Published in final edited form as:

Biometrics. 2015 December ; 71(4): 1111–1120. doi:10.1111/biom.12331.

Empirical Bayes Scan Statistics for Detecting Clusters of Disease Risk Variants in Genetic Studies

Kenneth J. McCallum and Iuliana Ionita-Laza*

Department of Biostatistics, Columbia University, New York, NY 10032, USA

Summary

Recent developments of high-throughput genomic technologies offer an unprecedented detailed view of the genetic variation in various human populations, and promise to lead to significant progress in understanding the genetic basis of complex diseases. Despite this tremendous advance in data generation, it remains very challenging to analyze and interpret these data due to their sparse and high-dimensional nature. Here we propose novel applications and new developments of empirical Bayes scan statistics to identify genomic regions significantly enriched with disease risk variants. We show that the proposed empirical Bayes methodology can be substantially more powerful than existing scan statistics methods especially so in the presence of many non-disease risk variants, and in situations when there is a mixture of risk and protective variants. Furthermore, the empirical Bayes approach has greater flexibility to accommodate covariates such as functional prediction scores and additional biomarkers. As proof-of-concept we apply the proposed methods to a whole-exome sequencing study for autism spectrum disorders and identify several promising candidate genes.

Keywords

Empirical Bayes; Next-generation sequencing; Rare variants; Scan statistics

1. Introduction

Recent progress in massively parallel sequencing technologies allows investigators to effectively obtain genetic information down to single base resolution on a genome-wide scale (Metzker, 2010). The resulting datasets are high dimensional and sparse, with most of the variants being observed only a small number of times (e.g. many of them are seen only once). This sparse nature of the data poses nontrivial statistical and computational difficulties (Ionita-Laza et al., 2013).

Both empirical and theoretical studies suggest that rare genetic variants can play an important role in influencing risk to disease (Kryukov et al., 2009). Over the past few years several statistical tests have been proposed to test for association with rare variants in a gene

*ii2135@columbia.edu.

Supplementary Materials. Web appendices, tables and figures referenced in sections 2.4, 2.6, 2.7, 2.8, 3.1, and 3.2, and source code for the eBScan software used here are available with this paper at the Biometrics website on Wiley Online Library.

(Li and Leal, 2008; Ionita-Laza et al., 2011; Wu et al., 2011; Lee et al., 2012; Ionita-Laza et al., 2013). These tests are based on the idea of grouping together variants in the gene, and testing for association at the gene rather than variant level. These tests are known as self-contained tests, i.e. they test for disease association with variants in a contained genetic region (e.g. a gene). The focus in this paper is on an alternative type of tests, namely competitive tests. Unlike self-contained tests, competitive tests test whether the association signal is stronger within a specific region compared with the association signal outside the region. Such competitive tests are of interest when association signal clusters in specific small regions, and one is interested in identifying the location of the cluster.

We describe here new cluster detection tests that test for significant clustering of rare disease risk variants (DRVs) in a larger genetic region. Unlike the self-contained tests, the goal here is to find the window within the larger region where DRVs cluster in a significant way compared with the rest of the region, i.e. the outside of the window. Clustering of disease mutations in small regions of a gene has been reported before for some Mendelian diseases (Robertson et al., 2003), and more recently for several complex traits as well. For example, a recent study focused on somatic cancer mutations found that disease-causing mutations cluster more often than random mutations, suggesting that mutation hotspots occur at the domain level (Yue et al., 2010). Furthermore, many large genomic regions (such as linkage regions, copy number variable regions or CNVRs) have been associated with various diseases, such as microdeletions on chromosome 22q11.2 strongly linked to schizophrenia risk (Liu et al., 2002). However these regions are large and tend to comprise many genes; the true underlying disease gene(s) in these regions have, for the most part, not been identified. For such large regions, it may be the case that only a small subregion (i.e. a gene within the CNVR) is associated with the disease, and here we discuss approaches to identify these small windows enriched in DRVs that reside in regions previously implicated in disease.

In general, significant clustering of events can be detected with scan statistics which use the maximum value of a test statistic in a sliding window approach (Naus, 1965). Scan statistics were originally developed for situations where the location of events were known with certainty (for example, in applications to find disease epidemics, an event is a disease case and can be observed directly); however, in our scenario it is not known whether a given genomic variant is a DRV or not. One approach is to use p-values from individual variant association tests and a threshold for significance to divide the dataset into two independent Poisson processes (Asimit et al., 2009). However, such a method does not work for our scenario since testing of individual variant association is underpowered for rare variants. A scan statistic based on the likelihood ratio test (LRT) has been proposed before in order to detect clusters of rare DRVs (Ionita-Laza et al., 2012, 2014). The LRT scan statistic aggregates variants in a scanning window. Empirical Bayes scan statistics are proposed here that seek to improve on the LRT by modeling the effect of variants with a beta distribution prior, this way allowing for heterogeneity of effects across variants.

