Published in final edited form as:

Genet Epidemiol. 2018 October ; 42(7): 673–683. doi:10.1002/gepi.22134.

A Subregion-based Burden Test for Simultaneous Identification of Susceptibility Loci and Sub-regions within

Bin Zhu¹, Lisa Mirabello¹, and Nilanjan Chatterjee^{2,3}

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA

²Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

³Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

Abstract

In rare variant association studies, aggregating rare and/or low frequency variants, may increase statistical power for detection of the underlying susceptibility gene or region. However, it's unclear which variants, or class of them, in a gene contribute most to the association. We proposed a subregion-based burden test (REBET) to simultaneously select susceptibility genes and identify important underlying sub-regions. The sub-regions are predefined by shared common biologic characteristics, such as the protein domain or functional impact. Based on a subset-based approach considering local correlations between combinations of test statistics of sub-regions, REBET is able to properly control the type I error rate while adjusting for multiple comparisons in a computationally efficient manner. Simulation studies show that REBET can achieve power competitive to alternative methods when rare variants cluster within sub-regions. In two case studies, REBET is able to identify known disease susceptibility genes, and more importantly pinpoint the unreported most susceptible sub-regions, which represent protein domains essential for gene function. R package REBET is available at <http://dceg.cancer.gov/tools/analysis/rebet>.

Keywords

burden test; disease susceptibility genes; rare variant association studies; subset-based approach

1 Introduction

There is an increasing focus on investigating the associations between rare variants and common complex diseases[Bansal, et al. 2010; Cirulli and Goldstein 2010] in the hope of explaining the “missing heritability”[Manolio, et al. 2009] that remains unexplained from genome-wide association studies (GWAS) of common variants. Advances in next-generation

Correspondences to: Bin Zhu PhD, 9609 Medical Center Drive, Room 7E618, Bethesda, MD 20892, Phone: 240-276-7420, bin.zhu@nih.gov.

Conflict of interest: The authors have nothing to disclose

sequencing technology, coupled with plummeting costs and improvements in bioinformatics tools and statistical analysis methods, enable examination of rare variants on a genome-wide scale. Recent studies have reported that rare variants contribute to the genetic susceptibility for a number of complex traits or diseases, including human adult height[Marouli, et al. 2017], lipid levels[Helgadottir, et al. 2016; Surakka, et al. 2015], autism[Krumm, et al. 2015], ischemic stroke[Auer, et al. 2015], prostate cancer[Mancuso, et al. 2016] and breast cancer[Cybulski, et al. 2015].

Detecting rare variant associations is statistically challenging due to the low minor allele frequency (MAF) and heterogeneous risk effects of rare variants. The power to detect a single rare variant, due to the sparsity of carries, is very low unless the sample size and/or effect size is extremely large[Derkach, et al. 2014; Moutsianas, et al. 2015]. Thus, aggregating rare variants in a gene or genomic region has been proposed to boost the power[Lee, et al. 2014], which however will negatively affect power due to heterogeneous effects. For example, rare variants in a gene could be deleterious, protective or neutral. Additionally, the aggregated rare variants can vary in their MAFs, effect sizes, genome-positions and functional consequences. To take these heterogeneous effects into account, some burden tests consider reweighting the effects of rare variants based on their MAFs or functional consequences[Madsen and Browning 2009; Morgenthaler and Thilly 2007]; variance component tests[Wu, et al. 2011] further allow both effect directions and magnitudes to vary across the gene. A limitation of these aggregated tests is that while they may have robust power for detecting a susceptibility gene containing clusters of causal variants, they do not identify which variants, or class of variants, in a gene contribute most to the association. It's possible to order rare variants by their marginally estimated effect sizes, but it's unclear how to make the cutoff to claim the significance of top ranking rare variants. For example, Sun and Wang [2014] proposed a post-hoc analysis after the aggregated test of a gene, to calculate a score as the weighted linear sum for all possible subsets of rare variant and it chooses the one with the maximum sample correlation between covariate-adjusted phenotypic outcomes and the calculated score. Although this method is able to rank susceptibility rare variants, the proposed procedure cannot be used to detect the significance of either subsets of variant or all variants in a gene as a whole.

Here, we propose a subREgion-based BurdEn Test (REBET) to simultaneously detect the rare variant association of a gene and identify the most susceptible sub-regions that drive the gene-level significant association. We first specify biologically meaningful sub-regions within a gene. The rare variants within each sub-region may share common biologic characteristics, such as functional domain or functional impact. REBET then searches all possible combinations of sub-regions, identifies the one with the strongest association signal through a selective subregion-based burden test, and assesses its statistical significance while adjusting for multiple tests involved in the sub-region search[Bhattacharjee, et al. 2012]. For detecting overall association for a gene, burden test is more powerful when risk effects are highly homogenous and variance component tests is more powerful when risk effects are highly heterogeneous; instead, REBET has robust power when risk effects are relatively homogeneous within sub-regions, but potentially heterogeneous across sub-regions. These scenarios are likely, as seen for two important cancer susceptibility genes, *TP53* and *BRCA1*

where deleterious variants are concentrated in specific protein functional domains [Petitjean, et al. 2007a; Rebbeck, et al. 2015].

We conducted simulation studies and two real data analyses to evaluate the performance of REBET. The simulation studies demonstrated that REBET properly controls the gene-level type I error rate and is more powerful than burden tests or variance component tests when risk effects indeed congregate in sub-regions. REBET has robust power even when the susceptible sub-regions include some neutral variants. We applied REBET to two real data applications: one focusing on the *ANGPTL4* gene in the Dallas Heart study [Romeo, et al. 2007] and the other on the *TP53* gene in the National Cancer Institute (NCI) sequencing studies of osteosarcoma [Mirabello, et al. 2015]. In both cases, REBET was capable of identifying the association between the gene and the disease to a level of significance similar to those from other competing methods. In addition, REBET pinpointed the unreported functional domains in which the rare variants drove the association. Although we demonstrated REBET for the testing of two targeted genes, we discussed how to choose coding and non-coding regions for a genome-wide analysis in discussion section.

