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[1,2], 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]Department of Disaster Public Health, 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 address: kuriyama@med.tohoku.ac.jp

Manuscript information:

# Word count for abstract: 147

# Word count for text: 3,375

# References: 51

# Tables: 3

# Figures: 3

# Supplementary Tables: 2

# Supplementary Figure: 1

# Supplementary Information: 3

Abstract

Autism spectrum disorder (ASD) has clinically and genetically heterogeneous characteristics. Here, we show a two-step genome-wide association study (GWAS). In the first step, we observed no significant associations in a GWAS including 597 cases and 370 controls. In the second step, we conducted a cluster analysis using k-means with 15 clusters based on Autism Diagnostic Interview-Revised (ADI-R) scores and history of vitamin treatment. We then conducted GWAS by each subgroup of cases vs all controls (cluster-based GWAS) and identified significant associations with 93 chromosomal loci that satisfied the genome-wide significance threshold of $P<5.0\times10^{-8}$. These loci included previously reported candidate genes for ASD: *CDH9*, *MED13L*, *SOX5*, *CADM2*, *CADM1*, *DAB1*, *SEMA5A*, *RORA*, *MED13*, *COBL*, *EPHA7*, *HIF1AN*, *ICE1*, *PML*, and *WNT7B*. We observed that clustering-based GWAS, even with a smaller sample size, revealed abundant significant associations. These findings suggest that clustering may successfully identify subgroups that are aetiologically more homogeneous.

Introduction

Autism spectrum disorder (ASD) has heterogeneous characteristics, in terms of both phenotypic features and genetics. Clinically, 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 identify relatively common genetic variants that are associated with an increased risk of ASDs[3], although several lines of evidence suggest strong genetic components contribute to the susceptibility to ASDs. There are higher concordance rates of ASDs in monozygotic twins (92%) than in dizygotic twins (10%)[4]. The sibling recurrence risk ratio (λs) is 22 for ASD[5]. The Human Gene module of the Simons Foundation Autism Research Initiative (SFARI) Gene serves as a comprehensive, up-to-date reference for all known human genes associated with ASD[6] and currently demonstrates ~1,000 genes that have potential links to ASD, 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 variables may increase the chances of identifying common genetic susceptibility factors. A simulation study demonstrated that analysis of case subsets could be a powerful strategy to uncover some of the hidden heritability of common complex disorders[8]. Several studies of ASD, Alzheimer's disease, neuroticism, or asthma indicated that items or symptoms were in some degree useful to identify more genetically homogeneous subgroups of these diseases than broadly defined ones[9-12]. In recent years, ASD has been investigated using machine learning methods[13,14]. Machine learning employs artificial intelligence techniques to discover useful masked patterns. Clustering algorithms of machine learning could make novel and potentially more homogeneous clusters, but these algorithms using phenotypic variables have not, to the best of our knowledge, been applied to subgrouping multifactorial diseases to date.

In the present study, we explored whether grouping persons with ASD using clustering algorithms with phenotypic and responses to treatment variables can be used to discriminate more genetically homogeneous ASD persons. We applied machine learning k-means[15] or affinity propagation (AP)[16] algorithms to cluster analysis. Based on these clusters, we conducted genome-wide association studies (GWASs). We used genetic data to evaluate whether our clusters identify biologically homogeneous subgroups.

Results

Clustering

We used phenotypic variables, history of treatment, and genome-wide genotypic data from the Simons Simplex Collection (SSC)[17], the largest cohort of autism simplex families amassed to date. The SSC is a core project and resource of the SFARI[6].

To classify persons with ASD into more homogeneous subgroups, we conducted cluster analyses using phenotypic variables of Autism Diagnostic Interview-Revised (ADI-R)[18] 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[19], we selected the variable history of vitamin treatment because we recently found that a cluster of persons with ASD is associated with potential responsiveness to vitamin B6 treatment[20,21]. 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 k-means[15] or AP[16] algorithms. The k-means algorithm requires cluster numbers determined by researchers. AP algorithms do not need a priori cluster numbers; rather, the algorithm itself finds the appropriate one. When using k-means algorithms, we chose 2, 3, 4, 5, 10, 15, and 20 clusters. Interestingly, we observed that the AP analysis classified the participants into 36 groups.

Cluster-based genome-wide association study

GWASs were applied to male ASD probands and their unaffected brothers. In the first step, we conducted GWAS for all 597 male probands vs all 370 unaffected brothers using the sib transmission/disequilibrium test (sib-TDT)[22]. We observed no significant associations (Fig. 1).

In the second step, we conducted GWAS by each subgroup of the probands vs unaffected brothers as controls without the brothers of the members of the subgroup being analysed (cluster-based GWAS) (Fig. 2) using k-means or AP algorithms. We applied the Cochran-Armitage trend test[23,24] and Fisher's exact test[25] to both algorithms. Notably, we observed that the number of genome-wide significant loci 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 three significant loci were observed for numbers of clusters between two and 36. 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[26,27]. 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. The quantile-quantile (Q-Q) plots of the cluster-based GWAS with 20 clusters by k-means using Fisher's exact test demonstrated that almost all observed p-values were high compared to the expected distribution of p-values. In addition, genomic inflation factor (λ) 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 to be a more appropriate method in the present cluster-based GWAS.

Regarding appropriate cluster numbers, we compared the Q-Q plots and λ values among the analyses and observed that as the number of clusters increased, the observed p-values were lower than the expected distribution of p-values. For instance, the Q-Q plots for the cluster-based GWAS with 20 clusters by k-means using the Cochran-Armitage trend test demonstrated that the observed p-values were very low compared to the expected distribution of p-values. In addition, λ values ranged from 1.022 to 1.093, and the average was 1.054 (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[28-30].

In contrast, inflation factor λ values of the cluster-based GWAS with 15 clusters by k-means ranged from 1.018 to 1.065, and the average was 1.043, which was below 1.050 (Table 1 and Fig. 3).

According to the above results, we considered the cluster-based GWAS with 15 clusters by k-means using the Cochran-Armitage trend test to be the most appropriate approach to the present dataset. The characteristics of each cluster are presented in Table 2.

Our results indicate that clustering by specific phenotypic variables might be informative and provide the best model for identifying aetiologically similar cases of ASD.

