BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 157–164
doi:10.1093/bioinformatics/btt654

Genome analysis

Advance Access publication November 20, 2013

On the simultaneous association analysis of large genomic regions: a massive multi-locus association test

Dandi Qiao[1,*], Michael H. Cho[2], Heide Fier[3], Per S. Bakke[4], Amund Gulsvik[4], Edwin K. Silverman[2] and Christoph Lange[1]

[1]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 20115, USA, [2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, [3]Department of Genomic Mathematics, University of Bonn, 53113 Bonn, Germany and [4]Department of Thoracic Medicine, Haukeland University Hospital and Section for Respiratory Medicine Institute of Medicine, University of Bergen, 5006 Bergen, Norway

Associate Editor: John Hancock

ABSTRACT

Motivation: For samples of unrelated individuals, we propose a general analysis framework in which hundred thousands of genetic loci can be tested simultaneously for association with complex phenotypes. The approach is built on spatial-clustering methodology, assuming that genetic loci that are associated with the target phenotype cluster in certain genomic regions. In contrast to standard methodology for multilocus analysis, which has focused on the dimension reduction of the data, our multilocus association-clustering test profits from the availability of large numbers of genetic loci by detecting clusters of loci that are associated with the phenotype.

Results: The approach is computationally fast and powerful, enabling the simultaneous association testing of large genomic regions. Even the entire genome or certain chromosomes can be tested simultaneously. Using simulation studies, the properties of the approach are evaluated. In an application to a genome-wide association study for chronic obstructive pulmonary disease, we illustrate the practical relevance of the proposed method by simultaneously testing all genotyped loci of the genome-wide association study and by testing each chromosome individually. Our findings suggest that statistical methodology that incorporates spatial-clustering information will be especially useful in whole-genome sequencing studies in which millions or billions of base pairs are recorded and grouped by genomic regions or genes, and are tested jointly for association.

Availability and implementation: Implementation of the approach is available upon request.

Contact: daq412@mail.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 18, 2013; revised on October 21, 2013; accepted on November 7, 2013

1 INTRODUCTION

In the search for disease susceptibility loci (DSLs), genome-wide association studies (GWAS) have been a successful instrument for the identification of replicable genetic associations (Hardy and Singleton, 2009; Manolio *et al*., 2008). For most complex diseases and phenotypes, they discovered numerous genetic associations that can be validated in independent populations, although the genetic effect sizes of the loci are relatively small. Despite the large number of detected loci, GWAS association signals are only able to explain a small fraction of the overall predicted heritability (Visscher *et al*., 2008), i.e. the phenomenon of 'missing heritability'. One possible explanation for this phenomenon is 'synthetic associations'(Dickson *et al*., 2010). Simulation studies, theoretical considerations and empirical evidence (Adzhubei *et al*., 2010; Cohen *et al*., 2006; Fearnhead *et al*., 2004; Kryukov *et al*., 2007; Nejentsev *et al*., 2009; Pritchard and Cox, 2002) suggest that genetic associations, as they are detected by GWAS, can be caused by multiple rare variants (RVs). Because common variants are poor proxies for RVs or are not in linkage disequilibrium (LD) with rare disease-causing variants, it may be difficult to identify or characterize rare DSLs in GWAS data.

Another plausible explanation for the phenomenon of 'missing heritability' is insufficient statistical power due to the multiple-testing problem. In a GWAS, millions of genetic loci are tested individually for association with the target phenotype, and the test results have to be adjusted for multiple comparisons, leading to extremely small *P*-value thresholds for overall statistical significance. The standard approach has been aimed to increase the sample size of GWAS as much as possible. For example, several meta-analyses of GWASs (Allen *et al*., 2010) have contained the data of >100 000 study subjects. However, such large sample sizes hold the danger of increased study heterogeneity and do not necessarily lead to increased statistical power.

The fundamental issue with the standard analysis approach to GWAS (single locus association testing and adjustment for multiple comparisons) is that an increase in genomic resolution, i.e. adding more and more genetic loci to the analysis, does not increase the probability to detect DSLs, but diminishes the statistical power of the approach. To address this issue, multilocus tests have been suggested. For example, gene-based analysis has been advocated (Neale and Sham, 2004) to complement allelic association analysis of single locus. This is motivated by the idea that causal variants for one disease tend to reside in proximity to each other and variants in adjacent regulatory regions are more likely to have functional relevance (Huang *et al*., 2011). PLINK (Purcell *et al*.,

*To whom correspondence should be addressed.