2 Method

2.1 The Burden Test for The Whole Region

We first introduce the notation, model and the burden test for the whole region (e.g. a given gene), which will set the stage for the selective REBET proposed in the next subsection. For I unrelated subjects, we denote Y_i , $i = 1, 2, \dots, I$, as the trait or disease status of the i th subject with $\mathbf{Y} = (Y_1, Y_2, \dots, Y_I)'$ the vector of Y_i 's. Let $G_{ij} \in \{0, 1, 2\}$ denote the number of minor allele on the j th locus, $j = 1, 2, \dots, J$, among total J candidate loci in the whole region with $\mathbf{G}_i = (G_{i1}, G_{i2}, \dots, G_{iJ})'$. Similar, $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iM})'$ denote M confounding covariates.

We assume that the phenotype Y_i is related to the rare variants $\mathbf{G}_i$ and covariates $\mathbf{X}_i$ through a generalized linear model

$$g\{E(Y_i)\} = \mathbf{X}_i' \boldsymbol{\alpha} + \mathbf{G}_i' \boldsymbol{\beta},$$

for which $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_M)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_J)'$ are coefficients for covariates and rare variants, respectively. For continuous phenotypes like the height and the glucose level in blood, the link function $g(\cdot)$ is specified as $g\{E(Y_i)\} = E(Y_i)$, while $g\{E(Y_i)\} = \text{logit}\{P(Y_i = 1)\}$ for the binary disease status.

To test the null hypothesis H_0 that the phenotype and rare variants are independent, i.e. $\boldsymbol{\beta} = \mathbf{0}$, score statistics $\mathbf{S}$ can be derived as $\mathbf{S} = \mathbf{G}'(\mathbf{Y} - \hat{\boldsymbol{\mu}})$, for which $\mathbf{S} = (S_1, S_2, \dots, S_J)'$, $\mathbf{G} = (\mathbf{G}_1', \mathbf{G}_2', \dots, \mathbf{G}_I')$ and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_I)'$ with $\hat{\mu}_i = g^{-1}(\mathbf{X}_i' \hat{\boldsymbol{\alpha}})$, $\hat{\boldsymbol{\alpha}}$ the estimate of coefficients $\boldsymbol{\alpha}$ under the H_0 and $g^{-1}(\cdot)$ the inverse function of $g(\cdot)$. Correspondingly, the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{S}$ is of the form $\boldsymbol{\Sigma} = \mathbf{G}'[\hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{X}(\mathbf{X}'\hat{\mathbf{D}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{D}}]\mathbf{G}$, for which $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \dots, \mathbf{X}_I')$ and $\hat{\mathbf{D}} = \text{diag}(\hat{d}_1^2, \hat{d}_2^2, \dots, \hat{d}_I^2)$ with $\hat{d}_i^2 = \sum_{j=1}^J (Y_{ij} - \hat{\mu}_i)^2 / I$ and $\hat{d}_i^2 = \hat{\mu}_i(1 - \hat{\mu}_i)$ for the continuous phenotype and the binary disease status, respectively. Burden test considers the aggregated

effect of rare variants and weights the score statistics $\mathbf{S}$ by a weight vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_J)'$, leading to the weighted score statistic of the form

$$W_{burden} = \boldsymbol{\omega}'\mathbf{S} = \sum_{j=1}^J \omega_j S_j$$

for which weights $\omega_j = \text{beta}(MAF_j, a, b)$, a beta distribution with parameters a and b . Morgenthaler and Thilly [2007] treated $\omega_j = \text{beta}(MAF_j, 1, 1) = 1$ and Madsen and Browning [2009] considered $\omega_j = \text{beta}(MAF_j, 0.5, 0.5) = 1/\sqrt{MAF_j(1-MAF_j)}$ among others, MAF_j the minor allele frequency of the j th rare variant.

Due to low linkage disequilibrium between rare variants, we impose off-diagonal entries of $\boldsymbol{\Sigma}$ as zeros to simplify the computation of the selective REBET in the next section. Correspondingly, the weighted burden test statistic for the whole region is given as

$$Z_{burden} = \frac{\sum_{j=1}^J \omega_j S_j}{\left(\sum_{j=1}^J \omega_j^2 \sigma_j^2\right)^{1/2}} = \sum_{j=1}^J \frac{1}{\sqrt{\pi_j}} Z_j$$

where σ_j is the j th diagonal element of $\boldsymbol{\Sigma}$, $\pi_j = \frac{\omega_j^2 \sigma_j^2}{\sum_{j=1}^J \omega_j^2 \sigma_j^2}$ and $Z_j = \frac{S_j}{\sigma_j}$.

2.2 The Selective Subregion-based Burden Test

In contrast to the weighted burden test statistic Z_{burden} involving all rare variants in the whole region, the REBET would consider rare variants in sub-regions and identify the specific combination of sub-regions with the largest test statistic, including the whole region as a special case. Note here the basic unit tested for the association in BEBET is neither single rare variants nor the all variants in a gene, but instead a subset of variants located in sub-regions with possible homogeneous biological function. In the following, we will derive the REBET test statistic and outline the hypothesis testing procedure.

