# 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], 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

## Author information

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

## [*]Corresponding author

Shinichi Kuriyama.

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

**Abstract**

Autism spectrum disorder (ASD) has phenotypically and genetically heterogeneous characteristics. Here, we show a two-step genome-wide association study (GWAS). We used two datasets: one genotyped with the Illumina Human Omni2.5 (Omni2.5) in the discovery stage, and the other genotyped with the Illumina BeadChip 1Mv3 (1Mv3) in the replication stage. In the first step in the discovery stage, we observed no significant associations in a GWAS of 597 probands and 370 controls. In the second step in the discovery stage, we conducted cluster analyses in the combined dataset of male probands using Omni2.5 and 1Mv3 using k-means with a cluster number of 15 based on Autism Diagnostic Interview-Revised (ADI-R) scores and history of vitamin treatment, and redivided it for the discovery and replication stages. We then conducted GWAS in each subgroup of probands vs controls without the brothers of the probands belonging to the subgroup being analysed (cluster-based GWAS) and identified 65 chromosomal loci, which included 30 intragenic loci located in 21 genes and 35 intergenic ones, 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*. Although we observed no consistent genes that displayed genome-wide significance between the results from the Omni2.5 and 1Mv3 datasets, we observed that cluster-based GWAS, even with a small sample size, revealed abundant significant associations. These findings suggest that clustering may successfully identify subgroups with relatively homogeneous disease aetiologies. Further studies are warranted to validate clusters and to replicate our findings in larger cohorts.

## 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 behaviours[1], but ASD also shows many other symptoms[2]. Regarding genetics, previous studies have not consistently identified relatively common 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 coincidence 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 the 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 phenotypic and responses to treatment variables may increase the chances of identifying genetic susceptibility factors. Traylor et al.[8] demonstrated that attempts to subgroup patients of a complex disease into more homogeneous ones could have more power to elucidate the hidden heritability in a simulation study. Several studies of ASD, Alzheimer's disease, neuroticism, or asthma indicated that items or symptoms were in some degree useful to identify more high-impact genetic factors compared to broadly defined diagnosis[9-12]. 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 above-mentioned 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 multifactorial diseases to date.

We therefore explored whether grouping persons with ASD using a clustering algorithm with phenotypic and responses to treatment variables can be used to discriminate more genetically homogeneous persons with ASD. We adopted a machine learning k-means[14] algorithm for cluster analysis. Based on these clusters, we conducted genome-wide association studies (GWASs). We regarded the GWAS results as an indicator of successful clustering in the present study.

## Results

### *Clustering*

We used phenotypic variables, history of treatment, and genotypic data from the Simons Simplex Collection (SSC)[15]. The SSC is a core project and resource of the SFARI[6]. We used two datasets from the SSC: one genotyped with the Illumina Human Omni2.5 (Omni2.5) dataset in the discovery stage, and the other genotyped with the Illumina BeadChip 1Mv3 (1Mv3) dataset in the replication stage. Although we used two datasets separately when conducting GWAS, we conducted cluster analyses in the combined dataset of male probands, and redivided it for the discovery and replication stages.

We conducted cluster analyses using phenotypic variables of Autism Diagnostic Interview-Revised (ADI-R)[16] scores and history of vitamin treatment. We chose these variables because the ADI-R is one of the most reliable estimates of ASD and has the ability to evaluate

substructure domains of ASD. Among the treatments[17], 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 used the k-means[14] algorithm. The k-means algorithm requires cluster numbers determined by researchers. We set a priori the cluster numbers of 2, 3, 4, 5, 10, 15, and 20.

***Cluster-based genome-wide association study***

In the first step we conducted GWAS in the Omni2.5 dataset, a total of 597 male probands and 370 unaffected brothers, using the sib transmission/disequilibrium test (sib-TDT)[20]. We observed no significant associations (Fig. 1).

In the second step, we conducted GWAS in each subgroup of probands vs controls without the brothers of the probands belonging to the subgroup being analysed (cluster-based GWAS) (Fig. 2). We applied the Cochran-Armitage trend test[21] and Fisher's exact test[22]. Notably, we observed that the number of genome-wide significant loci almost increased as the number of clusters increased when the Cochran-Armitage trend test was applied (Table 1). In contrast, when Fisher's exact test was applied, zero to five significant loci were observed for numbers of clusters between two and 20. Two reasons may explain the difference in the results between the two tests. The first is the difference in analysis methods for the genetic case-control data. The Cochran-Armitage trend test 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. Fisher's exact test

examines the allele frequency in cases and controls. The disease model and mode of inheritance may influence the difference, although those of ASD are largely unknown[23]. Our data might indicate that a case-control study of ASD should be analysed by genotype. The second is the conservative nature of Fisher's exact test. Genomic inflation factor ($\lambda$) values ranged from 0.615 to 0.738, and the average was 0.683, which was very small compared to one (Table 1). We therefore regarded the Cochran-Armitage trend test as a more appropriate method in the present cluster-based GWAS.

In contrast to the many previous studies in which genetically unrelated controls were used, we used the brothers of probands as controls. We first conducted the Cochran–Armitage trend test and Fisher's exact test in the whole data set, and found that the negative logarithms of P-values (-logP) were distributed downward compared with the expected values, as shown in Supplementary Fig. S1. We thus conducted the sib-TDT, a family-based association test, to take into account familial relationships among the participants.

We applied the sib-TDT to one subset of cluster 1 vs all the controls using k-means with 15 clusters, and found the observed -logP values were lower than expected, also as shown in Supplementary Fig. S1. Since the sib-TDT may efficiently work in a population consisting of substantial number of sibs, limited number of brothers of the probands in all the controls probably contributed to serious 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 and Fisher's exact test, as in many other studies. We therefore believe that, in the case of the dataset we used in the present study, the sib-TDT in the whole datasets GWAS and the Cochran–Armitage trend test in the cluster-based GWAS are the best methods to account for the relationships between participants.

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[24]. Although there are measures to evaluate the quality of the clusters[25], the number of clusters should also be determined according to the research purposes. We regarded the GWAS results as an indicator of successful clustering in the present study.

Regarding appropriate cluster numbers, in addition to the GWAS results, we compared λ values among the analyses. For instance, λ value for the cluster-based GWAS with 20 clusters by k-means using the Cochran-Armitage trend test demonstrated that λ values 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 of less than 1.050 is deemed safe for avoiding false positives[26-28].

In contrast, the inflation factor λ values of the cluster-based GWAS with a cluster number of 15 ranged from 1.017 to 1.091, and the average was 1.038, which was below 1.050 (Table 1 and Fig. 3).

According to the above results, we considered the cluster-based GWAS using the Cochran-Armitage trend test, coupled with k-means cluster analysis with a cluster number of 15, to be the most appropriate approach to the present dataset. The characteristics of each cluster are presented in Table 2. We further evaluated the validity of the cluster numbers with existing measures of the elbow method[25]. The elbow criterion potentially indicates the optimal number of clusters by identifying the point where the within-group sums of squares abruptly decrease. In our dataset, the point seemed to be approximately three or more (Supplementary Fig. S2); thus, the cluster number of 15 was included in the suggested range provided by the method.

Regarding sample size, if the data set consists of multiple heterogeneous subgroups, even a subgroup, which 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 Supplementary Fig. S3. As shown in this figure, in the conventional design in which a whole data set is involved, an actual effect of a variant would be "diluted" to a modest odds ratio (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. Only 30 aetiologically homogeneous probands and 300 controls can have a statistical power of approximately 0.98, calculated using the method proposed by Breslow and Day[29], to detect an associated variant with an OR of 29, and have a power of approximately 0.90 even for that with a lower OR of 20.

Our results indicate that clustering by specific phenotypic variables might provide a candidate example for identifying aetiologically similar cases of ASD.

***Gene interpretation***

Among the cluster-based GWAS, we mainly presented here the results with a cluster number of 15 and using the Cochran-Armitage trend test. In this cluster-based GWAS, we observed 65 chromosomal loci that satisfied the threshold of $P < 5.0 \times 10^{-8}$ (Table 1 and Fig. 3), and 30 out of the 65 loci were located within 21 genes and the remaining 35 intergenic loci (Table 3). Among them, 8 out of the 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.

In the SFARI Gene scoring system, ranging 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 categorised in a different category. Rare single gene variants, disruptions/mutations, and sub-microscopic deletions/duplications related to ASD are categorised as "Rare Single Gene Mutation". The relatively high correspondence between our results in part and the SFARI Gene scoring system indicates that the statistically significant loci we found may indeed be associated with ASD subgroups.

In addition to genes in the Human Gene module of the SFARI Gene, several important genes associated with ASD or other related disorders[30-33] from previous reports were included in our findings as follows: *CDH5* in Cluster 14, *DSCAM* in Cluster 8, *FOXK1* in Cluster 13, *GRIN2A* in Cluster 5, *NTM* in Cluster 8, and *SNCA* in Cluster 11 previously reported with ASD[34-39]; *PLCH2* in Cluster 11 previously reported with mental retardation[40]; *ARHGAP18* in Cluster 18, *CDC42BPA* in Cluster 3, *CXCL12* in Cluster 8, *HS3ST2* in Cluster 5 previously reported with schizophrenia[41-44]; *KCTD12* in Cluster 9, *PSAT1* in Cluster 8 previously reported with depressive disorder[45,46]; *ADAMTS1* in Cluster 10, *DOCK2* in Cluster 10, *HS3ST2* in Cluster 5, *NAMPT* in Cluster 5, *NAV* in Cluster 5 previously reported with Alzheimer's disease[47-51]; and *PEX10* in Cluster 11 previously reported with Down syndrome[52]. These findings suggest that the

statistically significant SNPs might explain autistic symptoms because these diseases are suggested to have shared aetiology, even in part, with ASD[30-33]. 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, although further confirmation is needed.

