Published in final edited form as:

Genet Epidemiol. 2017 May ; 41(4): 332–340. doi:10.1002/gepi.22040.

On the association analysis of genome-sequencing data: A spatial clustering approach for partitioning the entire genome into nonoverlapping windows

Heide Loehlein Fier^{1,2}, Dmitry Prokopenko³, Julian Hecker², Michael H. Cho³, Edwin K. Silverman³, Scott T. Weiss³, Rudolph E. Tanzi⁴, and Christoph Lange^{1,3}

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

²Working Group of Genomic Mathematics, University of Bonn, Bonn, Germany

³Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

⁴Genetics and Aging Research Unit, MassGeneral Institute for Neurodegenerative Disease, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts, United States of America

Abstract

For the association analysis of whole-genome sequencing (WGS) studies, we propose an efficient and fast spatial-clustering algorithm. Compared to existing analysis approaches for WGS data, that define the tested regions either by sliding or consecutive windows of fixed sizes along variants, a meaningful grouping of nearby variants into consecutive regions has the advantage that, compared to sliding window approaches, the number of tested regions is likely to be smaller. In comparison to consecutive, fixed-window approaches, our approach is likely to group nearby variants together. Given existing biological evidence that disease-associated mutations tend to physically cluster in specific regions along the chromosome, the identification of meaningful groups of nearby located variants could thus lead to a potential power gain for association analysis. Our algorithm defines consecutive genomic regions based on the physical positions of the variants, assuming an inhomogeneous Poisson process and groups together nearby variants. As parameters are estimated locally, the algorithm takes the differing variant density along the chromosome into account and provides locally optimal partitioning of variants into consecutive regions. An R-implementation of the algorithm is provided. We discuss the theoretical advances of our algorithm compared to existing, window-based approaches and show the performance and advantage of our introduced algorithm in a simulation study and by an application to Alzheimer's disease WGS data. Our analysis identifies a region in the ITGB3 gene that potentially harbors disease susceptibility loci for Alzheimer's disease. The region-based association signal of ITGB3 replicates in an independent data set and achieves formally genome-wide significance.

Correspondence: Heide Loehlein Fier, Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. heide.fier@gmail.com.

Software Implementation: An implementation of the algorithm in R is available at: https://github.com/heidefier/cluster_wgs_data.

Keywords

WGS data; clustering; genetic association analysis

1 INTRODUCTION

Whole-genome sequencing (WGS) data have become increasingly available for genetic association studies. Large-scale reference sequencing data sets such as the 1000 Genomes project have shown that most of these variants exhibit low minor allele frequencies (1000 Genomes Project Consortium, 2015; Sudmant et al., 2015), so that single marker association testing is likely to be underpowered (e.g., Ladouceur, Dastani, Aulchenko, Greenwood, & Richards, 2012; Lee, Abecasis, Boehnke, & Lin, 2014; Xu et al., 2014). The statistical methodology described and used for whole exome sequencing studies mainly has focused on single-marker tests and gene-based tests, where rare variants within a gene are pooled for association testing (e.g., He et al., 2014; Lange et al., 2014; Lohmueller et al., 2013). For whole genome sequencing data, these gene and gene exome borders vanish because not only the exomes of genes are sequenced, but also noncoding gene regions and intergenic regions. Recent works have shown that the vast majority of GWAS reported DSLs (disease susceptibility loci) are located in nonprotein coding regions, emphasizing the importance of a deep sequencing of intronic and intergenic regions (Hindorff et al., 2009; Kumar, 2013; Pennisi, 2011). As a result, restricting the association analysis for whole genome sequencing data to genes or gene exomes would result in a loss of potentially relevant information and would make the whole genome sequencing approach contradictory.