Suppose K non-overlapping sub-regions are pre-specified and the k th sub-region is denoted as R_k , $k = 1, 2, \dots, K$. The test statistic Z_{burden} for all J rare variants could be rewritten with

respect to the K sub-regions as $Z_{burden} = \sum_{k=1}^K \sqrt{\pi^k} Z^k$ with $\pi^k = \frac{\sum_{j \in R_k} \omega_j^2 \sigma_j^2}{\sum_{j=1}^J \omega_j^2 \sigma_j^2}$ and

$$Z^k = \sum_{j \in R_k} \frac{\omega_j S_j}{\left(\sum_{j \in R_k} \omega_j^2 \sigma_j^2\right)^{1/2}}, \text{ where } j \in R_k \text{ denotes the } j\text{th variant belonging to the sub-}$$

region R_k . There are $2^K - 1$ possible nonempty sub-region unions $\mathbf{U} = \{U_m; m = 1, 2, \dots, 2^K - 1\}$. Note that we consider not only K sub-regions but also their combinations as sub-region unions, since multiple sub-regions with different biological functional may contain the similar effect and aggregating over the relevant sub-regions would possibly enhance power.

For a sub-region union U_m , which includes multiple sub-regions R_k 's, denoted as $R_k \in U_m$, we derive a sub-region union burden test statistic $Z(U_m) = \sum_{R_k \in U_m} \sqrt{\pi^k(U_m)} Z^k$, in which $\pi^k(U_m) = \pi^k / \sum_{R_k \in U_m} \pi^k$ measuring the relative weight of π^k in the sub-region union U_m . When U_m includes all sub-regions, $Z(U_m)$ is equivalent to the Z_{burden} . In contrast to standard burden tests which consider all sub-regions only, REBET will calculate $Z(U_m)$ for each possible nonempty sub-region unions U and select the union of sub-regions with the maximal $Z(U_m)$ for which effect sizes of sub-regions are in the same direction. Specifically, to evaluate the null hypothesis that Y and G_j 's the variants located in a gene, are independent, we propose the absolute maximal REBET test statistic $Z_{max-REBET} = \max_{U_m \in \mathcal{U}} |Z(U_m)|$ as a one-sided test for sub-region unions with the same direction of effect. Note $Z_{max-REBET}$ is similar to the ASSET test statistic[Bhattacharjee, et al. 2012] with a key difference. Particularly, ASSET test statistic combines information on single SNP across different traits and then identify the traits with which the SNP is most significantly associated, while $Z_{max-REBET}$ combines information for multiple sub-regions in a gene for a single trait and then determines the sub-region union most significantly associated with the trait.

We use a subset-based approach[Bhattacharjee, et al. 2012] to evaluate the distribution of $Z_{max-REBET}$ and calculate its P value. The subset-based approach consists of two steps. It first obtains the multivariate normal distribution of test statistics $Z(U_m)$'s, $U_m \in \mathcal{U}$, with mean zeros, unit variance and covariance between sub-region union pairs. The subset-based approach then utilizes an analytic discrete local maxima (DLM) method[Taylor, et al. 2007] to estimate the tail probability of $Z_{max-REBET}$. DLM method exploits the local correlation between two neighboring sub-region unions, in which sub-regions involved are identical except for one sub-region. As a result, it achieves the computational efficiency and the accurate and conservative upper bound of P value, adjusted by multiple comparisons. More details and assumptions about the subset-based approach are given at Bhattacharjee et al[Bhattacharjee, et al. 2012]. Besides P value, the subset-based approach is able to find the most susceptible sub-region union of all possible $2^K - 1$ combinations or a restricted subset of them if further prior knowledge suggests so. When the number of sub-regions considered gets larger, the computation time would be exponentially increased. A fast subset-based approach has recently been developed to account for this computation issue by adding a pre-screening step and adjusting the inflated type I error rate.

3 Simulation Studies

3.1 Study Design

We conducted the simulation studies to evaluate the type I error of REBET and to compare REBET to other methods in terms of statistical power for detecting disease-gene associations (based on 10^5 replicates). The competing methods included two burden tests (CAST[Morgenthaler and Thilly 2007] and WSS[Madsen and Browning 2009]), two adaptive burden tests (aSUM[Han and Pan 2010] and VT[Price, et al. 2010]), one variance component test (SKAT[Wu, et al. 2011]) and two combined tests (SKAT-O[Lee, et al. 2012] and Mist[Sun, et al. 2013]). We also evaluated powers of alternative sub-region methods,

EBSCAN[McCallum and Ionita-Laza 2015], a scan statistic-based test, and application of SKAT to each sub-regions and pooling P-values by Fisher's method.

As illustrated in Table 1, eight scenarios were designed with various degree of clustering of risk variants within sub-regions of a gene. The hypothetical gene was assumed to consist of 5 sub-regions, each including 4 variants. For scenario A, all variants were risk variants with deleterious effects, which was in favor of (adaptive) burden tests; we then introduced strong clustering of risk variants, for example, where all variants in certain sub-regions were deleterious (sub-regions 1,3 and 5 in scenario B or sub-regions 1 and 5 in scenario C) and all variants in the remaining regions were neutral. Then we considered weaker clustering where some sub-regions contained mixture of deleterious and neutral variants (scenarios D and E). Additionally, we considered the case where some sub-regions contained a mixture of deleterious and neutral variants and others contained mixture of protective and neutral variants (scenario F). For evaluating the type-I error (based on 10^6 replicates), we considered the case when all variants in a gene were neutral in scenario G.

For each scenario, we investigated the REBET including the whole region without sub-regions selection and the REBET with sub-regions selection, respectively. Moreover, we tested the REBET by ignoring sub-region annotation and selecting variant singleton instead using a fast subset-based approach; we also carried out a sensitivity analysis of several popular choices of a and b for weights $\omega_j = \text{beta}(\text{MAF}_j, a, b)$ which might not be the same as the true ones. For Mist, we considered both the Mist ignoring sub-region annotation and the one incorporating sub-region annotation as the fixed effects, assuming that associated variants share some common biological characteristics specified by the sub-region. In contrast to REBET, Mist will not select the most significant sub-region union and does not require adjusting for multiple comparisons.

