Localizing Rare-Variant Association Regions via Multiple Testing Embedded in an Aggregation Tree

Xuechan Li, John Pura, Andrew Allen, Kouros Owzar, Matthew Harms, and Jichun Xie

November 8, 2022

Abstract

Rare variants (RVs) genetic association studies enable researchers to uncover the variation in phenotypic traits left unexplained by common variation. Traditional single-variant analysis lacks power; thus, researchers have developed various methods to aggregate the effects of RVs across genomic regions to study their collective impact. Some existing methods utilize a static delineation of genomic regions, often resulting in suboptimal effect aggregation, as neutral subregions within the test region will result in an attenuation of signal. To pinpoint genomic regions enriched for disease-associated RVs, we developed a novel method, DYNATE (DYNamic Aggregation TEsting). DYNATE dynamically and hierarchically aggregates smaller genomic regions into larger ones and performs multiple testing for disease associations with controlled weighted false discovery rate. Extensive numerical simulations demonstrate the superior performance of DYNATE under various scenarios compared to existing methods. Importantly, DYNATE-identified regions have higher enrichment levels of disease-associated RVs. We applied DYNATE to an amyotrophic lateral sclerosis (ALS) study and identified a new gene, *EPG5*, harboringpossibly pathogenic mutations.

Keywords: Rare-variant associations; aggregation tree; multiple testing

Introduction

Rare variant (RV) disease association analysis has been an important topic in genetics and genomics. Because single-variant analysis suffers low statistical power[2,16], a common approach to increase statistical power is to test the collective effect of all RVs within genomic prefixed regions. The testing methods include burden tests[1,19,26], variant component tests[28,36], and their combinations[17]. The prespecified regions could be genes[32,33], functional domains[11], or sliding fixed-length windows[4,13]. However, those prefixed-region-based methods are suboptimal when the pre-fixed regions are too long or short. For example, when the pre-fixed region is too long, it could harbor too many neutral variants. These neutral variants will attenuate the effect of any causal variants in the region. On the other hand, when the pre-fixed region is too short, even if the region contains some causal variants, the analysis may exclude adjacent regions containing additional causal RVs, leading to a loss of power. An extreme example of the latter is the single-variant association analysis.

Inspired by these considerations, varying-window methods[12,21] scan subregions with different lengths to optimize the power of detecting disease-associated RVs. Compared with fixed-region methods, varying-window methods are generally more powerful. However, in practice, the regions identified by the varying-window methods often include long stretches of sequence containing many neutral RVs; thus, they are less helpful in pinpointing the functionally important RVs in the downstream analysis. Moreover, searching regions with different lengths is computationally intensive, leading to long computing times.

This paper proposes a new method that pinpoints short regions containing enriched disease-associated RVs. The method, called DYNATE (DYNamic Aggregation TEsting), dynamically aggregates and tests genomic regions with varying lengths for disease association. DYNATE adopts a fine-to-coarse hierarchical testing strategy, so that, given the previous layer's testing results, DYNATE dynamically defines the testing regions, performs multiple testing on the current layer, and passes the results to the next layer. The hierarchical testing structure allows

DYNATE to explore both short and long regions efficiently, resulting in a fast computational algorithm.

We demonstrated that DYNATE outperforms other competing methods under various numerical settings: DYNATE has similar power as the state-of-the art fixed-region[11] and varying-window[21] methods in identifying disease-associated domains. However, its RV-level precision is much higher: the DYNATE-identified regions are shorter and highly enriched with pathogenic variants, while the competing methods often output extensive regions that cover too many neutral variants. We applied DYNATE to an amyotrophic lateral sclerosis (ALS) study to pinpoint the regions containing ALS-associated RVs. DYNATE successfully identified regions in well-known ALS genes *SOD1* (MIM: 147450) and *TARDBP* (MIM: 605078) while also identifying a new region in *EPG5* (MIM: 615068), which could not be identified by competing methods. Further, we compared the identified region lengths. The DYNATE-identified regions were much shorter, suggesting a higher disease-associated RV specificity. Thus, DYNATE provides more specific targets for the downstream functional pathway analysis.

Materials and Methods

Notations

DYNATE adopts a dynamic and hierarchical testing structure. It splits the whole-exome or genome into small regions called leaves, tests their disease associations, hierarchically aggregates these regions into larger regions, and then tests those larger regions. Here, we provide notations.

Leaves and nodes

Consider a case-control study with N_0 controls, N_1 cases, and J observed qualifying RVs in D homology-based protein domains[15]. DYNATE has multiple layers. From now on, we use the superscript ℓ to represent the notation defined on layer ℓ.

