustering, we combined the two datasets of male probands, one genotyped using the Omni2.5 array and the other genotyped using the 1Mv3 array. After clustering, we redivided the new dataset according to the SNP arrays used. In the discovery stage, we used the Omni2.5 dataset and the 1Mv3 dataset in the replication stage.

Genotype data and quality control

We used the SSC dataset, in which probands and unaffected brothers had already been genotyped in other previous studies (16, 26). In the discovery stage, we used the dataset genotyped by the Omni2.5 array, which has 2,383,385 probes. We excluded SNPs with a minor allele frequency (MAF) < 0.01, call rate < 0.95, and Hardy-Weinberg equilibrium test P < 0.000001.

In the replication study, where we used the dataset genotyped using the 1Mv3 array, we applied the same cutoff values for quality control as those used in the discovery stage. The 1Mv3 array includes 1,147,689 SNPs. The Omni2.5 array and the 1Mv3 array shared 675,923 SNPs.

Statistical analysis

As a preliminary study, we conducted a conventional GWAS in the whole Omni2.5 dataset, with a total of 597 male probands and 370 unaffected brothers. Here, we used the brothers of the cases as controls, in contrast to many previous studies in which genetically unrelated controls were used. We thus adopted the sib transmission disequilibrium test (sib-TDT) (27), a family-based

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

association test, to take into account familial relationships among the participants. In the second step, in the discovery stage, we conducted cluster-based GWAS in each subgroup of the cases, which had been divided using the k-means (14) algorithm, and the controls. As mentioned above, the controls were the brothers of the cases, and we then excluded the unaffected brothers of the cases belonging to the subgroup being analyzed. Details of the study design are shown in Figure 1. We applied the Cochran-Armitage trend test (28), which examines the risk of disease in those who do not have the allele of interest, those who have a single copy, and those who are homozygous.

We further tested the significantly associated loci found in the discovery studies in the replication stage. The level of significance for association was set as $P < 0.05$ in the replication studies.

Association analyses were performed with the PLINK software package (29). The detected SNPs were subsequently annotated using ANNOVAR (30). Manhattan plots and Q-Q plots were generated using the 'qqman' package in R version 3.0.2.

Data availability

All the data used in the study are available only to those granted access by the Simons Foundation.

Results

Cluster-based genome-wide association study

As a preliminary study, we conducted a conventional GWAS with the Omni2.5 dataset using the sib-TDT. We observed no significant associations (Figure 2). Although we adopted the sib-TDT

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

here because we used the brothers of the cases as controls, we also used the Cochran-Armitage trend test and found that the $-\log_{10}P$ values were distributed downward compared with the expected values, as shown in Figure S2.

We also applied the sib-TDT to cluster 1, which was obtained by dividing all the cases using k-means with k of 15, and all the controls and found that the observed -logP values were lower than expected, as shown in Figure S2. Since the sib-TDT may efficiently work in a population consisting of a substantial number of sibs, a limited number of brothers of the probands among all the controls probably contributed to a substantial loss of power. Thus, we excluded the brothers of the probands in each subset from the controls so that each subset of probands has no genetic relations with the rest of the controls and conducted the Cochran-Armitage trend test, as in many other studies. In the present study, therefore, we applied the sib-TDT to the GWAS of the whole dataset, whereas in the cluster-based GWAS, we excluded in turn the unaffected brothers of the cases belonging to the subgroup being analyzed and used the Cochran-Armitage trend test to account for the relationships between participants.

Under the hypothesis that ASD consists of hundreds of subgroups (16), we compared λ values giving larger numbers of clusters as priority. The λ for the cluster-based GWAS with a k of 20 ranged from 1.015 to 1.107, and the average was 1.053 (Table 1), indicating that the rate of false positives was relatively high. Several lines of evidence suggest that regarding an appropriate threshold of inflation factor λ, empirically, a value less than 1.050 is deemed safe for avoiding false positives (31-33).

In contrast, λ with k of 15 ranged from 1.017 to 1.091, and the average was 1.038, which was below 1.050 (Table 1 and Figure 3).

According to the above results, we considered the cluster-based GWAS using the

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Cochran-Armitage trend test, coupled with k-means cluster analysis with k of 15, to be the most appropriate approach to the present dataset. The characteristics of each cluster are presented in Table 2.

Gene interpretation

We observed 65 chromosomal loci that satisfied the threshold of $P < 5.0 \times 10^{-8}$ (Table 1 and Figure 3); 30 out of the 65 loci were located within 21 genes, and the remaining 35 loci were intergenic (Table 3). Among them, 8 loci were located within or near the genes associated with the Human Gene module of the SFARI Gene scoring system (6); *GABBR2* (score 4, Rare Single Gene Mutation, Syndromic, Functional) in Cluster 1; *CNTNAP5* (score 4, Rare Single Gene Mutation, Genetic Association) in Cluster 3; *ITPR1* (score 4, Rare Single Gene Mutation) in Cluster 5; *DNAH17* (score 4, Rare Single Gene Mutation) in Cluster 7; *SDK1* (score none, Rare Single Gene Mutation, Genetic Association) in Cluster 13; *SRRM4* (score 5, Rare Single Gene Mutation, Functional) in Cluster 13; *CNTN5* (score 3, Rare Single Gene Mutation, Genetic Association) in Cluster 14; and *DPP10* (score 3, Rare Single Gene Mutation) in Cluster 15.

The SFARI Gene scoring system ranges from “Category 1”, which indicates “high confidence”, through “Category 6”, which denotes “evidence does not support a role”. Genes of a syndromic disorder (e.g., fragile X syndrome) related to ASD are categorized in a different category. Rare single gene variants, disruptions/mutations, and submicroscopic deletions/duplications related to ASD are categorized as “Rare Single Gene Mutation”.

In addition to genes in the Human Gene module of the SFARI Gene, several important genes associated with ASD or other related disorders (34-37) from previous reports were included in our findings as follows: *CDH5* in Cluster 14, *DSCAM* in Cluster 8, *FOXK1* in Cluster

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13, *GRIN2A* in Cluster 5, *NTM* in Cluster 8, and *SNCA* in Cluster 11 previously reported with ASD (38-43); *PLCH2* in Cluster 11 previously reported with mental retardation (44); *ARHGAP18* in Cluster 18, *CDC42BPA* in Cluster 3, *CXCL12* in Cluster 8, and *HS3ST2* in Cluster 5 previously reported with schizophrenia (45-48); *KCTD12* in Cluster 9 and *PSAT1* in Cluster 8 previously reported with depressive disorder (49, 50); *ADAMTS1* in Cluster 10, *DOCK2* in Cluster 10, *HS3ST2* in Cluster 5, *NAMPT* in Cluster 5, and *NAV* in Cluster 5 previously reported with Alzheimer's disease (51-55); and *PEX10* in Cluster 11 previously reported with Down syndrome (56).

Replication study

We conducted replication studies with another independent dataset that included a total of 712 male probands and 354 unaffected brothers and had been genotyped using the 1Mv3 array. As mentioned before, we had previously carried out cluster analyses in the combined dataset genotyped with either Omni2.5 or 1Mv3 and then redivided it according to the SNP arrays used. The characteristics of each of the 15 clusters in the 1Mv3 dataset are presented in Table S1.

Among the 65 genome-wide significant chromosomal loci found in the discovery study, seven chromosomal loci were included in the 1Mv3 array. Of these loci, rs11064685, within *SRRM4* in Cluster 13, had a significantly different distribution ($p = 0.03$) in cases vs. controls in the replication cohort (Table 4).