2007) provides 'set-based' tests using the average single nucleotide polymorphism (SNP) statistic across the set of SNPs to implement this idea. Moreover, other tests such as the minSNP test, the Bayesian imputation-based association mapping (BIMBAM) test (Servin and Stephens, 2007), the versatile gene-based test (VEGAS) test (Liu *et al*., 2010) and the LASSO regression method for GWAS (Wu *et al*., 2009) have been proposed. Later on, Huang *et al*. (2011) proposed a gene-wide significance (GWiS) test, which estimates the number of independent effects within a gene. For next-generation sequencing data, methods that aggregate over a set of RVs to search for associated genomic regions with the disease status are shown to be more powerful than single locus approaches, e.g. the cohort allelic sums test (CAST) (Morgenthaler and Thilly, 2007), the combined multivariate and collapsing (CMC) method (Li and Leal, 2008), the weighted sum statistic by Madsen and Browning (Madsen and Browning, 2009), the kernel-based adaptive-clustering (KBAC) test (Liu and Leal, 2010), the sequence kernel association test (SKAT) (Wu *et al*., 2011), replication-based test (RBT) (Ionita-Laza *et al*., 2011) and so forth. There are several advantages of such gene-based tests over single loci tests. First, collapsing the small effects across the variants within a gene could give larger effect size to detect the association. Second, due to the smaller number of genes to be tested, the multiple-testing problem is reduced. Moreover, the associations of genes across different populations can be directly compared even though there could be different LD patterns within the genes across the populations (Huang *et al*., 2011).

However, most of the approaches can handle only a limited number of genetic loci, i.e. typically <100. None of them is able to incorporate the information about the physical location of the loci and their clustering. In this article, we are proposing a novel approach called the Bin test that can test a large genomic region for association with the target phenotype by taking into account the physical location of the variants that show evidence for association and their physical clustering. The genomic region can refer to one gene, a specified segment of the genome, a pathway, an entire chromosome or the complete genome. The approach is computationally fast and applicable to binary and complex phenotypes. The methodology is evaluated in simulation studies using a GWAS dataset from the COPDGene study and is applied to several collaborating chronic obstructive pulmonary disease (COPD) genetic studies. The simulation studies and the application results suggest that the approach has sufficient power to test simultaneously all genotyped loci on the entire genome or a specific chromosome.

2 METHODS

The proposed test assesses whether there is significant clustering of causal variants within a specified region. We consider both the level of associations between the variants and the trait, and the location of the variants. The degree of association between a variant and the phenotype is represented by the association *P*-values, which is easy to obtain from any dataset and allows the application of our method to both quantitative traits and dichotomous traits. To put this into a one-dimensional clustering problem, we need to consider four aspects of the test:

(1) What distance measure to use: the physical distance between two variants or a newly defined distance measure.

(2) Which single nucleotide variants (SNVs) to look at: a *P*-value cutoff to select the variants. Note that we use SNV to refer to all the variants, including variants with allele frequency <1%, and use SNP to refer to variants with allele frequency >1%.

(3) Whether to look at the distance to the nearest neighbor or the distances to the neighboring variants, and how many neighboring variants should be considered in the calculation of distances.

(4) How to quantify the difference between the distribution of the observed distances and the distribution of distances under the null, i.e. what test to use.

2.1 Distance measure

The first three aforementioned questions refer to the choice of distance distribution. Considering the absolute size of the physical distances between the variants and the *P*-values obtained from the association tests, our goal is to have a distance measure such that the distance between two variants is small if the 'average' *P*-value of the two variants is small, and if the physical distance between the two variants is small, relative to the other variants. Thus, we consider the multiplication of the physical distance with the association information rather than the addition of the two values to avoid the situation where the 'average *P*-value' is overwhelmed by the physical distance. To obtain the 'average' degree of association of the two variants, multiplication of the two *P*-values is also more suitable than addition, as one large *P*-value would dominate a much smaller *P*-value. We define a new distance measure D between two variants that combines the *P*-value with the physical distance between the variants:

$$D_{i,j} = dist_{i,j} * \sqrt{S_i S_j}$$

where the subscript i and j refer to any two variants in the region of interest. The distance measure is motivated by the fact that this distance equals the area below the geometric average of the *P*-values of the two variants. We use the square root here to have the absolute value of $D_{i,j}$ to lie in a reasonable range.

2.2 Cutoff values

There are two parameters that can be varied in the test: a cutoff value for the *P*-values—P, such that only variants with *P*-values below P are considered in the test of clustering; and the number of neighboring variants around each variant for calculating the distances—R. We could use a *P*-value of 1 to include all the variants and consider the distances from one variant to all the other variants in the region, but simulations suggest that this is computationally costly and has relatively low power comparing with including only variants with *P*-values below a threshold. Thus, a threshold on the *P*-value for selecting variants is used. The nearest-neighbor method is commonly used in clustering analysis, and it requires less computational cost. However, it does not give much information on the second, third or higher-level neighbors. Thus we consider both the distance to the nearest neighbor and the distances to a predefined R number of neighboring variants in the region.

In our analysis, this threshold of neighboring variants R and the cutoff value of *P*-values P are set to be the values that correspond to specified quantiles of all the variants in the region of interest. For example, we may specify the cutoff quantile for the *P*-values to be 0.1%, which means the top 0.1% variants with the smallest *P*-values are included in the analysis. If we specify the quantile threshold of neighboring variants R to be 1%, it means that the number of neighboring variants used to calculate the distances from each variant is 1% times N, where N is the total number of variants.