Most of the current methodology on grouping algorithms for whole genome data has focused on window-based approaches (e.g., Morrison et al., 2013, Panoutsopoulou, Tachmazidou, & Zeggini, 2013, Yazdani, Yazdani, & Boerwinkle, 2015). The intuition behind window-based approaches is similar to gene-based tests, namely that variants that lie within defined windows are jointly tested for association. There are several window-based approaches that one could think of, such as sliding windows, partially overlapping windows, or consecutive windows. Although these window-based approaches serve the purpose of reducing the number of statistical tests and therefore address the multiple testing problem, they are usually very sensitive toward the chosen window-size. Thus, the same genetic variants that show a significant association within a 10 kb sized window might not show a significant association in two consecutive windows of 5 kb size, if the association signal is split up by two consecutive windows. Although sliding windows and partially overlapping windows might address this problem to a certain degree, they have the disadvantage that they introduce a stronger dependence structure between the test statistics compared to single variant tests and reduce the number of statistical tests less effectively compared to consecutive window-based approaches.

Consequently, one challenge for whole genome sequencing data association analysis is to find meaningful grouping algorithms for rare variants. In this communication, we introduce an algorithm for sequence data that groups variants along the chromosome into consecutive, nonoverlapping windows based on their spatial distribution.

From a biological point of view, there are several arguments, why variants in a distinct genetic region could be part of the same disease mechanism: First, variants within the same protein functional domain are likely to be located in close proximity in the DNA sequence and could have similar impact on disease risk (Krebs, Goldstein, & Kilpatrick, 2014). Second, gene regulatory elements tend to cluster in certain genomic locations, such as the promoter regions (Raab & Kamakaka, 2010). Finally, most recently it has been shown that local recombination rates affect the accumulation of disease-associated mutations along the chromosome (Hussin et al., 2015). Given these findings, the identification of groups of variants that are closely located to each other (i.e., cluster) and similarly affect the disease susceptibility is essential for WGS analysis, as suitable sets of variants are required for region-based association analyses (Fier et al., 2012, Lee et al., 2014, Lin & Tang, 2011, Wu et al., 2011).

There exist a number of methods for the optimal grouping of one-dimensional spatial data (Evans 1977, Fisher, 1958, Jenks, 1977, Jenks & Caspall, 1971, Wang & Song, 2011). Most of these methods try to arrange the data in a way where the variance of the observations is minimized within intervals and maximized between intervals. However, all of those described clustering algorithms for one-dimensional spatial data have in common that they are computationally exhaustive with large data sets and even more important require a default number of intervals for optimization. Because the underlying spatial structure of large sequencing data sets is unknown, an a priori determination of a meaningful number of clusters is difficult to achieve.

We overcome these shortcomings of existing clustering algorithms by introducing an algorithm that wriggles along the chromosome and iteratively estimates the mutation intensity using a nonhomogeneous Poisson point process. It then divides the chromosome into regions at those points where consecutive variants share larger physical distances to each other than expected under the estimated exponential distributions.

The approach is intuitive and computationally efficient, that is, it slices one chromosome containing several million variants into nonoverlapping, consecutive windows within less than 5 min (Intel Core i5 2.67GHz, 8Gb RAM). We apply the algorithm to a whole genome sequencing family study for Alzheimer's disease to show the practicability and advances of our approach.

2 METHODOLOGY

A genetic region (e.g., a chromosome or gene) has been sequenced (e.g., WGS) and the physical positions of a total of N variants have been sorted and recorded in an ascending order. We denote the physical position of the i th variant as p_i with $i = 1, \dots, N$.

Under the assumption that the occurrence of genetic variations along the chromosome is not entirely random but varies spatially, the genetic region can be described by an inhomogeneous Poisson process, where the physical length of the chromosomal region investigated can be interpreted as the total time line investigated, the variants can be interpreted as process occurrences, and the spatial distance between genetic variants

represent the waiting times between process occurrences. Joyce and Tavaré (1995) have shown theoretically that, for a sufficiently large sample, allele counts approximately follow a Poisson process whose rate is determined by the overall mutation rate. For an inhomogeneous stochastic process, the process intensity λ varies along the time and given recent findings¹⁷, it is reasonable to assume for our methodological framework that the spatial distribution of variants also varies for different regions of the chromosome.

In order to estimate spatially varying exponential rates for our genetic region, we apply a sliding window approach. We denote w_j to be the j th sliding window with $j = 1, \dots, M$ containing K consecutively located genetic variants (e.g., $K = 100$). M represents the total number of sliding windows and depends on K (which is user-defined) so that $M \frac{N}{K}$.