In this study, as in many previous studies, the threshold for genome-wide significance was set at a P-value of $5.0 \times 10^{-8}$. Since the probands were categorised into 15 clusters, we also applied a P-value of $3.3 \times 10^{-9}$, which was obtained by dividing $5.0 \times 10^{-8}$ by 15, and found that 16 loci still survived the more stringent threshold (Table 3).

***Replication study***

We conducted replication studies in another independent dataset of 1Mv3, a total of 712 male probands and 354 unaffected brothers, which had been genotyped using the 1Mv3 array. In the first step, we conducted GWAS in the whole dataset using the sib-TDT, and we observed no significant associations.

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 15 clusters in the 1Mv3 dataset are presented in Supplementary Table S1. In the second step, we conducted cluster-based GWAS in the 1Mv3 dataset using the Cochran-Armitage trend test[21]. We observed that the number of genome-wide significant loci slightly increased as the number of clusters increased (Supplementary Table S2), as observed with the Omni2.5 dataset. In the cluster-based GWAS with a cluster number of 15, a total of 7 chromosomal loci, which included one intragenic locus located in *THSD4* and 6

intergenic loci, satisfied the threshold of $P < 5.0 \times 10^{-8}$ (Supplementary Table S3).

Between the results from the Omni2.5 and 1Mv3 datasets, we observed no consistent genes that displayed genome-wide significance, although a consistent increase in the number of genome-wide significant loci as the number of clusters increased was observed. Several explanations may be possible: First, the loci that showed genome-wide significance in the Omni2.5 and the 1Mv3 datasets might be almost false positives. Second, 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. Third, the replication study used different controls from those used in the discovery study; thus, the difference in characteristics in the two groups of controls may also have affected the results of the replication study. Finally, the extremely heterogeneous features of ASD might affect the results of the replication study. If ASD is actually associated with more than 1,000 genes[6], the aetiological mechanism involving the substantial number of genes must be highly complex, making it unlikely that we would identify consistently associated loci among subgroups.

## Discussion

To the best of our knowledge, this is the first study to demonstrate that grouping persons with ASD using clustering algorithms is useful to discriminate more genetically homogeneous ASD persons. We observed many statistically significant SNPs, which is consistent with the findings from previous studies, and significant high ORs and corresponding reasonable lambda values, indicating that our results indeed have reasonable validity.

Previous studies regarding ASD, Alzheimer's disease, neuroticism, or asthma found that

items or symptoms showed, to some degree, increased ORs between the cases' loci and controls' loci compared to those from previous studies using broadly defined disease diagnoses[9-12]. These findings may indicate that GWAS with 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. Clustering algorithms could make essentially homogeneous clusters. To the best of our knowledge, these algorithms using phenotypic variables have not been applied for subgrouping multifactorial diseases to date. The present study demonstrates that clustering is one of the successful approaches to identifying more homogeneous subgroups.

Selection of variables is a critical issue in conducting clustering analysis. In this study, we focused on ADI-R variables and treatment, which have been indicated as candidates in previous studies[16,18,19]. We believe this protocol is an appropriate way of identifying subgroups of ASD. Nevertheless, further clustering utilizing other variables is warranted because ASD is highly heterogeneous and there are many variables for evaluating ASD symptoms. We can obtain many types of clusters from various views, and the ultimate cluster is the individuals themselves because every person has different genetic factors; however, we believe that one of the goals of clustering is the identification of subgroups based on treatment responsiveness, which may indicate the implementation of precision medicine for ASD.

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 successfully identified subgroups that are aetiologically more homogeneous (Supplementary fig. S3) To date, genetic studies have been conducted with very large sample sizes and have found modest to moderate impacts of genetic factors on multifactorial diseases,

called missing heritability[53]. The present study indicates that the reason for the observed modest effects in previous genetic studies may be disease heterogeneity because we observed several significantly high ORs. Our approach using clustering algorithms in machine learning methods might be a breakthrough approach for dealing with the issue of missing heritability. GWAS with a larger sample size is useful, but our data indicate that another strategy, such as clustering by phenotype, may also be useful.

Our data highlights the relevance of cluster-based GWAS as a means to identify more homogeneous subgroups of ASD than broadly defined one. The present study may provide clues to elucidate the aetiologies of ASD as well as that of other multifactorial diseases.

## Methods

We conducted the present study in accordance with the guidelines of the Declaration of Helsinki[54] 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 from all participants was obtained by the Simons Foundation Autism Research Initiative (SFARI)[15]. For participants under the age of 18 years, they obtained informed consent from a parent and/or legal guardian. Additionally, for participants 10 to 17 years of age, they obtained informed assent from the individuals.

### *Datasets*

We used phenotypic variables, history of treatment, and genotypic data from the Simons Simplex Collection (SSC)[15]. 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 about ADI-R scores and vitamin treatment 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[55,56] for the genotype data. Based on the PCA analyses, we excluded data beyond 4 standard deviations of principle components 1 or 2 (Supplementary Fig. S4). 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 about ADI-R scores and 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. Therefore, we used data from 712 probands and 354 unaffected brothers in the replication study.

### *Clustering*

In the cluster analysis, we used phenotypic variables of the ADI-R score and treatment[16]. Among 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 histories of treatments, the use of vitamins, though it does not guarantee effectiveness, was also included in the preprocessed dataset because we recently found that a cluster of persons with ASD is associated with potential responsiveness to vitamin B6 treatment[19].

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 phenotype 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)[57].

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 it 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[15,58]. 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 cut-off 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 shored 675,923

SNPs.

***Statistical analysis***

In the discovery studies and in the replication studies, GWAS were applied to ASD probands and unaffected brothers. In the first step, we conducted a GWAS for all male probands vs all unaffected brothers using the sib-TDT analyses. The first step association test was the sib-TDT for all probands and controls. In the second step, we conducted a GWAS by each subgroup of the male probands vs unaffected brothers without the brothers of the members of the subgroup being analysed (cluster-based GWAS) using the k-means[14] algorithm. We applied the Cochran-Armitage trend test[21] and Fisher's exact test[22] to both algorithms. Fig. 2 details the study design.

Association analyses were performed in the PLINK software package[59]. The detected SNPs were subsequently annotated using ANNOVAR[60]. Manhattan plots and the quantile-quantile (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.

## References

1. American Psychological Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)* (American Psychological Association (APA), 2018).
2. Rapin, I. Autism. *N. Engl. J. Med.* **337**, 97-104 (1997).
3. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol.* **14**, 1109-1120 (2015).
4. Bailey, A. *et al*. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol. Med.* **25**, 63-77 (1995).
5. Lauritsen, M. B., Pedersen, C. B. & Mortensen, P. B. 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 (2005).
6. Gene, S. Gene scoring. https://gene.sfari.org/database/gene-scoring/ (2018).
7. Eissa, N. *et al*. Current enlightenment about etiology and pharmacological treatment of autism spectrum disorder. *Front. Neurosci.* **12**, 304 (2018).
8. Traylor, M., Markus, H. & Lewis, C. M. Homogeneous case subgroups increase power in genetic association studies. *Eur. J. Hum. Genet.* **23**, 863-869 (2015).
9. Chaste, P. *et al*. 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 (2015).
10. Mukherjee, S. *et al*. Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Mol. Psychiatry* (2018). doi: 10.1038/s41380-018-0298-8.
11. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & van der Sluis, S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat. Commun.* **9**, 905 (2018).

12. Lavoie-Charland, E., Berube, J. C., Boulet, L. P. & Bosse, Y. Asthma susceptibility variants are more strongly associated with clinically similar subgroups. *J. Asthma* **53**, 907-913 (2016).

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

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

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

16. Beggiato, A. *et al*. Gender differences in autism spectrum disorders: divergence among specific core symptoms. *Autism. Res.* **10**, 680-689 (2017).

17. Sharma, S. R., Gonda, X. & Tarazi, F. I. Autism spectrum disorder: classification, diagnosis and therapy. *Pharmacol. Ther.* **190**, 91-104 (2018).

18. Kuriyama, S. *et al*. Pyridoxine treatment in a subgroup of children with pervasive developmental disorders. *Dev. Med. Child Neurol.* **44**, 284-286 (2002).

19. Obara, T. *et al*. Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. *Sci. Rep.* **8**, 14840 (2018).

20. Spielman, R. S. & Ewens, W. J. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450-458 (1998).

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

22. Fisher, R. A. On the interpretation of $\chi^2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85**, 87-94 (1922).

23. Emily, M. Power comparison of cochran-armitage trend test against allelic and genotypic tests in large-scale case-control genetic association studies. *Stat. Methods Med. Res.* **27**, 2657-2673 (2018).

24. Cutting, D. R., Karger, D. R., Pedersen, J. O. & Tukey, J. W. 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* 318-329 (Association for Computing Machinery (ACM), 1992).

25. Ketchen, D. J., & Shook, C. L. The application of cluster analysis in strategic management research: an analysis and critique. *Strateg. Manag. J.* **17**, 441-458 (1996).

26. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459-463 (2010).

27. Zeng, P. *et al*. Statistical analysis for genome-wide association study. *J. Biomed. Res.* **29**, 285-297 (2015).

28. Wang, Y. *et al*. Genome-wide association study of piglet uniformity and farrowing interval. *Front. Genet.* **8**, 194 (2017).