Gene interpretation

Among the cluster-based GWASs, we mainly presented here the results using the Cochran-Armitage trend test by k-means with 15 clusters. In this cluster-based GWAS, we identified significant associations with 93 chromosomal loci that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$ (Table 1 and Fig. 3), and this cluster-based GWAS demonstrates that a total of 93 single nucleotide polymorphisms (SNPs), including 45 intragenic and 48 intergenic SNPs, satisfied the genome-wide significance threshold (Table 3). Among them, 9 genes corresponded to the Human Gene module of the SFARI Gene scoring system[6]; *CDH9* (score 4) in Cluster 3; *MED13L* (score 2, Rare Single Gene Mutation, Syndromic) in Clusters 7 and 13; *SOX5* (Rare Single Gene Mutation, Syndromic, Genetic Association) in Cluster 9; *CADM2* (score 4) in Cluster 9; *CADM1* (score 4, Rare Single Gene Mutation) in Cluster 10; *DAB1* (score 5) in Cluster 11; *SEMA5A* (score 3) in Cluster 12; *RORA* (Rare Single Gene Mutation, Syndromic, Genetic Association, Functional) in Cluster 13; and *MED13* (score 2, syndromic) 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 predisposing to autism in the context of a syndromic disorder (e.g., fragile X syndrome) are placed in a separate category. Rare single gene variants, disruptions/mutations, and sub-microscopic deletions/duplications directly linked to ASD are placed in “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[31,32] from previous reports were included in our findings as follows: *COBL* in Cluster 12, *EPHA7* in Cluster 3, *HIF1AN* in Cluster 4, *ICE1* in Cluster 2, *PML* in Cluster 15, and *WNT7B* in Cluster 8 previously reported with ASD[33-38]; *LHPP* in Cluster 7 previously reported with depression[39]; *KIDINS220* in Cluster 7 previously reported with intellectual disability[40]; *ALPL* in Cluster 6 previously reported with deleterious neurological outcome[41]; and *PAX2* in Cluster 4 previously reported with development of the central nervous system[42]. These findings suggest that the statistically significant SNPs might explain autistic symptoms because these diseases are suggested to share common aetiology, even in part, with ASD[31,32]. 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.

Replication study

To further validate the associations identified in the GWASs, we performed replication studies on another independent dataset from SSC, 1Mv3. In the first step, we conducted GWAS for all 712 male probands vs all 354 unaffected brothers using the sib-TDT test, and we observed no significant associations.

In the second step, we classified the male probands by k-means into 15 clusters and conducted GWAS for each subgroup vs the unaffected brothers as controls without the siblings of the members of the subgroup being analysed using the Cochran-Armitage trend test[23,24]. We observed that the number of genome-wide significant loci slightly increased as the number of clusters increased (Supplementary Table S1), as observed with the Omni2.5 data set. In this cluster-based GWAS using the Cochran-Armitage trend test by k-means with 15 clusters, we identified significant associations at 8 chromosomal loci that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$. Furthermore, this cluster-based GWAS demonstrated that a total of 8 SNPs, including 5 intragenic and 3 intergenic SNPs, satisfied the genome-wide significance threshold (Supplementary Table S2).

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 numbers of clusters increased was observed. One possible explanation might be the extremely heterogeneous features of the ASD genotype. If the genotype has more than 1,000 genes[6], each analysis with a sample size of less than one hundred vs hundreds with 15 clusters could find different genes.

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 odds ratios and corresponding reasonable lambda values, indicating 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, larger odds ratios of the odds among cases' loci to the odds among controls' loci compared to that 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 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 demonstrate 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[18,20,21]. 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 kinds 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.

AP is a relatively recently developed unsupervised machine learning clustering algorithm that identifies clusters of similar points using a set of points and a set of similarity values between the points and provides a representative example, called an exemplar, for each cluster[16]. We identified 36 clusters and 1,253 significant loci using the AP analysis, but our data also showed that the lambda values ranged from 1.032 to 1.093, with an average lambda value of 1.076 (Table 1). Although AP is a useful algorithm to identify clusters, the lambda values exceeded the appropriate threshold, i.e., less than 1.050, necessary to avoid false positives[28-30].

Therefore, the observed significant loci might include both true positives and false positives and we selected here the Cochran-Armitage trend test.

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. To date, genetic studies have been conducted with huge sample sizes and have found modest to moderate impacts of genetic factors on multifactorial diseases, called missing heritability[43]. 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 odds ratios. Our approach using clustering algorithms in machine learning methods may be a breakthrough approach for dealing with the issue of missing heritability and for identifying disease architectures. 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 strongly highlights the relevance of cluster-based GWAS as a means to identify more homogeneous subgroups of ASD than broadly defined ASD. The present study may provide clues to discover 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[44] 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)[17]. For participants under the age of 18 year, 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 genome-wide genotypic data from the Simons Simplex Collection (SSC)[17,] the largest cohort of autism simplex families amassed to date. The SSC establishes a repository of genetic samples from 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 child or adult probands who have no missing information about ADI-R scores and vitamin treatment and 391 unaffected brothers for whom Omni2.5 array data 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[45,46]. We also performed PCA for the genotype data in our study. Based on the PCA analyses, we excluded data beyond 4 standard deviations of principle components 1 or 2 (Supplementary Fig. S1). Therefore, we used data from 597 probands and 370 unaffected siblings.

In the replication study, we used the SSC 1Mv3 dataset. In the dataset, data from 735 affected male child or adult probands with no missing information about ADI-R scores and vitamin treatment and 387 unaffected child or adult male siblings 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 siblings in the replication study.

Cluster analysis

In the cluster analysis, we used phenotypic variables of the Autism Diagnostic Interview-Revised (ADI-R) score and treatment[18]. Among ADI-R scores, “The total score for the Verbal Communication Domain on the ADI-R algorithm minus the total score for the Nonverbal Communication Domain on the ADI-R algorithm”, “The total score for the Nonverbal Communication Domain on the ADI-R algorithm”, “The total score for the Restricted, Repetitive, and Stereotyped Patterns of Behavior Domain on the ADI-R algorithms”, and “The total score for the Reciprocal Social Interaction Domain on the ADI-R algorithms” 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[21].

We applied machine learning k-means[15] or affinity propagation (AP)[16] algorithms 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. AP algorithms do not need a priori cluster numbers, as the algorithm itself finds the appropriate number. When using k-means algorithms, we chose 2, 3, 4, 5, 10, 15, and 20 clusters. The ordinary k-means algorithm was first applied to the preprocessed dataset to divide the participants into more homogeneous subgroups[15]. Then, we used the relatively recently developed AP algorithm[16]. AP is an unsupervised clustering analysis using a message-passing-based algorithm. In the present study, AP was performed without diagonal components using a dumping factor of 0.9. These analyses were performed with the scikit-learn toolkit in Python 2.7 (Supplementary Information S1, Supplementary Information S2 and Supplementary Information S3)[47].

The cluster analyses described above were performed in the replication study as well.

Genotype data and quality control

We used the SSC dataset, in which probands and unaffected siblings had already been genotyped in other previous studies[17,48]. In the discovery-stage genome-wide association study (GWAS), all members of each family were analysed on the same array version, the Illumina HumanOmni2.5, which has approximately 2,450,000 probes. We excluded SNPs with a minor allele frequency (MAF) < 0.01, call rate < 0.95, and Hardy-Weinberg equilibrium test P < 0.000001 and obtained genotype data for 1000 participants in SSC.