On layer 1, we partition these D homology-based protein domains into $m^{(1)}$ leaves to increase resolution and localize disease-associated RVs. We define a *leaf* as a genic region where M subjects carry qualifying rare variants. Here, M denotes the leaf size. It is a much smaller number compared to N_0 and N_1. When $M = 1$, only one subject carries qualifying RVs in each leaf. In practice, we recommend $M \geq 5$.

With protein domain annotations, we further optimize the definition of leaves by forcing the breakdown of leaves at the beginning and the end of the domains. Thus, a leaf does not spread across two domains. If fewer than M subjects carry qualifying variants in the last leaf of a domain, the last leaf is combined with the preceding leaf in the same domain if it exists. For a domain with less than M qualifying variants, the whole domain region forms a leaf.

Suppose the partition ends up with $m^{(1)}$ leaves, denoted by $\{1\}, \ldots, \{m^{(1)}\}$. On higher layers (layer $\ell \geq 2$), it aggregates the accepting nodes from the previous layer to form new nodes. A node S is a union of one or more leaves. A node represents a genome region with qualifying RVs. Thus, leaves are nodes on layer 1. For any node S, the set of its qualifying RVs is denoted by $\mathcal{V}(S)$, $\mathcal{V}(S) \subset [J]$.

On all layers, we define qualified nodes as those newly formed on the current layer and have not been tested before. On layer 1, all leaves are qualified nodes. On layer ℓ ($\ell \geq 2$), a qualified node is a union of two or more nodes on layer $\ell - 1$. The set of all qualified nodes on layer ℓ is denoted by $\mathcal{B}^{(\ell)}$. Specifically, $\mathcal{B}^{(1)} = \{\{i\} : i \in [m^{(1)}]\}$, where $[m^{(1)}]$ denotes the set $\{1, \ldots, m^{(1)}\}$.

Node hypotheses and P-values

We call leaf $\{i\}$ a null leaf if

for all $j \in \mathcal{V}(\{i\})$, the qualified RV j is not associated with the disease;

otherwise, we call it alternative. Denote the set of null leaves by $\mathcal{B}_0^{(1)}$ and the set of alternative leaves by $\mathcal{B}_1^{(1)}$.

Similarly, for any node S, we denote the node hypothesis by H_S. The null is

$$H_{S,0} : \forall\, i \in S, \{i\} \in \mathcal{B}_0^{(1)} \tag{1}$$

For simplicity, for leaf hypothesis $H_{\{i\}}$, we denote them as H_i. H_i is null if $\{i\} \in \mathcal{B}_0^{(1)}$ and alternative if $\{i\} \in \mathcal{B}_1^{(1)}$.

To test these hypotheses, we need to summarize their P-values. DYNATE is flexible to accommodate existing RV test statistics including burden tests[1,19,25], variance-component tests[29,36], and omnibus tests[17]. For fast computation, we mainly used the following two statistics: 1) Lancaster's mid-P correction for the Fisher's exact test (FL), and 2) efficient score statistics with saddle point approximation (SS). More details are provided in Supplementary Note 1.

After leaf P-values are calculated, we use Stouffer's method to calculate node P-values. For node S, the P-value is

$$T_S = \bar{\Phi}(\sum_{i \in S} \bar{\Phi}^{-1}(T_i)/\sqrt{|S|}), \tag{2}$$

where $\bar{\Phi}(\cdot)$ is the complementary cumulative density function of the standard Gaussian distribution. Alternatively, we may consider using other aggregation approaches, such as Fisher's combination[10], chi-square aggregation, or Cauchy aggregation[23].

Type I Error criterion: node-FDR

When a node is the union of multiple leaves, we might not be able to dichotomize its null and alternative status. For example, suppose node $S = \{1, 2\}$ with leaf $\{1\}$ is null and leaf $\{2\}$

is alternative. Then S is $1/2$ alternative. More generally, for any node S, let $S_1 = S \cap \{i : H_i \text{ is alternative}\}$. Then $\theta_S = {|S_1|}/{|S|}$. We refer to S as θ_S-alternative or $(1-\theta_S)$-null.

To address this fractional null/alternative status of nodes, we introduce a new type I error measure called node-false-discovery-rate (node-FDR), which is the expectation of node-false-discovery-proportion (node-FDP)

$$\text{node-FDP} = \frac{\sum_{S\in\mathcal{R}}\{1-\theta_S\}}{|\mathcal{R}| \vee 1}, \quad \text{node-FDR} = \mathrm{E}(\text{node-FDP}), \tag{3}$$

where $\mathcal{R}$ is the set of the rejected nodes. Rejecting node S will contribute $1-\theta_S$ to the numerator and 1 to the denominator.

When identifying disease-associated regions, DYNATE aims to asymptotically control node-FDR. Thus, it prioritizes identifying the nodes containing higher proportions of alternative leaves. For example, when the alternative leaves are concentrated in a shorter region, DYNATE will prioritize identifying this short region rather than a more extended region containing too many null leaves. Because these shorter regions usually contain a higher proportion of disease associated RVs, they are more likely to be pathogenic or protective.