Discussion

One of the most important findings of our study was that reasonably decreasing the sample size could increase the statistical power. A plausible explanation is that our clustering may have

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

successfully identified subgroups that are etiologically more homogeneous. At least three reasons could reduce the possibility of false positives of the present results of statistically significant SNPs in cluster-based GWAS. First, the present study validated the usefulness and feasibility of the concept of a previous simulation study (8), which indicated that homogeneous case subgroups increase power in genetic association studies by Traylor and colleagues, using measurement data in the real world. Second, a substantial number of statistically significant SNPs in cluster-based GWAS observed in the present study were located within or near previously reported candidate genes for ASD (6, 38-43). Third, we calculated λ of Q-Q plots of logarithm of P-values to base 10 ($-\log_{10}P$) for each cluster number and compared them for the validity of clustering. Genomic control is widely used to control false positive signals due to population stratification in GWAS. λ is commonly used for genomic control.

We observed many statistically significant SNPs in cluster-based GWAS: *CDH5*, *CNTN5, CNTNAP5, DNAH17, DPP10, DSCAM*, *FOXK1*, *GABBR2, GRIN2A5*, *ITPR1, NTM, SDK1, SNCA* and *SRRM4.* In particular, loci within the *SRRM4* gene had significantly different distributions in the cases vs. controls in the replication cohort. Previous studies indicate that *SRRM4* is strongly associated with ASD, indicating that our results may be valid to some degree. The gene regulates neural microexons. In the brains of individuals with ASD, these microexons are frequently dysregulated (57). Additionally, nSR100/SRRM4 haploinsufficiency in mice induced autistic features such as sensory hypersensitivity and altered social behavior and impaired synaptic transmission and excitability (58).

In addition to *SMMR4*, we observed several genes located within or near previously reported candidate genes for ASD. The relatively high correspondence between our results in part and the SFARI Gene scoring system (6) indicates that the statistically significant loci we

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

found may be associated with ASD subgroups (Table 1 and Figure 3). We also observed several important genes associated with ASD and other related disorders (34-37) from previous reports. These findings suggest that the statistically significant SNPs might explain autistic symptoms because these diseases are suggested to have shared etiology, even in part, with ASD (34-37). Associations at the remaining significant loci that were not in the SFARI module or described above have not been previously reported, and to the best of our knowledge, some of them might be novel findings. These results might suggest that novel genetic loci of ASD could be found by identifying better defined subgroups, although further confirmation is needed in future cohorts with larger sample sizes.

Previous studies regarding Alzheimer's disease, neuroticism, or asthma found that items or symptoms showed, to some degree, increased ORs between the case loci and control loci compared to those from previous studies using broadly defined disease diagnoses (9-11). These findings may indicate that GWAS based on a symptom or an item could identify genetically more homogeneous subgroups and let us hypothesize that a relatively reasonable combination of symptoms or items could identify more genetically homogeneous subgroups.

In contrast, Chaste and colleagues showed that stratifying children with ASD based on the phenotype only modestly increased power in GWAS (12). The discrepancy between their findings and ours might be explained by at least two reasons. First, Chaste and colleagues used one item or symptom alone, whereas we used combinations of them with a machine learning method. DeMichele-Sweet and colleagues reported that subgrouping only by having psychosis could lead to the identification of limited loci that had small effects (59), but Mukherjee and colleagues found a substantial number of suggestive loci that had extreme ORs after categorizing persons with Alzheimer's disease based on relative performance across cognitive domains by

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

modern psychometric approaches (9). It may be necessary to utilize an appropriate combination of data to reveal masked patterns of data sets. Second, the number of subgroups was quite different between Chaste and colleagues' study and ours. Chaste and colleagues divided their participants mainly into two subgroups, but we divided ours into 15. In fact, when we employed a cluster number of two in our study, we observed no significant loci (Table 1). If ASD consists of more than hundreds of subgroups (6), grouping with a sufficient number of clusters may be necessary.

Validation of clusters is essential. In the present study, we selected the k-means algorithm, focused on ADI-R items and treatment as variables, and determined cluster numbers based on the λ of the Q-Q plots. Although we believe this approach is one of the relevant ways, selection of variables, selection of algorithms and selection of cluster numbers still remain to be considered in future mathematical and biological cluster validation studies because controversies surrounding evaluation of the quality of the clusters are important issues and are still ongoing and because validated clusters may lead to elucidate the genetic architectures of ASD (24).

The present study has a limitation to be noted. Substantial differences in the two genotyping platforms may have affected the results of the replication study. The Omni2.5 array includes 2,383,385 autosomal SNPs, whereas the 1Mv3 array includes 1,147,689 SNPs, with 675,923 shared SNPs between the two. Of the 65 statistically significant chromosomal loci in the discovery data, only seven chromosomal loci were shared between the two arrays.

Our study demonstrated that if the data set consists of multiple heterogeneous subgroups, even a subgroup that includes a much smaller number of homogeneous individuals could detect high-impact genetic factors. Hypothetical examples of the concept of cluster-based GWAS are shown in Figure S3. As shown in this figure, in the conventional design in which a whole data set

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is involved, an actual effect of a variant would be "diluted" to a modest OR of, e.g., 1.5, and at least thousands or tens of thousands of individuals would be required to detect it as a significantly associated variant. In contrast, cluster-based GWAS would be more likely than the conventional design to detect associated variants, without their effects being diluted, and with much higher ORs. As shown in Table 3, only 30 etiologically homogeneous probands and 300 controls can have a statistical power of approximately 1.00, calculated using the method based on the results in Nam's study (60). Although the integral model, which assumes many genetic variants have a small effect, may contribute to the formation of some subgroups of ASD, our results indicate that clustering by specific phenotypic variables may provide a candidate example for identifying etiologically similar cases of ASD.

Our data indicate the relevance of cluster-based GWAS as a means to identify more homogeneous subgroups of ASD than broadly defined subgroups. Future investigation of cluster validation and replication with a larger sample size is therefore warranted. Such studies will provide clues to elucidate the genetic structures and etiologies of ASD and facilitate the development of precision medicine for ASD.

Acknowledgments

We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the staff at the Simons Foundation Autism Research Initiative (SFARI). The present study was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI Grant Numbers 19390171 and 16H05242. MEXT had no role in the design or execution of the study.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Disclosures

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. American Psychological Association (2013): *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*. Washington: American Psychological Association.
2. Rapin I (1997): Autism. *N Engl J Med* 337: 97-104.
3. Geschwind DH, State MW (2015): Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* 14: 1109-1120.
4. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E, *et al.* (1995): Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* 25: 63-77.
5. Lauritsen MB, Pedersen CB, Mortensen PB (2005): Effects of familial risk factors and place of birth on the risk of autism: a nationwide register-based study. *J Child Psychol Psychiatry* 46: 963-971.
6. Gene S (2018): Gene scoring. https://gene.sfari.org/database/gene-scoring/.
7. Eissa N, Al-Houqani M, Sadeq A, Ojha SK, Sasse A, Sadek B (2018): Current enlightenment about etiology and pharmacological treatment of autism spectrum disorder. *Front Neurosci* 12: 304.
8. Traylor M, Markus H, Lewis CM (2015): Homogeneous case subgroups increase power in genetic association studies. *Eur J Hum Genet* 23: 863-869.
9. Mukherjee S, Mez J, Trittschuh EH, Saykin AJ, Gibbons LE, Fardo DW, *et al.* (2018): Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Mol Psychiatry*. doi: 10.1038/s41380-018-0298-8.
10. Nagel M, Watanabe K, Stringer S, Posthuma D, van der Sluis S (2018): Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat Commun* 9: 905.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

11. Lavoie-Charland E, Berube JC, Boulet LP, Bosse Y (2016): Asthma susceptibility variants are more strongly associated with clinically similar subgroups. *J Asthma* 53: 907-913.

12. Chaste P, Klei L, Sanders SJ, Hus V, Murtha MT, Lowe JK, *et al.* (2015): A genome-wide association study of autism using the Simons simplex collection: does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biol Psychiatry* 77: 775-784.

13. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, *et al.* (2016): Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* 19: 1454-1462.

14. MacQueen J (1967): Some methods for classification and analysis of multivariate observations. In: *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, pp 281-297.

15. World Medical Association (2013): World medical association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 310: 2191-2194.