3.2 Evaluation and Comparison of Powers

Table 1 showed powers for various competing approaches under different scenarios. When all the variants were deleterious risk variants in scenario A, it was not surprising that the burden tests and the REBET for the whole region had the highest power, since they assumed all rare variants being associated with the trait under the alternative hypothesis, consistent with scenario A. In contrast, the alternative hypotheses of the SKAT, SKAT-O, Mist and REBET with sub-regions selection assumed the direction and distribution of risk effects different from the ones in scenario A, i.e. part of risk effects being deleterious while others being protective for the SKAT, SKAT-O and Mist, or the risk effects being clustered in sub-regions for the REBET with sub-regions selection. Consequently, their powers were reduced, most dramatically by the SKAT.

In scenarios B and C, the gene contained mixture of neutral and deleterious variants with the two types of variants perfectly clustering within sub-regions. Here, powers of burden tests and the REBET for whole region decreased noticeably. Although the powers of SKAT-O and Mist were less impacted, their powers were clearly lower than that by the REBET with sub-regions selection, which successfully considered the possible pattern of clustered risk variants and improved the power even with the adjustment of multiple comparisons.

Similarly, incorporating sub-region annotation would also improve the power by Mist but not as powerful as the REBET and could not pinpoint the most significant sub-region union.

When neutral and deleterious variants in the sub-regions clustered strongly but not perfectly (scenario D), the power of REBET with sub-region selection was higher compared to the other methods. When the clustering was weak (scenario E and F), the power of REBET was comparable or slightly lower relative to the other methods that were most powerful in the respective scenarios. Note that in scenario E, it included two sub-regions (the 2nd and 4th sub-region) with the same effect direction as the effect direction of the 1st and 5th sub-region and led to higher power compared to scenario F in which the effect direction of the 2nd and 4th sub-region was opposite to the effect direction of the 1st and 5th sub-region. Finally, when all variants were neutral, the REBET was able to properly control the type I error rate at the targeted level (scenario G).

As for the alternative sub-region methods, application of SKAT to each sub-regions and pooling P-values led to similar powers as applying SKAT to the whole region (Supplementary Table 1). Scan statistic-based test, EBSCAN, however performed poorly in our simulation setting. It's not surprising, since scan statistic-based tests tested if the association signal within a sub-region was much higher than the association in other sub-regions, while REBET compared the association signals between case and control group. Take scenario A as an example. Although rare variants were enriched in cases, the level of enrichment of rare variants in any sub-region was identical to other sub-regions and hence scan statistic-based tests would not be significant.

We also evaluated the power of REBET when individual variants were chosen as the selection unit, or additional prior constraint on the number of combinations was imposed (Table 1). First, the simulations demonstrated the necessity of selection on sub-regions instead of individual variants as singletons. In particular, selection on singletons led to very low power; this is not surprising given the low frequency of rare variants requires aggregating their effects in a gene to boost the power in the first place; in addition, considering all possible combinations of singleton would drastically increase the number of multiple comparisons to be adjusted and further restrained the power. Second, a noteworthy feature of REBET is the ability of imposing the constraint on the search of possible combinations to reduce the number of multiple comparisons and consequently improve the power. For example, when sub-regions represent exons, based on the a priori knowledge of the gene transcript isoforms, only a subset of exon combinations of a gene is possible to be transcribed to the mRNA as transcripts. To demonstrate the benefits of adding the constraint, in scenario B we specified a constraint that only cumulatively combining sub-regions 1, 3, 5, 2 and 4 sequentially was allowed. As a result, the possible number of combination of sub-regions was reduced from $2^5 - 1 = 31$ to 5 and the power increased from 0.909 to 0.936.

Finally, the simulations (results in Supplementary Table 2) illustrated the robustness of the choices of a and b for weights $\omega_j = \text{beta}(MAF_j, a, b)$. When the a and b used in the model were different from the ones simulating data ($a=0.5$ and $b=0.5$), the power would not be damped much compared to the one using the true a and b .

3.3 Evaluation of Sub-region Selection

A unique feature of REBET is the ability to select the most susceptible sub-regions. Table 2 summarizes the frequencies of individual sub-regions selected under various scenarios when REBET claimed a significant association (P -value < 0.001). We also reported the frequencies of sub-region union selected for various scenarios.

When all variants were deleterious in scenario A, each sub-region (denoted as {SR1}, {SR2}, {SR3}, {SR4} or {SR5}) had around 80% chance, i.e. among 80% of replicates, to be selected. Closer inspection of the selection of sub-region unions revealed more intriguing patterns. There were 37.34%, 40.32%, 17.87%, 4.04% and 0.44% chances for which REBET would select 5, 4, 3, 2 or 1 sub-regions, respectively. Altogether, REBET selected three or more sub-regions for 95.52% replicates. Even when one of five regions was unselected, the chance of missing the first sub-region {SR1}, the one with the strongest effect size, was 17.32%, lower than 21.08% chance of missing the sub-region {SR5} with the weakest effect. In scenario B, the individual deleterious sub-region ({SR1}, {SR3} or {SR5}) would be selected over 93% chance. In 69.21% of replicates, REBET selected the true deleterious sub-region union ({SR1, SR3, SR5}) while in 13.79% and 13.46% of replicates REBET would miss only one deleterious sub-region ({SR3, SR5}, {SR1, SR5} or {SR1, SR3}) or identify one additional neutral sub-region ({SR1, SR2, SR3, SR5} or {SR1, SR3, SR4, SR5}), respectively. When the number of deleterious sub-regions reduced to two ({SR1} and {SR5}) in scenario C, REBET selected individual deleterious sub-regions ({SR1} or {SR5}) over 97% chance and selected the true deleterious sub-region union ({SR1, SR5}) with 68.79% chance, which decreased to 33.02%, 14.59% and 65.04% in the following three scenarios (D, E and F). In scenarios D and E, the deleterious sub-regions 1 and 5 included neutral variants, and in scenario F the additional protective variants existed at neutral sub-regions 2 and 4. Note that when the neutral sub-regions contained one protective instead of deleterious variant, the chance of selecting the true deleterious sub-region union ({SR1, SR5}) was significantly improved from 14.59% (scenario E) to 65.04% (scenario F). This is due to the low frequencies of selecting sub-regions 2 and/or 4 in scenario F. In short, these simulations illustrated that REBET was able to select most of the deleterious sub-regions, particularly for the ones with stronger effects, and the chance of selecting the true deleterious sub-region union depends on the proportion of deleterious variants in the deleterious sub-regions and the presence of protective or deleterious variants in the neutral sub-regions.