29. Breslow, N. E. & Day, N. E. *Statistical Methods in Cancer Research*. (Oxford University Press, 1993).

30. Anttila, V. *et al*. Analysis of shared heritability in common disorders of the brain. *Science* **360**, pii: eaap8757 (2018).

31. Orru, S. *et al*. 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.* **1**, 1-5

(2019).

32. Sragovich, S., Merenlender-Wagner, A., & Gozes, I. ADNP plays a key role in autophagy: from autism to schizophrenia and alzheimer's Disease. *Bioessays.* **39**, doi: 10.1002/bies.201700054. (2017).

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

34. Redies, C., Hertel, N., & Hübner, C.A. Cadherins and neuropsychiatric disorders. *Brain Res.* **1470**, 130-144 (2012).

35. Varghese, M. *et al*. Autism spectrum disorder: neuropathology and animal models. *Acta. Neuropathol.* **134**, 537-566 (2017).

36. Atsem, S. *et al*. Paternal age effects on sperm FOXK1 and KCNA7 methylation and transmission into the next generation. *Hum. Mol. Genet.* **22**, 4996-5005 (2016).

37. Barnby, G. *et al*. 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 (2005).

38. Minhas, H.,M. *et al*. 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 (2013).

39. Abou-Donia, M. B., Suliman, H. B., Siniscalco, D., Antonucci, N. & ElKafrawy, P. *De novo* blood Biomarkers in autism: autoantibodies against neuronal and glial proteins. *Behav. Sci (Basel).* **9**, pii: E47. doi: 10.3390/bs9050047. (2019).

40. Lo Vasco, V. R. Role of phosphoinositide-specific phospholipase C η2 in isolated and

syndromic mental retardation. *Eur., Neurol.* **65**, 264-269 (2011).

41. Potkin, S. G. *et al*. Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia. *Mol. Psychiatry.* **14**, 416-428 (2009).

42. Konopaske, G. T. *et al*. Dysbindin-1 contributes to prefrontal cortical dendritic arbor pathology in schizophrenia. *Schizophr. Res.* **201**, 270-277. (2018).

43. Openshaw, R., L. *et al*. JNK signalling mediates aspects of maternal immune activation: importance of maternal genotype in relation to schizophrenia risk. *J. Neuroinflammation.* **16**, 18 (2019).

44. Ikeda, M. *et al*. 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 (2010).

45. Teng, X. *et al*. KCTD: a new gene family involved in neurodevelopmental and neuropsychiatric disorders. *CNS Neurosci. Ther.* **25**, 887-902 (2019).

46. Lin, C. H., Huang, M. W., Lin, C. H., Huang, C. H. & Lane, H. Y. 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. (2019).

47. Kunkle, B. W. *et al*. 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 (2019).

48. Cimino, P. J., Sokal, I., Leverenz, J., Fukui, Y. & Montine, T. J. DOCK2 is a microglial specific regulator of central nervous system innate immunity found in normal and Alzheimer's disease brain. *Am. J. Pathol.* **175**, 1622-1630. (2009).

49. Sepulveda-Diaz, J. E. *et al*. HS3ST2 expression is critical for the abnormal phosphorylation of tau in Alzheimer's disease-related tau pathology. *Brain.* **138**, 1339-1354 (2015).

50. Ghosh, D., Levault, K. R. & Brewer, G. J. 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. (2014).

51. Zong, Y. *et al*. miR-29c regulates NAV3 protein expression in a transgenic mouse model of Alzheimer's disease. *Brain Res.* **1624**, 95-102 (2015).

52. Lu, J. *et al*. Global hypermethylation in fetal cortex of down syndrome due to DNMT3L overexpression. *Hum. Mol. Genet.* **25**, 1714-1727 (2016).

53. Manolio, T. A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

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

55. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).

56. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

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

58. Sanders, S. J. *et al*. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**,

863-885 (2011).

59. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

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

**Acknowledgements**

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.

**Author contributions**

A.N., M.N., S.M., S.O. G.T. and S.K. designed the study. M.N. and S.K. conducted the clustering analyses. A.N., M.N., S.M., S.O. and G.T. conducted GWAS. A.N., M.N., S.M., S.O. G.T. and S.K. drafted the manuscript. M.U., R.S., S.M., T.O., M.I., C.Y., H.M., Y.K., K.M., T.K., M.K., T.U., H.O., A.H., M.K., H.M., and S.K. helped with the interpretation of data. A.N., M.N., S.M., S.O., G.T., M.U., R.S., S.M., T.O., M.I., C.Y., H.M., Y.K., K.M., T.K., M.K., T.U., H.O., A.H., M.K., H.M., S.K., and S.K. edited the manuscript and gave intellectually critical contributions to it.

**Competing interests**

The authors declare no competing interests.

## Figure legends

**Fig. 1. 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 SSC 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 $P < 5.0 \times 10^{-8}$.

**Fig. 2. Details of the cluster-based GWAS in the present study.**

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.

**Fig. 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 the 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, labelled in the figure, that satisfied the threshold of $P < 5.0 \times 10^{-8}$. The red lines indicate the threshold for genome-wide significance ($P < 5.0 \times 10^{-8}$).

## Tables

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

| Test methods | No. of clusters | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 10 | 15 | 20 |
| Sibling-based transmission disequilibrium test | 0 | - | - | - | - | - | - | - |
| λ value | 1.032 | - | - | - | - | - | - | - |
| Cochran-Armitage trend test | - | 0 | 2 | 5 | 1 | 26 | 65 | 211 |
| Mean λ value (min-max) | - | 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) |
| Fisher's exact test | - | 0 | 2 | 5 | 0 | 0 | 0 | 4 |
| Mean λ value (min-max) | - | 0.889 (0.889-0.890) | 0.874 (0.856-0.886) | 0.855 (0.845-0.861) | 0.835 (0.822-0.848) | 0.773 (0.701-0.810) | 0.718 (0.667-0.786) | 0.680 (0.602-0.737) |

**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 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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.