2.3 Test on the distance distribution

To test whether there is clustering of small *P*-values, the distribution of the distances between the variants needs to be compared with the

distribution of the distances under the null hypothesis in some way. The most popular non-parametric method to compare the empirical distribution of one sample with a specified distribution, or to compare the empirical distributions of two samples, is the Kolmogorov–Smirnov (KS) statistic, defined as follows:

$$D_{n_1,n_2} = \sup_x |F_{1,n}(x) - F_{2,n}(x)|$$

where $F_{1,n}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$ is the observed cumulative distribution function of the first sample, and similarly $F_{2,n}(x)$ is the observed cumulative distribution function of the second sample. The first sample, in our case, refers to the observed distances between the variants. The second sample refers to the distances between the variants obtained under the null hypothesis using permutations.

We also considered an alternative approach, called the Bin test statistic that extends the idea in Kowalski *et al.* (2002) and Olson *et al.* (2005). The Bin test is a permutation test that compares the observed proportions of distances in 10 given intervals to the expected proportions of distances using the M statistic (referred to as the Bin test):

$$M = (Prop - E(Prop))^T S^{(-1)} (Prop - E(Prop))$$

The distances between the variants obtained using permutations under the null are ordered and put into 10 bins with equal size, therefore there are 10% of all the distances in each of the 10 bins. Thus, E(*Prop*) is set to be a 10×1 vector of 10% in this statistic, i.e. (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1). Then, the minimum and maximum of the distances in each bin give the interval of distance of each bin. *Prop* is then the 10×1 vector of the proportions of the observed distances in these 10 intervals. $S^{(-1)}$ is the 10×10 Moore–Penrose generalized inverse of the variance covariance matrix of the proportions of distances in the 10 intervals from each permutation under the null. The number of equally spaced bins could be varied, and unequally spaced bins could be used, as discussed in White *et al.* (2009). We chose 10 equally spaced bins here to simplify the problem, but further investigation is needed to evaluate the performance of the statistic with other choices.

For both the KS and the Bin tests, the null distribution of distances is obtained by permuting the case and control status among the subjects, which conserves the LD between the variants.

Other distribution tests could also be used here, such as the Ansari–Bradley test. From a limited number of simulations, the Ansari–Bradley test gives a moderate power that is higher than the KS test, but does not perform as good as the Bin test (data not shown here).

2.4 Summary of the method

Here is a summary of the procedure of the method:

(1) Choose a *P*-value cutoff P and a cutoff R for the number of neighboring variants to be included in the calculation.

(2) Calculate the single-variant association *P*-values and include only variants with single-variant association *P*-value that is smaller than the cutoff P.

(3) Calculate the new distance measure for each variant with their neighboring variants within the cutoff number R. The distance measures form a distribution of observed distances.

(4) By permuting the case and control status, using the same cutoffs, we get a different distribution of the distances under the null for each permutation.

(5) By putting all the distances obtained under the null together to form the null distribution, the Bin statistic (or the KS statistic) can be calculated for the observed distance distribution.

(6) Similarly, for each permutation, the Bin statistic (or the KS statistic) can be calculated.

(7) Compute the *P*-value of the test by comparing the Bin statistic (or the KS statistic) of the observed distance distribution to the distribution of the Bin statistics (or the KS statistic) obtained from permutations under the null.

3 RESULTS

We assessed the performance of the KS test and the Bin test using simulations based on the genotypes of the African American (AA) samples in the GWAS dataset of the COPDGene study (Regan *et al.*, 2011). Also, the Bin test was applied to the COPD status of several collaborating COPD genetic studies to look for COPD susceptibility loci in the application section.

3.1 Simulation results

3.1.1 Simulation results on entire chromosome Simulations were done using the genotypes of the AA samples from the COPDGene study. There are approximately 700k SNPs included for the 2570 AA samples after the quality control (QC) steps for this study. Variants with minor allele frequency (MAF) <0.01, high missing rate (>5% for SNPs with MAF $\geq$ 5%, and >2% for SNVs with MAF $\leq$ 5%), Hardy-Weinberg equilibrium (HWE) $P < 10e-3$ and concordance rate $<99\%$ using 205 duplicated samples were removed. Samples with call rate $<98.5\%$, and mismatched gender and race were also excluded. Autosomal SNPs with HWE $P > 0.01$, $MAF > 0.05$ and markers represented in Hapmap III were used for principal component analysis. EIGENSOFT 3.0 was used to obtain the PCs to adjust for population substructure for the AA samples. We used 2569 samples in our simulations due to missing information for one case.