In the replication study, we used genotyping data generated using the Illumina BeadChip in the SSC 1Mv3 datasets. We applied the same quality control criteria as those used in the discovery-stage GWAS.

Statistical analysis

In the discovery studies and in the replication studies, GWAS were applied to ASD probands and unaffected siblings. In the first step, we conducted a GWAS for all male probands vs all unaffected male siblings using sib-TDT analyses. The first step association test was the sib-TDT for all cases and controls. In the second step, we conducted a GWAS by each subgroup of the male probands vs unaffected male siblings without the siblings of the members of the subgroup being analysed (cluster-based GWAS) using k-means[15] or AP[16] algorithms. We applied the Cochran-Armitage trend test[23,24] and Fisher’s exact test[25] to both algorithms. Details of the study design are also indicated in Fig. 2.

Association analyses were performed in PLINK version 1.07[49] and 1.9[50]. The detected SNPs were subsequently annotated using ANNOVAR[51]. 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.

References

1. American Psychological Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)* (American Psychological Association, Washington, DC, 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. Thabtah, F. Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. *Inform. Health Soc. Care* 1-20 (2018). doi: 10.1080/17538157.2017.1399132

15. 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, Berkeley, 1967).

16. Frey, B. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972-976 (2007).

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

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

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

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

21. 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).

22. 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).

23. 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).

24. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885 (2007).

25. Fisher, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85**, 87-94 (1922).

26. Sasieni, P. D. From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253-1261 (1997).

27. 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).

28. 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).

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

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

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

32. Bralten, J. et al. Autism spectrum disorders and autistic traits share genetics and biology.

Mol. Psychiatry **23**, 1205-1212 (2018).

33. Griswold, A. J. et al. Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Hum. Mol. Genet.* **21**, 3513-3523 (2012).

34. Wurzman, R., Forcelli, P. A., Griffey, C. J. & Kromer, L. F. Repetitive grooming and sensorimotor abnormalities in an ephrin-a knockout model for autism spectrum disorders. *Behav. Brain. Res.* **278**, 115-128 (2015).

35. Martinez-Noel, G. et al. Identification and proteomic analysis of distinct UBE3A/E6AP protein complexes. *Mol. Cell. Biol.* **32**, 3095-3106 (2012).

36. Zhang, B. et al. Multigenerational autosomal dominant inheritance of 5p chromosomal deletions. *Am. J. Med. Genet. A* **170**, 583-593 (2016).

37. Silva, A. E., Vayego-Lourenco, S. A., Fett-Conte, A. C., Goloni-Bertollo, E. M. & Varella-Garcia, M. Tetrasomy 15q11-q13 identified by fluorescence in situ hybridization in a patient with autistic disorder. *Arq. Neuropsiquiatr.* **60**, 290-294 (2002).

38. Darbandi, S. F. et al. Neonatal Tbr1 dosage controls cortical layer 6 connectivity. *Neuron* **S0896-S6273**, 30829-30838 (2018).

39. Cui, L. et al. Relationship between the LHPP gene polymorphism and resting-state brain activity in major depressive disorder. *Neural Plast.* **2016**, 9162590 (2016).

40. Josifova, D. J. et al. Heterozygous KIDINS220/ARMS nonsense variants cause spastic paraplegia, intellectual disability, nystagmus, and obesity. *Hum. Mol. Genet.* **25**, 2158-2167 (2016).

41. Hofmann, C. et al. Compound heterozygosity of two functional null mutations in the ALPL gene associated with deleterious neurological outcome in an infant with

hypophosphatasia. *Bone* **55**, 150-157 (2013).

42. Namm, A., Arend, A. & Aunapuu, M. Expression of Pax2 protein during the formation of the central nervous system in human embryos. *Folia Morphol.* **73**, 272-278 (2014).

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

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

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

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

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

48. 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).

49. 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).

50. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

51. 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 males' probands vs all males' unaffected siblings using the sib transmission/disequilibrium test.

We conducted GWAS for all 597 male probands vs all 370 unaffected brothers using the sib transmission/disequilibrium test (sib-TDT). We observed no significant associations in this GWAS. The dotted line indicates the threshold for genome-wide significance ($P < 5.0 \times 10^{-8}$).

Fig. 2. Methods of GWAS according to each subgroup of the probands vs the unaffected brothers as controls without the brothers of the members of the subgroup being analysed in the present study.

We call GWAS according to each subgroup of the probands vs the unaffected brothers as controls without the brothers of the members of the subgroup as "Cluster-based GWAS". This panel shows the detailed methods of Cluster-based GWAS in the present study.

Fig. 3. Manhattan plots (a) and corresponding quantile-quantile plots (b) in GWAS for cluster-based males' probands and males' unaffected siblings who did not include corresponding probands by k-means algorithms with 15 clusters using Cochran-Armitage trend test.

We conducted GWAS according to each subgroup of the probands vs the unaffected brothers as controls without the brothers of the members of the subgroup being analysed (cluster-based GWAS) using the k-means with 15 clusters and the Cochran-Armitage trend test. Among 15 clusters, significant associations were observed in 14 clusters. In total, we identified significant associations in 93 chromosomal loci that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$. The genetic loci that were previously reported candidate genes for ASD and satisfied the genome-wide significance threshold are labelled. The dotted line indicates the threshold for genome-wide significance ($P < 5.0 \times 10^{-8}$).

Tables

Table 1. Number of genome-wide significant loci for each clustering algorithm and test method using the Omni2.5 dataset with MAF <0.01 deleted

Clustering algorithm		k-means							Affinity propagation
No. of clusters	1	2	3	4	5	10	15	20	36
Test method									
Sibling-based transmission disequilibrium test	0	-	-	-	-	-	-	-	-
λ value	1.025	-	-	-	-	-	-	-	-
Cochran-Armitage trend test	-	0	0	1	5	24	93	267	1,253
Mean λ value (min-max)	-	1.055 (1.048-1.061)	1.044 (1.033-1.058)	1.039 (1.035-1.044)	1.023 (1.015-1.030)	1.020 (1.005-1.031)	1.043 (1.018-1.065)	1.054 (1.022-1.093)	1.076 (1.032-1.093)
Fisher’s exact test	-	0	0	2	2	0	2	3	1
Mean λ value (min-max)	-	0.893 (0.885-0.900)	0.871 (0.854-0.893)	0.868 (0.862-0.875)	0.840 (0.828-0.847)	0.772 (0.734-0.806)	0.720 (0.681-0.778)	0.683 (0.615-0.738)	0.601 (0.474-0.708)

Table 2. The characteristics of each of 15 clusters by k-means in the Omni2.5 dataset