4 Application

4.1 The Dallas Heart Study

The first application of REBET was to test the association between triglyceride level and rare variants of the *ANGPTL4* gene (MIM 605910) in the population-based Dallas heart study [Romeo, et al. 2007]. Although the previous analyses have reported a gene-level significant rare variant association, little is known about which sub-group of rare variants with shared functional characteristics contributes most to this significant association. In our analysis, we focused on rare and less frequent variants ($MAF < 0.05$) in subjects with European ancestry ($n=1043$). A total of 19 functional variants (missense, nonsense or

frameshift) were included and located in four functional domains [Zhu, et al. 2012]: signal peptide domain, coiled-coil domain, linker domain and fibrinogen-like domain, defined as sub-regions of *ANGPTL4* gene. The triglyceride level was treated as either a continuous trait or a binary trait with a low-triglyceride group (triglyceride level less than or equal to the 25th percentile, n=269) and a high-triglyceride group (triglyceride level more than or equal to the 75th percentile, n=258). Other methods, including burden test, SKAT, SKAT-O and Mist with or without considering functional domains, were also applied. We adjusted for gender (Female=544 and Male=499) and age (mean=44.73 years, standard deviation=9.99 years) in all methods.

For both triglyceride level and group, REBET with sub-region selection, SKAT, SKAT-O and Mist without functional domain annotation all led to a P-value less than 0.01 (Table 3). Application of SKAT to sub-regions and pooling P-values showed P-value 0.005 for triglyceride level and P-value was 0.012 for triglyceride group. P-values by EBSCAN were not significant (EBSCAN1=0.07, EBSCAN2=0.05, LRT=0.05) for triglyceride group while P-values were significant for triglyceride level (EBSCAN1=0.03, EBSCAN2=0.03, LRT=0.01). A closer examination of the frequency and distribution of rare variants for triglyceride group revealed that nine unique variants reached $MAF < 0.05$ and were located in three domains (Figures 1a and 1b). Notably, one variant E40K ($MAF = 1.23\%$) was observed in 12 subjects in the low-triglyceride group and in one subject in the high-triglyceride group as a protective variant ($Z_j = -3.13$), and likely contributed heavily to the significant association results. Indeed, when E40K was excluded, none of the tests would show a P-value < 0.4 . Among the remaining eight variants, four of them were protective (Z_j of E167K, G223R, S302fd and G361S were -1.91 , -0.79 , -0.75 and -0.75 , respectively) and the others were deleterious (Z_j of E190Q, P251T, V308M and R336C were 0.84 , 0.84 , 0.80 and 0.07 , respectively). The existence of both protective and deleterious variants was in favor of the model assumption of SKAT, SKAT-O and Mist with smaller P values, but not the assumption of REBET for the whole region and burden test. Despite generating significant P values, these methods could not select which variants or group of them contributed most or estimate the corresponding significance by a P value.

REBET was able to investigate one step deeper into sub-regions of the gene. We treated functional domains as sub-regions. The variants in the same functional domain, instead of the gene, more likely hold the homogeneous effect, while effect of variants in the different functional domains might be heterogeneous. Three functional domain involved in *ANGPTL4* for triglyceride group; the coiled-coil domain and linker domain were protective ($Z^k = -3.13$ and -1.19 , $\pi^k = 0.54$ and 0.04 while the fibrinogen-like domain was slightly deleterious ($Z^k = 0.11$, $\pi^k = 0.42$), thus there was a mixture of both protective and deleterious variants in these domains. REBET was able to consider all possible domain combinations and selected the coiled-coil domain as the most significant domain with a protective effect (domain odds ratio (OR)= 0.014, 95% confidence interval (CI): [0.001, 0.349]) for low- and high-triglyceride groups. This result is consistent with the current literature about the function of *ANGPTL4*. It has been shown that *ANGPTL4* inhibits lipoprotein lipase which in turn regulates glucose and lipoprotein metabolism [Yoshida, et al. 2002]. Among all functional domains of *ANGPTL4*, it's the coiled-coil domain that binds transiently to lipoprotein lipase and inactivates it [Sukonina, et al. 2006].

4.2 NCI Osteosarcoma Sequencing Study

The NCI osteosarcoma sequencing study investigated the rare exonic germline variants to identify genes with a higher burden of rare variants in 334 European (EUR) ancestry matched cases and 987 EUR ancestry controls. Here, we focused on one known cancer predisposition gene (CPG), *TP53* (MIM 191170), which was demonstrated to have a high prevalence of pathogenic mutations[Mirabello, et al. 2015] on an examination of the association between its rare variants ($MAF < 1\%$) and case-control status.