16. Fischbach GD, Lord C (2010): The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron* 68: 192-195.

17. Beggiato A, Peyre H, Maruani A, Scheid I, Rastam M, Amsellem F, *et al.* (2017): Gender differences in autism spectrum disorders: divergence among specific core symptoms. *Autism Res* 10: 680-689.

18. Kuriyama S, Kamiyama M, Watanabe M, Tamahashi S, Muraguchi I, Watanabe T, *et al.* (2002): Pyridoxine treatment in a subgroup of children with pervasive developmental disorders. *Dev Med Child Neurol* 44: 284-286.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

19. Obara T, Ishikuro M, Tamiya G, Ueki M, Yamanaka C, Mizuno S, *et al.* (2018): Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. *Sci Rep* 8: 14840.

20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006): Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904-909.

21. Patterson N, Price AL, Reich D (2006): Population structure and eigenanalysis. *PLoS Genet* 2: e190.

22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* (2011): Scikit-learn: machine learning in Python. *J Mach Learn Res* 12: 2825-2830.

23. Cutting DR, Karger DR, Pedersen JO, Tukey JW (1992): Scatter/gather: a cluster-based approach to browsing large document collections. In: *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: Association for Computing Machinery (ACM), pp 318-329.

24. Raykov YP, Boukouvalas A, Baig F, Little MA (2016): What to do when K-means clustering fails: a simple yet principled alternative algorithm. *PLoS One* 11: e0162259.

25. Guo G, Chen L, Ye Y, Jiang Q (2017): Cluster validation method for determining the number of clusters in categorical sequences. *IEEE Trans Neural Netw Learn Syst* 28: 2936-2948.

26. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, *et al.* (2011): Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70: 863-885.

27. Spielman RS, Ewens WJ (1998): A sibship test for linkage in the presence of association:

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the sib transmission/disequilibrium test. *Am J Hum Genet* 62: 450-458.

28. Freidlin B, Zheng G, Li Z, Gastwirth JL (2002): Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 53: 146-152.

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al.* (2007): PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-575.

30. Wang K, Li M, Hakonarson H (2010): ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164.

31. Price AL, Zaitlen NA, Reich D, Patterson N (2010): New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11: 459-463.

32. Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, *et al.* (2015): Statistical analysis for genome-wide association study. *J Biomed Res* 29: 285-297.

33. Wang Y, Ding X, Tan Z, Ning C, Xing K, Yang T, *et al.* (2017): Genome-wide association study of piglet uniformity and farrowing interval. *Front Genet* 8: 194.

34. Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, Duncan L, *et al.* (2018): Analysis of shared heritability in common disorders of the brain. *Science* 360: eaap8757.

35. Orru S, Papoulidis I, Siomou E, Papadimitriou DT, Sotiriou S, Nikolaidis P, *et al.* (2019): Autism spectrum disorder, anxiety and severe depression in a male patient with deletion and duplication in the 21q22.3 region: a case report. *Biomed Rep* 10: 354-358.

36. Sragovich S, Merenlender-Wagner A, Gozes I (2017): ADNP plays a key role in autophagy: from autism to schizophrenia and Alzheimer's disease. *Bioessays* 39: 1700054.

37. Oxelgren UW, Myrelid A, Anneren G, Ekstam B, Goransson C, Holmbom A, *et al.*

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2017): Prevalence of autism and attention-deficit-hyperactivity disorder in down syndrome: a population-based study. *Dev Med Child Neurol* 59: 276-283.

38. Redies C, Hertel N, Hubner CA (2012): Cadherins and neuropsychiatric disorders. *Brain Res* 1470: 130-144.

39. Varghese M, Keshav N, Jacot-Descombes S, Warda T, Wicinski B, Dickstein DL, *et al.* (2017): Autism spectrum disorder: neuropathology and animal models. *Acta Neuropathol* 134: 537-566.

40. Atsem S, Reichenbach J, Potabattula R, Dittrich M, Nava C, Depienne C, *et al.* (2016): Paternal age effects on sperm FOXK1 and KCNA7 methylation and transmission into the next generation. *Hum Mol Genet* 25: 4996-5005.

41. Barnby G, Abbott A, Sykes N, Morris A, Weeks DE, Mott R, *et al.* (2005): Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. *Am J Hum Genet* 76: 950-966.

42. Minhas HM, Pescosolido MF, Schwede M, Piasecka J, Gaitanis J, Tantravahi U, *et al.* (2013): An unbalanced translocation involving loss of 10q26.2 and gain of 11q25 in a pedigree with autism spectrum disorder and cerebellar juvenile pilocytic astrocytoma. *Am J Med Genet A* 161a: 787-791.

43. Abou-Donia MB, Suliman HB, Siniscalco D, Antonucci N, ElKafrawy P (2019): De novo blood biomarkers in autism: autoantibodies against neuronal and glial proteins. *Behav Sci (Basel)* 9: E47.

44. Lo Vasco VR (2011): Role of phosphoinositide-specific phospholipase C η2 in isolated and syndromic mental retardation. *Eur Neurol* 65: 264-269.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. Potkin SG, Turner JA, Fallon JA, Lakatos A, Keator DB, Guffanti G, *et al.* (2009): Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia. *Mol Psychiatry* 14: 416-428.

46. Konopaske GT, Balu DT, Presti KT, Chan G, Benes FM, Coyle JT (2018): Dysbindin-1 contributes to prefrontal cortical dendritic arbor pathology in schizophrenia. *Schizophr Res* 201: 270-277.

47. Openshaw RL, Kwon J, McColl A, Penninger JM, Cavanagh J, Pratt JA, *et al.* (2019): JNK signalling mediates aspects of maternal immune activation: importance of maternal genotype in relation to schizophrenia risk. *J Neuroinflammation* 16: 18.

48. Ikeda M, Tomita Y, Mouri A, Koga M, Okochi T, Yoshimura R, *et al.* (2010): Identification of novel candidate genes for treatment response to risperidone and susceptibility for schizophrenia: integrated analysis among pharmacogenomics, mouse expression, and genetic case-control association approaches. *Biol Psychiatry* 67: 263-269.

49. Teng X, Aouacheria A, Lionnard L, Metz KA, Soane L, Kamiya A, *et al.* (2019): KCTD: a new gene family involved in neurodevelopmental and neuropsychiatric disorders. *CNS Neurosci Therapeutics* 25: 887-902.

50. Lin CH, Huang MW, Lin CH, Huang CH, Lane HY (2019): Altered mRNA expressions for N-methyl-D-aspartate receptor-related genes in WBC of patients with major depressive disorder. *J Affect Disord* 245: 1119-1125.

51. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, *et al.* (2019): Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* 51: 414-430.

52. Cimino PJ, Sokal I, Leverenz J, Fukui Y, Montine TJ (2009): DOCK2 is a microglial

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specific regulator of central nervous system innate immunity found in normal and Alzheimer's disease brain. *Am J Pathol* 175: 1622-1630.

53. Sepulveda-Diaz JE, Naini SMA, Huynh MB, Ouidja MO, Yanicostas C, Chantepie S, *et al.* (2015): HS3ST2 expression is critical for the abnormal phosphorylation of tau in Alzheimer's disease-related tau pathology. *Brain* 138: 1339-1354.

54. Ghosh D, Levault KR, Brewer GJ (2014): Relative importance of redox buffers GSH and NAD(P)H in age-related neurodegeneration and Alzheimer disease-like mouse neurons. *Aging Cell* 13: 631-640.

55. Zong Y, Yu P, Cheng H, Wang H, Wang X, Liang C, *et al.* (2015): miR-29c regulates NAV3 protein expression in a transgenic mouse model of Alzheimer's disease. *Brain Res* 1624: 95-102.

56. Lu J, McCarter M, Lian G, Esposito G, Capoccia E, Delli-Bovi LC, *et al.* (2016): Global hypermethylation in fetal cortex of down syndrome due to DNMT3L overexpression. *Hum Mol Genet* 25: 1714-1727.

57. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, *et al.* (2014): A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159: 1511-1523.