Cluster	Verbal score from ADI-R				Non-verbal score from ADI-R				RRB score from ADI-R				Social score from ADI-R				Treatment with vitamin B6 (%)
	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	
1 (n = 63)	8.0 (1.5)	8 (7-9)	4	11	7.0 (1.7)	7 (6-8)	1	10	8.6 (1.6)	8 (7-10)	6	12	18.3 (1.3)	18 (17-19)	15	21	61.9
2 (n = 41)	7.9 (1.5)	8 (7-9)	4	11	12.2 (1.5)	13 (11-14)	9	14	4.5 (1.3)	4 (4-6)	2	7	25.3 (1.2)	26 (25-26)	23	28	65.9
3 (n = 28)	5.1 (1.9)	5 (4.5-6)	0	8	4.7 (1.8)	5 (3-6)	1	8	4.2 (1.9)	4 (3-5)	1	9	9.5 (1.1)	9 (9-10)	8	12	53.6
4 (n = 48)	7.3 (1.8)	7 (6-8)	3	11	12.4 (1.2)	12.5 (12-13)	10	14	6.3 (1.4)	6 (6-7)	3	9	21.1 (1.4)	21 (20-22)	18	23	54.2
5 (n = 35)	7.1 (1.5)	7 (6-8)	3	10	8.4 (1.6)	8 (7-9)	6	13	6.0 (2.2)	6 (5-7)	1	12	12.3 (1.6)	12 (11-14)	9	15	62.9
6 (n = 50)	9.2 (1.3)	9 (8-10)	6	12	12.1 (1.5)	12.5 (11-13)	9	14	9.5 (1.3)	10 (8-10)	7	12	25.2 (1.3)	25 (24-26)	23	27	62.0
7 (n = 40)	6.0 (1.5)	6 (5-7)	3	10	10.9 (1.7)	11 (9.5-12)	8	14	5.6 (1.7)	6 (5-6)	3	10	16.5 (1.1)	16.5 (16-17)	14	19	67.5
8 (n = 29)	5.4 (1.8)	5 (4-7)	2	9	7.1 (1.8)	7 (6-8)	4	11	4.1 (1.4)	4 (3-5)	1	7	20.8 (1.5)	21 (20-21)	18	24	44.8
9 (n = 35)	9.3 (1.4)	9 (8-10)	6	12	6.9 (1.6)	7 (5-8)	4	9	9.6 (1.7)	10 (8-11)	6	12	23.1 (1.3)	23 (22-24)	21	27	60.0
10 (n = 61)	7.6 (1.7)	8 (6-9)	4	12	8.5 (1.3)	9 (8-9)	6	10	6.2 (1.5)	6 (6-7)	3	9	23.3 (1.6)	23 (22-24)	21	27	49.2
11 (n = 45)	9.2 (1.4)	9 (8-10)	6	12	12.7 (1.4)	13 (12-14)	9	14	7.6 (1.5)	7 (7-8)	5	12	28.2 (1.2)	28 (27-29)	26	30	75.6
12 (n = 29)	4.7 (1.5)	5 (4-6)	2	8	4.4 (1.7)	4 (4-6)	1	7	6.0 (1.9)	6 (5-7)	3	10	14.4 (1.5)	14 (13-15)	12	17	69.0
13 (n = 32)	8.9 (1.6)	9 (8-10)	5	11	3.6 (1.7)	4 (2-5)	0	6	7.4 (2.1)	7.5 (6-8)	3	12	12.1 (2.1)	13 (10.5-14)	8	15	59.4
14 (n = 34)	8.1 (1.5)	8 (7-9)	5	12	6.2 (1.7)	6 (5-8)	3	10	3.6 (1.2)	3.5 (3-4)	2	6	16.5 (1.5)	16 (15-18)	14	20	41.2
15 (n = 27)	9.5 (1.6)	10 (9-11)	5	12	11.0 (1.4)	11 (10-12)	9	14	10.5 (1.2)	10 (10-12)	8	12	20.2 (1.6)	20 (19-22)	17	22	66.7

ADI-R, Autism Diagnostic Interview-Revised; RRB, repetitive and restricted behaviours; SD, standard deviation; p, percentile.

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

Cluster no.	ID	Chr	hg19	Minor/ Major	MAF (%)	RR	95% CI	P	GENESYMBOL	Function
1	rs55928054	13	113425260	A/G	4.18	5.46	2.87-10.41	1.75E-08	ATP11A	Intronic
1	rs75194052	14	88001041	A/G	1.83	15.1	4.89-46.68	2.09E-08	LINC01148, LINC02330	Intergenic
2	rs77656540	5	5496528	A/G	1.15	29.79	6.29-141.03	2.81E-11	ICE1, LINC02145	Intergenic
2	rs74733575	8	77518812	A/G	3.21	7.86	3.71-16.64	2.67E-10	LINC01111, ZFHX4-AS1	Intergenic
2	rs16875609	5	5470026	G/A	1.28	19.86	5.24-75.32	5.19E-10	ICE1	Intronic
2	rs3806873	5	5462607	G/A	1.28	19.86	5.24-75.32	5.19E-10	ICE1	Exonic
2	rs16875597	5	5460434	A/G	1.29	19.8	5.22-75.11	5.52E-10	ICE1	Intronic
2	rs3806874	5	5462620	A/G	1.29	19.8	5.22-75.11	5.52E-10	ICE1	Exonic
2	rs4702269	5	5447607	G/A	1.17	17.15	4.37-67.25	2.39E-08	ICE1	Exonic
2	rs74752669	3	88127756	A/C	1.16	16.98	4.33-66.6	2.89E-08	CGGBP1, ZNF654	Intronic
3	rs76626536	16	80471427	A/G	4.64	6.43	3.4-12.16	2.04E-10	LOC102724084	ncRNA_Intronic
3	rs77884662	11	13911075	A/C	1.03	21.49	5.27-87.6	9.64E-10	FAR1, LOC101928132	Intergenic
3	rs62528479	8	104219144	C/A	2.96	8.29	3.75-18.3	1.01E-09	BAALC	Intronic
3	rs74384601	11	118449313	G/A	1.03	21.43	5.26-87.36	1.02E-09	ARCN1	Intronic
3	rs78513244	1	2360342	A/G	3.11	7.67	3.52-16.74	3.66E-09	PEX10, PLCH2	Intergenic
3	rs11023007	11	13928677	A/G	1.16	16.07	4.44-58.17	1.42E-08	FAR1, LOC101928132	Intergenic
3	rs16895575	5	26394185	G/A	1.16	16.07	4.44-58.17	1.42E-08	LINC02211, CDH9	Intergenic
3	rs4707805	6	94294685	G/A	1.16	16.03	4.43-58.01	1.50E-08	EPHA7, TSG1	Intergenic
3	rs10581	1	202910318	A/G	2.75	8.43	3.66-19.41	2.38E-08	ADIPOR1	UTR3
3	rs11106191	12	78060912	A/G	1.67	11.05	3.84-31.77	3.25E-08	E2F7, NAV3	Intergenic
3	rs58365105	8	110971624	A/G	3.08	7.74	3.54-16.88	4.65E-08	SYBU, KCNV1	Intergenic
4	rs74785766	20	18706805	A/G	9.57	3.36	2.17-5.18	1.96E-08	DTD1	Intronic
4	rs112633050	20	18718066	A/G	9.62	3.34	2.16-5.15	2.29E-08	DTD1	Intronic
4	rs10882708	10	97764915	C/A	1.64	11.63	3.88-34.84	2.69E-08	ENTPD1-AS1	ncRNA_Intronic
4	rs117112406	10	24240962	A/G	1.01	21.75	4.45-106.23	3.69E-08	KIAA1217	Intronic
4	rs118085556	10	102407984	A/G	1.02	21.62	4.43-105.62	4.07E-08	HIF1AN, PAX2	Intergenic
6	rs76324396	1	21841196	A/G	1.27	16.1	4.23-61.25	3.32E-08	ALPL	Intronic
6	rs9621415	22	32629026	A/G	1.27	16.1	4.23-61.25	3.32E-08	SLC5A4	Intronic
7	rs75262399	2	8920178	C/A	1.14	17.75	4.53-69.61	1.29E-08	KIDINS220	Intronic
7	rs77055713	2	8939140	C/G	1.14	17.7	4.51-69.41	1.36E-08	KIDINS220	Intronic
7	rs1782772	10	126147677	A/G	1.89	10.17	3.79-27.31	1.47E-08	NKX1-2, LHPP	Intergenic
7	rs77507687	2	26939229	G/A	1.89	10.17	3.79-27.31	1.47E-08	KCNK3	Intronic
7	rs11067544	12	115786013	A/G	5.82	4.73	2.7-8.29	1.84E-08	TBX3, MED13L	Intergenic
8	rs13437654	7	12360883	A/G	2.11	12.07	4.7-30.98	9.13E-10	TMEM106B, VWDE	Intergenic
8	rs10272812	7	12322548	A/G	2.11	12.07	4.7-30.98	9.13E-10	TMEM106B, VWDE	Intergenic
8	rs7788409	7	12300659	G/A	2.12	12.03	4.69-30.89	9.72E-10	TMEM106B,	Intergenic