Note that K should be chosen sufficiently large so that the parameters of a nonnormal distribution can be approximated but small enough so that varying spatial denseness along the chromosome can be detected. Given these requirements, we recommend to chose 50 K 200. By definition, the sliding step of the windows is half of the window size, so that except for the start and the end of the chromosomal region, all variants are covered exactly twice. Of course, it would also be possible to select a different sliding window step (e.g., 1 variant, 5 variants, 10 variants ...). However, it is important to note that the number of identified consecutive regions will grow with the number of tested windows. Similarly, a smaller value for K will result in a larger number of tested windows. As a consequence, if a greater number of the more windows with fewer variants are tested, the more likely it is that the algorithm will be more likely to will detect more breaks due to outliers, because as the fitting of the exponential distribution becomes more sensitive toward outliers. We

recommend to set 50 K 200 and the sliding step to $\frac{K}{2}$ and discuss later, how identified, consecutive regions can be further filtered.

The physical distances between the ordered positions of the sequenced variants within window w_l can be described by the distance vector d_l with $l = 1, \dots, M$.

In the Poisson process framework, the waiting times between process occurrences are exponentially distributed and therefore we assume that the components of d_l can be interpreted as independent draws from an exponential distribution with common exponential rate λ_l . Under this assumption, the exponential rate λ_l can be estimated using the mean-

based estimator $\lambda_{l,\text{mean}} = \frac{1}{\bar{d}_l}$ or the median-based estimator $\lambda_{l,\text{median}} = \frac{1}{\bar{d}_l} \ln 2$. Divergences in the estimation results of $\lambda_{l,\text{mean}}$ and $\lambda_{l,\text{median}}$ pinpoint to a skewed distribution of the components of d_l . Because the estimate of $\lambda_{l,\text{mean}}$ is sensitive toward outliers, a single large distance in window d_l will produce a a estimation result of $\lambda_{l,\text{mean}}$, which is larger than the estimation result of $\lambda_{l,\text{median}}$. The same holds the other way around, if window d_l contains several small distances, the estimation result of $\lambda_{l,\text{median}}$ will be larger than the estimation result of $\lambda_{l,\text{mean}}$. Because we aim to identify clusters of nearby located variants however, we remove only those components of d_l that cause the estimation result of $\lambda_{l,\text{mean}}$ to be larger than the estimation result of $\lambda_{l,\text{median}}$. We therefore iteratively remove the largest distances out of

window d_l until $\lambda_{l,\text{mean}} = \lambda_{l,\text{median}}$. It is intuitive to see that depending on the skewness of the distances within one window, there might be no outlying distances that are removed within one window or, if the distance distribution contains several long outlier distances, one or more distances are being removed. The removed distances from the single windows then define the spatial breaks within the data and variants that lie between these spatial breaks are tested together for association. As a result, we receive consecutive, nonoverlapping regions that can be tested for association.

To summarize, the algorithm works as follows:

```

1. Divide the given chromosomal region into  $M$  sliding windows with  $K$  variants per window and a sliding step of  $K/2$ 


---


2. for each sliding window  $w_l$  do


---


3. estimate  $\frac{1}{\bar{d}_l}$  and  $\frac{\ln 2}{\tilde{d}_l}$ 


---


4. if  $\left(\frac{1}{\bar{d}_l} < \frac{\ln 2}{\tilde{d}_l}\right)$  do


---


5. return(NULL)


---


6. end if


---


7. else do


---


8. define index  $t \leftarrow 0$ 


---


9. while  $\left(\frac{1}{\bar{d}_l} > \frac{\ln 2}{\tilde{d}_l}\right)$  do


---


10. sort distances  $d_l$  in an decreasing order


---


11. remove  $d_l[t]$ 


---


12. determine the chromosomal position of the removed gap:  $p[d_l[t]]$ 


---


13.  $t \leftarrow t + 1$ 


---


14. end while


---


15. return  $\{p[d_l[t]]\}$ 


---


16. end else


---


17. end for


---


18. sort  $\{p[d_l[t]]\}$  in an increasing order and group variants according to spatial breaks


---



```

We provide a ready to use R-implementation of the algorithm.