DYNATE: Dynamic and Hierarchical Testing

To identify disease-susceptible regions, DYNATE adopts a fine-to-coarse strategy. First, it tests the leaf-level disease associations, where leaves are the smallest testing units. Then, it hierarchically aggregates leaves into nodes (corresponding to longer regions) and tests the node-level disease associations. The general algorithm is listed in Algorithm 1. To illustrate the implementation of Algorithm 1, we provided a toy example in Supplementary Figure S1.

Algorithm 1 includes multiple steps: leaf P-value cutoff, aggregation, and node P-value cutoffs. Another key step is to choose tuning parameters. Here, we provide a summary of each step; the details were provided in Supplementary Note 2.

- *Leaf P-value cutoff.* This step is to choose a proper cutoff $\hat{t}^{(1)}$ to control the node-FDR

Algorithm 1: Dynamic hierarchical testing

Input: Maximum Layer L, leaf P-values $(T_1, \ldots, T_{m^{(1)}})$, and the desired FDR level α
Output: The set of rejected nodes $\mathcal{R}$.
Derive the leaf P-value cutoff $\hat{t}^{(1)}$; `// Leaf P-value cutoff`
Set the rejection set: $\mathcal{R} = \{\{i\} : T_i < \hat{t}^{(1)}, i \in [m^{(1)}]\}$;
for $\ell \in \{2, \ldots, L\}$ **do**
  Derive the aggregated node set $\mathcal{B}^{(\ell)}$; `// Aggregation`
  for $S \in \mathcal{B}^{(\ell)}$ **do** Derive node P-values as in (2);
  Derive the node P-value cutoff $\hat{t}^{(\ell)}$; `// Node P-value cutoffs`
  Update the rejection set: $\mathcal{R} = \mathcal{R} \cup \{S \in \mathcal{B}^{(\ell)} : T_S < \hat{t}^{(\ell)}\}$;
end

on layer 1. We noticed that on layer 1, the node-FDR is the same as the FDR for all leaf hypotheses; thus, using traditional FDR cutoff (such as the cutoff for the classic Benjamini and Hochberg procedure[3]) works well on layer 1.

- *Aggregation.* This step is to aggregate the accepted nodes on previous layers into larger nodes on the current layer. Here, "accepted" nodes refer to those not rejected on the previous layer. The purpose is to aggregate neighboring accepted nodes that are likely to be co-null or co-alternative; if they are co-alternative, aggregation will increase their chance of being rejected.
- *Node P-value cutoffs.* This step is to choose a proper cutoff $\hat{t}^{(\ell)}$ to control the node-FDR on layer ℓ with $\ell \geq 2$. Because a node can be partially null, we hope that we could roughly estimate nodes' overall null level to provide a control of node-FDR. We use the sampling technique to get this overall estimate.
- *Tuning parameter selection.* The layer number L and leaf size M are tuning parameters. The default tuning parameters were set to optimize the RV-level specificity in the findings and the power. Our default setting was validated by the numerical studies shown in **Simulation Studies**.

Simulation Studies

We evaluated the performance of DYNATE under a wide range of simulation settings. We mimicked the structure of RVs and domains in the ALS dataset, which contains $263,789$ qualified variants located in $20,666$ qualified functional domains. Please see **Application to an Amyotrophic Lateral Sclerosis (ALS) Study** for how we pre-processed the data to obtain the qualified variants and domains. We randomly selected $2,000$ domains with $25,622$ RVs.

Out of the 2000 domains, we randomly selected 20 domains with more than 30 RVs to be dense pathogenic domains (DPDs) and 20 other domains to be sparse pathogenic domains (SPDs). A DPD is a domain harboring a pathogenetic region with $\min(20, L_d/2)$ observed RVs, where L_d is the number of observed RVs in the domain. Within this pathogenic region, π of them are randomly selected as pathogenic RVs. In the simulation, we vary $\pi \in \{20\%, 40\%, 60\%, 80\%\}$ to evaluate the performance of DYNATE under different settings. An SPD is a domain harboring only one pathogenic RV; this RV location is randomly selected.

Let $\boldsymbol{G}_k = (G_{k,1}, \ldots, G_{k,J})^T$ denote the observed RV vector for individual k where $G_{k,j}$ is the indicator of whether individual k carries the mutation of RV j. We assumed that $G_{k,j} \sim \mathrm{Bern}(\rho_j)$ and simulated $\log_{10} \rho_j$ from $\mathrm{Unif}(-3, -1.7)$ so that ρ_j ranges from 0.001 to 0.02. Here, $\mathrm{Bern}(\rho)$ denote a Bernoulli distribution with parameter ρ, and $\mathrm{Unif}(a, b)$ denote a uniform distribution over the interval $[a, b]$ respectively. When ρ_j is small, it is close to the minor allele frequency.