									VWDE	
8	rs10227871	7	12366459	A/G	2.12	12.03	4.69-30.89	9.72E-10	TMEM106B, VWDE	Intergenic
8	rs10247702	7	12323757	C/A	2.12	12.03	4.69-30.89	9.72E-10	TMEM106B, VWDE	Intergenic
8	rs60756657	20	18569645	G/A	1.46	14.44	4.54-45.89	3.06E-09	DTD1	Intronic
8	rs11905972	20	18514464	G/A	1.46	14.4	4.53-45.76	3.25E-09	SEC23B	Intronic
8	rs117859793	8	106870076	A/G	1.06	20.11	4.93-82.07	3.67E-09	ZFPM2-AS1	ncRNA_Intronic
8	rs1079506	7	12342641	A/G	2.25	10.7	4.29-26.68	3.83E-09	TMEM106B, VWDE	Intergenic
8	rs17569054	12	58381546	A/T	3.05	7.71	3.49-17.06	5.25E-09	ATP23, LINC02403	Intergenic
8	rs72582233	7	12240705	G/A	2.37	9.66	3.96-23.52	1.24E-08	THSD7A, TMEM106B	Intergenic
8	rs72582242	7	12249139	G/A	2.37	9.66	3.96-23.52	1.24E-08	THSD7A, TMEM106B	Intergenic
8	rs78193076	7	12319434	A/C	2.37	9.66	3.96-23.52	1.24E-08	TMEM106B, VWDE	Intergenic
8	rs28459566	7	12260090	G/A	2.39	9.57	3.93-23.32	1.47E-08	TMEM106B	Intronic
8	rs28550800	22	46282759	A/C	9.1	3.95	2.45-6.36	2.41E-08	ATXN10, WNT7B	Intergenic
8	rs10251962	7	12398651	G/A	2.11	9.39	3.63-24.29	2.85E-08	VWDE	Intronic
8	rs10270435	7	12418602	G/A	2.11	9.39	3.63-24.29	2.85E-08	VWDE	Intronic
8	rs77271688	7	12461020	G/A	2.12	9.36	3.62-24.22	3.01E-08	VWDE, LOC102725191	Intergenic
8	rs10231277	7	12320020	C/A	2.52	8.73	3.65-20.84	4.23E-08	TMEM106B, VWDE	Intergenic
8	rs73807820	4	37572186	A/G	1.19	15.09	4.16-54.65	4.46E-08	C4orf19	Intronic
9	rs12322120	12	24245580	A/C	27.98	2.35	1.86-2.97	1.02E-09	SOX5	Intronic
9	rs11831634	12	24232157	A/G	26.63	2.35	1.83-3.01	1.66E-09	SOX5	Intronic
9	rs115282974	3	85649418	G/A	2.08	10	3.87-25.82	6.09E-09	CADM2	Intronic
10	rs72997986	11	115396003	A/G	4.46	5.03	2.69-9.4	4.35E-08	CADM1, LOC101928985	Intergenic
10	rs73000027	11	115433485	A/G	5.35	4.43	2.5-7.83	4.49E-08	CADM1, LOC101928985	Intergenic
11	rs74036338	16	84633030	G/A	1.52	15.56	4.78-50.62	9.34E-10	COTL1	Intronic
11	rs72676911	1	57881845	G/A	2.15	9	3.56-22.72	2.03E-08	DAB1	Intronic
11	rs9956246	18	54960100	G/A	2.52	7.82	3.35-18.28	2.18E-08	BOD1L2, ST8SIA3	Intergenic
12	rs7724569	5	9457341	A/G	2.34	9.79	4.02-23.86	1.30E-09	SEMA5A	Intronic
12	rs76094962	11	28485359	G/A	2.49	9.19	3.86-21.93	2.43E-09	METTL15, MIR8068	Intergenic
12	rs76015064	7	51975132	A/C	1.04	20.4	5-83.24	2.78E-09	COBL, POM121L12	Intergenic
12	rs77285841	6	135142226	A/G	1.56	12.24	4.08-36.76	1.57E-08	LOC101928304, ALDH8A1	Intergenic
12	rs60004245	6	13432162	A/G	1.56	12.24	4.08-36.76	1.57E-08	GFOD1	Intronic
12	rs57510388	6	6788339	G/A	3.65	6.8	3.29-14.04	1.66E-08	LY86, RREB1	Intergenic
12	rs77201757	12	108614952	A/G	1.17	15.3	4.22-55.44	3.48E-08	WSCD2	Intronic
12	rs114018272	3	177909410	G/A	1.17	15.3	4.22-55.44	3.48E-08	LINC02015, LINC01014	Intergenic
13	rs78771643	4	48729665	T/A	1.16	22.31	5.71-87.12	1.08E-10	FRYL	Intronic
13	rs114358580	8	75211161	A/G	2.71	8.34	3.65-19.05	3.06E-09	JPH1	Intronic
13	rs1993471	15	61040025	A/C	1.79	11.19	4.05-30.9	6.50E-09	RORA	Intronic
13	rs9651906	12	116245161	A/G	2.95	7.19	3.24-15.96	2.53E-08	TBX3, MED13L	Intergenic
13	rs7312889	12	116245839	G/A	2.95	7.19	3.24-15.96	2.53E-08	TBX3, MED13L	Intergenic