Overview of the empirical Bayes approach

In large scale genetic association studies, allele counts for cases and controls are observed for numerous variants. Information about the association between each variant and the phenotype of interest is contained within these allele counts. However, for most variants the counts are small and so making inferences based on individual variants is difficult if not impossible. Our previous LRT approach pools information across variants in a region, essentially assuming that each variant in the region has the same effect. The motivation for the empirical Bayes approach is that it offers a compromise, pooling information from across variants while preserving information about individual variants. Specifically, the strength of the association between a variant and the phenotype is considered to be a random quantity governed by a prior distribution. The empirical Bayes approach uses the full set of variants to estimate this prior.

2. Methods

2.1 Notations

We assume that we have sequencing data on individuals in two distinct groups. The case group consists of individuals who have been diagnosed with a particular disease, such as autism spectrum disorder, while the control group consists of random individuals from the population (or possibly screened not to have the disease). For each individual i we observe

$$\mathbf{D}_i = \{s_i, \mathbf{G}_i\},$$

where s_i is an indicator variable for the phenotypic condition (i.e. case or control) of individual i and $\mathbf{G}_i = (g_{i1}, \dots, g_{iv})$ is a vector of variables g_{ik} for each of the v variants detected across all subjects. Define $g_{ik} = 0$ if individual i is homozygous for the major allele at variant k , $g_{ik} = 1$ if the individual is heterozygous, and $g_{ik} = 2$ if the individual is homozygous for the minor allele. The data can be summarized by the vectors

$$\mathbf{N} = (N_1, \dots, N_v) \quad \text{and} \quad \mathbf{Y} = (Y_1, \dots, Y_v),$$

where N_k is the sum of the g_{ik} over all i (i.e. number of minor alleles across all individuals), and Y_k is the sum of the g_{ik} over all i such that $s_i = 1$ (i.e. number of minor alleles for cases only). Although we focus the presentation on the case-control design, we note here that the extension to trio design is straightforward. For a trio design, N_k is the number of heterozygous parents at variant k , and Y_k is the number of times the minor allele is being transmitted to the offspring from a heterozygous parent (Ionita-Laza et al., 2014).

A scan statistic will be used to search for clusters of disease risk variants in a genomic region of interest where there is a strong indication that DRVs might cluster (e.g. a known CNVR, linkage region, or even a single gene). Let $\mathbf{R} = \{k : 1 \leq k \leq L\}$ be a genomic region consisting of L consecutive positions. For $1 \leq i \leq j \leq L$, define the window $\mathbf{W}_{ij} = \{k : i \leq k \leq j\}$ and its complement $\mathbf{W}_{ij}^c = \mathbf{R} \setminus \mathbf{W}_{ij}$. We want to test the hypotheses

H_0 : No clustering of disease risk variants occurs

vs

H_1 : For some i, j there is a cluster of disease risk variants in window $\mathbf{W}_{ij}$.

More precise hypotheses are given in the next section. For now it is important to note the distinction between a cluster of variants and a cluster of disease risk variants. A cluster of variants would indicate a region with an unusually large number of variants given its physical size. In contrast, a cluster of disease risk variants is a region with an unusually large number of DRVs given the total number of variants in the region. A cluster of either type does not imply the existence of a cluster of the other type. Clusters of DRVs are the subject of this study.

2.2 The Likelihood Ratio Test

A scan statistic for clusters of DRVs based on the likelihood ratio test (LRT) was previously proposed in (Ionita-Laza et al., 2012), and it is briefly described here. Given N_k , the distribution of Y_k is modeled as

$$Y_k | N_k, p_k \sim \text{Binom}(N_k, p_k), \quad (1)$$

where the binomial probability parameter p_k is a measure of the association between variant k and the phenotypic trait of interest. If C_1 is the number of cases and C_2 the number of controls in the sample, then $p_k = C_1 / (C_1 + C_2)$ would indicate no association, while $p_k > C_1 / (C_1 + C_2)$ indicates a positive association.

Let $n_G = \sum_k N_k$ and $y_G = \sum_k Y_k$. For the given window $\mathbf{W}_{ij}$ let $n_{ij} = \sum_{k \in \mathbf{W}_{ij}} N_k$ and $y_{ij} = \sum_{k \in \mathbf{W}_{ij}} Y_k$. If $p_k = p_g$ is constant across $k \in \mathbf{R}$, then the global rate can be estimated by

$$\hat{p}_g = y_G / n_G.$$

However, if a cluster occurs then for some window $\mathbf{W}_{ij}$, corresponding to that cluster, we expect the rate parameters to differ between variants inside and outside of $\mathbf{W}_{ij}$. Let p_{ij} denote the rate in $\mathbf{W}_{ij}$ and q_{ij} the rate outside $\mathbf{W}_{ij}$. These are estimated by

$$\hat{p}_{ij} = y_{ij} / n_{ij} \quad \text{and} \quad \hat{q}_{ij} = (y_G - y_{ij}) / (n_G - n_{ij}).$$