Depending on the data set analyzed and the asymptotic properties of the test statistic used for association analysis, the identified regions might be filtered in a last step, for example, only regions containing at least a certain number of variants are tested for association or very large regions might be split up in subregions of equal sizes.

Due to the likely presence of linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs) in one region, it is important to use a suitable association test that accounts for LD between the SNPs, for example, for family data the region-based rare variant family-based association test (RV-FBAT) (De, Yip, Ionita-Laza, & Laird, 2013), and for case-control data Sequence Kernel Association Test (SKAT) (Ionita-Laza, Lee, Makarov, Buxbaum, & Lin, 2013; Wu et al., 2011).

3 SIMULATION STUDY

We performed extensive simulations to assess the efficiency of our grouping strategy and to compare it with other possible testing strategies. We selected SKAT-O (Lee et al., 2012) as the association test statistic and modified the SKAT-O simulation workflow. We generated 10,000 haplotypes for a 1 Mb region by using the coalescent model implemented in COSI (Schaffner et al., 2005). We used the best-fit setting that follows the LD patterns and population history for European population. We considered only rare variants with a MAF 5% and removed singletons. We then created genotypes by randomly matching two haplotypes.

For the null scenario, we extract 1,000 cases and 1,000 controls by applying the following model, where the parameter α_0 is selected so that we obtain a prevalence of 1%, $X_1 \sim (0, 1)$ is a continuous covariate and $X_2 \sim \text{Bernoulli}(0.5)$ is a binary covariate:

$$\text{Logit}(P(y=1)) = \alpha_0 + 0.5X_1 + 0.5X_2.$$

In order to evaluate the power of the association test and specificity, we considered two scenarios. In the first scenario which we refer to as the “scenario without gaps,” we chose for every replication one region with the 10 most densely located consecutive rare variants (MAF 3%) along the simulated region to be causative. In the second scenario, the “scenario with gaps,” we additionally remove variants around the 10 most densely located causative variants. The radius for removing variants was twice the maximum distance between all variants in the region. This was done to show the strength of our grouping strategy in a spatially clustered setting. The binary phenotypes for 2,000 individuals were extracted from:

$$\text{Logit}(P(y=1)) = \alpha_0 + 0.5X_1 + 0.5X_2 + \sum_{i=1}^{10} \beta_i G_i^c,$$

where X_1 and X_2 are specified as above, G_i^c are the genotypes of the 10 causative variants

and $|\beta_i| = \frac{\ln 5}{4} |\log_{10} \text{MAF}_i|$.

For every replication, we divided the 1 MB region into smaller regions with two different methods: a standard consecutive approach, where always 30 consecutive variants define one region, and our proposed spatial clustering strategy. We used a sliding window step of 100 variants and an overlap of 50 variants as parameters for our clustering algorithm.

We then computed the association P values with SKAT-O for all windows, applied Bonferroni correction, and evaluated the most significant window. In Table 1, the simulation results in terms of statistical power and locus specificity are shown for all scenarios. The first power measure represents the fraction of region-based P values of the most significant windows over all replications that were smaller than the given alpha-level. This is basically the statistical power of the association test to detect the causative signal in the region. The second measure is the specificity of the most significant regions identified, that is, how many regions from the first measure contained at least one causative variant. We denote the power of the consecutive window approach as “power.con” and the specificity of the consecutive window approach as “spec.con.” Similarly, we refer to the power and specificity of our cluster approach as “power.cluster” and “spec.cluster.” Both testing strategies maintained the selected alpha-levels in the null scenario. All simulation results in Table 1 are under the alternative hypothesis and are based on 1,000 replications.

In the scenario without gaps, the power and specificity of the consecutive approach was slightly higher compared to our cluster approach. This observation was expected, since we simulated no spatial breaks in the data that were related to the causative SNPs. In the scenario with gaps, the power and specificity gain of our approach compared to the consecutive variants testing strategy is clearly visible and amounts to more than 100% for some alpha levels.

4 APPLICATION TO ALZHEIMER’S WGS DATA

We applied our outlined approach to whole genome sequencing data from the NIA (National Institute of Aging) study of the Alzheimer’s Disease Sequencing Project (ADSP). In order to compare its performance to a standard approach, we also applied a standard analysis approach in which always 50 consecutive variants were tested for association.