58. Quesnel-Vallieres M, Dargaei Z, Irimia M, Gonatopoulos-Pournatzis T, Ip JY, Wu M, *et al.* (2016): Misregulation of an activity-dependent splicing network as a common mechanism underlying autism spectrum disorders. *Mol Cell* 64: 1023-1034.

59. DeMichele-Sweet MAA, Weamer EA, Klei L, Vrana DT, Hollingshead DJ, Seltman HJ, *et al.* (2018): Genetic risk for schizophrenia and psychosis in Alzheimer disease. *Mol Psychiatry* 23: 963-972.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

60. Nam JM (1987): Simple approximation for calculating sample sizes for detecting linear trend in proportions. *Biometrics* 43: 701-705.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure/Table legends

Figure 1. Details of the cluster-based GWAS in the present discovery stage in the Simons Simplex Collection dataset.

In the present study, a GWAS using each subgroup of the probands vs the unaffected brothers as controls without the brothers of the members of the subgroup was designated a "cluster-based GWAS". This panel shows the detailed methods of the cluster-based GWAS in the discovery stage.

Figure 2. Manhattan plots (a) and corresponding quantile-quantile plots (b) in GWAS for all male probands vs their unaffected brothers using the sib transmission/disequilibrium test.

We conducted a GWAS in the Simons Simplex Collection dataset of 597 male probands and 370 unaffected brothers genotyped by the Illumina Human Omni2.5 array using the sib transmission/disequilibrium test (sib-TDT). We observed no significant associations in this GWAS with the genome-wide threshold of $P < 5.0 \times 10^{-8}$.

Figure 3. Manhattan plots (a) and corresponding quantile-quantile plots (b) for cluster-based GWASs with a cluster number of 15.

We performed cluster analysis using k-means with a cluster number of 15 and conducted cluster-based GWAS. Among 15 clusters, significant associations were observed in 14 clusters. In total, we observed 65 chromosomal loci, labeled in the figure, that satisfied the threshold of $P < 5.0 \times 10^{-8}$. The red lines indicate the threshold for genome-wide significance ($P < 5.0 \times 10^{-8}$).

Table 1. Number of genome-wide significant loci and λ values according to the number of clusters using the k-means algorithm and the Omni2.5 dataset with MAF <0.01 deleted.

Table 2. Characteristics of each of 15 k-means clusters in the Omni2.5 dataset.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3. Association table of the cluster-based GWAS with 15 k-means clusters in the Omni2.5 dataset.

Table 4. Replication results for 65 statistically significant chromosomal loci in the discovery studies by an independent dataset of 1Mv3.

Tables

Table 1. Number of genome-wide significant loci and λ values according to the number of clusters using the k-means algorithm and the Omni2.5 dataset with MAF <0.01 deleted.

No. of Clusters	1	2	3	4	5	10	15	20
Number of genome-wide significant loci	0	0	2	5	1	26	65	211
Mean λ value (min-max)	1.032	1.057 (1.052-1.062)	1.036 (1.027-1.045)	1.035 (1.029-1.041)	1.021 (1.012-1.033)	1.024 (1.000-1.042)	1.038 (1.017-1.091)	1.053 (1.015-1.107)

λ values were calculated from the sibling-based transmission disequilibrium test in the number of clusters 1 or the Cochran-Armitage trend test in the number of clusters from 2 to 20.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2. Characteristics of each of 15 k-means clusters in the Omni2.5 dataset.

Cluster No.	n	Verbal score from ADI-R				Nonverbal score from ADI-R				Restricted and repetitive patterns of behavior score from ADI-R				Social score from ADI-R				Vitamin B6 treatment (%)
		Mean (SD)	Median (p25-p75)	Min	Max	Mean (SD)	Median (p25-p75)	Min	Max	Mean (SD)	Median (p25-p75)	Min	Max	Mean (SD)	Median (p25-p75)	Min	Max	
All	597	7.7 (2.1)	8.0 (6.0-9.0)	0	12	8.9 (3.3)	9.0 (6.0-12.0)	0	14	6.8 (2.5)	7.0 (5.0-8.0)	1	12	19.8 (5.3)	20.0 (16.0-24.0)	8	30	59.6
1	33	7.4 (2.2)	7.0 (6.0-10.0)	3	11	4.4 (1.6)	4.0 (3.0-6.0)	1	7	8.5 (1.6)	8.0 (7.0-10.0)	6	12	14.0 (1.5)	14.0 (13.0-15.0)	11	17	60.6
2	49	8.9 (1.3)	9.0 (8.0-10.0)	6	12	12.3 (1.5)	12.0 (11.0-14.0)	9	14	6.2 (1.3)	6.0 (6.0-7.0)	3	8	27.1 (1.3)	27.0 (26.0-28.0)	24	30	79.6
3	45	6.0 (1.9)	6.0 (5.0-7.0)	2	10	8.8 (1.5)	9.0 (8.0-10.0)	6	12	5.0 (1.5)	5.0 (4.0-6.0)	2	7	16.8 (1.1)	17.0 (16.0-18.0)	15	19	64.4
4	59	9.0 (1.5)	9.0 (8.0-10.0)	6	12	8.1 (1.5)	8.0 (7.0-9.0)	4	10	8.8 (1.9)	8.0 (8.0-10.0)	5	12	23.8 (1.4)	24.0 (23.0-25.0)	21	27	57.6
5	28	7.3 (1.1)	7.0 (6.5-8.0)	5	9	9.1 (1.7)	9.0 (8.0-10.0)	7	13	6.1 (2.3)	6.0 (5.0-7.0)	1	12	12.7 (1.7)	13.0 (12.0-14.0)	9	15	60.7
6	29	7.7 (1.9)	8.0 (7.0-9.0)	2	12	4.6 (1.8)	5.0 (4.0-6.0)	0	8	4.0 (1.1)	4.0 (3.0-5.0)	2	6	15.8 (1.4)	16.0 (15.0-17.0)	14	19	44.8
7	37	6.5 (1.8)	6.0 (5.0-8.0)	3	11	12.5 (1.3)	12.0 (12.0-14.0)	10	14	5.6 (1.4)	6.0 (5.0-7.0)	3	8	19.4 (1.8)	20.0 (18.0-21.0)	15	22	56.8
8	23	8.3 (1.6)	8.0 (7.0-10.0)	5	11	4.2 (2.1)	4.0 (3.0-6.0)	0	8	5.9 (1.9)	6.0 (4.0-8.0)	3	10	9.7 (1.1)	10.0 (9.0-11.0)	8	12	60.9
9	46	9.0 (1.3)	9.0 (8.0-10.0)	5	12	12.4 (1.3)	13.0 (11.0-13.0)	10	14	9.2 (1.8)	9.0 (8.0-10.0)	6	12	22.7 (1.4)	22.5 (22.0-24.0)	20	25	69.6
10	43	6.6 (1.4)	7.0 (6.0-7.0)	4	9	11.7 (1.5)	12.0 (10.0-13.0)	9	14	5.0 (1.5)	5.0 (4.0-6.0)	2	8	24.1 (1.3)	24.0 (23.0-25.0)	22	26	55.8
11	34	4.4 (1.6)	5.0 (3.0-6.0)	0	7	4.9 (1.8)	5.0 (4.0-6.0)	1	9	4.1 (1.7)	4.0 (3.0-5.0)	1	9	10.9 (1.9)	10.5 (9.0-13.0)	8	14	55.9

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12	38	8.8 (1.6)	9.0 (8.0-10.0)	5	12	9.7 (1.5)	9.0 (8.0-11.0)	8	13	9.2 (1.3)	9.0 (8.0-10.0)	7	12	18.1 (1.3)	18.0 (17.0-19.0)	15	20	65.8
13	52	7.1 (2.0)	7.0 (5.5-8.5)	3	12	7.4 (1.5)	7.0 (6.0-9.0)	4	10	4.6 (1.5)	4.5 (3.5-6.0)	1	7	22.0 (1.7)	22.0 (21.0-23.0)	19	27	44.2
14	46	7.9 (1.5)	8.0 (7.0-9.0)	4	11	6.2 (1.6)	6.0 (5.0-7.0)	1	9	8.4 (1.7)	8.0 (7.0-10.0)	5	12	19.4 (1.4)	19.0 (18.0-20.0)	17	22	58.7
15	35	9.5 (1.4)	9.0 (8.0-11.0)	7	12	12.7 (1.4)	13.0 (12.0-14.0)	9	14	9.6 (1.1)	10.0 (9.0-10.0)	8	12	27.5 (1.5)	27.0 (26.0-29.0)	25	30	54.3

ADI-R: Autism Diagnostic Interview-Revised.
SD: standard deviation.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3. Association table of the cluster-based GWAS with 15 k-means clusters in the Omni2.5 dataset.