The likelihood ratio for window $\mathbf{W}_{ij}$ can then be defined as

$$S_{ij} = \left(\frac{\hat{p}_{ij}}{\hat{p}_G} \right)^{y_{ij}} \left(\frac{1 - \hat{p}_{ij}}{1 - \hat{p}_G} \right)^{n_{ij} - y_{ij}} \left(\frac{\hat{q}_{ij}}{\hat{p}_G} \right)^{y_G - y_{ij}} \left(\frac{1 - \hat{q}_{ij}}{1 - \hat{p}_G} \right)^{(n_G - n_{ij}) - (y_G - y_{ij})}. \quad (2)$$

If windows are limited to the form $\mathbf{W}_{i,i+w}$ for some fixed window size w and $i \in \mathbf{R}$, then a scan statistic based on the likelihood ratio is given by

$$S = \max_i S_{i,i+w}. \quad (3)$$

The null distribution of S is determined by repeatedly permuting the case-control labels and recalculating S .

The empirical Bayes methodologies described in the next section extend this model by allowing the binomial success parameters to vary within windows so that for each variant i , p_i is a random variable with a distribution that depends on whether or not the window is part of a disease risk variant cluster.

2.3 The Empirical Bayes Model

As above, given N_k , the distribution of Y_k is modeled as

$$Y_k | N_k, p_k \sim \text{Binom}(N_k, p_k), \quad (4)$$

where the binomial probability parameter p_k is a measure of the association between variant k and the phenotypic trait of interest. The strength and direction (i.e. risk or protective) of the association for each variant is modeled as

$$p_k \sim \text{Beta}(\alpha, \beta). \quad (5)$$

A useful feature of the beta-binomial distribution is that it can assume bimodal forms. If $\alpha, \beta < 1$, then probability mass accumulates around the extremes at 0 and 1. Because of this, the empirical Bayes model can capture a mixture of risk and protective variants in the DRVs. Since this depends only on the parameter estimates, the presence or absence of protective variants in the model is determined automatically when the model is trained.

Given (4) and (5), the distribution of Y_k can be written in terms of α and β as

$$\begin{aligned} P(Y_k | N_k, \alpha, \beta) &= \int_0^1 \binom{N_k}{Y_k} p_k^{Y_k} (1-p_k)^{N_k-Y_k} \frac{p_k^{\alpha-1} (1-p_k)^{\beta-1}}{B(\alpha, \beta)} dp_k \\ &= \binom{N_k}{Y_k} \frac{B(Y_k + \alpha, N_k - Y_k + \beta)}{B(\alpha, \beta)}, \end{aligned} \quad (6)$$

the beta-binomial distribution, where $B(\cdot, \cdot)$ is the beta function.

The parameters control the distribution of p_k and so (5) implies that all of the variants are drawn from a common population. We would like to model the scenario in which variants fall into one of two distinct subpopulations: disease risk variants (DRVs) which have a causal association with the phenotype, and non-DRVs. This can be done by replacing (5) with a mixture of beta distributions

$$p_k \sim \pi \text{Beta}(\alpha_1, \beta_1) + (1 - \pi) \text{Beta}(\alpha_2, \beta_2), \quad (7)$$

where $0 < \pi < 1$ is the mixture probability. The observations are then modeled as

$$P(Y_k|N_k, \theta) = \pi \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_1, N_k - Y_k + \beta_1)}{B(\alpha_1, \beta_1)} + (1 - \pi) \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_2, N_k - Y_k + \beta_2)}{B(\alpha_2, \beta_2)}, \quad (8)$$

where $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \pi)$.

In scan statistics the null hypothesis is that the data follows a single distribution across the entire region, while the alternative is that for some subregion $\mathbf{W}_{ij}$ the distribution in $\mathbf{W}_{ij}$ is different from that in $\mathbf{W}_{ij}^c$. In other words, under the null hypothesis that no clustering occurs, θ is constant across all variants. If the pairs (Y_k, N_k) are assumed to be independent across k , then the null distribution of the data is given by

$$P(\mathbf{Y}|\mathbf{N}, \theta) = \prod_k \left[\pi \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_1, N_k - Y_k + \beta_1)}{B(\alpha_1, \beta_1)} + (1 - \pi) \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_2, N_k - Y_k + \beta_2)}{B(\alpha_2, \beta_2)} \right]. \quad (9)$$

Under the alternative hypothesis that clustering occurs, the region can be partitioned into two sets $\mathbf{W}$, a window of consecutive base pairs, and $\mathbf{W}^c$ such that variants are more likely to be DRVs inside $\mathbf{W}$ than in $\mathbf{W}^c$. This can be modeled by allowing π to have different values between the two sets so that the alternative distribution is given by

$$\begin{aligned} P(\mathbf{Y}|\mathbf{N}, \theta, \mathbf{W}) &= \prod_{k \in \mathbf{W}} \left[\pi_1 \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_1, N_k - Y_k + \beta_1)}{B(\alpha_1, \beta_1)} + (1 - \pi_1) \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_2, N_k - Y_k + \beta_2)}{B(\alpha_2, \beta_2)} \right] \\ &\times \prod_{k \notin \mathbf{W}} \left[\pi_2 \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_1, N_k - Y_k + \beta_1)}{B(\alpha_1, \beta_1)} + (1 - \pi_2) \binom{N_k}{Y_k} \frac{B(Y_k + \alpha_2, N_k - Y_k + \beta_2)}{B(\alpha_2, \beta_2)} \right]. \end{aligned} \quad (10)$$