In our simulations, we used the genetic data on chromosome 7 from the COPDGene study, but generated the case and control status according to our disease model. There are in total 36 726 SNPs on chromosome 7. Two different scenarios were considered. First, we selected nine SNPs on chromosome 7 as the causal variants that reside close to each other, and considered both protective and deleterious effects of the variants and different effect size. The MAF of the nine SNVs are (0.1740, 0.4914, 0.1734, 0.1244, 0.4673, 0.2552, 0.1098, 0.0309, 0.0728). The physical distances between the adjacent variants are 530, 1564, 1011, 813, 1087, 249, 685 and 707. The LDs between the nine causal variants in the dataset are shown in Figure 1. Two sets of effect sizes are simulated for this scenario. For scenario 1, the odds ratio of the nine SNVs are (0.8, 1.1, 0.9, 1.2, 0.9, 1.2, 1.2, 1.5, 1.5) and the average disease rate of our disease model for this population is 0.145; for scenario 2, the odds ratio of the nine

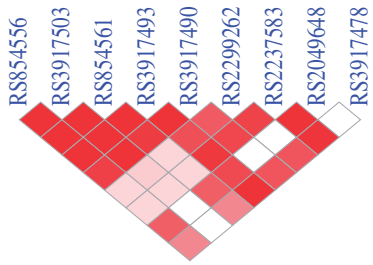

Fig. 1. The LD plot of the nine causal SNVs used in the simulation

Table 1. The power of the tests

Power	Effect size 1	Effect size 2	Effect size 3
Bin test	0.920	0.990	0.274
KS test	0.620	0.845	0.179

Note: The power of the tests for three scenarios, obtained from 200 simulations with 2000 permutations in each permutation set. The power is the number of simulations with $P < 0.05$. The effect sizes (odds ratio) of the nine SNVs with the intercept at the front are effect size 1: (0.135, 0.8, 1.1, 0.9, 1.2, 0.9, 1.2, 1.2, 1.5, 1.5) and effect size 2: (0.135, 0.8, 1.1, 0.8, 1.3, 0.9, 1.3, 1.2, 1.5, 1.5) for the first two columns. For effect size 3, 100 SNVs were chosen within a random segment on the chromosome and are assigned with randomly generated effect sizes from a normal distribution with mean 1 and SD = 0.05 with an intercept odds of 1 in each simulation.

SNVs are (0.8, 1.1, 0.8, 1.3, 0.9, 1.3, 1.2, 1.5, 1.5) and the average disease rate of our disease model for this population is 0.150. Then given the effect size and the genotypes of the samples in the COPDGene study, we generated the case and control status accordingly. Second, we considered the possibility of having a lot of causal variants with small effect size. Thus, 100 causal variants were chosen in proximity to each other on the chromosome by randomly selecting 100 variants in a randomly selected region on the chromosome. The effect sizes (odds ratio) were generated using a normal distribution with mean 1 and SD = 0.05.

Sensitivity analysis was done to assess the effects of the *P*-value cutoff P, and the number of neighboring variants R, on the power and type I error of the Bin test for different disease models and sample size. The results and discussions are in the Supplementary Material. According to the analysis, the optimal cutoff values are similar with different disease models and sample sizes we considered, and the power is mostly robust to the cutoff values in general, as long as extreme cutoff values are not used. Here to achieve a good power, SNVs with *P*-value in the top 0.5% percentile were included in the analysis, and $0.1\% * N$ neighboring SNVs next to each SNV were used in the tests, where N is the total number of SNVs in the dataset. Owing to the computational limitation, 2000 permutations were used in each permutation set to maintain the type I error, as explained in our sensitivity analysis (Supplementary Material). For each scenario, 200 simulations were generated to obtain the estimated power and the type I error rate. The power of the test is the percentage of simulations in which the permutation *P*-value is <0.05. The results of the Bin test are shown in Table 1, as well as the power of the KS test, as a comparison. We observed a higher power of the Bin test comparing with the KS test in all the scenarios. Therefore, the Bin test is recommended and is used in the calculation of the association *P*-values of the chromosomes in the application section.

We also computed the type I error rate of the Bin test on three different autosomal chromosomes by randomly generating the probability of having the disease for each individual using a uniform distribution $Unif(0, 0.5)$, and then randomly generated the disease status for each sample using a Bernoulli distribution with these probabilities. It is shown in Table 2 that the type I error rate is well maintained with different LD patterns on different chromosomes.

Table 2. The type I error rate of the test on chromosome 7, 10 and 22

Type I Error	Chromosome 7	Chromosome 10	Chromosome 22
Bin test	0.030	0.065	0.020
KS test	0.035	0.025	0.040

Note: 2000 permutations in each permutation set were used and 200 replicates were generated to compute the type I error rate.

3.2 Application results

3.2.1 Results on each chromosome The test was applied to a case-control cohort from Bergen, Norway with 863 cases and 808 controls (Cho *et al*., 2012) from the GenKOLS study (Pillai *et al*., 2009) to see if there is any chromosome that is significantly clustered with variants associated with COPD status. Based on the dataset that passed the QC steps from Cho *et al*. (2012), any SNVs with MAF <0.01, call rate <98% and HWE $P < 0.000001$ were also removed, and we were left with 495 829 SNVs for the autosomal chromosomes. Population outliers were further removed, and we were left with 854 cases and 805 controls. We also applied the test to the first 1000 subjects from the COPDGene study (Cho *et al*., 2012). Based on the dataset that passed the QC steps from Cho *et al*. (2012) and after additional QC steps as for the GenKOLS dataset, 797 218 SNVs were left and 496 cases and 498 controls were used for the association tests. Similarly to the simulations, the cutoff value for the association *P*-value percentiles was set to be 0.5%, and the quantile of neighboring SNVs around each SNV to be included in the analysis was set to be 0.1%. The *P*-values for the test of clustering on each chromosome were obtained using 2000 permutation set.