Cluster No.	ID	Chr	hg19	Minor/ major	MAF (%)	OR	95% CI	P	GENESYMBOL	Function	Power
1	rs111629286	11	130,152,136	A/G	1.80	13.42	4.38-41.17	1.36×10^{-8}	ZBTB44	Intronic	0.997
1	rs115140946	6	37,891,923	C/A	1.03	21.07	4.79-92.77	2.87×10^{-8}	ZFAND3	Intronic	0.878
1	rs9462391	6	38,123,030	A/G	1.03	21.07	4.79-92.77	2.87×10^{-8}	ZFAND3	Downstream	0.878
1	rs10217283	9	101,423,675	A/G	1.42	15.51	4.45-54.12	2.95×10^{-8}	GABBR2	Intronic	0.976
1	rs114109395	6	38,005,546	A/G	1.03	21.01	4.77-92.51	3.02×10^{-8}	ZFAND3	Intronic	0.877
2	rs115621412	9	74,366,033	C/A	7.89	4.42	2.48-7.87	8.13×10^{-9}	CEMIP2	Intronic	1.000
3	rs77507687	2	26,939,229	G/A	2.00	12.43	4.37-35.36	6.10×10^{-9}	KCNK3	Intronic	1.000
3	rs76880969	1	227,711,506	G/A	1.00	27.15	5.30-139.20	8.20×10^{-9}	CDC42BPA, ZNF678	Intergenic	0.865
3	rs115483919	2	125,010,267	A/G	1.00	27.15	5.30-139.20	8.20×10^{-9}	CNTNAP5	Intronic	0.865
5	rs16965293	16	9,551,490	A/G	2.31	14.04	5.00-39.45	3.83×10^{-10}	LINC01195, GRIN2A	Intergenic	1.000
5	rs77489014	9	106,962,281	A/G	1.41	19.47	5.51-68.82	6.69×10^{-10}	SMC2, LOC105376194	Intergenic	0.991
5	rs117473168	9	106,848,270	A/G	1.55	16.90	5.02-56.93	2.64×10^{-9}	SMC2	ncRNA exonic	0.991
5	rs7199670	16	22,875,238	A/G	11.28	5.28	2.76-10.10	4.98×10^{-9}	HS3ST2	Intronic	1.000
5	rs73142209	12	77,859,299	G/A	1.54	16.18	4.82-54.31	5.33×10^{-9}	E2F7, NAV3	Intergenic	0.989
5	rs118167078	15	65,723,796	A/G	1.54	16.18	4.82-54.31	5.33×10^{-9}	IGDCC4, DPP8	Intergenic	0.989
5	rs11919513	3	4,841,384	G/A	3.22	10.18	3.99-26.03	8.92×10^{-9}	ITPR1	Intronic	1.000
5	rs13332627	16	22,874,928	G/A	9.23	6.10	2.96-12.57	1.22×10^{-8}	HS3ST2	Intronic	1.000
5	rs111920363	7	143,656,906	A/G	1.15	19.46	4.89-77.39	1.29×10^{-8}	OR2F1	Upstream	0.933
5	rs9939816	16	22,876,408	A/C	9.25	6.08	2.95-12.53	1.30×10^{-8}	HS3ST2	Intronic	1.000
5	rs76096239	14	97,193,704	A/G	1.67	13.79	4.27-44.54	3.25×10^{-8}	PAPOLA, LINC02299	Intergenic	0.986
5	rs1054028	16	22,927,214	G/A	14.36	5.02	2.62-9.61	3.32×10^{-8}	HS3ST2	UTR3	1.000
5	rs78486970	7	106,127,612	G/A	6.87	5.46	2.67-11.18	3.68×10^{-8}	NAMPT, CCDC71L	Intergenic	1.000

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

6	rs148617803	1	76,136,228	G/A	1.32	22.57	5.95-85.64	2.77×10^{-10}	SLC44A5, ACADM	Intergenic	0.988
6	rs55985845	10	25,163,664	T/A	2.51	11.71	4.26-32.20	7.18×10^{-9}	PRTFDC1	Intronic	1.000
6	rs73094424	12	39,840,397	A/G	2.11	12.09	4.12-35.52	2.70×10^{-8}	KIF21A, ABCD2	Intergenic	0.998
6	rs58845693	3	122,804,247	G/A	1.18	18.07	4.55-71.72	4.24×10^{-8}	PDIA5	Intronic	0.915
6	rs11709496	3	122,809,400	G/A	1.18	18.07	4.55-71.72	4.24×10^{-8}	PDIA5	Intronic	0.915
6	rs199531954	12	95,064,359	C/A	1.19	17.92	4.52-71.10	4.92×10^{-8}	TMCC3, MIR492	Intergenic	0.913
7	rs79033134	17	76,473,288	A/G	1.53	16.29	4.87-54.44	4.24×10^{-9}	DNAH17	Intronic	0.995
7	rs57127555	17	76,475,811	C/A	1.54	16.24	4.86-54.28	4.49×10^{-9}	DNAH17	Intronic	0.995
7	rs75382702	11	81,149,755	A/G	1.28	16.94	4.54-63.23	3.18×10^{-8}	LINC02720, MIR4300HG	Intergenic	0.961
8	rs73149247	3	100,864,047	G/A	2.21	11.41	3.88-33.54	5.80×10^{-11}	ABI3BP, IMPG2	Intergenic	1.000
8	rs12418400	11	131,263,123	G/A	1.56	20.88	6.09-71.57	6.68×10^{-11}	NTM	Intronic	0.996
8	rs78323783	10	45,084,432	A/G	1.17	24.79	6.13-100.30	2.28×10^{-10}	CXCL12, TMEM72	Intergenic	0.997
8	rs72991663	6	130,143,713	A/G	2.85	13.30	4.84-36.53	5.51×10^{-10}	ARHGAP18, TMEM244	Intergenic	1.000
8	rs74922057	21	41,595,011	A/G	1.31	19.67	5.22-74.14	3.13×10^{-9}	DSCAM	Intronic	0.962
8	rs115035406	21	41,580,474	G/A	1.42	16.53	4.61-59.31	1.97×10^{-8}	DSCAM	Intronic	0.957
8	rs114994877	4	136,731,494	A/G	1.42	16.53	4.61-59.31	1.97×10^{-8}	LINC02485, LINC00613	Intergenic	0.957
8	rs117008682	9	103,245,053	G/A	1.43	16.48	4.59-59.15	2.08×10^{-8}	MSANTD3	Intronic	0.957
8	rs117772706	9	81,338,445	G/A	1.43	16.44	4.58-58.98	2.19×10^{-8}	PSAT1, LOC101927450	Intergenic	0.957
9	rs4885429	13	77,400,673	G/A	2.14	13.69	4.91-38.16	4.67×10^{-10}	LMO7DN, KCTD12	Intergenic	1.000
9	rs45618836	7	73,480,258	G/A	2.26	11.94	4.43-32.18	2.30×10^{-9}	ELN	Intronic	1.000
9	rs7299395	12	41,714,602	A/G	3.27	8.52	3.65-19.89	1.15×10^{-8}	PDZRN4	Intronic	1.000
9	rs55772967	7	73,448,499	G/A	2.89	8.91	3.66-21.66	2.09×10^{-8}	ELN	Intronic	1.000
10	rs72799348	2	22,637,443	A/G	2.31	12.84	4.74-34.77	6.57×10^{-10}	LINC01822, LINC01884	Intergenic	1.000
10	rs76159464	5	169,446,509	A/G	1.02	28.05	5.47-144.00	5.03×10^{-9}	DOCK2	Intronic	0.877