More precisely, if each variant k is considered to have a distribution characterized by $P(Y_k|N_k, \theta_k)$, then the hypotheses are

$$H_0: \forall k, \theta_k = \theta_0$$

vs

$$H_1: \text{For some } \mathbf{W}_{ij}, \theta_k = \theta_1 \text{ if } k \in \mathbf{W}_{ij} \text{ and } \theta_k = \theta_2 \text{ otherwise, } \theta_1 \neq \theta_2.$$

Two distinct models are considered. In EB-SCAN1 the null hypothesis is that $\forall k$

$$\theta_k = \theta_0 = (\alpha_1, \beta_1, \alpha_2, \beta_2, 0)$$

which is equivalent to saying that the data follows the distribution in (6) with parameters α_2 and β_2 . The alternative hypothesis is that for some $\mathbf{W}_{ij}$

$$\theta_k = \begin{cases} \theta_1 = (\alpha_1, \beta_1, \alpha_2, \beta_2, 1) & \text{if } k \in \mathbf{W}_{ij} \\ \theta_2 = (\alpha_1, \beta_1, \alpha_2, \beta_2, 0) & \text{otherwise} \end{cases}$$

This still assumes the data follows (6), but with different parameter values in and out of the cluster. This model assumes that under the null all variants are drawn from the same population and that under the alternative the cluster boundaries exactly separate the DRVs and non-DRVs. Note that under this model three separate pairs of values exist for (α, β) , one for each of $\theta_0, \theta_1, \theta_2$.

In EB-SCAN2 the null hypothesis is that $\forall k$

$$\theta_k = \theta_0 = (\alpha_1, \beta_1, \alpha_2, \beta_2, \pi)$$

while the alternative hypothesis is that for some $\mathbf{W}_{ij}$

$$\theta_k = \begin{cases} \theta_1 = (\alpha_1, \beta_1, \alpha_2, \beta_2, \pi_1) & \text{if } k \in \mathbf{W}_{ij} \\ \theta_2 = (\alpha_1, \beta_1, \alpha_2, \beta_2, \pi_2) & \text{otherwise} \end{cases}$$

where $\pi_1 \neq \pi_2$. This assumes that a mixture of DRVs and non-DRVs exists across the entire region and only the mixture proportion varies between the cluster and non-cluster variants. Specifically, (α, β) is assumed to be the same across $\theta_0, \theta_1, \theta_2$ (unlike for the first model), while π varies across the three parameter sets.

2.4 Parameter Estimation

The prior parameters for both the null and alternative distributions are estimated from the data. For EB-SCAN1 parameter estimation is relatively straightforward. The estimation procedure for EB-SCAN2 is slightly more complicated since it assumes a two component mixture, with one component corresponding to DRVs and one to non-DRVs. Full details are given in Web-based Appendix A.

2.5 Hypothesis Testing

For a given window $\mathbf{W}$ the empirical Bayes factor can be calculated as

$$S_{\mathbf{W}} = \frac{P(\mathbf{Y} | \mathbf{N}, \hat{\theta}_1, \hat{\theta}_2, \mathbf{W})}{P(\mathbf{Y} | \mathbf{N}, \hat{\theta}_0)}, \quad (11)$$

where $\hat{\theta}_0, \hat{\theta}_1$, and $\hat{\theta}_2$ are the estimates of the parameters under the null and alternative hypothesis respectively. For any pair of genomic coordinates $i \leq j$ we can define a window $\mathbf{W}_{ij} = \{k : i \leq k \leq j\}$. In practice, limiting the form of windows reduces the computational burden with little loss of power. Windows of the form $\mathbf{W}_{i, i+w}$ are considered, where w is integer valued, and represents an a priori fixed window size, with $1 \leq i \leq L - w$.

Let $S_{i,w}$ be the empirical Bayes factor for window $\mathbf{W}_{i,i+w}$. The test statistic for fixed window size w is then $S_w = \max_i S_{i,w}$. The distribution of S_w will be approximated using a permutation method described below.

2.6 Window Size

Both the empirical Bayes tests and the LRT compare the distribution of observed variants inside a sliding window to those outside the window and will have the greatest power when the size of the sliding window is approximately equal to the size of the true cluster. Since cluster sizes are not known a priori, it is not possible to pick a single optimal window size; however, multiple window sizes, $\mathbf{W} = w_1, \dots, w_m$, can be used and $S = \max_{w \in \mathbf{W}} S_w$ can be employed as the test statistic. More discussion on the choice of possible window sizes can be found in Web-based Appendix A.