Association *P*-values of the Armitage trend test for the SNVs, adjusted for ancestry, were computed for the samples and are plotted in Figures 2 and 3 later in the text for the two cohorts. We included age, sex and pack-years of smoking as the covariates in our analyses. For the GenKOLS cohort, the order of the SNVs according to their significance magnitude is mostly the same as in the original article (Pillai *et al*., 2009), and the Manhattan plot is shown in Figure 2 later in the text. No SNV reached the genome-wide significance level (5×10^{-8}) in our analysis. The original article identified two SNVs at the a-nicotinic acetylcholine receptor (*CHRNA3/5*) locus on chromosome 15 that were replicated using other datasets. Other studies have indicated that loci near *HHIP* may be related to COPD (Pillai *et al*., 2009; Wilk *et al*., 2009). It has also been found from previous studies that the *FAM13A* locus on chromosome 4 includes a disease susceptibility locus for COPD (Cho *et al*., 2010).

The results of our clustering method applied on each chromosome for each dataset are shown in Table 3. The significance level we used here is $0.05/22 = 0.00227$. Surprisingly, we observed a strong signal on chromosome 10 in the GenKOLS cohort in which there is no significant indication of causal SNVs from the single variant association tests. From the COPDGene dataset, there is no chromosome with significant *P*-value, but chromosome 10 has a small *P*-value around 0.05. By looking at only the Manhattan plot, the significant *P*-value in the

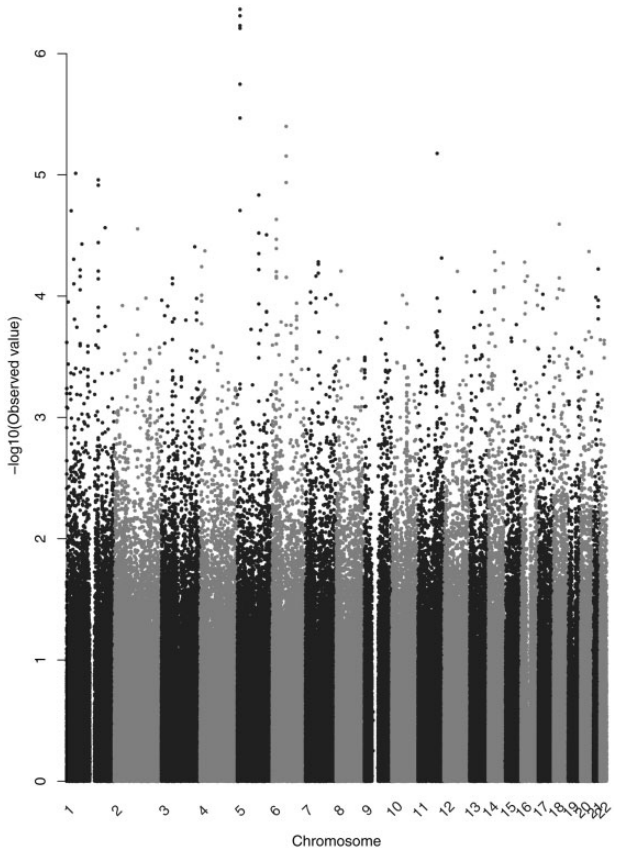

Fig. 2. The Manhattan plot of the adjusted *P*-values of the SNVs in the GenKOLS study

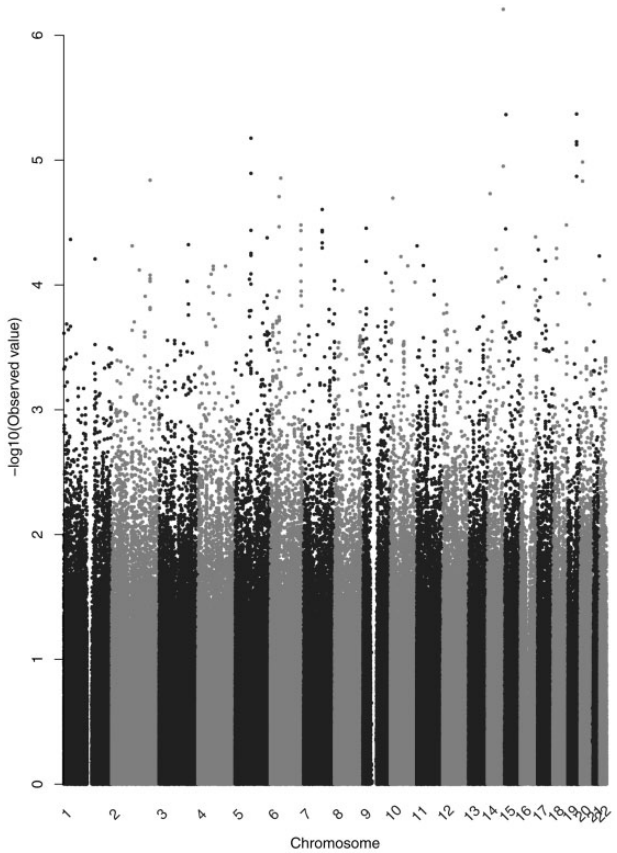

Fig. 3. The Manhattan plot of the adjusted *P*-values of the SNVs in the COPDGene dataset