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

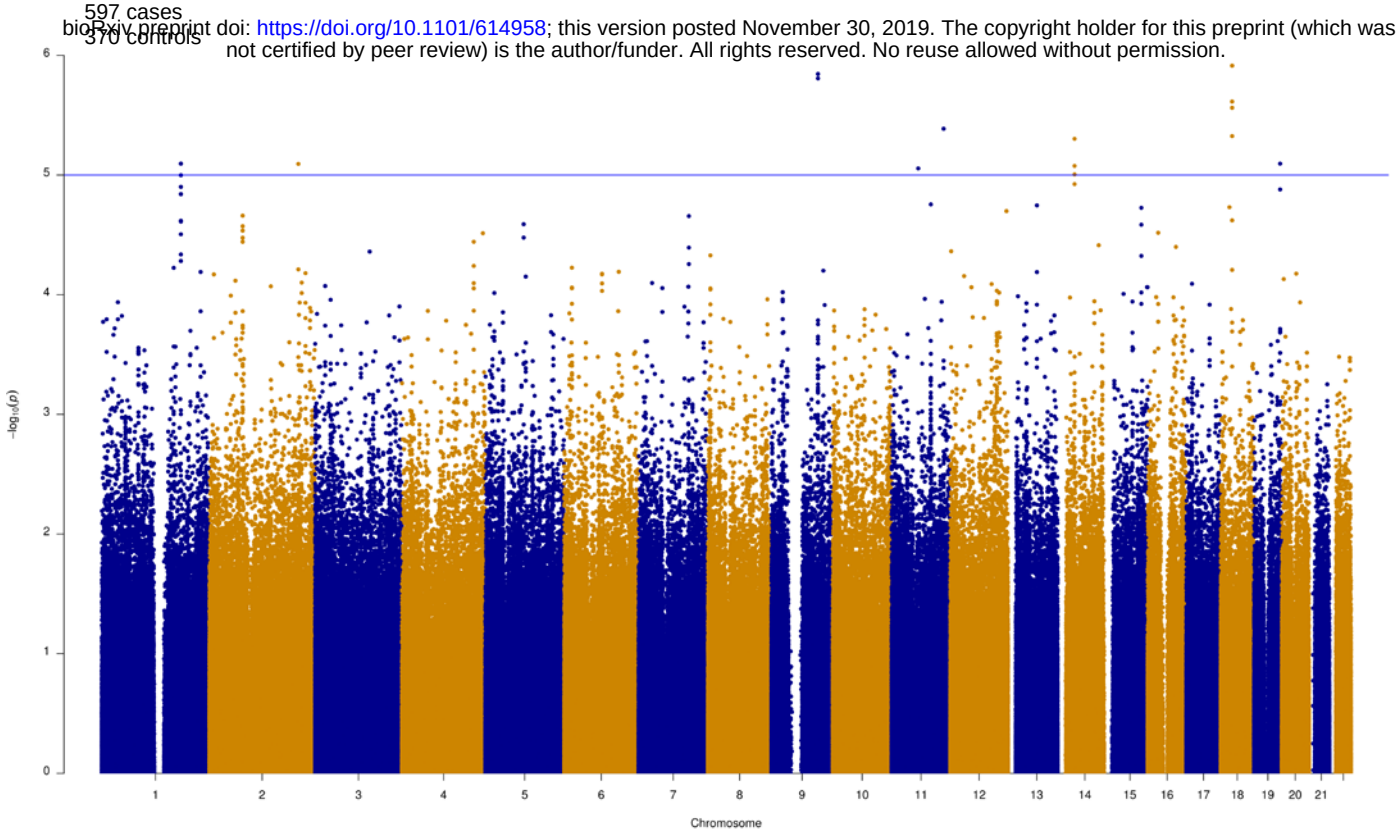
| Cluster No. | ID | Chr | hg19 | Minor/ major | MAF (%) | OR | 95% CI | P | GENESYMBOL | Function |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs111629286 | 11 | 130,152,136 | A/G | 1.80 | 13.42 | 4.38-41.17 | $1.36 \times 10^{-8}$ | ZBTB44 | Intronic |
| 1 | rs115140946 | 6 | 37,891,923 | C/A | 1.03 | 21.07 | 4.79-92.77 | $2.87 \times 10^{-8}$ | ZFAND3 | Intronic |
| 1 | rs9462391 | 6 | 38,123,030 | A/G | 1.03 | 21.07 | 4.79-92.77 | $2.87 \times 10^{-8}$ | ZFAND3 | Downstream |
| 1 | rs10217283 | 9 | 101,423,675 | A/G | 1.42 | 15.51 | 4.45-54.12 | $2.95 \times 10^{-8}$ | GABBR2 | Intronic |
| 1 | rs114109395 | 6 | 38,005,546 | A/G | 1.03 | 21.01 | 4.77-92.51 | $3.02 \times 10^{-8}$ | ZFAND3 | Intronic |
| 2 | rs115621412 | 9 | 74,366,033 | C/A | 7.89 | 4.42 | 2.48-7.87 | $8.13 \times 10^{-9}$ | TMEM2 | Intronic |
| 3 | rs77507687 | 2 | 26,939,229 | G/A | 2.00 | 12.43 | 4.37-35.36 | $6.10 \times 10^{-9}$ | KCNK3 | Intronic |
| 3 | rs76880969 | 1 | 227,711,506 | G/A | 1.00 | 27.15 | 5.30-139.20 | $8.20 \times 10^{-9}$ | CDC42BPA, ZNF678 | Intergenic |
| 3 | rs115483919 | 2 | 125,010,267 | A/G | 1.00 | 27.15 | 5.30-139.20 | $8.20 \times 10^{-9}$ | CNTNAP5 | Intronic |
| 5 | rs16965293 | 16 | 9,551,490 | A/G | 2.31 | 14.04 | 5.00-39.45 | $3.83 \times 10^{-10}$ | LINC01195, GRIN2A | Intergenic |
| 5 | rs77489014 | 9 | 106,962,281 | A/G | 1.41 | 19.47 | 5.51-68.82 | $6.69 \times 10^{-10}$ | SMC2LOC105376194 | Intergenic |
| 5 | rs117473168 | 9 | 106,848,270 | A/G | 1.55 | 16.90 | 5.02-56.93 | $2.64 \times 10^{-9}$ | SMC2 | ncRNA exonic |
| 5 | rs7199670 | 16 | 22,875,238 | A/G | 11.28 | 5.28 | 2.76-10.10 | $4.98 \times 10^{-9}$ | HS3ST2 | Intronic |
| 5 | rs73142209 | 12 | 77,859,299 | G/A | 1.54 | 16.18 | 4.82-54.31 | $5.33 \times 10^{-9}$ | E2F7, NAV3 | Intergenic |
| 5 | rs118167078 | 15 | 65,723,796 | A/G | 1.54 | 16.18 | 4.82-54.31 | $5.33 \times 10^{-9}$ | IGDCC4, DPP8 | Intergenic |
| 5 | rs11919513 | 3 | 4,841,384 | G/A | 3.22 | 10.18 | 3.99-26.03 | $8.92 \times 10^{-9}$ | ITPR1 | Intronic |
| 5 | rs13332627 | 16 | 22,874,928 | G/A | 9.23 | 6.10 | 2.96-12.57 | $1.22 \times 10^{-8}$ | HS3ST2 | Intronic |
| 5 | rs111920363 | 7 | 143,656,906 | A/G | 1.15 | 19.46 | 4.89-77.39 | $1.29 \times 10^{-8}$ | OR2F1 | Upstream |
| 5 | rs9939816 | 16 | 22,876,408 | A/C | 9.25 | 6.08 | 2.95-12.53 | $1.30 \times 10^{-8}$ | HS3ST2 | Intronic |
| 5 | rs76096239 | 14 | 97,193,704 | A/G | 1.67 | 13.79 | 4.27-44.54 | $3.25 \times 10^{-8}$ | PAPOLA, LINC02299 | Intergenic |
| 5 | rs1054028 | 16 | 22,927,214 | G/A | 14.36 | 5.02 | 2.62-9.61 | $3.32 \times 10^{-8}$ | HS3ST2 | UTR3 |
| 5 | rs78486970 | 7 | 106,127,612 | G/A | 6.87 | 5.46 | 2.67-11.18 | $3.68 \times 10^{-8}$ | NAMPT, CCDC71L | Intergenic |
| 6 | rs148617803 | 1 | 76,136,228 | G/A | 1.32 | 22.57 | 5.95-85.64 | $2.77 \times 10^{-10}$ | SLC44A5, ACADM | Intergenic |
| 6 | rs55985845 | 10 | 25,163,664 | T/A | 2.51 | 11.71 | 4.26-32.20 | $7.18 \times 10^{-9}$ | PRTFDC1 | Intronic |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | rs73094424 | 12 | 39,840,397 | A/G | 2.11 | 12.09 | 4.12-35.52 | $2.70 \times 10^{-8}$ | KIF21A, ABCD2 | Intergenic |
| 6 | rs58845693 | 3 | 122,804,247 | G/A | 1.18 | 18.07 | 4.55-71.72 | $4.24 \times 10^{-8}$ | PDIA5 | Intronic |
| 6 | rs11709496 | 3 | 122,809,400 | G/A | 1.18 | 18.07 | 4.55-71.72 | $4.24 \times 10^{-8}$ | PDIA5 | Intronic |
| 6 | rs199531954 | 12 | 95,064,359 | C/A | 1.19 | 17.92 | 4.52-71.10 | $4.92 \times 10^{-8}$ | TMCC3, MIR492 | Intergenic |
| 7 | rs79033134 | 17 | 76,473,288 | A/G | 1.53 | 16.29 | 4.87-54.44 | $4.24 \times 10^{-9}$ | DNAH17 | Intronic |
| 7 | rs57127555 | 17 | 76,475,811 | C/A | 1.54 | 16.24 | 4.86-54.28 | $4.49 \times 10^{-9}$ | DNAH17 | Intronic |
| 7 | rs75382702 | 11 | 81,149,755 | A/G | 1.28 | 16.94 | 4.54-63.23 | $3.18 \times 10^{-8}$ | LOC101928944, MIR4300HG | Intergenic |
| 8 | rs73149247 | 3 | 100,864,047 | G/A | 2.21 | 11.41 | 3.88-33.54 | $5.80 \times 10^{-11}$ | ABI3BP, IMPG2 | Intergenic |
| 8 | rs12418400 | 11 | 131,263,123 | G/A | 1.56 | 20.88 | 6.09-71.57 | $6.68 \times 10^{-11}$ | NTM | Intronic |
| 8 | rs78323783 | 10 | 45,084,432 | A/G | 1.17 | 24.79 | 6.13-100.30 | $2.28 \times 10^{-10}$ | CXCL12, TMEM72 | Intergenic |
| 8 | rs72991663 | 6 | 130,143,713 | A/G | 2.85 | 13.30 | 4.84-36.53 | $5.51 \times 10^{-10}$ | ARHGAP18, TMEM244 | Intergenic |
| 8 | rs74922057 | 21 | 41,595,011 | A/G | 1.31 | 19.67 | 5.22-74.14 | $3.13 \times 10^{-9}$ | DSCAM | Intronic |
| 8 | rs115035406 | 21 | 41,580,474 | G/A | 1.42 | 16.53 | 4.61-59.31 | $1.97 \times 10^{-8}$ | DSCAM | Intronic |
| 8 | rs114994877 | 4 | 136,731,494 | A/G | 1.42 | 16.53 | 4.61-59.31 | $1.97 \times 10^{-8}$ | LINC02485, LINC00613 | Intergenic |
| 8 | rs117008682 | 9 | 103,245,053 | G/A | 1.43 | 16.48 | 4.59-59.15 | $2.8 \times 10^{-8}$ | MSANTD3 | Intronic |
| 8 | rs117772706 | 9 | 81,338,445 | G/A | 1.43 | 16.44 | 4.58-58.98 | $2.19 \times 10^{-8}$ | PSAT1, LOC101927450 | Intergenic |
| 9 | rs4885429 | 13 | 77,400,673 | G/A | 2.14 | 13.69 | 4.91-38.16 | $4.67 \times 10^{-10}$ | LMO7DN, KCTD12 | Intergenic |
| 9 | rs45618836 | 7 | 73,480,258 | G/A | 2.26 | 11.94 | 4.43-32.18 | $2.30 \times 10^{-9}$ | ELN | Intronic |
| 9 | rs7299395 | 12 | 41,714,602 | A/G | 3.27 | 8.52 | 3.65-19.89 | $1.15 \times 10^{-8}$ | PDZRN4 | Intronic |
| 9 | rs55772967 | 7 | 73,448,499 | G/A | 2.89 | 8.91 | 3.66-21.66 | $2.9 \times 10^{-8}$ | ELN | Intronic |
| 10 | rs72799348 | 2 | 22,637,443 | A/G | 2.31 | 12.84 | 4.74-34.77 | $6.57 \times 10^{-10}$ | LINC01822, LINC01884 | Intergenic |
| 10 | rs76159464 | 5 | 169,446,509 | A/G | 1.02 | 28.05 | 5.47-144.00 | $5.03 \times 10^{-9}$ | DOCK2 | Intronic |
| 10 | rs12483301 | 21 | 28,070,591 | G/A | 1.92 | 11.89 | 3.94-35.92 | $6.74 \times 10^{-9}$ | CYYR1, ADAMTS1 | Intergenic |
| 10 | rs72883714 | 18 | 23,987,552 | A/G | 2.17 | 11.25 | 4.08-31.06 | $1.59 \times 10^{-8}$ | TAF4B, LINC01543 | Intergenic |
| 10 | rs1876769 | 2 | 22,678,191 | A/G | 2.17 | 11.25 | 4.08-31.06 | $1.59 \times 10^{-8}$ | LINC01822, LINC01884 | Intergenic |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | rs17043765 | 2 | 22,656,804 | A/G | 2.17 | 11.25 | 4.08-31.06 | $1.59 \times 10^{-8}$ | LINC01822, LINC01884 | Intergenic |
| 11 | rs74645195 | 4 | 48,330,367 | G/A | 2.71 | 10.26 | 3.95-26.66 | $1.34 \times 10^{-8}$ | TEC, SLAIN2 | Intergenic |
| 11 | rs78513244 | 1 | 2,360,342 | A/G | 3.25 | 9.33 | 3.79-22.98 | $1.35 \times 10^{-8}$ | PEX10, PLCH2 | Intergenic |
| 11 | rs10027938 | 4 | 90,242,059 | A/G | 16.93 | 4.48 | 2.49-8.05 | $2.29 \times 10^{-8}$ | GPRIN3, SNCA | Intergenic |
| 12 | rs117647850 | 8 | 79,156,756 | A/G | 3.08 | 10.88 | 4.44-26.68 | $5.10 \times 10^{-11}$ | LOC102724874, PKIA | Intergenic |
| 12 | rs4131532 | 1 | 3,540,256 | A/G | 1.54 | 15.63 | 4.68-52.14 | $8.61 \times 10^{-9}$ | MEGF6, TPRG1L | Intergenic |
| 12 | rs77964987 | 4 | 183,685,432 | G/A | 4.77 | 7.06 | 3.21-15.53 | $4.97 \times 10^{-8}$ | TENM3 | Intronic |
| 13 | rs117954350 | 7 | 4,440,757 | A/G | 1.02 | 52.73 | 6.34-438.60 | $4.00 \times 10^{-10}$ | SDK1, FOXK1 | Intergenic |
| 13 | rs11064685 | 12 | 119,590,881 | G/A | 6.14 | 5.15 | 2.66-9.97 | $4.46 \times 10^{-8}$ | SRRM4 | Intronic |
| 14 | rs77983358 | 12 | 82,393,237 | G/A | 1.52 | 21.71 | 5.50-85.76 | $1.29 \times 10^{-10}$ | LINC02426, CCDC59 | Intergenic |
| 14 | rs7118821 | 11 | 96,876,267 | C/A | 1.01 | 26.18 | 5.11-134.00 | $1.50 \times 10^{-8}$ | LOC105369443 | Intergenic |
| 14 | rs7122015 | 11 | 96,950,548 | G/A | 1.01 | 26.18 | 5.11-134.00 | $1.50 \times 10^{-8}$ | LOC105369443, CNTN5 | Intergenic |
| 14 | rs7106102 | 11 | 96,885,969 | A/G | 1.01 | 26.10 | 5.10-133.70 | $1.58 \times 10^{-8}$ | LOC105369443 | Intergenic |
| 14 | rs7189512 | 16 | 66,324,048 | A/G | 3.28 | 7.26 | 3.13-16.88 | $4.62 \times 10^{-8}$ | LINC00922, CDH5 | Intergenic |
| 15 | rs77311527 | 2 | 5,516,750 | G/A | 2.45 | 11.87 | 4.44-31.79 | $2.19 \times 10^{-9}$ | LINC01249, LINC01248 | Intergenic |
| 15 | rs276833 | 2 | 114,769,078 | A/G | 1.29 | 18.00 | 4.81-67.43 | $1.25 \times 10^{-8}$ | LINC01191, DPP10 | Intergenic |