2.7 Permutation Tests

Permutation tests can be conducted in either of two ways. One is to permute the case-control labels and the other is to permute the coordinates of variants. The two approaches test different null hypotheses. More details are given in the Web-based Appendix A. In our simulation studies coordinate permutation failed to control the type I error rate at the appropriate level, regardless of whether DRVs were included in the simulation or not. For this reason, detailed results given below are for the case-control permutation only.

2.8 Incorporation of information on functional predictions

There exist numerous variant functional annotations that can predict the prior likelihood that a variant is a DRV. For example, the PolyPhen-2 score (Adzhubei et al., 2010) estimates the probability that a non-synonymous variant will be damaging based on models of protein structure and function. Evolutionary conservation score systems such as GERP++ (Davydov et al., 2010) help determine whether variants are within regions that are evolutionarily conserved. Incorporating such rich functional information into the model may allow for both a better model fit and improved detection of DRVs clusters. The models proposed here are flexible and could be extended to incorporate such information. A full discussion of the proposed method is included in Web Appendix A.

3. Results

3.1 Simulation Studies

We performed simulation studies to estimate the size and power of the empirical Bayes tests, and compare with the existing LRT method, and several self-contained tests as well (SKAT and Burden tests (Lee et al., 2012)). The COSI (Schaffner et al., 2005) software package was used to simulate variants in a 1 Mb genomic region for 10,000 haplotypes. The model used in the simulations was calibrated for the European population. Subregions of size 50kb were sampled from the 1 Mb region.

For the power simulations, 10kb clusters were chosen with random start positions in the 50kb regions. Outside the cluster boundaries variants with minor allele frequency (MAF) < 1% were designated as DRVs with probability 0% or 1%. Inside the clusters variants were

designated as DRVs with probabilities of 5%, 10%, 30% or 50%. For each scenario 1000 simulations were run to assess power. For the type 1 error simulations, we simulated a scenario with no DRVs. For estimates of type I error rate 10,000 simulations were used.

The disease risk d_i for individual i was calculated as

$\text{logit}(d_i) = \beta_0 + \sum_{j \in \{DRV_s\}} g_{ij} c |\log_{10} MAF_j|$, where summation is over DRVs, g_{ij} is equal to the number of copies of variant j possessed by individual i , $c = 0.549$ (corresponding to a relative risk of 5.19 for a DRV with MAF 0.001), and MAF_j is the minor allele frequency for variant j . DRVs were chosen among variants with MAF less than 0.01, as explained above. Phenotype status was assigned as a Bernoulli random variable with probability d_i , and individuals were generated until the desired number of cases and controls was obtained.

Datasets with a mixture of protective and deleterious variants were also simulated. The procedure was similar to that described above, except that the sign of g_{ij} was switched from positive to negative for a randomly selected subset of 20% and 50% of DRVs.

As discussed before, power depends on the length of the windows used to scan for clusters. We use window sizes $\mathbf{W} = \{5\text{kb}, 10\text{kb}, 15\text{kb}\}$ to scan for clusters in the 50kb regions.

Power and type 1 error—The type I error rate is estimated from simulations that assume no DRVs exist in the region. All three scan statistics methods control type I error at the nominal level under this null hypothesis. Results are summarized in Web Appendix B table 1.

Results for power from the simulations at level $\alpha = 0.01$ are shown in figure 1. EB-SCAN1 is more powerful than either of the other scan statistic methods in every configuration of the simulations. This advantage is substantial except in the cases where all methods have power greater than 0.9. EB-SCAN1 does particularly well when the proportion of DRVs in a region is small (5%-10%) and when there is a mixture of risk and protective variants, as we expect based on theoretical grounds. In most cases, the LRT is the second most powerful method. When the percentage of causal variants in the cluster is large and no DRVs occur outside the cluster, EB-SCAN2 has a slight advantage over the LRT; however, it has a weak performance overall. This may be a result of inaccurate estimates of the mixture model components. Simulation studies have shown that component parameter estimates are biased for mixtures of normal distributions (Lourens et al., 2013), and the same is likely true for the mixture of beta-binomials used in EB-SCAN2. In EB-SCAN1, all model parameters are re-estimated for each window. In contrast, for EB-SCAN2 the component distributions are initially fit to the full data, and then only the mixing proportions are re-estimated for each window. If the initial component distributions are dominated by neutral variants, then it may be that there is no mixing proportion that yields a good fit to the window containing the DRV cluster, reducing the power of EB-SCAN2.