13	rs7304809	12	116244393	G/A	2.95	7.19	3.24-15.96	2.53E-08	TBX3, MED13L	Intergenic
13	rs7140271	14	101483629	C/A	4.5	5.82	3.04-11.14	2.96E-08	MEG8, MIR379	Intergenic
14	rs16849132	1	201575549	G/A	1.18	20.41	5.22-79.79	7.78E-10	RPS10P7, NAV1	Intergenic
14	rs117486297	11	23647862	G/A	1.57	14.29	4.66-43.8	1.06E-09	MIR8054, LUZP2	Intergenic
15	rs117647850	8	79156756	A/G	2.86	11.02	4.99-24.33	4.11E-12	LOC102724874, PKIA	Intergenic
15	rs116747981	3	168859282	A/G	3.38	8.29	3.95-17.37	7.88E-11	MECOM	Intronic
15	rs6808748	3	122672821	A/G	1.04	22.1	5.43-90.01	5.28E-10	SEMA5B	Intronic
15	rs1930850	13	71023752	G/A	1.18	17.85	4.95-64.4	1.78E-09	ATXN8OS, LINC00348	Intergenic
15	rs12939556	17	60325665	G/A	8.46	4.32	2.64-7.06	2.32E-09	MED13, TBC1D3P2	Intergenic
15	rs117925398	8	78908718	C/A	1.83	13.15	4.79-36.12	2.53E-09	LOC102724874, PKIA	Intergenic
15	rs79758193	13	69880137	A/G	1.56	13.26	4.43-39.72	3.14E-09	LINC00383	ncRNA_Intronic
15	rs12936559	17	60325222	A/G	8.59	4.23	2.59-6.91	3.58E-09	MED13, TBC1D3P2	Intergenic
15	rs78052401	4	159836336	A/G	1.17	16.57	4.58-59.93	7.90E-09	C4orf45	Exonic
15	rs115132435	5	172928190	A/G	1.17	16.57	4.58-59.93	7.90E-09	MIR8056, LOC285593	Intergenic
15	rs2325297	13	71028791	G/A	1.17	16.57	4.58-59.93	7.90E-09	ATXN8OS, LINC00348	Intergenic
15	rs9564696	13	71031795	A/C	1.17	16.57	4.58-59.93	7.90E-09	ATXN8OS, LINC00348	Intergenic
15	rs79566457	4	159706995	G/A	1.17	16.53	4.57-59.76	8.33E-09	FNIP2	Intronic
15	rs77930743	15	74288796	A/G	3.64	7.37	3.58-15.16	8.97E-09	PML	Intronic
15	rs76964192	15	74300156	A/G	3.64	7.37	3.58-15.16	8.97E-09	PML	Intronic

Association tests were carried out using Cochran-Armitage test.

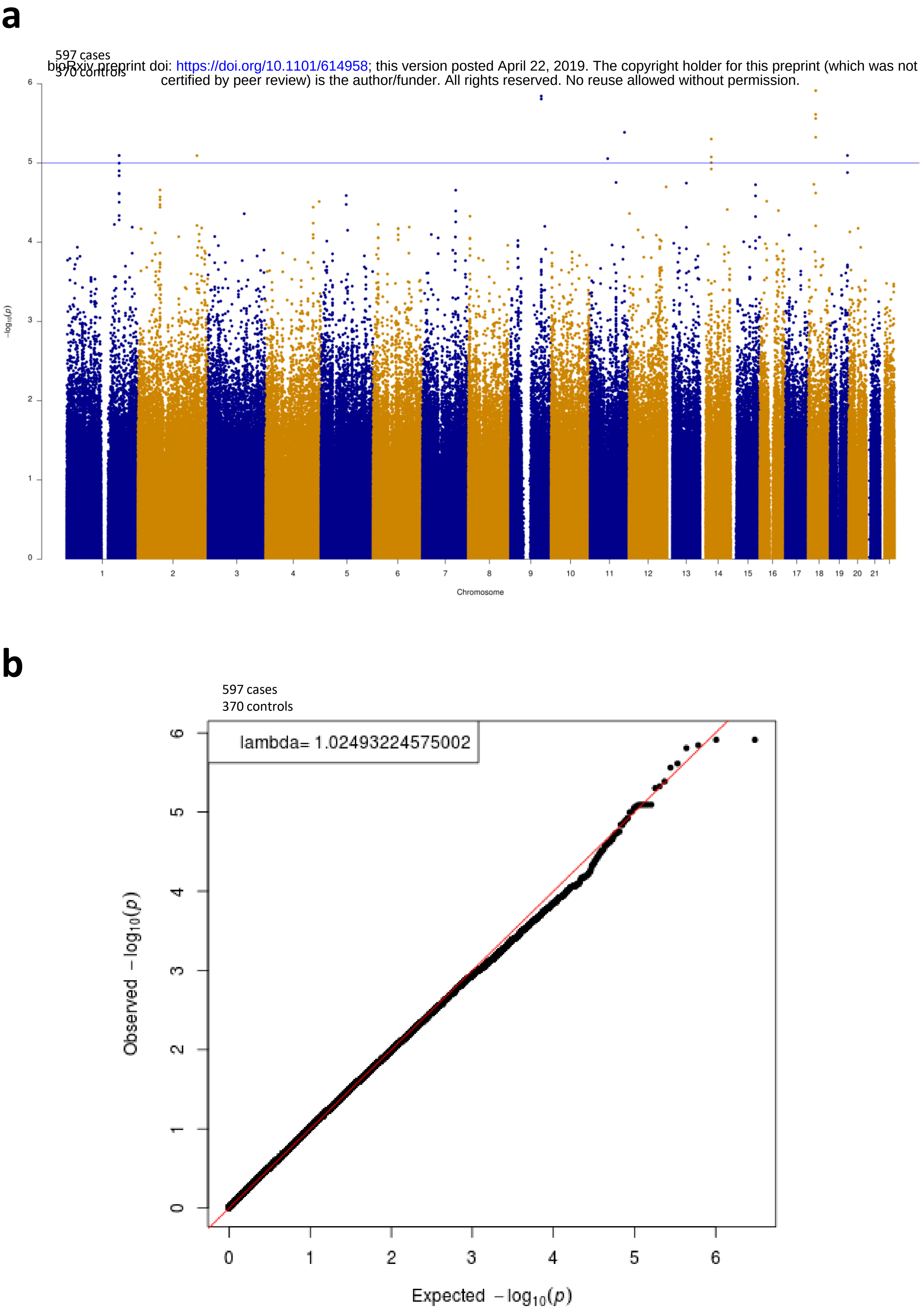


Fig. 1. Manhattan plots (**a**) and corresponding quantile-quantile plots (**b**) in GWAS for all males' probands vs all males' unaffected siblings using the sib transmission/disequilibrium test.