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10	rs12483301	21	28,070,591	G/A	1.92	11.89	3.94-35.92	6.74×10^{-9}	CYYR1, ADAMTS1	Intergenic	1.000
10	rs72883714	18	23,987,552	A/G	2.17	11.25	4.08-31.06	1.59×10^{-8}	TAF4B, LINC01543	Intergenic	1.000
10	rs1876769	2	22,678,191	A/G	2.17	11.25	4.08-31.06	1.59×10^{-8}	LINC01822, LINC01884	Intergenic	1.000
10	rs17043765	2	22,656,804	A/G	2.17	11.25	4.08-31.06	1.59×10^{-8}	LINC01822, LINC01884	Intergenic	1.000
11	rs74645195	4	48,330,367	G/A	2.71	10.26	3.95-26.66	1.34×10^{-8}	TEC, SLAIN2	Intergenic	1.000
11	rs78513244	1	2,360,342	A/G	3.25	9.33	3.79-22.98	1.35×10^{-8}	PEX10, PLCH2	Intergenic	1.000
11	rs10027938	4	90,242,059	A/G	16.93	4.48	2.49-8.05	2.29×10^{-8}	GPRIN3, SNCA	Intergenic	1.000
12	rs117647850	8	79,156,756	A/G	3.08	10.88	4.44-26.68	5.10×10^{-11}	LOC102724874, PKIA	Intergenic	1.000
12	rs4131532	1	3,540,256	A/G	1.54	15.63	4.68-52.14	8.61×10^{-9}	MEGF6, TPRG1L	Intergenic	0.994
12	rs77964987	4	183,685,432	G/A	4.77	7.06	3.21-15.53	4.97×10^{-8}	TENM3	Intronic	1.000
13	rs117954350	7	4,440,757	A/G	1.02	52.73	6.34-438.60	4.00×10^{-10}	SDK1, FOXK1	Intergenic	0.635
13	rs11064685	12	119,590,881	G/A	6.14	5.15	2.66-9.97	4.46×10^{-8}	SRRM4	Intronic	1.000
14	rs77983358	12	82,393,237	G/A	1.52	21.71	5.50-85.76	1.29×10^{-10}	LINC02426, CCDC59	Intergenic	0.999
14	rs7118821	11	96,876,267	C/A	1.01	26.18	5.11-134.00	1.50×10^{-8}	LINC02737	Intergenic	0.847
14	rs7122015	11	96,950,548	G/A	1.01	26.18	5.11-134.00	1.50×10^{-8}	LINC02737, CNTN5	Intergenic	0.847
14	rs7106102	11	96,885,969	A/G	1.01	26.10	5.10-133.70	1.58×10^{-8}	LINC02737	Intergenic	0.845
14	rs7189512	16	66,324,048	A/G	3.28	7.26	3.13-16.88	4.62×10^{-8}	LINC00922, CDH5	Intergenic	1.000
15	rs77311527	2	5,516,750	G/A	2.45	11.87	4.44-31.79	2.19×10^{-9}	LINC01249, LINC01248	Intergenic	1.000
15	rs276833	2	114,769,078	A/G	1.29	18.00	4.81-67.43	1.25×10^{-8}	LINC01191, DPP10	Intergenic	0.970

OR: Odds ratio.
CI: Confidence interval.
Powers were calculated using the method based on the results in Nam's study (60).

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4. Results of replication studies in the 1Mv3 dataset for statistically significant chromosomal loci in the discovery studies.

Cluster No.	ID	Chr	hg19	Minor/ major	MAF (%)	OR	95% CI	P	GENESYMBOL	Function
5	rs13332627	16	22,874,928	G/A	10.0	0.50	0.18-1.45	0.195	HS3ST2	intronic
5	rs7199670	16	22,875,238	A/G	12.2	0.51	0.20-1.33	0.1629	HS3ST2	intronic
5	rs1054028	16	22,927,214	G/A	15.0	0.51	0.22-1.21	0.121	HS3ST2	UTR3
10	rs1876769	2	22,678,191	A/G	1.4	NA	-	0.1822	LINC01822, LINC01884	intergenic
13	rs11064685	12	119,590,881	G/A	8.2	1.89	1.06-3.37	0.02858	SRRM4	intronic
14	rs7189512	16	66,324,048	A/G	3.5	2.16	0.83-5.67	0.1085	LINC00922, CDH5	intergenic
15	rs276833	2	114,769,078	A/G	1.3	0.71	0.09-5.75	0.75	LINC01191, DPP10	intergenic

OR: Odds ratio.
CI: Confidence interval.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