Our data set that was obtained through dbGaP contained whole genome sequence data from 127 families, containing between two and five offspring. Within the nuclear families, there was at least one confirmed affected and one confirmed unaffected family member. After checking for Hardy-Weinberg equilibrium (cut-off P -value for the test on Hardy-Weinberg equilibrium: 10^{-05}), the data set contained 27.8 million variants.

We further filtered the variants and kept only those variants that occurred in at least two independent families, thus removing variants that were private to individuals and/or to single families. After this step, 17.8 million variants were left for association analysis.

The application of our clustering approach (with $K = 100$ variants per sliding window) to the filtered WGS data of the NIA study identified 431,904 regions. Due to the small sample size and the dominance of variants with small allele frequencies, we decided to keep all regions that contained at least five variants for association analysis. This limited the number of

regions to 356,415. Table 2 shows descriptive statistics regarding the distribution of the number of variants in the clustered regions. On average, a region consisted out of 49.5 variants and was 6,550 base pairs long. The largest identified region spanned over 156,804 base pairs and contained 902 variants. The identified regions of our proposed clustering approach were tested for association using the region-based RV-FBAT approach for rare variants (De et al., 2013).

For comparison, the consecutive testing approach with a window size of 50 variants resulted in 356,366 regions. These regions were also tested for association using the region-based RV-FBAT. For the consecutive approach, the mean region/window size was about 7,686 base pairs. This is roughly 1 kb larger compared to the regions in our clustering approach.

As the region-based RV-FBAT approach is based on the empirical genetic variance/covariance matrix, it can be numerically instable. To avoid such numerical artifacts, we excluded those regions from both approaches in which the smallest P -value for any single variant in the respective region is 100,000 times greater than the region-based FBAT P -value. This cleaning step removed one window from the consecutive approach on chromosome 1. The tested window on chromosome 1 that was removed yielded a region-based P -value of 1.51×10^{-9} , however when we took a closer look at the tested SNPs within this window, we saw that the majority of the SNPs showed low allele frequencies and only two of the SNPs showed a nominal significant, but weak association signal. The distribution of the P values for this region is shown in Figure 1 and strongly supports the assumption that the region-based association P -value for this region is a numerical artifact. Figure 2 shows the QQ plots for both tested approaches, the quantile-distribution of the test statistics and the mean inflation factor lambda. For both approaches, we do not observe a general inflation of the test statistics.

We did not observe any genome-wide significant regions for both testing strategies. This was expected given the small sample size. In our clustering approach, two regions achieved an association P -value of less than 5×10^{-6} , while for the consecutive testing approach we did not observe any association P values that were smaller than 5×10^{-6} . Moreover, the clustering approach identified 12 independent regions with an association P -value of less than 5×10^{-5} (all regions with a P -value of less than 5×10^{-5} that were located less than 1 MB apart were considered to belong to one region). Using the same definition for independent regions, the consecutive approach only identified eight independent regions with such an association P -value cutoff. This could suggest that our clustering approach has more statistical power than a consecutive window approach with fixed size. A description of the most significant regions from the consecutive approach is reported in Appendix Table A1.

Table 3 lists the most significantly associated regions from the clustering approach. The most associated region on chromosome 17 is located within the ITGB3 gene. The gene has previously been implicated in the process of cognitive decline of Alzheimer's disease patient (Stellos et al., 2010).

4.1 Replication using IGAP data

We decided to follow-up the two regions from chromosomes 2 and 17, which showed an association P -value of less than 5×10^{-6} . We therefore tried to replicate our findings, using association summary statistics from imputed genotype data of the International Genomics of Alzheimer's Project (IGAP) consortium (Lambert et al., 2013). The publicly available stage 1 summary statistics from IGAP are from a meta-analysis of four different imputed GWAS data sets, comprising the information of roughly 17,000 cases and 37,000 controls all with European ancestry (for details on how the meta-analysis was performed please refer to Lambert et al., 2013). The IGAP summary statistics were available for variants with a lower MAF $\geq 1\%$. We extracted all summary statistics for variants that are located in our two identified regions on chromosome 2 and 17 and used VEGAS 2 (Mishra & Macgregor, 2015) to calculate association P -values for the two regions. The replication results in Table 4 show that the region on chromosome 17 achieves a nominal significant association P -value (P -value 0.0138) providing further support for the relevance of the identified region on chromosome 17 for Alzheimer's disease.