We have also compared power of the scan statistic approaches against that of self-contained tests such as SKAT and Burden. Specifically, using the same sliding window approach as used for the scan statistics, we calculate the score statistics for SKAT and Burden tests for each window, and the resulting p values, and take as our test statistic the minimum p value

over all scanning windows. The statistical significance of the minimum p value is assessed using permutations. Results of these power comparisons are shown in figure 1. Overall, EB-SCAN1 has better power than SKAT and Burden when the percentage of DRVs within the cluster region is higher (30% or 50%) and the percentage of protective variants is small (0% or 20%), while the SKAT test tends to become slightly more powerful as the percentage of DRVs in the cluster region decreases (5% or 10%), and as the percentage of protective variants also increases (50%). Furthermore, we evaluated the overlap between the estimated cluster window and the true (simulated) cluster window for both the scan statistic approaches and the self-contained tests, and noticed that the scan statistic methods have much better performance in estimating the location of the underlying cluster compared with the self-contained tests (table 1 and Web Appendix B table 2).

3.2 Applications to Sequencing data on Autism Spectrum Disorders

We have applied the two empirical Bayes tests and the LRT to a whole exome sequencing dataset with a combined sample size of 861 subjects, 488 autism spectrum disorder (ASD) cases and 373 controls of European ancestry. This dataset has been sequenced as part of the ARRA Autism Sequencing Collaboration (AASC), and is publicly available through dbGAP. Whole-exome sequencing was performed at the Broad Institute using an Illumina HiSeq2000 platform. Data was processed with Picard (DePristo et al., 2011) and BWA (Li and Durbin, 2010) to map reads to hg19. Variants were called using the Genome Analysis Toolkit (McKenna et al., 2010) and only those variants that passed standard quality control filters were analyzed.

We are interested in applying the proposed methods to 163 ASD candidate genes and recurrent copy number variable regions (CNVRs) implicated in autism by previous studies; each of these regions is treated as a separate region of interest for testing purposes. The precise genes within these CNVRs that influence risk are for the most part unknown because these large CNVRs encompass many genes. Furthermore, so far these CNVRs have not been comprehensively investigated using sequencing data. We apply the proposed scan-statistic methods to this WES study for ASD to elucidate the disease gene(s) within these CNVRs. The rationale is that although the frequency of the CNVs in these regions is low (most of them are *de novo*) and the individuals in the sequencing study are known to be non-carriers of CNVs in these regions, it is reasonable to expect that with the availability of whole-exome sequencing data rare (single nucleotide) variants within these CNVRs will be identified in additional cases. Only variants that were categorized as non-synonymous, splice-site, or stop gained/lost were included in the analysis. The regions were scanned using variable window sizes ranging from 5kb to 30kb.

Several regions could not be analyzed due to size constraints. Regions where the distance between the first and last variant was less than 5kb, the smallest scanning window size, were omitted. Furthermore, some large regions were omitted because the analysis was too computationally intensive, with the largest region included in the study being approximately 19Mb. Finally, some of the regions had substantial overlap, effectively duplicating one another. Consequently, of the 163 regions only 121 were analyzed. Tables summarizing results for the significant regions are contained in Web Appendix B (tables 3, 4, and 5). The

LRT finds 7 of them to have significant clusters at the 0.05 level, while EB-SCAN1 finds 8 and EB-SCAN2 finds 7. A total of 12 significant, distinct regions are found. Of those, 3 are found by all three methods, 1 is unique to EB-SCAN1, 2 are unique to EB-SCAN2, and 2 are unique to the LRT. Figure 2 summarizes the overlap between the methods and lists the genes covered by the significant clusters for each method.

Several of the genes identified in these regions using the scan statistic approaches are interesting. For example, *RNF135* is detected by all three methods, and mutations in this gene have been previously identified in families with learning disability, dysmorphic features and overgrowth (Douglas et al., 2007). Figure 3 shows the log of the Bayes factors and likelihood ratio for each 5kb window of the region containing *RNF135*. As shown, the highest scoring window is the same for both the LRT and the empirical Bayes approaches. Similarly, mutations in *SYNGAP1*, a gene identified by both LRT and EB-SCAN1, have been found in individuals with intellectual disability and autism (Berryer et al., 2013). These results illustrate the ability of the proposed methods to identify known disease associated genes, and serve as proof-of-concept for application to larger studies.

4. Discussion

Here we have proposed empirical Bayes scan statistics to identify genomic regions where disease risk variants cluster in a significant way. The proposed methods can be applied to any genomic region where DRVs are expected to cluster in small windows (e.g. CNVRs). The empirical Bayes framework is natural and more flexible than the LRT framework, since it can allow for incorporation of variant covariates, such as functional scores for individual variants, and it can accommodate effect heterogeneity across variants, including both risk and protective variants. Compared to the LRT test, the empirical Bayes approach tends to be more powerful when there is a large number of null variants, and also when there is a mixture of risk and protective variants. This is likely due to the fact that the empirical Bayes methods make use of the likelihood of allele counts at individual variants, rather than just the mean across all variants in a window. Because of this, a small number of variants that have relatively little effect on the mean can still significantly influence the Bayes factor. Similarly, a mix of protective and risk variants that cancel out in the mean is still detectable in the empirical Bayes model.