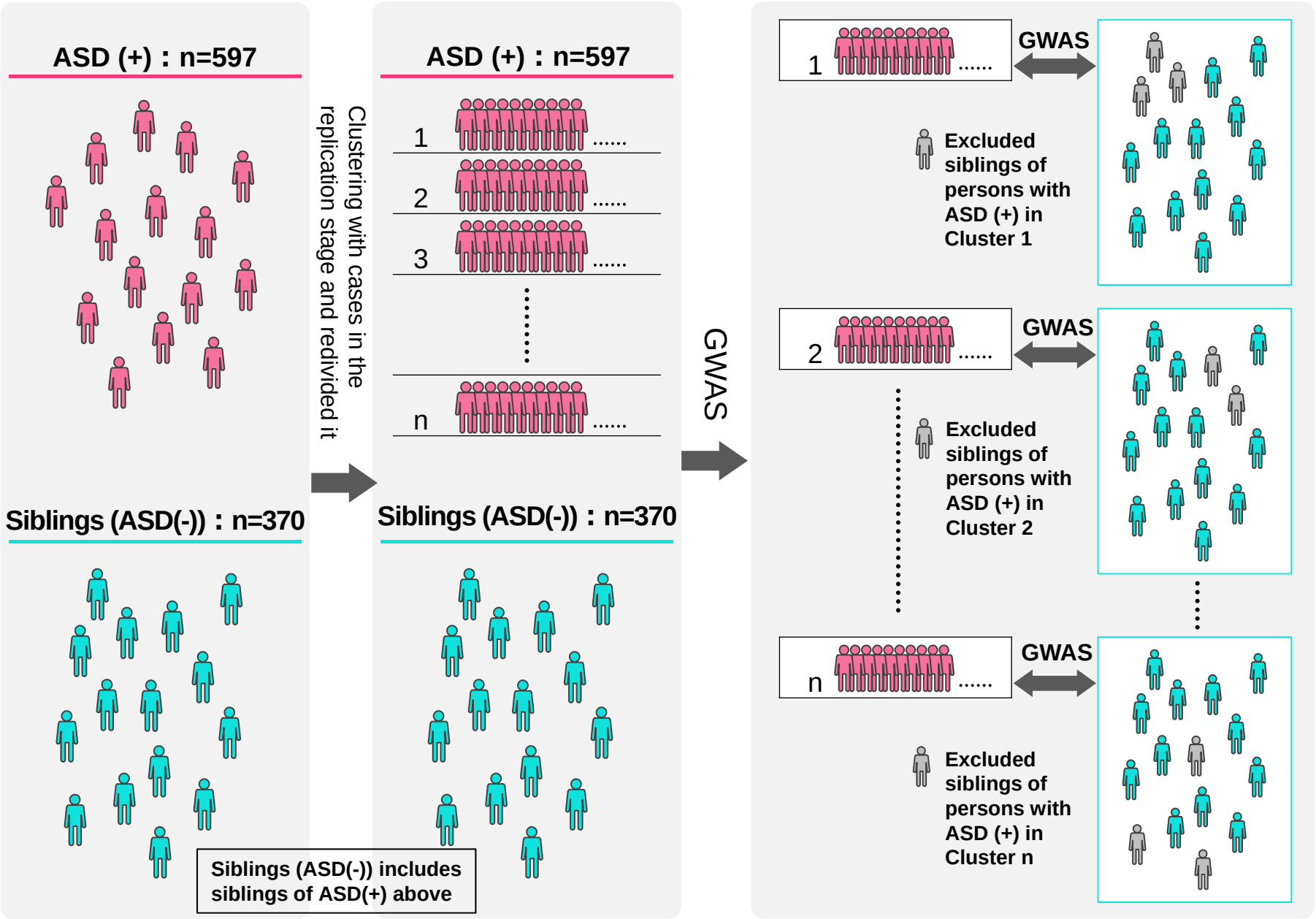
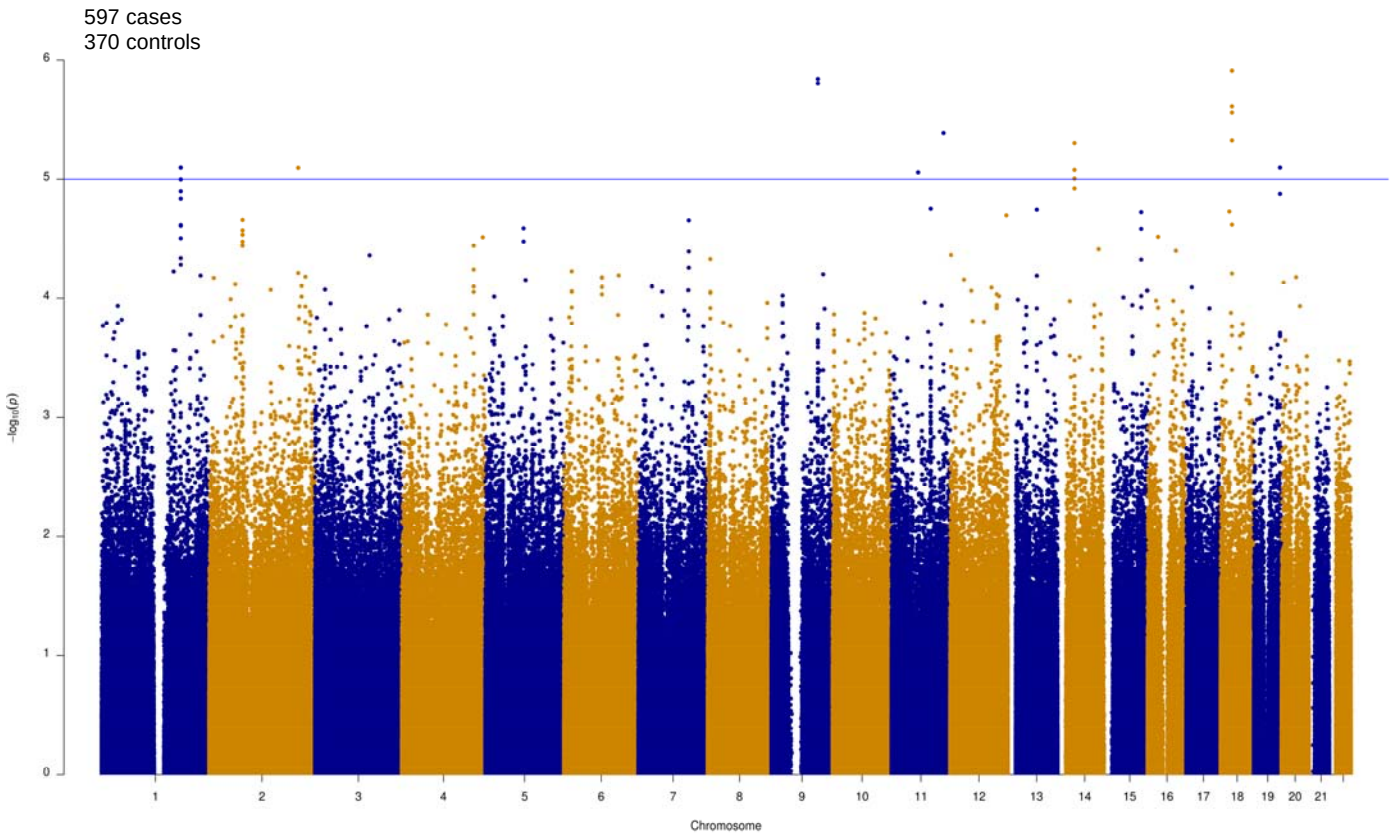


Figure 1. Details of the cluster-based GWAS in the present discovery stage in the Simons Simplex Collection dataset.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a