5 CONCLUSION

In this communication, we presented a new testing strategy for NGS data. Based on the hypotheses that nearby rare variants are likely to be part of the same functional process that causes an increased disease susceptibility, we developed a spatial grouping algorithm for sequencing data. One advantage of our algorithm is that it takes into account the locally varying density of genetic variation along the genome. We show the practicability of our algorithm in a simulation study and by an application to Alzheimer's disease WGS data and compared its performance to a standard approach, where groups of consecutive variants with a fixed variant size are tested for association. Although both testing strategies did not identify any genome-wide significant regions, our clustering approach, although the number of tested windows is almost the same as for the consecutive approach, identifies more potentially associated regions. Although both window strategies suggested association in a region on chromosome 17 that contains the ITGB3 gene, the clustering approach identified a window that spanned over 3,696 base pairs and contained 53 variants for which the association P -value was 1.37×10^{-6} . The window in the same region on chromosome 17 that was identified by the standard consecutive window approach achieved a P -value of 7.857×10^{-6} . We tried to replicate the identified region on chromosome 17 using association summary statistics from imputed data of the IGAP consortium and found a nominal significant association P -value of the region in the replication data.

Acknowledgments

The project was supported by Cure Alzheimer's fund, by the National Institute of Mental Health (R01MH081862, R01MH087590), by the National Heart, Lung and Blood Institute (R01HL089856, R01HL089897), by the National Human Genome Institute (R01HG008976), and the BONFOR Programme of the University of Bonn Deutsche Forschungsgemeinschaft (DFG).

Grant sponsor: Cure Alzheimer; Grant sponsor: National Institute of Mental Health; Grant numbers: R01MH081862 and R01MH087590; Grant sponsor: National Heart, Lung and Blood Institute; Grant numbers: R01HL089856 and R01HL089897; Grant sponsor: National Human Genome Institute; Grant number: R01HG008976; Grant sponsor: Deutsche Forschungsgemeinschaft (DFG).

APPENDIX

TABLE A1

Description of the most significantly associated regions from the consecutive testing approach (association P -value $< 5 \times 10^{-6}$)

Association P -Value ** $< 5 \times 10^{-6}$ $* < 5 \times 10^{-5}$	Chromosomal Position	Nearest Gene	No. of Tested Variants	FBAT's Association Z-Score/ P -Value
*	Chr1: 149074879–149085248		50	$-4.249/2.146 \times 10^{-5}$
*	Chr1: 189077787–189084176		50	$-4.436/9.150 \times 10^{-6}$
*	Chr1: 197113573–197124999		50	$-4.078/4.541 \times 10^{-5}$
*	Chr6: 31279600–31286209		50	$-4.258/2.062 \times 10^{-5}$
*	Chr13: 69253862–69263495		50	$-4.312/1.619 \times 10^{-5}$
*	Chr14: 62593736–62602349		50	$4.230/2.333 \times 10^{-5}$
*	Chr15: 83917801–83933427		50	$-4.095/4.215 \times 10^{-5}$
*	Chr17: 17:45317233–45321270		50	$4.469/7.857 \times 10^{-6}$
	Chr17: 45350393–45358131		50	$4.073/4.634 \times 10^{-5}$

References

- 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015; 526(7571):68–74. [PubMed: 26432245]
- De G, Yip WK, Ionita-Laza I, Laird N. Rare variant analysis for family-based design. *PloS One*. 2013; 8(1):e48495. [PubMed: 23341868]
- Evans IS. The selection of class intervals *Transactions of the Institute of British Geographers*. 1977:98–124.
- Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, Lange C. Location, location, location': A spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. *Bioinformatics*. 2012; 28(23):3027–3033. [PubMed: 23044548]
- Fisher WD. On grouping for maximum homogeneity. *Journal of the American statistical Association*. 1958; 53(284):789–798.
- He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, Santos-Cortez RLP, Shendure J. Rare-variant extensions of the transmission disequilibrium test: Application to autism exome sequence data. *American Journal of Human Genetics*. 2014; 94(1):33–46. [PubMed: 24360806]
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*. 2009; 106(23):9362–9367.
- Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, Awadalla P. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nature Genetics*. 2015; 47(4):400–404. [PubMed: 25685891]
- Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*. 2013; 92(6):841–853. [PubMed: 23684009]
- Jenks GF. Optimal Data Classification for Choropleth Maps (Occasional Paper 2). Lawrence: University of Kansas, Department of Geography; 1977.
- Jenks GF, Caspall FC. Error on choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers*. 1971; 61(2):217–244.