An important application of these methods is to large and rare CNVRs that are strongly associated with various psychiatric traits, and for which the underlying disease genes are not known (Ionita-Laza et al., 2012, 2014). Here, we have applied the proposed methods to a whole exome sequencing dataset for autism, and identified several interesting findings. For example, genes *SYNGAP1* and *RNF135* are both strong candidate genes for autism and have been identified by both the LRT and EB-SCAN1 tests. Although the sample size of this study is modest, the application of the proposed methods to larger studies for autism and other complex traits has the potential to identify strong candidate genes for these diseases.

Singletons (variants seen only once) can play an important role in influencing risk to disease. For such variants the beta-binomial distribution is equivalent to a binomial distribution with $p = \alpha/(\alpha + \beta)$. Because of this the utility of the empirical Bayes tests

depends on the number of variants that are not singletons. As an extreme example, if all of the variants in a region are singletons then the empirical Bayes model is equivalent to the LRT.

The null hypotheses for the self-contained test and the clustering test are different. In competitive tests, such as the LRT and empirical Bayes tests discussed here, the null hypothesis being tested is that the association signal at the variants within a window matches that outside the window. This contrasts with self-contained tests, such as Burden and SKAT, which test the null hypothesis that there is no association within a specific window. Therefore, the alternative hypothesis for the competitive test implies the alternative hypothesis for the self-contained test, while the converse is not true. Although this may imply that the self-contained test will tend to have better power than a competitive test, this is not always the case, as we show in our simulation studies. In particular, this assumes that the self-contained test is at least as powerful as the competitive test when both alternative hypotheses are true, which is not necessarily true. Furthermore, an advantage that the competitive test has is that it is possible to specify the null distribution empirically, based on the data outside the scanning window. Because of this, it may be possible for the competitive test to have greater power when both alternative hypotheses are true. Furthermore, there may be situations where an investigator wants specifically to find regions where DRVs cluster, and in those cases the competitive tests are appropriate, and have better precision in localizing the true underlying cluster region, as we show in our simulation studies.

An important confounding factor in genetic association studies is population stratification. Although the proposed methods do not easily allow for covariate adjustment, once significant clustering has been detected in a window, it is possible to test for association with variants in the window using a regression-based test such as Burden or SKAT, with adjustment for covariates of interest. Furthermore, for population stratification, because the competitive test compares the distribution within a window with that outside the window it is likely that any biases caused by population substructure in the region would affect the whole region, rather than a small ‘cluster’ window.

There are several improvements that might be made to the empirical Bayes methodology in future work. As noted in the methods, it would be possible to model the parameters of the prior distributions as functions of covariates, allowing additional information (e.g. functional genomics annotations) to be incorporated into the model. The empirical Bayes method also has the advantage that it calculates the posterior probability that each variant is a DRV during the model training stage. These posterior probabilities could potentially be used to prioritize variants in the region of interest as disease risk variants. Additionally, the proposed methods search for a single contiguous cluster, and cannot detect multiple separate clusters of DRVs (e.g. multiple genes in a CNVR). In such cases, Hidden Markov Models could potentially be used to segment the data into regions enriched for DRVs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We gratefully acknowledge support by NSF grant DMS-1100279 and NIH grant R01MH095797 (II-L). We thank Joseph Buxbaum for helpful comments on the ASD application.

References

- Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nature Methods*. 2010; 7:248–249. [PubMed: 20354512]
- Asimit J, Yoo Y, Waggott D, Sun L, Bull S. Region-based analysis in genome-wide association study of framingham heart study blood lipid phenotypes. *BMC Proceedings*. 2009; 3:S127. [PubMed: 20017993]
- Berryer M, Hamdan F, Klitten L, Mazller R, Carmant L, Schwartzentruber J, et al. Mutations in *syngap1* cause intellectual disability, autism, and a specific form of epilepsy by inducing haploinsufficiency. *Human Mutation*. 2013; 34:385–394. [PubMed: 23161826]
- Davydov E, Goode D, Sirota M, Cooper G, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using *gerp++*. *PLoS computational Biology*. 2010; 6:e1001025. [PubMed: 21152010]
- DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. a framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*. 2011; 43:491–498. [PubMed: 21478889]
- Douglas J, Cilliers D, Coleman K, Tatton-Brown K, Barker K, Bernhard B, et al. Mutations in *rnf135*, a gene within the *nf1* microdeletion region, cause phenotypic abnormalities including overgrowth. *Nature Genetics*. 2007; 39:963–965. [PubMed: 17632510]
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics*. 2011; 7:e1001289. [PubMed: 21304886]
- Ionita-Laza I, Cho M, Laird N. Statistical challenges in sequence-based association studies with population-and family-based designs. *Statistics in Biosciences*. 2013; 5:54–70.
- Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*. 2013; 92:841–853. [PubMed: 23684009]
- Ionita-Laza I, Makarov V, ARRA Autism Consortium, Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in *lrp2*, a gene linked and associated with autism spectrum disorders, in three datasets. *American Journal of Human Genetics*. 2012; 90:1002–1013. [PubMed: 22578327]
- Ionita-Laza I, Xu B, Makarov V, Buxbaum J, Roos J, Gogos J, et al. A scan-statistic based analysis of exome sequencing data identifies *fan1* at 15q13.3 as a susceptibility gene for schizophrenia and autism. *Proceedings of the National Academy of Science*. 2014; 111:343–348.
- Kryukov G, Shpunt A, Stamatoyannopoulos J, Sunyaev S. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Science*. 2009; 106:3871–3876.
- Lee S, Wu M, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012; 13:762–775. [PubMed: 22699862]
- Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*. 2008; 83:311–321. [PubMed: 18691683]
- Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*. 2010; 26:589–595. [PubMed: 20080505]
- Liu H, Abecasis G, Heath S, Knowles A, Demars S, Chen Y, et al. Genetic variation in the 22q11 locus and susceptibility to schizophrenia. *Proceedings of the National Academy of Science*. 2002; 99:16859–16864.