OR; Odds ratio.
CI: Confidence interval.

**a**




**b**

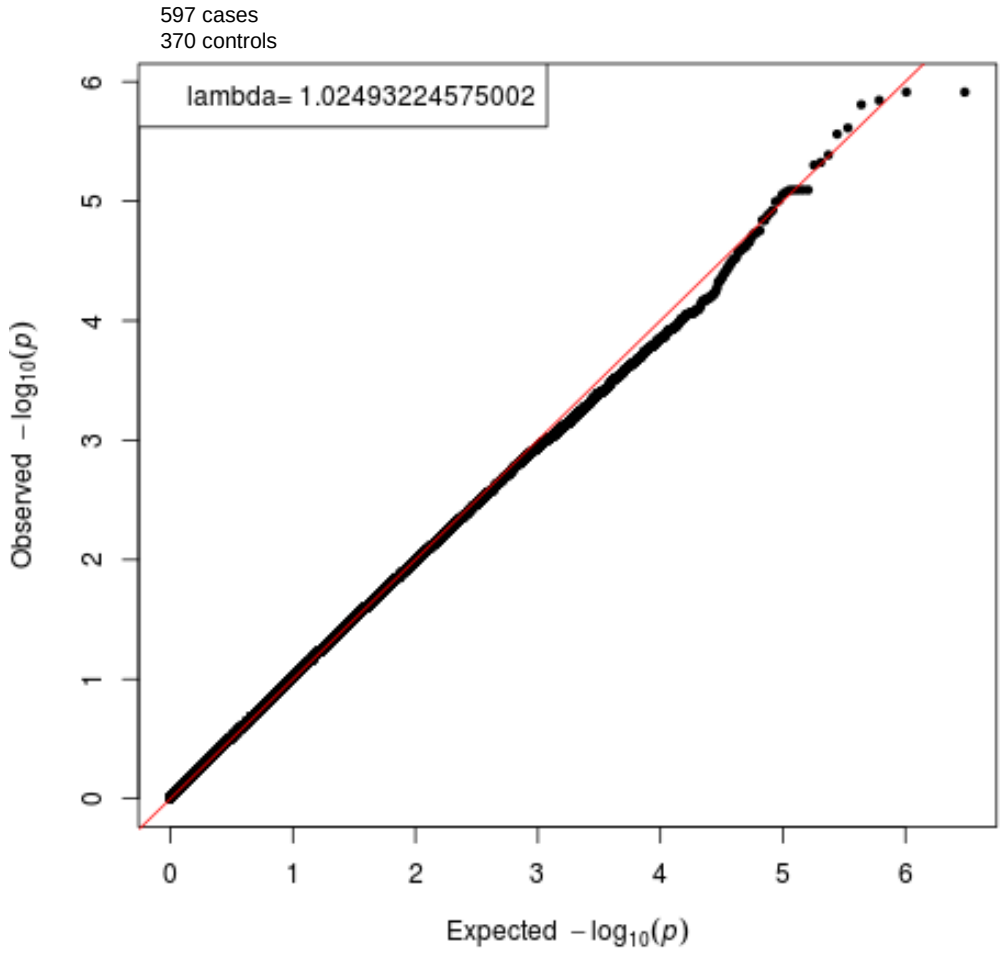



**Fig. 1.** Manhattan plots (a) and corresponding quantile-quantile plots (b) in GWAS for all male probands vs their unaffected brothers using the sib transmission/disequilibrium test.