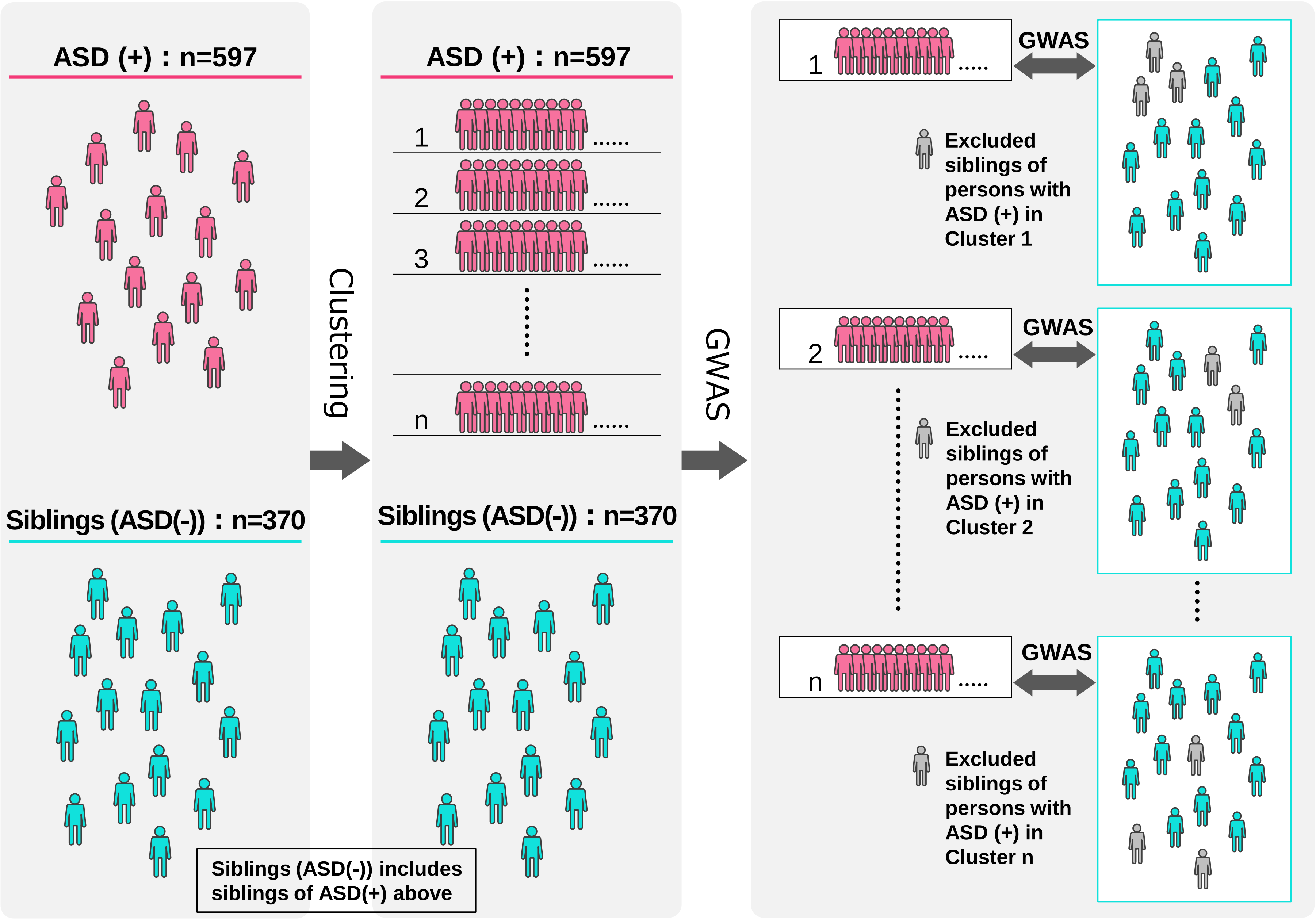
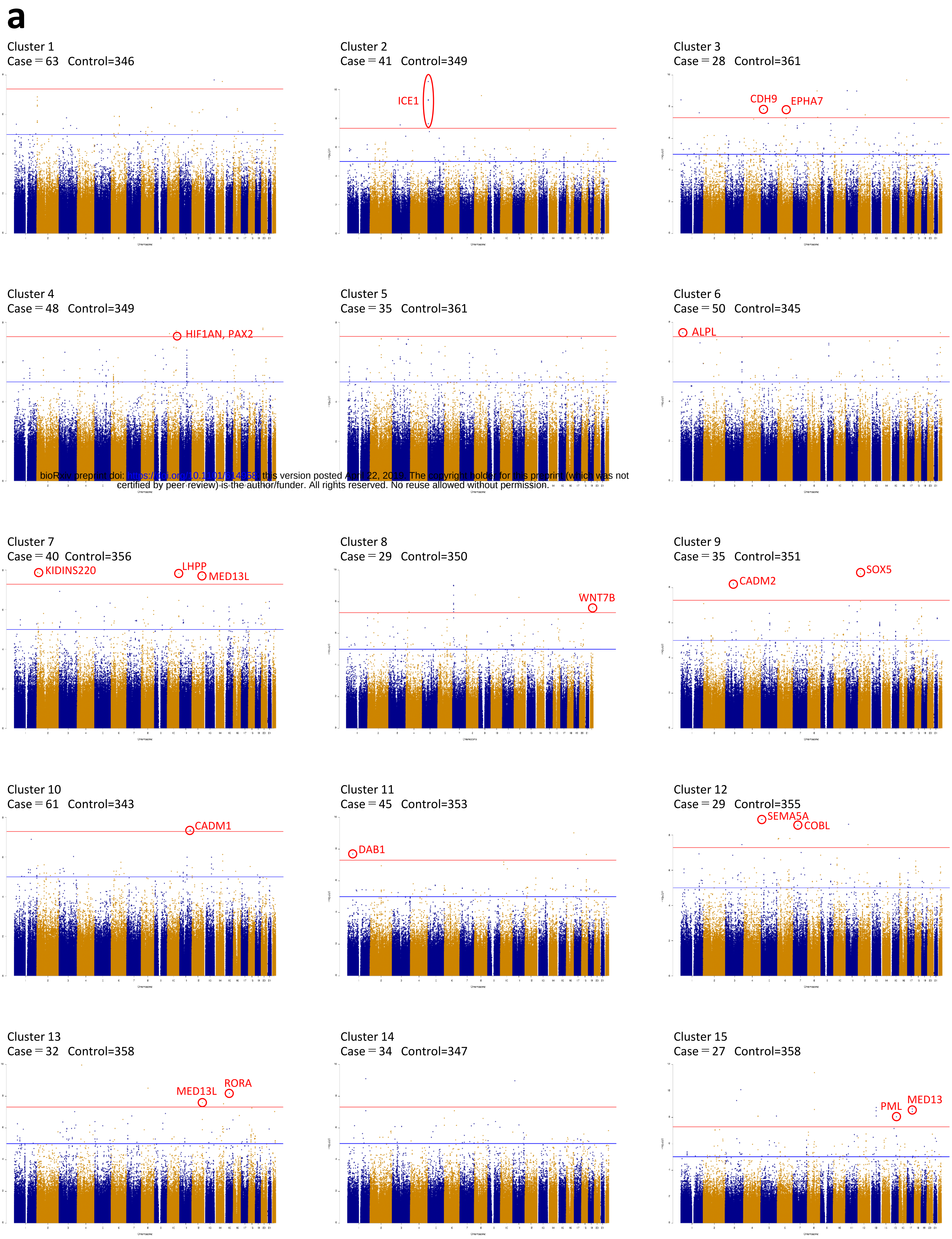