GenKOLS cohort seems to be caused by the clustering of many nearby SNVs with *P*-values around $10^{-3} \sim 10^{-4}$. We also applied the test on the combined dataset that includes subjects from the COPDGene study, the GenKOLS COPD cohort and subjects from Normative Aging study (NAS) and National Emphysema Treatment Trial (NETT) (Cho *et al.*, 2012). Note that the GenKOLS cohort is a more homogeneous population than the other two cohorts. As we know from previous studies that there are causal variants residing on Chromosome 4 and 15, the *P*-values from our clustering method for the two chromosomes are significant. Chromosome 10 also has a relatively small *P*-value after combining the three cohorts.

3.2.2 More insights for chromosome 10 The distribution of the new distances D for chromosome 10 of the GenKOLS cohort is shown in Figure 4. Our test compares this observed distribution with the distance distribution obtained using permutations under the null (8000 permutations were used here). Each bin of the histogram contains 10% of the distances obtained using permutation under the null, therefore we can see that the significant difference between the distributions comes from the first bin, in which there is a much larger proportion of the observed distances. The 10% quantile of the distances obtained under the null is 1447. We are interested to see which pairs of SNVs contribute the most to this difference between the two distributions. Figure 5 shows the physical positions of the SNVs with their *P*-values on the *y*-axis for the observed COPD status. The SNVs are colored according to the number of distances D to the neighboring SNVs that are less than 1447. The more contribution the SNV makes to the difference between the observed distribution and the null distribution of distances, the deeper the red color. There is apparently one cluster on chromosome 10. Most of the SNVs in this cluster reside on *C10orf11*. This result is interesting as it has been shown in a large-scale GWAS that this gene is associated with lung function (Artigas *et al.*, 2011). The original GWAS includes 48 201 individuals of European ancestry with an additional 46 411 individuals in the follow-up study, whereas in our dataset there are less than 1700 subjects included. Thus this result shows that our method could have much higher power in detecting the causal variants by considering the physical locations of SNVs and treating it as a clustering problem.

We further applied our method on this interesting gene *C10orf11* and obtained a $P < 0.0001$ (10000 permutations have been done here). Note that the *P*-value should be compared with the significance level after adjusting for all the genes on the genome. For the application on genes, we used all the SNVs on the gene to calculate the test statistic, i.e. no *P*-value cutoff or threshold for the number of neighboring SNVs. The LD plot and the association plot of this locus are shown in Figure 6. We also applied the SKAT (Wu *et al.*, 2011) and its optimal version (SKAT-O) (Lee *et al.*, 2012) to this gene, and obtained *P*-values of 0.558 and 0.733, respectively. We expected to observe such

Table 3. The *P*-values of the 22 autosomal chromosomes for three datasets

Chromosome *P*-value	GenKOLS	COPDGene	Combined three cohorts
1	0.37333	0.40402	0.41136
2	0.77598	0.82509	0.72507
3	0.26704	0.70237	0.51206
4	0.57603	0.45009	**0.00050**
5	0.47846	0.56363	0.07276
6	**0.05783**	0.89963	0.20125
7	**0.04598**	0.95485	0.44038
8	0.22121	0.70352	0.63613
9	0.25456	0.18883	0.38433
10	**0.00010**	**0.05563**	**0.02539**
11	0.10232	0.07616	0.43665
12	0.35209	0.40256	0.74019
13	0.73528	0.26468	0.65424
14	0.20009	0.46675	0.43448
15	**0.04807**	0.78198	**0.00050**
16	0.92608	0.12645	0.33425
17	0.85643	0.58766	0.20875
18	0.72968	0.47390	0.08625
19	0.61114	0.091421	0.62746
20	0.61280	0.59677	0.44517
21	0.90056	0.51515	0.95108
22	0.99166	0.18927	0.08608

Note: The *P*-values of the 22 autosomal chromosomes for the GenKOLS cohort, the COPDGene study and the combined dataset that includes the GenKOLS subjects, the subjects from the COPDGene study and the subjects from Normative Aging Study and National Emphysema Treatment Trial. With Bonferroni correction, the *P*-values should be compared with 0.00227. *P*-values < 0.05 are shown in bold font.