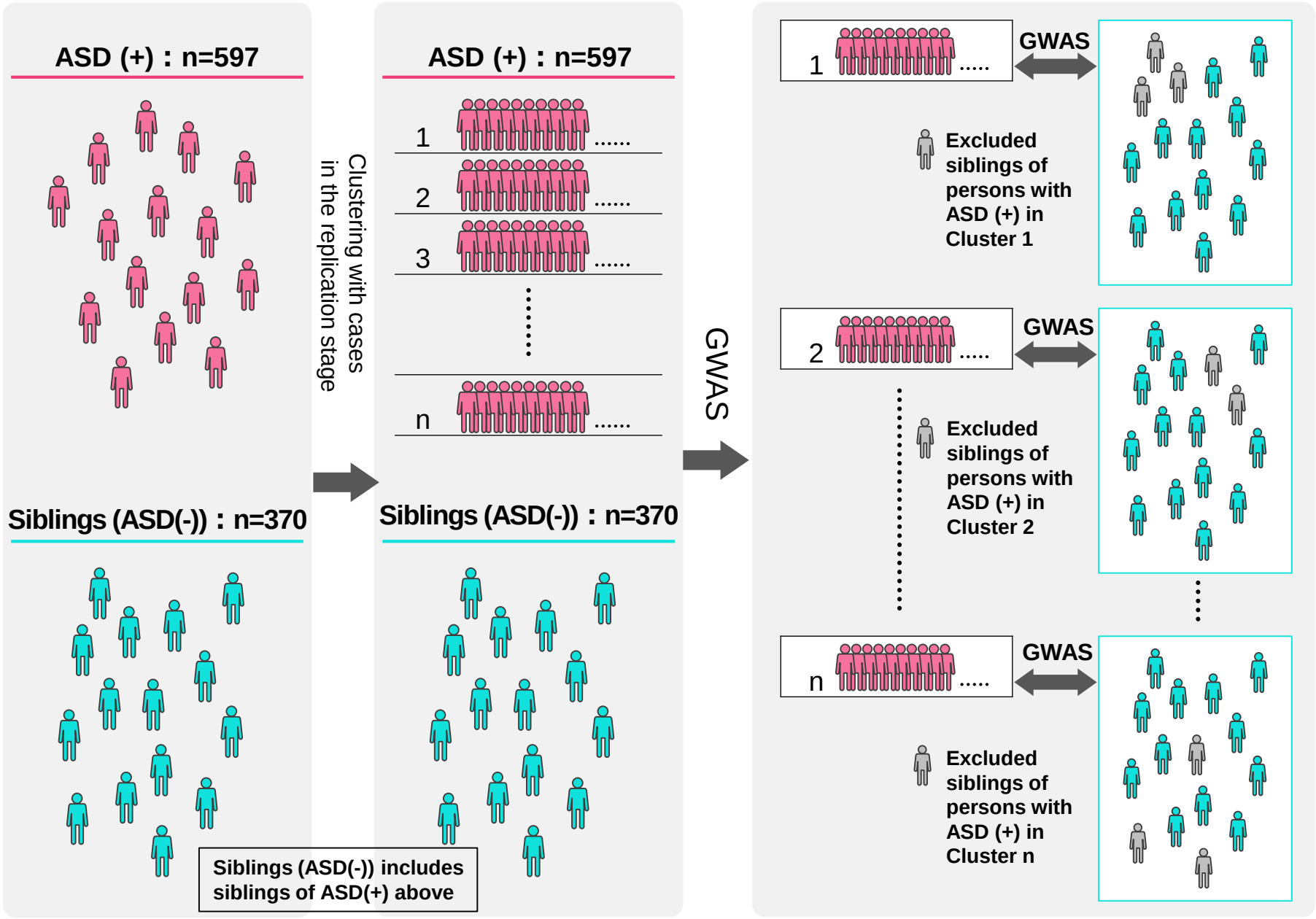
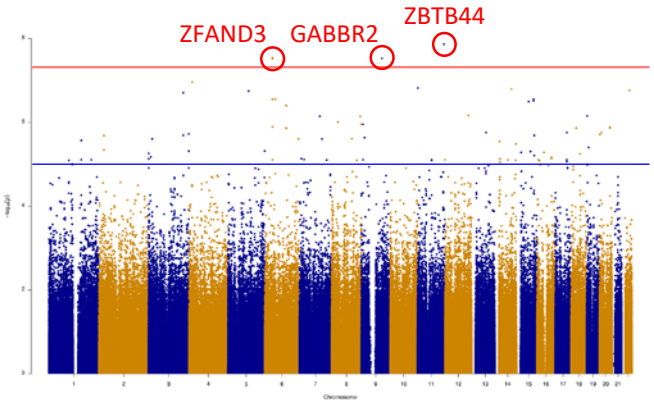


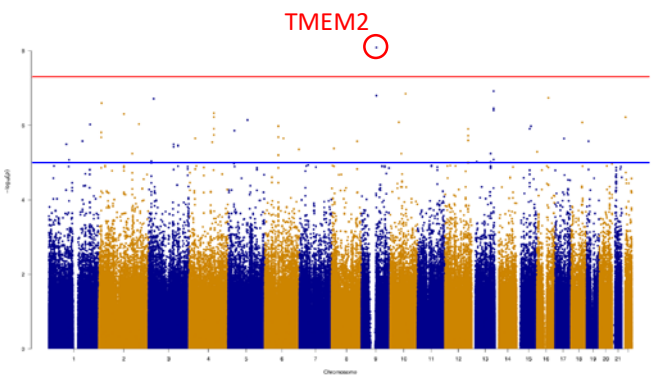

Fig. 2. Details of the cluster-based GWAS in the discovery stage.