ASD (+) : n=597
Siblings (ASD(-)) : n=370
Clustering
ASD (+) : n=597
1
2
3
n
Siblings (ASD(-)) : n=370
Siblings (ASD(-)) includes siblings of ASD(+) above
GWAS
1
GWAS
Excluded siblings of persons with ASD (+) in Cluster 1
2
GWAS
Excluded siblings of persons with ASD (+) in Cluster 2
n
GWAS
Excluded siblings of persons with ASD (+) in Cluster n

Fig. 2. GWAS according to each subgroup of the probands vs the unaffected brothers as controls without the brothers of the members of the subgroup being analysed in the present study.


a
Cluster 1
Case = 63 Control=346
Cluster 2
Case = 41 Control=349
ICE1
Cluster 3
Case = 28 Control=361
CDH9
EPHA7
Cluster 4
Case = 48 Control=349
HIF1AN, PAX2
Cluster 5
Case = 35 Control=361
Cluster 6
Case = 50 Control=345
ALPL
Cluster 7
Case = 40 Control=356
KIDINS220
LHPP
MED13L
Cluster 8
Case = 29 Control=350
WNT7B
Cluster 9
Case = 35 Control=351
CADM2
SOX5
Cluster 10
Case = 61 Control=343
CADM1
Cluster 11
Case = 45 Control=353
DAB1
Cluster 12
Case = 29 Control=355
SEMA5A
COBL
Cluster 13
Case = 32 Control=358
MED13L
RORA
Cluster 14
Case = 34 Control=347
Cluster 15
Case = 27 Control=358
PML
MED13

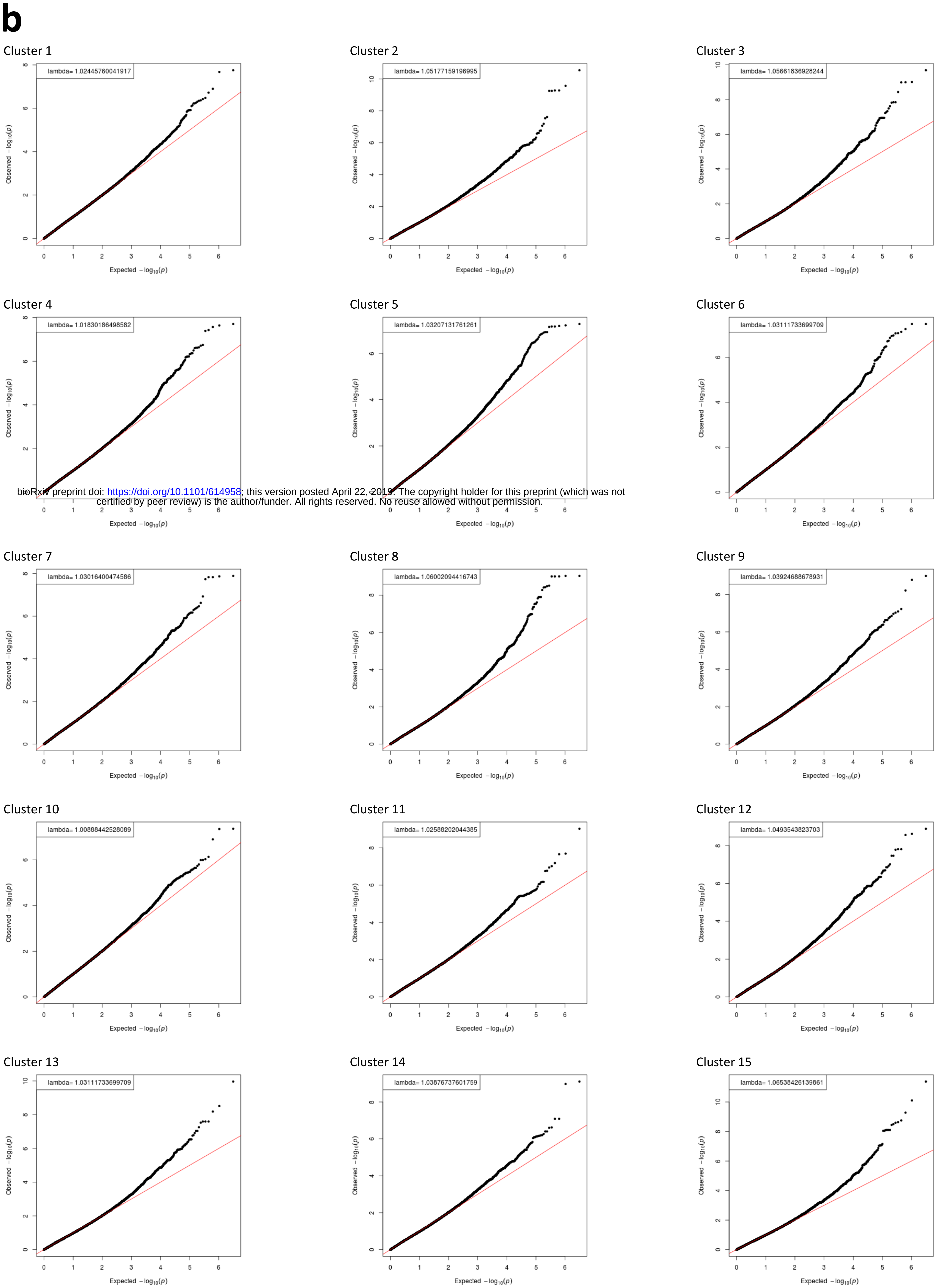


Fig. 3. Manhattan plots (**a**) and corresponding quantile-quantile plots (**b**) in GWAS for cluster-based males' probands and males' unaffected siblings who did not include corresponding probands by k-means algorithms with 15 clusters using Cochran-Armitage trend test.