Unlike the significance of the *ANGPTL4* triglyceride level association driven largely by one less frequent variant (E40K), there were nineteen variants within *TP53* that were all extremely rare with sixteen singletons ($MAF = 0.038\%$) and three variants occurring twice ($MAF = 0.076\%$). These variants were spread over four functional domains[Joerger and Fersht 2008]: P53 DNA-binding domain (14 unique variants), linker domain (2 unique variants), tetramerization domain (2 unique variants) and extreme C terminus (1 unique variant) (Figure 1c and 1d). Most variants were located in the P53 DNA-binding domain in the cases. Most variants were deleterious (14/19), including one variant N235S observed in both cases and controls. The remaining variants in the controls were all protective variants. Given the prevalence of deleterious rare variants, it's not surprising that the burden test and REBET for the whole region, assuming a homogeneous effect direction across variants, was more significant than the P values estimated by SKAT and Mist (Table 3). Application of SKAT to sub-regions and pooling P-values gave P-value = 0.003. P-values by EBSCAN were not significant (EBSCAN1=0.24, EBSCAN2=0.05, LRT=0.06) possible due to EBSCAN comparing between sub-regions instead of case and control group.

Among the four domains, the extreme C terminus was protective ($Z^k = -0.82$, $\pi^k = 0.09$) while the P53 DNA-binding domain, linker domain and tetramerization domain were all deleterious ($Z^k = 4.58, 0.80$ and 2.43 , $\pi^k = 0.73, 0.09$ and 0.09). REBET selected the P53 DNA-binding domain and tetramerization domain as the most significant domains with deleterious effects (OR = 289.28, 95% CI: [23.63, 3541.30]). The P53 DNA-binding domain includes an immunoglobulin-like β -sandwich which provides sequence-specific DNA binding to regulate gene transcription in response to various cell stress[Eldeiry, et al. 1992]. Additionally, as one of the most frequent tumor suppressor genes for many cancers, most of the somatic rare variants have been shown to be located in the P53 DNA-binding domain in the International Agency for Research on Cancer (IARC) TP53 database[Petitjean, et al. 2007b], which further highlights its functional importance. Although likely not as functionally critical as the P53 DNA-binding domain, the tetramerization domain has been shown to be important for DNA binding and protein-protein interactions among other functions[Chene 2001].

5 Discussion

We present REBET as a flexible, powerful and biological interpretable approach to select the most susceptible rare variants in sub-regions and test their associations with complex traits. In contrast to the other methods that do not select functionally informative sub-regions within a gene, REBET attempts to optimize power of association test statistics when susceptibility variants are likely to cluster within sub-regions. The sub-regions are

predefined such that the variants within a sub-region share the same biological function or have similar functional impact. REBET examines all possible combinations of sub-regions and finds the combination that shows the strongest evidence of association. As a result, REBET improves the resolution of significant results from the gene level, as typically performed by other gene-based tests, to the sub-region level, e.g. at the level of the functional domains. REBET reconciles the conflict between the limited power of testing individual rare variants and the challenges of sorting out the most likely individual susceptibility variants following gene-level tests.

Our simulation studies show that for overall detection of a susceptibility gene, REBET can have either higher, comparable or slightly lower power compared to the other methods we evaluated in a variety of scenarios. The simulation studies also demonstrate that REBET is able to identify the sub-regions containing the clustered variants, even when the identified sub-regions include half of variants as neutral ones. In both case studies, REBET can identify the most susceptible domains, which are essential for the gene function as supported by the existing literature. In short, the simulation and case-studies demonstrate that REBET can effectively incorporate annotation information of variants for boosting power for detection of overall genetic association and simultaneously identify sub-regions of likely functional importance, offering a clearer biological interpretation of results.

REBET is computationally efficient by utilizing a subset-based approach considering the local correlation between two neighboring sub-regions. It achieves the accurate and conservative upper bound of P value estimate, and does not require permutation to calculate the P value to improve the computational efficiency. Additionally, the REBET test statistic is based on the score test calculated based on the null hypothesis only. The analysis of triglyceride levels in the Dallas heart study and case-control analysis of the NCI osteosarcoma sequencing study only took 0.04 and 0.07 seconds on a laptop with a 2.5 GHz CUP.

A number of extensions are possible. In this paper, we focused on coding variants annotated by the protein functional domain, which is reliable and a natural choice of sub-region within a gene. A comprehensive data-base of protein families and domains is available at Pfam[Finn, et al. 2014]. In addition, REBET could easily adapt other annotations for both coding [Adzhubei, et al. 2010; Ng and Henikoff 2003] and non-coding variants[Alexander, et al. 2010] with various types of functional score. For example, the sub-regions for coding variants would be defined by the groups of variants with similar functional scores. If sample size permits, various forms of annotation could be combined and sub-regions could be annotated on a finer scale; possibly the various forms of annotation would be organized in a hierarchical structure to restrain the number of possible combinations and multiple comparisons. With the increasing utilization of whole genome sequencing and the increasing interest in non-coding variants, REBET may become more useful when the definition of regions of interest are less clear and there is a need to search a number of possible intergenic regions and their combinations, such as enhancer, promoter and 5'UTR. It's of interest to extend the REBET for somatic mutations to identify the driver genes, for which the tumors with silent mutations are treated as controls and tumors with non-silent mutations as cases. The frequencies of somatic mutation are extremely low which requires very large sample size. Finally, we have demonstrated the usage of REBET to identify the most susceptible

domains with either homogeneous protective or deleterious effect. Following ASSET[Bhattacharjee, et al. 2012], it's straightforward for REBET to evaluate the susceptibility of a combination of protective domains and deleterious domains by considering two one-sided tests and combining the resulting P values using the Fisher's method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. (<http://biowulf.nih.gov>).