# a

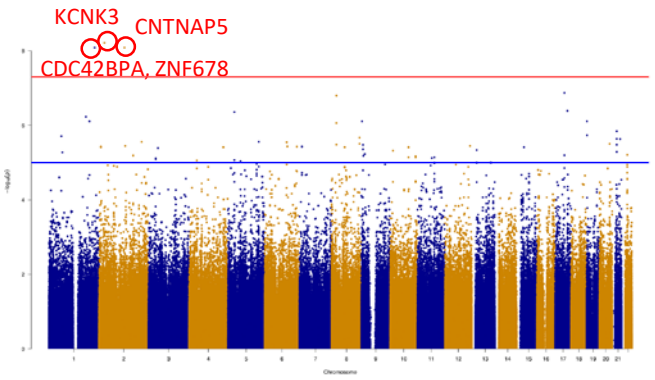
Cluster 1
Case = 33 Control = 357



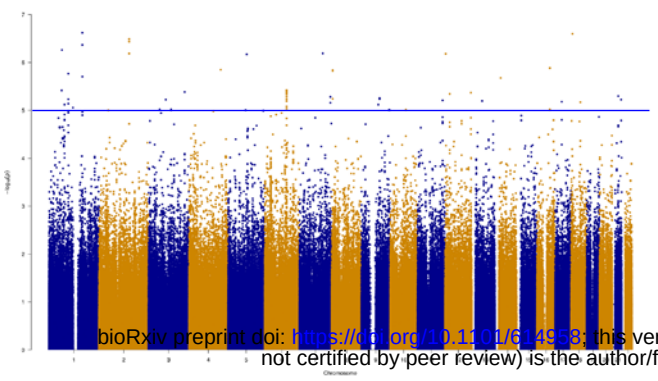

Cluster 2
Case = 49 Control = 350



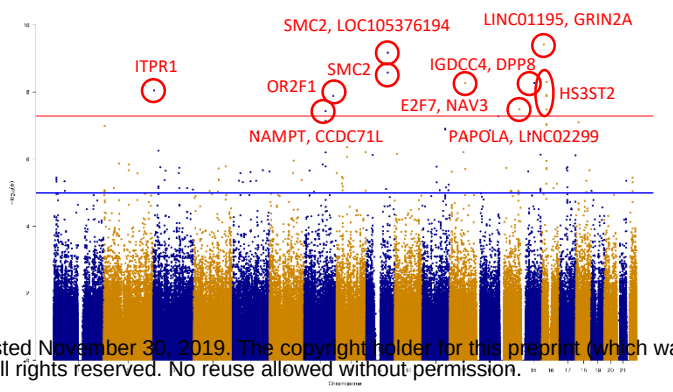

Cluster 3
Case = 45 Control = 355



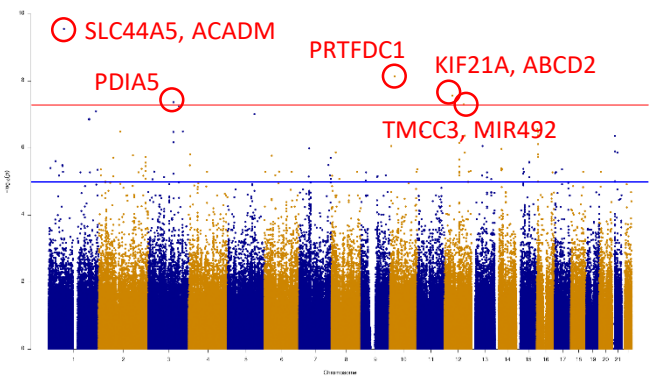

Cluster 4
Case = 59 Control = 339



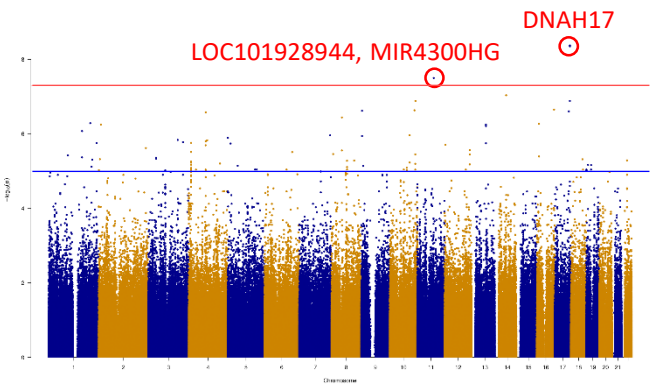
Cluster 5
Case = 28 Control = 362



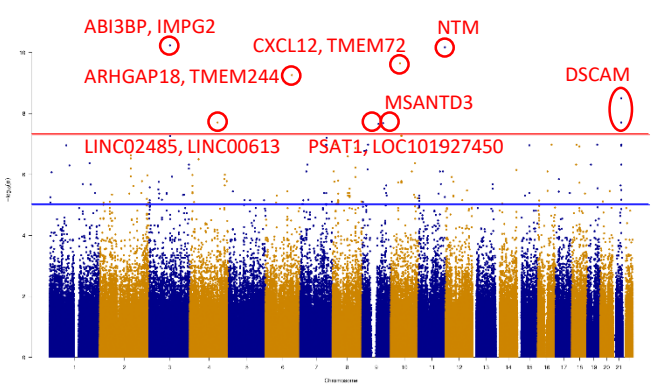

Cluster 6
Case = 29 Control = 351



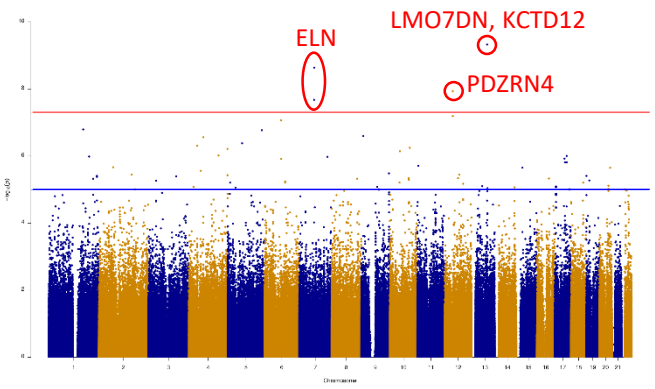

Cluster 7
Case = 37 Control = 354



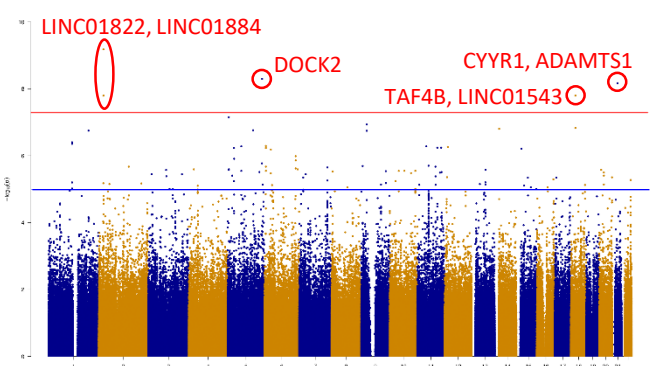

Cluster 8
Case = 23 Control = 363



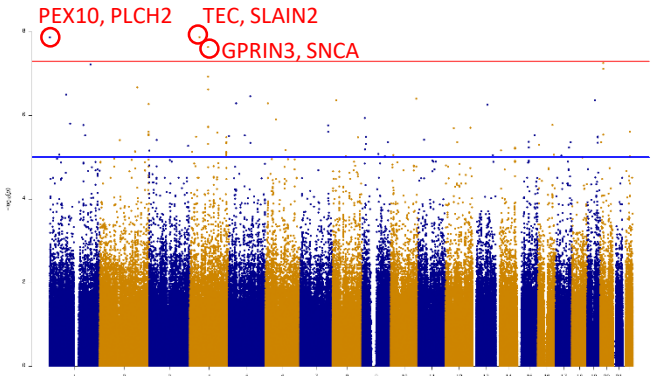

Cluster 9
Case = 46 Control = 352



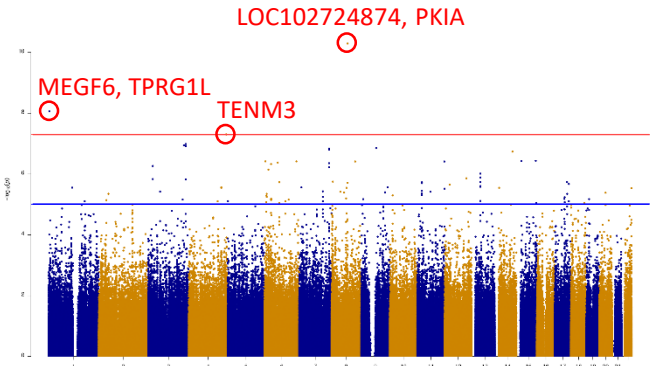

Cluster 10
Case = 43 Control = 348



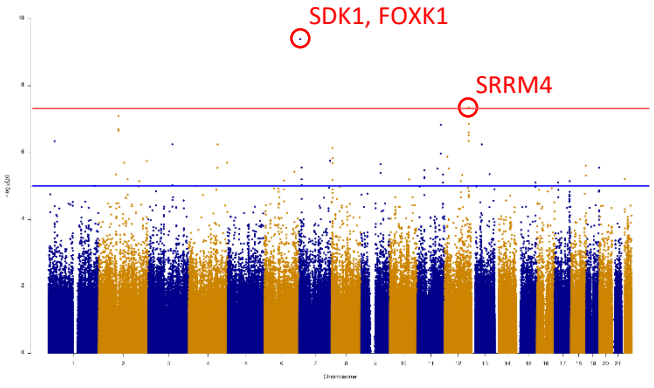

Cluster 11
Case = 34 Control = 354



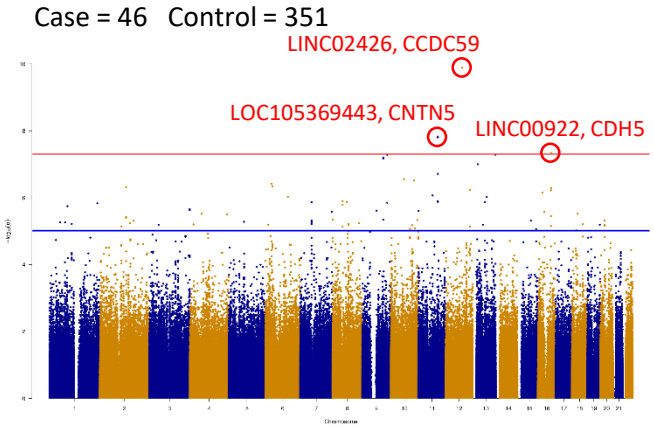

Cluster 12
Case = 38 Control = 352



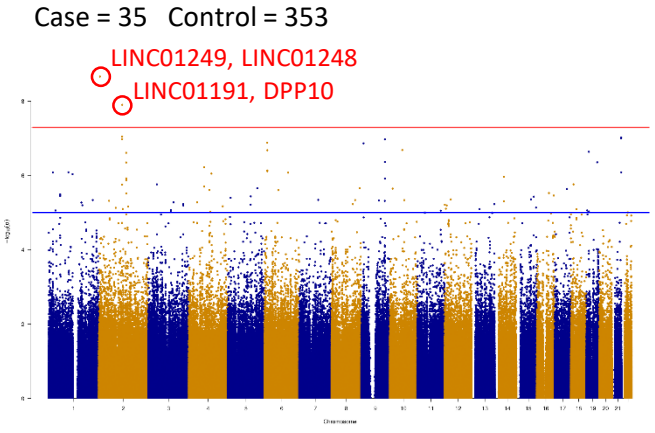

Cluster 13
Case = 52 Control = 341




Cluster 14