- Joyce P, Tavaré S. The distribution of rare alleles. *Journal of Mathematical Biology*. 1995; 33(6):602–618. [PubMed: 7608640]
- Krebs, JE., Lewin, B., Goldstein, ES., Kilpatrick, ST. *Lewin's genes XI*. Jones & Bartlett Publishers; 2014.
- Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, Hofker MH. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet*. 2013; 9(1):e1003201. [PubMed: 23341781]
- Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB. The empirical power of rare variant association methods: Results from sanger sequencing in 1,998 individuals. *PLoS Genet*. 2012; 8(2):e1002496. [PubMed: 22319458]
- Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, Grenier-Boley B. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*. 2013; 45(12):1452–1458. [PubMed: 24162737]
- Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang ZZ, Jun G. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *American Journal of Human Genetics*. 2014; 94(2):233–245. [PubMed: 24507775]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*. 2012; 91(2):224–237. [PubMed: 22863193]
- Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*. 2014; 95(1):5–23. [PubMed: 24995866]
- Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics*. 2011; 89(3):354–367. [PubMed: 21885029]
- Lohmueller KE, Sparsø T, Li Q, Andersson E, Korneliussen T, Albrechtsen A, Kilpeläinen TO. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *American Journal of Human Genetics*. 2013; 93(6):1072–1086. [PubMed: 24290377]
- Mishra A, Macgregor S. VEGAS2: Software for more flexible gene-based testing. *Twin Research and Human Genetics*. 2015; 18(01):86–91. [PubMed: 25518859]
- Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, Bis J. Whole genome sequence-based analysis of a model complex trait, high density lipoprotein cholesterol. *Nature genetics*. 2013; 45(8):899. [PubMed: 23770607]
- Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. *Human Molecular Genetics*. 2013; 22(R1):R16–R21. [PubMed: 23922232]
- Pennisi, Elizabeth. Disease risk links to gene regulation. *Science*. 2011; 332(6033):1031–1031. [PubMed: 21617055]
- Raab JR, Kamakaka RT. Insulators and promoters: Closer than we think. *Nature Reviews Genetics*. 2010; 11(6):439–446.
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*. 2005; 15(11):1576–1583. [PubMed: 16251467]
- Stellos K, Panagiota V, Kögel A, Leyhe T, Gawaz M, Laske C. Predictive value of platelet activation for the rate of cognitive decline in Alzheimer's disease patients. *Journal of Cerebral Blood Flow & Metabolism*. 2010; 30(11):1817–1820. [PubMed: 20717123]
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Konkel MK. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015; 526(7571):75–81. [PubMed: 26432246]
- Wang H, Song M. Ckmeans. 1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *R Journal*. 2011; 3(2):29–33. [PubMed: 27942416]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rarevariant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*. 2011; 89(1):82–93. [PubMed: 21737059]

- Xu C, Tachmazidou I, Walter K, Ciampi A, Zeggini E, Greenwood CM. Estimating genome-wide significance for whole-genome sequencing studies. *Genetic Epidemiology*. 2014; 38(4):281–290. [PubMed: 24676807]
- Yazdani A, Yazdani A, Boerwinkle E. Rare variants analysis using penalization methods for whole genome sequence data. *BMC Bioinformatics*. 2015; 16(1):405. [PubMed: 26637205]

FIGURE 1.

Distribution of Should be 356,387 tested regions and 356,365 tested regions in figure 1 the P-values for chromosome 1 from the consecutive approach (FBAT P-value 1.51×10^{-9})

FIGURE 2.