Funding

This work was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, MD, USA.

References

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nature Methods*. 2010; 7(4): 248–249. [PubMed: 20354512]
- Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. *Nature Reviews Genetics*. 2010; 11(8):559–571.
- Auer PL, Nalls M, Meschia JF, Worrall BB, Longstreth WT, Seshadri S, Kooperberg C, Burger KM, Carlson CS, Carty CL, et al. Rare and Coding Region Genetic Variants Associated With Risk of Ischemic Stroke The NHLBI Exome Sequence Project. *Jama Neurology*. 2015; 72(7):781–788. [PubMed: 25961151]
- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*. 2010; 11(11):773–785.
- Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, Yeager M, Chung CC, Chanock SJ, Chatterjee N, et al. A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. *American Journal of Human Genetics*. 2012; 90(5):821–835. [PubMed: 22560090]
- Chene P. The role of tetramerization in p53 function. *Oncogene*. 2001; 20(21):2611–2617. [PubMed: 11420672]
- Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*. 2010; 11(6):415–425.
- Cybulski C, Carrot-Zhang F, Kluzniak W, Rivera B, Kashyap A, Wokolorczyk D, Giroux S, Nadaf J, Hamel N, Zhang SY, et al. Germline RECQL mutations are associated with breast cancer susceptibility. *Nature Genetics*. 2015; 47(6):643–646. [PubMed: 25915596]
- Derkach A, Lawless JF, Sun L. Pooled Association Tests for Rare Genetic Variants: A Review and Some New Results. *Statistical Science*. 2014; 29(2):302–321.
- Eldeiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B. Definition of a Consensus Binding-Site for P53. *Nature Genetics*. 1992; 1(1):45–49. [PubMed: 1301998]
- Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014; 42(Database issue):D222–30. [PubMed: 24288371]
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*. 2010; 70(1):42–54. [PubMed: 20413981]

- Helgadóttir A, Gretarsdóttir S, Thorleifsson G, Hjartarson E, Sigurdsson A, Magnúsdóttir A, Jonasdóttir A, Kristjánsson H, Sulem P, Oddsson A, et al. Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nature Genetics*. 2016; 48(6):634. [PubMed: 27135400]
- Joerger AC, Fersht AR. Structural biology of the tumor suppressor p53. *Annual Review of Biochemistry*. 2008; 77:557–582.
- Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, Raja A, Coe BP, Stessman HA, He ZX, et al. Excess of rare, inherited truncating mutations in autism. *Nature Genetics*. 2015; 47(6):582–588. [PubMed: 25961944]
- Lee S, Abecasis GR, Boehnke M, Lin XH. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*. 2014; 95(1):5–23. [PubMed: 24995866]
- Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012; 13(4):762–75. [PubMed: 22699862]
- Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *Plos Genetics*. 2009; 5(2)
- Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, Mallick S, Li H, Stram A, Sheng X, et al. The contribution of rare variation to prostate cancer heritability. *Nature Genetics*. 2016; 48(1):30. [PubMed: 26569126]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. [PubMed: 19812666]
- Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, Fine RS, Lu Y, Schurmann C, Highland HM, et al. Rare and low-frequency coding variants alter human adult height. *Nature*. 2017; 542(7640):186–190. [PubMed: 28146470]
- McCallum KJ, Ionita-Laza I. Empirical Bayes scan statistics for detecting clusters of disease risk variants in genetic studies. *Biometrics*. 2015; 71(4):1111–20. [PubMed: 26033425]
- Mirabello L, Yeager M, Mai PL, Gastier-Foster JM, Gorlick R, Khanna C, Patino-Garcia A, Sierrasesumaga L, Lecanda F, Andrulis IL, et al. Germline TP53 variants and susceptibility to osteosarcoma. *J Natl Cancer Inst*. 2015; 107(7)
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis*. 2007; 615(1–2):28–56. [PubMed: 17101154]
- Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, McVean G, Boehnke M, Altshuler D, et al. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *Plos Genetics*. 2015; 11(4)
- Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*. 2003; 31(13):3812–3814. [PubMed: 12824425]
- Petitjean A, Achatz MIW, Borresen-Dale AL, Hainaut P, Olivier M. TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*. 2007a; 26(15):2157–2165. [PubMed: 17401424]
- Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. *Human Mutation*. 2007b; 28(6):622–629. [PubMed: 17311302]
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*. 2010; 86(6):832–8. [PubMed: 20471002]
- Rebbeck TR, Mitra N, Wan F. Association of Type and Location of BRCA1 and BRCA2 Mutations With Risk of Breast and Ovarian Cancer (vol 313, pg 1347, 2015). *Jama-Journal of the American Medical Association*. 2015; 314(6):628–628.
- Romeo S, Pennacchio LA, Fu YX, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature Genetics*. 2007; 39(4):513–516. [PubMed: 17322881]

- Sukonina V, Lookene A, Olivecrona T, Olivecrona G. Angiopoietin-like protein 4 converts lipoprotein lipase to inactive monomers and modulates lipase activity in adipose tissue. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(46):17450–17455. [PubMed: 17088546]
- Sun H, Wang S. A power set-based statistical selection procedure to locate susceptible rare variants associated with complex traits with sequencing data. *Bioinformatics*. 2014; 30(16):2317–23. [PubMed: 24755303]
- Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol*. 2013; 37(4):334–44. [PubMed: 23483651]
- Surakka I, Horikoshi M, Magi R, Sarin AP, Mahajan A, Lagou V, Marullo L, Ferreira T, Miraglio B, Timonen S, et al. The impact of low-frequency and rare variants on lipid levels. *Nature Genetics*. 2015; 47(6):589–597. [PubMed: 25961943]
- Taylor JE, Worsley KJ, Gosselin F. Maxima of discretely sampled random fields, with an application to ‘bubbles’. *Biometrika*. 2007; 94(1):1–18.
- 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. *Am J Hum Genet*. 2011; 89(1):82–93. [PubMed: 21737059]
- Yoshida K, Shimizugawa T, Ono M, Furukawa H. Angiopoietin-like protein 4 is a potent hyperlipidemia-inducing factor in mice and inhibitor of lipoprotein lipase. *Journal of Lipid Research*. 2002; 43(11):1770–1772. [PubMed: 12401877]
- Zhu PC, Goh YY, Chin HFA, Kersten S, Tan NS. Angiopoietin-like 4: a decade of research. *Bioscience Reports*. 2012; 32(3):211–219. [PubMed: 22458843]