Case = 46 Control = 351

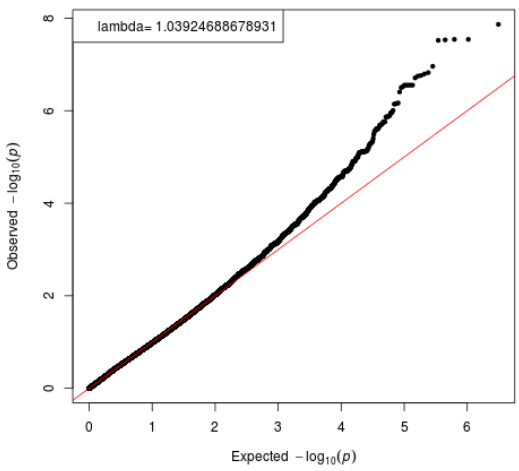



Cluster 15
Case = 35 Control = 353

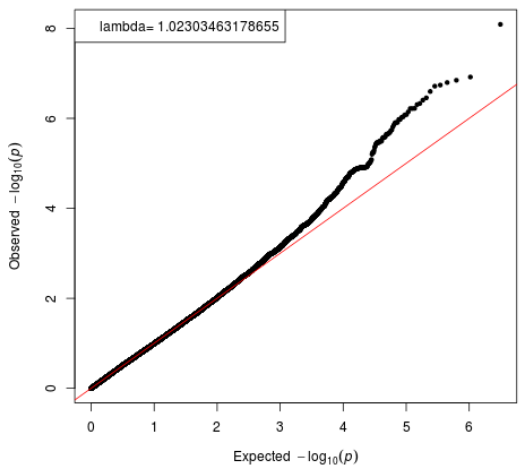


**b**

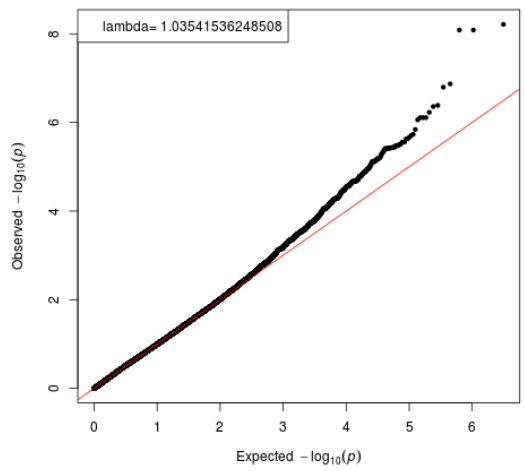
Cluster 1



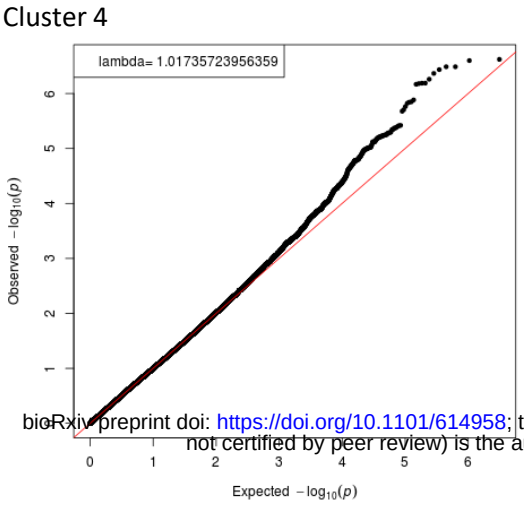

Cluster 2



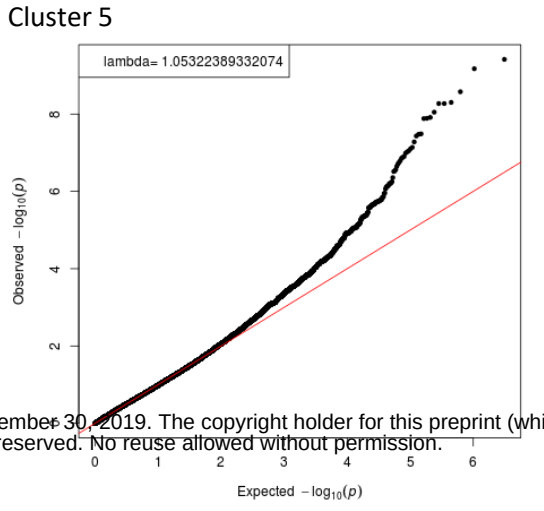

Cluster 3



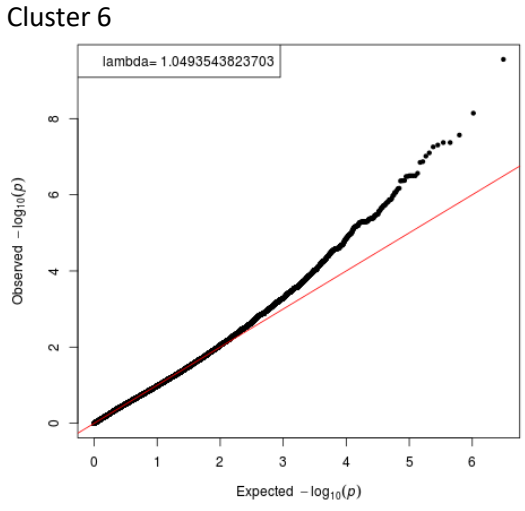

Cluster 4




Cluster 5

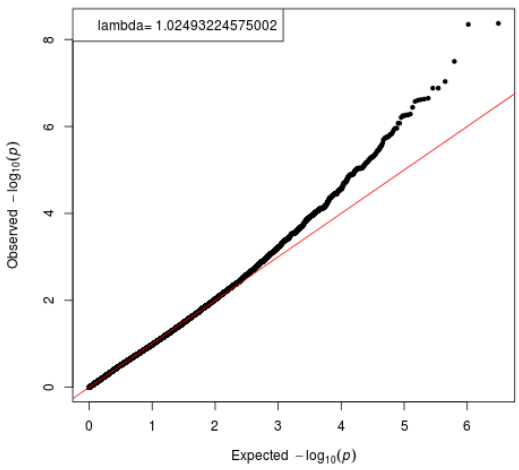



Cluster 6

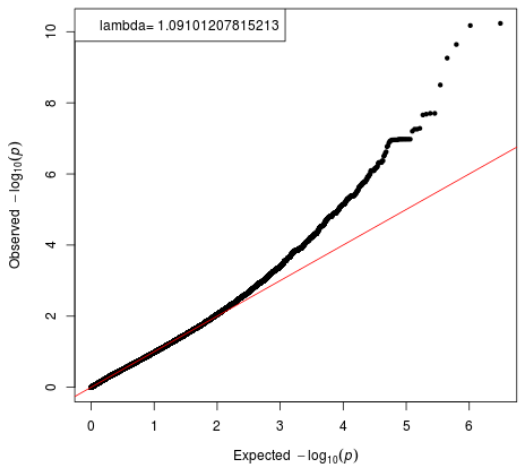




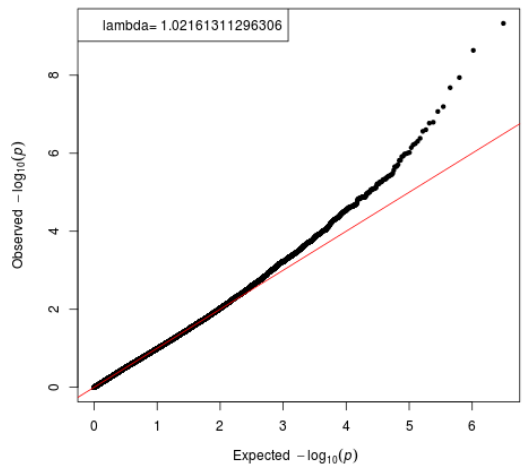

Cluster 7



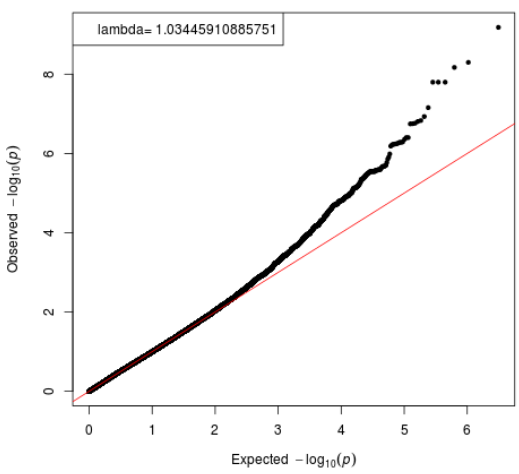

Cluster 8



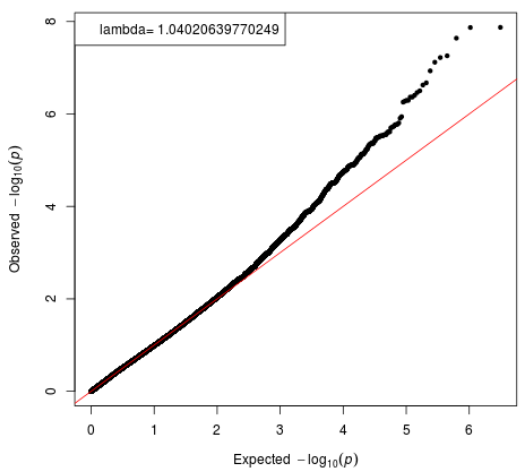

Cluster 9



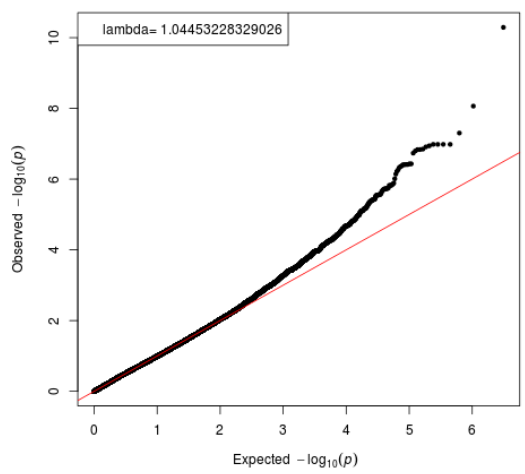

Cluster 10



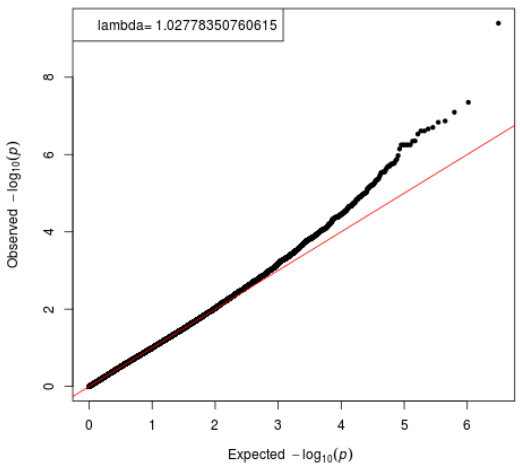

Cluster 11



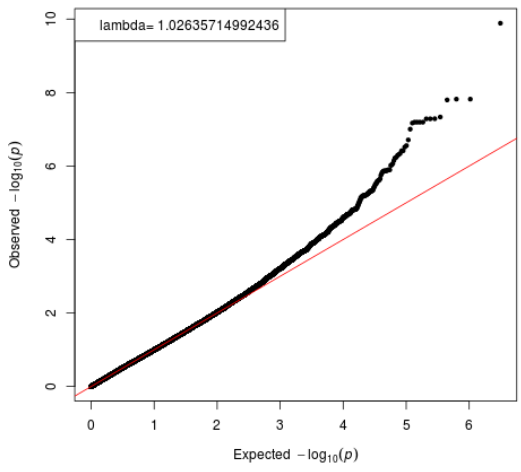

Cluster 12




Cluster 13




Cluster 14




Cluster 15

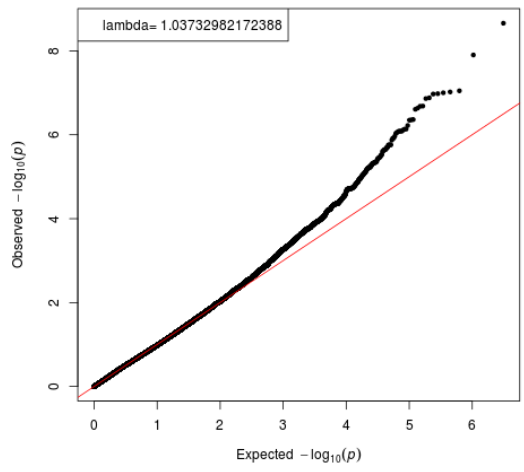



**Fig. 3.** Manhattan plots (a) and corresponding quantile-quantile plots (b) for cluster-based GWASs with a cluster number of 15.