- Lourens S, Zhang Y, Long J, Paulsen J. Bias in estimation of a mixture of normal distributions. *Journal of Biometrics and Biostatistics*. 2013
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*. 2010; 20:1297–1303. [PubMed: 20644199]
- Metzker M. Sequencing technologies - the next generation. *Nature Review Genetics*. 2010; 11:31–46.
- Naus JI. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*. 1965; 60:532–538.
- Robertson S, Twigg S, Sutherland-Smith A, Biancalana V, Gorlin R, Horn D, et al. Localized mutations in the gene encoding the cytoskeletal protein filamin a cause diverse malformations in humans. *Nature Genetics*. 2003; 33:487–491. [PubMed: 12612583]
- Schaffner S, Foo C, Gabriel S, Reich D, Daly M, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*. 2005; 15:1576–1583. [PubMed: 16251467]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*. 2011; 89:82–93. [PubMed: 21737059]
- Yue P, Forrest W, Kaminker J, Lohr S, Zhang Z, Cavet G. Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human Mutation*. 2010; 31:264–271. [PubMed: 20052764]

Figure 1.

Estimated power from simulations is shown for all three scan statistic methods, and the two self-contained tests: SKAT and Burden. The top row shows results with no DRVs outside of the cluster and the bottom row with 1% DRVs outside the cluster. From left to right, simulations had 0%, 20%, or 50% protective DRVs. The percentage of causal variants in each simulation was 5%, 10%, 30% or 50%. Power is based on the variable window size approach with window sizes of 5kb, 10kb, and 15kb. All simulations use regions of size 50kb and clusters of size 10kb.

Figure 2.

Results of applying the three scan statistic methods to CNVRs previously implicated in ASD. Genes overlapping with significant clusters from the ASD data are listed for each of the three models.

Figure 3.

Log Bayes factors and log likelihood ratio are plotted against the midpoint of each window for CNVR chr17: 29.165.874–30.215.887. Results are shown for scanning windows of size 5 kb. The highest scoring window for all three methods overlaps the *RNF135* gene.

Table 1

Jaccard index from simulations for the scan statistic methods and the two self-contained tests. Estimates are based on 1000 simulations for each scenario. 50kb regions are used with 10kb clusters. The percentage of variants inside the cluster that are DRVs ranges from 5% to 50%, while the percentage outside clusters is either 0% or 1%.

% Inside	% Outside	% Protective	EB-SCAN1	EB-SCAN2	LRT	SKAT	Burden
5	0	0	0.223	0.336	0.281	0.114	0.116
5	0	20	0.190	0.283	0.224	0.105	0.114
5	0	50	0.157	0.217	0.164	0.110	0.109
5	1	0	0.193	0.260	0.225	0.116	0.123
5	1	20	0.177	0.242	0.199	0.105	0.113
5	1	50	0.143	0.188	0.165	0.124	0.115
10	0	0	0.372	0.465	0.391	0.119	0.120
10	0	20	0.310	0.398	0.318	0.118	0.126
10	0	50	0.235	0.284	0.196	0.125	0.130
10	1	0	0.314	0.407	0.326	0.113	0.124
10	1	20	0.272	0.352	0.279	0.122	0.122
10	1	50	0.213	0.241	0.180	0.111	0.120
30	0	0	0.714	0.768	0.661	0.126	0.126
30	0	20	0.645	0.655	0.512	0.143	0.114
30	0	50	0.480	0.376	0.241	0.129	0.122
30	1	0	0.688	0.738	0.633	0.133	0.123
30	1	20	0.606	0.627	0.486	0.136	0.112
30	1	50	0.452	0.329	0.233	0.135	0.127
50	0	0	0.853	0.900	0.810	0.145	0.125
50	0	20	0.788	0.785	0.626	0.138	0.127
50	0	50	0.645	0.417	0.282	0.138	0.112
50	1	0	0.844	0.887	0.783	0.146	0.136
50	1	20	0.767	0.775	0.602	0.155	0.124
50	1	50	0.650	0.416	0.277	0.132	0.104