Figure 1. Lollipop plots illustrating locations of rare variant in *ANGPTL4* gene for low and high triglyceride group (A and B) in the Dallas heart study and *TP53* gene for the case and control group (C and D) in the NCI osteosarcoma sequencing study. The horizontal bars represent the gene colored by the protein functional domains; the lollipops indicate changes in the amino acids by the rare variant, for example, E40K denoting that the corresponding rare variant changes the glutamic acid to lysine in the 40th amino acid of the *ANGPTL4* protein.

Performance of detection of the association between rare variants and traits. For scenario A to F, we evaluated the power when part of variants were risk variants; in scenario G, we evaluated the type I error rate when all variants were neutral. We applied REBET whole region, REBET subregion-based, REBET variant-based and other methods, including two burden tests (CAST and WSS), two adaptive burden tests (aSUM and VT), one variance-component test (SKAT) and two combined tests (SKAT-O and Mist). Mist is able to incorporate sub-region annotations (Mist-cov).

REBET whole region	REBET subregion-based 0.5/0.5	REBET variant-based 0.5/0.5	CAST	WSS	aSUM	VT	SKAT	SKAT-O	MIST	MIST-cov
A: 1111/1111/1111/1111/1111	0.872	0.730	0.608	0.795	0.668	0.693	0.148	0.716	0.649	0.413
B: 1111/0000/1111/0000/1111	0.790	0.909 (0.936) [±]	0.327	0.712	0.590	0.530	0.341	0.674	0.671	0.705
C: 1111/0000/0000/0000/1111	0.492	0.874	0.051	0.391	0.475	0.373	0.280	0.424	0.448	0.651
D: 1110/1000/0000/0001/0111	0.490	0.570	0.061	0.397	0.409	0.320	0.283	0.429	0.453	0.433
E: 1100/1000/0000/0001/0011	0.214	0.186	0.022	0.172	0.152	0.111	0.150	0.202	0.228	0.173
F: 1100/-1000/0000/000-1/0011	0.014	0.089	0.001	0.013	0.068	0.032	0.104	0.076	0.082	0.126
G: 0000/0000/0000/0000/0000/	1.001 [*]	0.957	0.914	1.069	1.092	1.057	0.848	0.897	0.916	0.918

A cohort of two thousand cases and two thousand controls with binary disease status was simulated. Denote the probability of having disease as $P(Y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$, for which

$$\eta_{i_1} = \beta_0 + \sum_{j=1}^{20} \beta_j X_{ij} \omega_i^{I_c} = \text{beta}(\text{MAF}_j, 0.5, 0.5) = 1/\sqrt{\text{MAF}_j(1 - \text{MAF}_j)}.$$

∅ We considered five sub-regions separated by forward slashes (/), each including four rare variants, their effect denoted as the deleterious ($J_j = 1$), absent ($J_j = 0$) or protective effect ($J_j = -1$). Minor allele frequencies (MAF) gradually increase from 0.0005 to 0.01 by 0.0005 each.

[†]The number of combination of sub-regions is constrained by prior knowledge.

* The type I error rate is divided by the targeted level 0.001. Hence value one indicated achieving the targeted type I error rate.

The frequency (%) of sub-regions (SR) selected as part of the most susceptible sub-region union using REBET subregion-based method in table 1.

Table 2

Scenarios	SR1	SR2	SR3	SR4	SR5
A: 1111/1111/1111/1111/1111	84.01	82.03	81.70	81.12	81.22
B: 1111/0000/1111/0000/1111	96.10	8.64	93.71	8.87	93.25
C: 1111/0000/0000/0000/1111	98.69	6.69	6.63	6.73	97.17
D: 1110/1000/0000/0001/0111	97.04	39.41	12.46	38.33	94.98
E: 1100/1000/0000/0001/0011	91.13	55.77	19.00	54.47	89.27
F: 1100/-1000/0000/0000-1/0011	92.39	4.40	16.46	3.65	92.61

Table 3

The P-values of rare variant association tests applied to the Dallas heart study and NCI osteosarcoma sequencing study. The applied methods included REBET whole region, REBET subregion-based, burden test, SKAT, SKAT-O, Mist and Mist considering covariates (Mist-Cov). The REBET subregion-based will select the most significant functional domains listed in the parentheses.

Study	Trait	REBET whole region	REBET region-based	Burden	SKAT	SKAT-O	MIST	MIST-cov
The Dallas heart study	Triglyceride level	3.31e-2	8.69e-3 (CC)	3.00e-2	1.51e-3	2.01e-3	4.58e-3	1.76e-2
	High/low triglyceride group	1.34e-2	9.25e-3 (CC)	1.20e-2	2.31e-3	3.20e-3	3.70e-3	4.74e-2
NCI osteosarcoma sequencing study		3.63e-6	9.23e-6 (P53C/TET)	3.04e-6	1.78e-3	1.09e-5	1.10e-5	4.87e-5