b

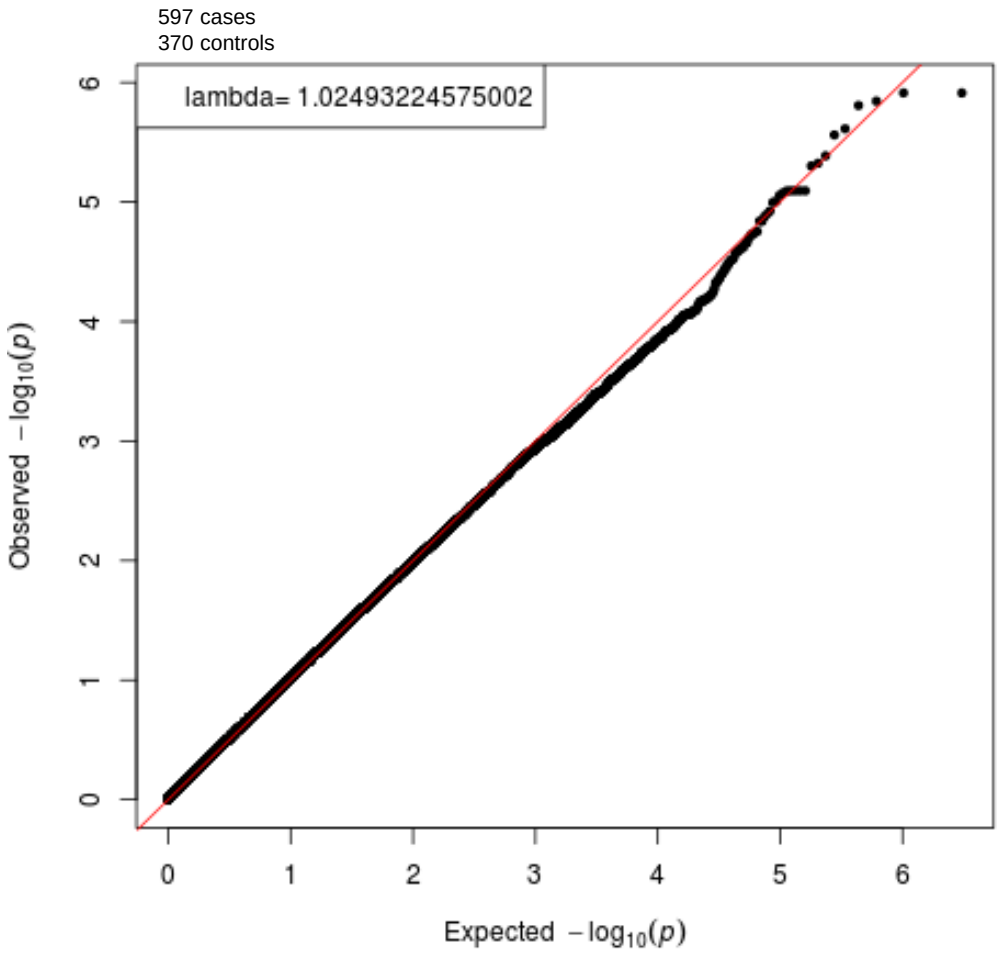
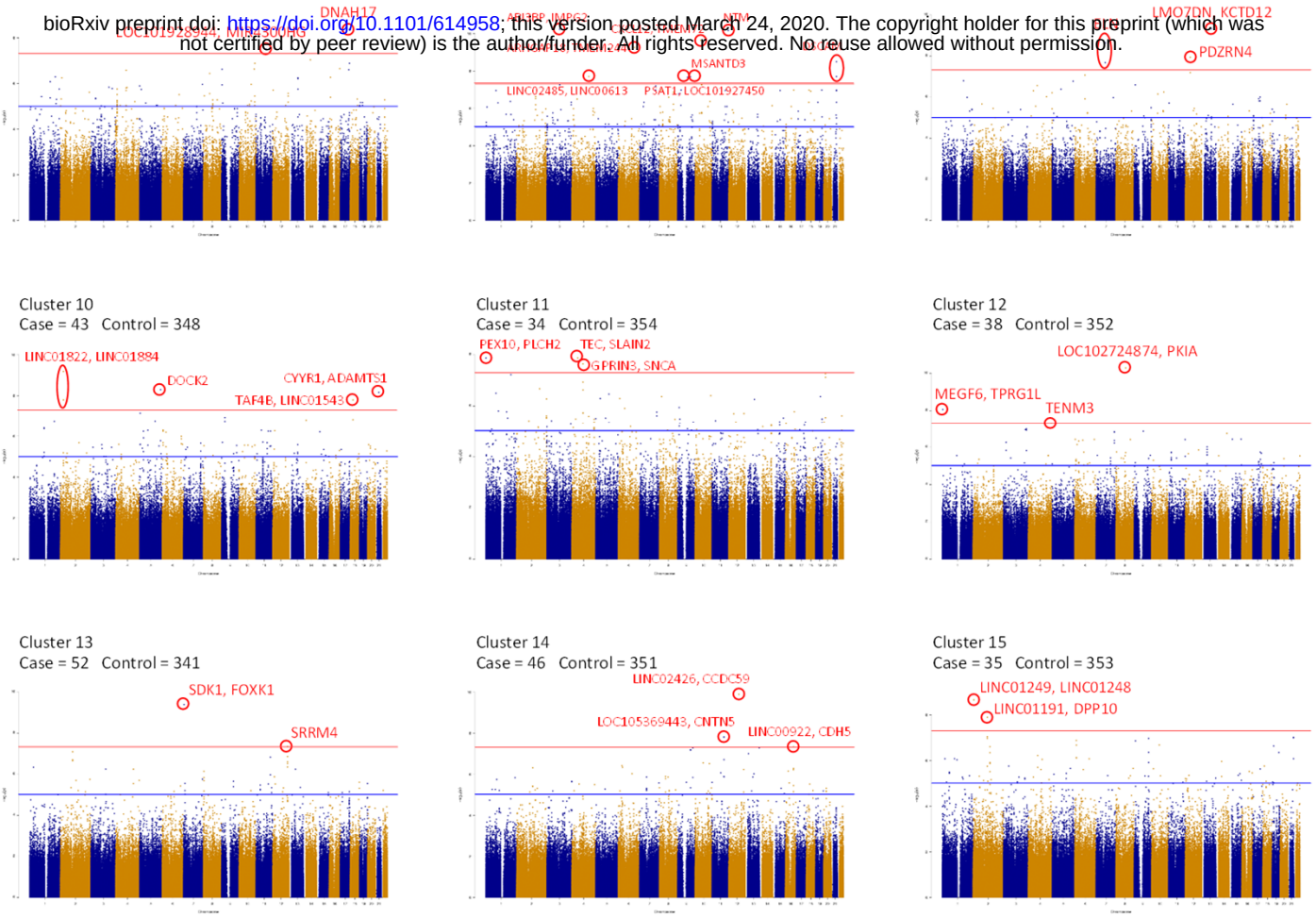
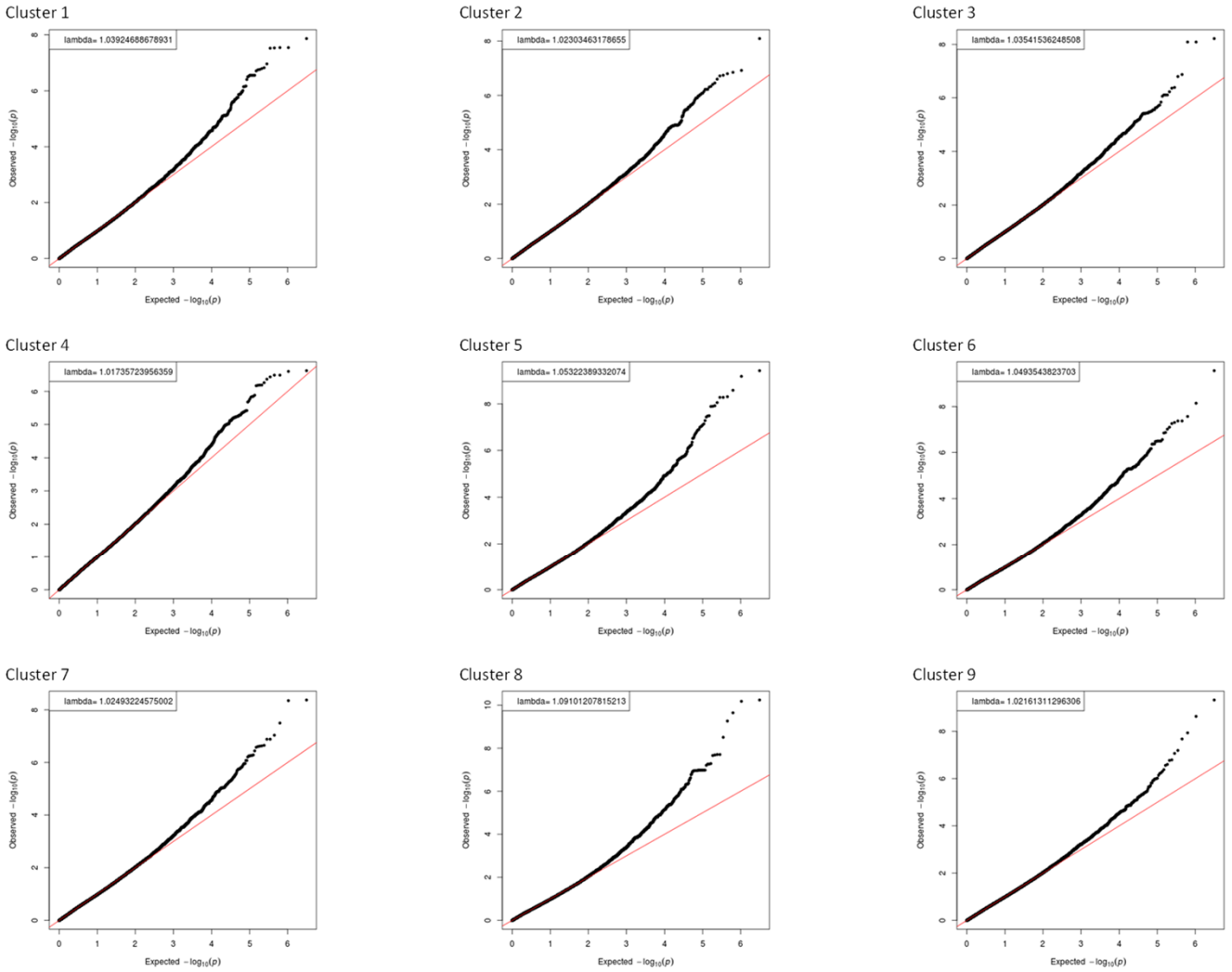


Figure 2. Manhattan plots (a) and corresponding quantile-quantile plots (b) in GWAS for all male probands vs their unaffected brothers using the sib transmission/disequilibrium test.

bioRxiv preprint doi: https://doi.org/10.1101/614958; this version posted March 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



b