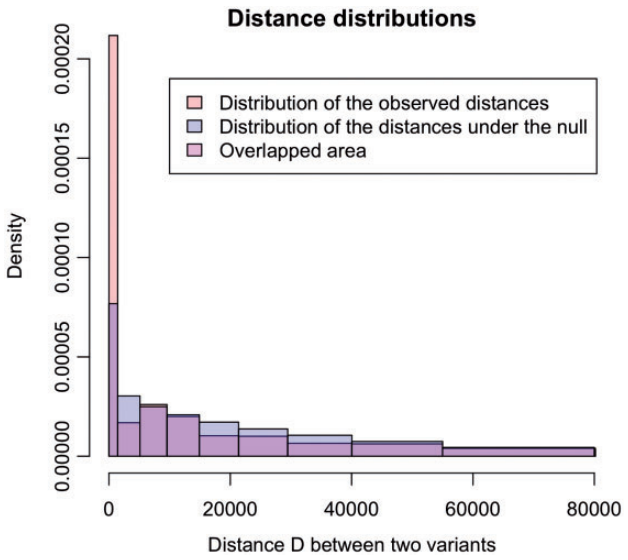

Fig. 4. The distance distributions of the observed distances. The distance distributions of the observed distances D between two variants, and the distance distribution of the distances obtained using 8000 permutations under the null. Each bin contains 10% of the distances obtained using permutations under the null and is colored with light blue. The observed distances are then assigned to each of the bins and are colored with red. The last bin (largest 10% of distances) is not shown in the histogram for clearer visualization

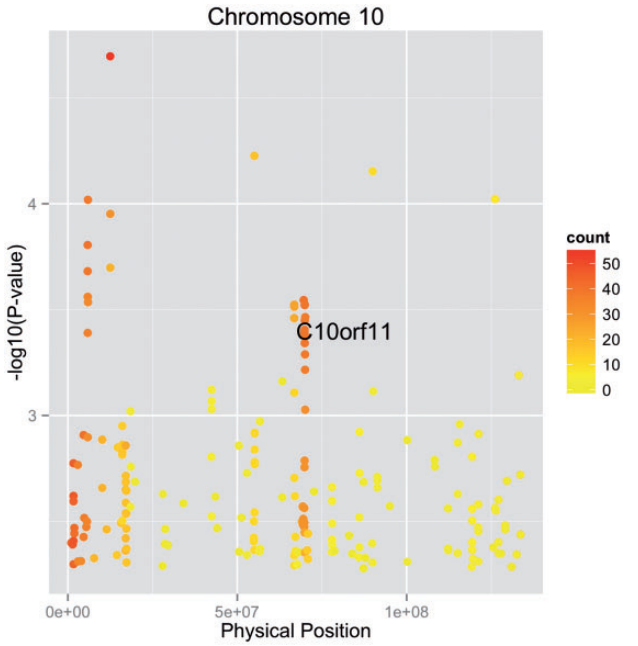

Fig. 5. The *P*-values of the SNVs versus their physical positions on chromosome 10. The points are colored from yellow to red according to the number of distances D that are less than 1447 (10% quantile of distances under the null) between each SNV to their neighboring SNVs. The spectrum on the right shows the corresponding counts of such neighboring distances for each SNV. The region that contributes the most to the distribution difference overlaps with C10orf11

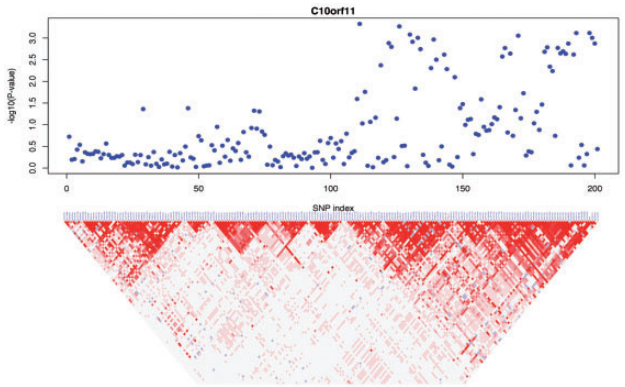

Fig. 6. The *P*-values for SNVs in C10orf11 and the LD structure. The upper plot shows the *P*-values of the SNVs on *C10orf11* with respect to their index in the dataset. The lower plot shows the LD structure of the SNVs in the GenKOLS cohort. The two plots are matched by SNV index

difference in the *P*-values, as this gene satisfies the hypothesis of our tests; therefore our test should have more power under such situation. This observation is also supported by the simulation results shown in Supplementary Table S4. Some of the SNVs in the cluster on chromosome 10 seem to lie in the gene *KCNMA1*, therefore we applied our test on this gene, which is next to *C10orf11*. It turns out that the *P*-value is 0.67732, showing no sign of association. We also applied our test on the candidate

genes *FAM13A* and *CHRNA3/5* locus, the *P*-values are 0.01159 and 0.07147787, respectively, for the GenKOLS cohort. For the COPDGene cohort, *C10orf11* has a *P*-value of 0.06943, which is not significant, possibly due to low power.

3.2.3 Result on the entire genome We have also applied the test to the entire genome to see if there is any region in the genome that is clustered with the causal variants. Same *P*-value cutoff and threshold for calculating the distances to the neighboring SNV are used. With 1500 permutations set, the *P*-value is 0.069315 for the entire genome, showing weak significance of association between the genome and the phenotype. This application shows the potential of our method for testing large genomic regions when no significant association is found for univariate tests. One possible way to search for the associated loci with the phenotype is to conduct a binary search using our method. Interested readers could refer to the Supplementary Material for some discussion about the procedure.