QQ plots and mean inflation factor (lambda) of the test statistics for the clustering and consecutive testing approach

TABLE 1

Simulation results for the different scenarios, based on 1000 replications

Alpha	Scenario without Gaps				Scenario with Gaps			
	power.con	spec.con	power.cluster	spec.cluster	power.con	spec.con	power.cluster	spec.cluster
1×10^{-10}	0.035	0.257	0.034	0.235	0.034	0.294	0.044	0.636
1×10^{-9}	0.039	0.282	0.04	0.275	0.036	0.333	0.049	0.633
1×10^{-8}	0.046	0.326	0.044	0.295	0.044	0.364	0.065	0.677
1×10^{-7}	0.061	0.361	0.056	0.357	0.065	0.431	0.102	0.755
1×10^{-6}	0.079	0.392	0.063	0.333	0.071	0.423	0.145	0.821
1×10^{-5}	0.1	0.460	0.078	0.397	0.096	0.458	0.195	0.851
1×10^{-4}	0.119	0.487	0.099	0.404	0.11	0.445	0.255	0.867
0.001	0.165	0.491	0.149	0.409	0.159	0.472	0.342	0.874
0.01	0.233	0.502	0.217	0.392	0.231	0.468	0.47	0.864
0.1	0.37	0.430	0.356	0.346	0.397	0.401	0.616	0.826

TABLE 2

Descriptive statistics of the regions identified by the clustering approach

	Minimum	First Quantile	Median	Mean	Third Quantile	Maximum
Number of variants	5	13	28	49.5	61	902

TABLE 3

Description of the most significantly associated regions from the variant clustering approach (association p-value $< 5 \times 10^{-5}$)

Association P-Value ** 5×10^{-6} * 5×10^{-5}	Chromosomal Position of the Tested Region	Name of Nearest Gene (± 100 kb to the Chromosomal Position of the Region, * Denotes That the Region Is Located within the Gene, N/A if No Gene is within 100 kb Radius to the Region Borders)	No. of Tested Variants	FBAT's Association Z-Score/P-Value
*	Chr1:189070813-189093704	N/A	215	$-4.103/4.08 \times 10^{-5}$
**	<i>Chr2: 122970146-122973576</i>	N/A	22	$-4.801/1.53 \times 10^{-6}$
*	Chr3: 4933881-4940125	<i>BHLHE40-AS1</i>	48	$-4.177/2.96 \times 10^{-5}$
*	Chr5: 6223650-6224033	<i>FLJ33360</i>	9	$4.169/3.05 \times 10^{-5}$
*	Chr5: 165891216-165892449	N/A	13	$4.305/1.67 \times 10^{-5}$
*	Chr6: 119335244-119364536	<i>FAM184A*</i>	204	$4.134/3.57 \times 10^{-5}$
*	Chr7: 22551545-22555223	<i>STEAP1B</i>	35	$4.075/4.60 \times 10^{-5}$
*	Chr11: 132539474-132540443	<i>OPCML*</i>	6	$4.089/4.34 \times 10^{-5}$
*	Chr13: 69238183-69265900	N/A	167	$-4.406/1.05 \times 10^{-5}$
*	Chr14: 22360508-22364979	<i>TCRA*</i>	58	$-4.143/3.43 \times 10^{-5}$
**	<i>Chr17: 45318080-45321776</i>	<i>ITGB3</i>	53	$4.829/1.37 \times 10^{-6}$
*	Chr17: 45332734-45339475	<i>ITGB3*</i>	46	$4.134/3.57 \times 10^{-5}$
*	Chr17: 45340266-45356782	<i>ITGB3*</i>	103	$4.083/4.45 \times 10^{-5}$
*	Chr21:38399788-38400584	<i>RIPPLY3</i>	13	$-4.192/2.77 \times 10^{-5}$

Note: Associated regions with a P-value of less than 5×10^{-6} are printed cursive; all positions are based on HG19.

TABLE 4

Replication results of the two most significantly associated regions on chromosomes 2 and 17 using the IGAP data

Chromosomal Position	IGAP	
	<i>P</i> -Value	Number of Tested Variants
Chr2: 122970146–122973576	0.309	11
Chr17: 45318080–45321776	0.0138	34