4 DISCUSSION

In summary, we proposed here an approach for the detection of clustering of causal variants in a genomic region of any size. Many existing methods collapse the effects of variants across a region or a gene, however, few of them use the physical location of these variants and many would suffer loss of power when too many variants are included. Simulations and application results suggest that our approach provides sufficient power to detect associated genomic regions with complex disease, especially when the causal variants reside relatively close to each other, even with small effect size. Also, the increase in statistical power allows analyses with a smaller sample size, which enables the possibility to compare more extreme phenotypes.

The same idea of testing for clustering could be applied to sequencing data, where thousands of variants would be available for each gene. For variants that are extremely rare, the univariate *P*-value may include only random noise, thus other measures of the association at each variant need to be considered to apply our method, i.e. standard analysis approaches for RV analysis. Moreover, the genomic region could refer to the genes in the same pathway (Wang *et al*., 2007), thus whether there is significant clustering of small *P*-values in each pathway could be examined.

However, there are several drawbacks we need to consider. Because permutation is used to obtain the *P*-value of the test statistic, there is extensive computational cost if the test is applied to a large number of small regions, which requires more number of permutations. From our experience, calculating the *P*-value for a dataset containing ~35k SNVs and 2570 subjects with a quantile *P*-value cutoff 0.5% (which is equivalent to including about 180 SNVs) and number of neighboring variants cutoff 0.1% (which is equivalent to 35 neighboring variants for each SNV), and with 5000 permutations, takes ~22 h using a 800 MHz AMD Phenom II X4 910e CPU. To obtain the *P*-value for different regions, clusters can be used to parallel the work. Second, the power may also be compromised if the regions are extremely small, limiting the possibility of clusters and their detection, and if the number of regions to be tested are extremely large due to multiple-testing problem. Right now the method is limited to population-based studies because permutation of the affections status is used to evaluate the *P*-values, but because the associations are represented by *P*-values, which could be obtained from either population-based association tests or family-based association tests, there is potential to extend the approach to family-based association studies.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the generous support from the Department of Biostatistics, Harvard School of Public Health. They also would like to thank the Channing Division of Network Medicine, Department of Medicine in Brigham and Women's Hospital, affiliated with Harvard Medical School for providing them with the COPD data used in the simulations and applications, and their insightful comments about the article.

Funding: National Institutes of Health [R01HL089856, R01HL089897 and R01HL113264].

Conflict of Interest: none declared.

REFERENCES

Adzhubei,I. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Allen,H. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Artigas,M. *et al.* (2011) Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat. Genet.*, **43**, 1082–1090.

Cho,M.H. *et al.* (2010) Variants in fam13a are associated with chronic obstructive pulmonary disease. *Nat. Genet.*, **42**, 200–202.

Cho,M.H. *et al.* (2012) A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum. Mol. Genet.*, **21**, 947–957.

Cohen,J. *et al.* (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl Acad. Sci. USA*, **103**, 1810.

Dickson,S. *et al.* (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol*, **8**, e1000294.

Fearnhead,N. *et al.* (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA*, **101**, 15992.

Hardy,J. and Singleton,A. (2009) Genomewide association studies and human disease. *N. Engl.J. Med.*, **360**, 1759–1768.

Huang,H. *et al.* (2011) Gene-based tests of association. *PLoS Genet.*, **7**, e1002177.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Kowalski,J. *et al.* (2002) A nonparametric test of gene region heterogeneity associated with phenotype. *J. Am. Stat. Assoc.*, **97**, 398–408.

Kryukov,G. *et al.* (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727–739.

Lee,S. *et al.* (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

Li,B. and Leal,S. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Liu,D.J. and Leal,S.M. (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, e1001156.

Liu,J. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139.

Madsen,B. and Browning,S. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Manolio,T. *et al.* (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.*, **118**, 1590.
Morgenthaler,S. and Thilly,W. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutat. Res.*, **615**, 28–56.
Neale,B. and Sham,P. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.*, **75**, 353–362.
Nejentsev,S. *et al.* (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387.
Olson,K.L. *et al.* (2005) Real time spatial cluster detection using interpoint distances among precise patient locations. *BMC Med. Inform. Decis. Mak.*, **5**, 19.
Pillai,S. *et al.* (2009) A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.*, **5**, e1000421.
Pritchard,J. and Cox,N. (2002) The allelic architecture of human disease genes: common disease-common variant . . . or not? *Hum. Mol. Genet.*, **11**, 2417.
Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
Regan,E. *et al.* (2011) Genetic epidemiology of COPD (COPDGene) study design. *COPD*, **7**, 32–43.
Servin,B. and Stephens,M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.*, **3**, e114.
Visscher,P. *et al.* (2008) Heritability in the genomics eraconcepts and misconceptions. *Nat. Rev. Genet.*, **9**, 255–266.
Wang,K. *et al.* (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.
White,L. *et al.* (2009) The choice of the number of bins for the M statistic. *Comput. Stat. Data Anal.*, **53**, 3640–3649.
Wilk,J.B. *et al.* (2009) A genome-wide association study of pulmonary function measures in the Framingham heart study. *PLoS Genet.*, **5**, e1000429.
Wu,M. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
Wu,T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.