The binary phenotype of the individual k was independently generated based on the model

$$\text{logit}\, \mathrm{P}(Y_k = 1 | Z_{k,1}, Z_{k,2}, \boldsymbol{G}_k) = \gamma_0 + \eta Z_{k,1} + \eta Z_{k,2} + \boldsymbol{G}_k^T \boldsymbol{\beta}.$$

Here, $Z_{k,1}$ and $Z_{k,2}$ are two non-genetic covariates with $Z_{k,1} \sim N(0, 1)$ and $Z_{k,2} \sim \mathrm{Bern}(0.5)$, respectively. Their coefficients are η, varying in different settings with $\eta \in \{0, 0.5\}$. The other set of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$ represents the RV effect size; we set β_j as a decreasing function of ρ_j, $\beta_j = -c \log_{10} \rho_j$, where $c = 2.5$ for pathogenic RVs in the DPDs and $c = 7.5$ for those in the SPDs. We set the intercept γ_0 so that the disease prevalence is 1%.

Based on the 6 different combinations of $\pi \in \{20\%, 40\%, 60\%\}$ and $\eta \in \{0, 0.5\}$ (covariate), we generated 6 large population, each with 2 million individuals. We conduct the numerical studies with 100 repetitions under each population. In each repetition, we randomly draw N_1 cases and N_0 controls. We let $N_1 = 1000$ and vary $N_0 \in \{1000, 2000, 3000\}$ to check the performance of DYNATE when the number of controls increases.

When applying DYNATE to the simulated datasets, we used both the FL and SS statistics (See **Node hypotheses and P-values**). Under all scenarios, we set the layer number $L = 3$ and leaf size $M = 11$ according to the criteria described in **Tuning parameter selection**. In order to justify the default tuning parameters selection and check the robustness of DYNATE, we also apply DYNATE with different M and L settings: $M \in \{7, 9, 11, 13\}$; for $M \in \{7, 9, 11\}$, $L = 3$; for $M = 13$, $L = 2$.

We compared the performance of the DYNATE with competing methods, including the domain-based collapsing (DC) method[11] with the FL (DC-FL) and SS (DC-SS) statistics and the SGANG[21] with their suggested burden (SCANG-B), SKAT (SCANG-S), and SKAT-O (SCANG-O) statistics and the default parameter setting. The desired domain-FDR for the DC method, familywise error rate (FWER) for SCANG, and node-FDR for the DYNATE are all set at 5%. The results yielded from these methods are evaluated using the average node-FDP, RV-FDP, RV-sensitivity, domain-FDP, and domain-sensitivity across repetitions. The node-FDP is defined in (3), and other measures are defined below:

$$\begin{aligned}
\text{RV-FDP} &= \frac{\text{Number of the identified neutral RVs}}{\text{Total number of the identified RVs}};\\
\text{RV-Sensitivity} &= \frac{\text{Number of the identified pathogenic RVs}}{\text{Total number of the pathogenic RVs}};\\
\text{Domain-FDP} &= \frac{\text{Number of the identified neutral domains}}{\text{Total number of the identified domains}};\\
\text{Domain-Sensitivity} &= \frac{\text{Number of the identified pathogenic domains}}{\text{Total number of the pathogenic domains}}.
\end{aligned}$$

Here, an RV is called “identified” if it is located in one of the identified regions; a domain is called “identified” if any of its RVs, regions, or itself is called significant by the algorithms.

We also used RV-level and domain-level F1 scores as overall accuracy measurements: a high F1 score indicates high accuracy. The RV-level and domain-level F1 scores are defined below:

$$\text{RV-F1} = \frac{2\{1-(\text{RV-FDP})\} \times \text{RV-Sensitivity}}{1-(\text{RV-FDP}) + \text{RV-Sensitivity}};$$
$$\text{Domain-F1} = \frac{2\{1-(\text{Domain-FDP})\} \times \text{Domain-Sensitivity}}{1-(\text{Domain-FDP}) + \text{Domain-Sensitivity}}.$$

We used domain-level, node-level, and RV-level measurements to systematically evaluate the performance of DYNATE and the competing methods. Among them, DC[11] is a domain-collapsing method; it can only provide domain-level testing. Thus, the comparison between DYNATE and DC was conducted on the domain-level. SCANG[21] is a varying-window scanning method; it can identify disease-associated regions with various lengths. Thus, the comparsion between DYNATE and SCANG was conducted on both region-level. Furthermore, we used the RV-level measurements to evaluate these methods' abilities to pinpoint highly disease-susceptible genomic regions and narrow down the collection of possibly disease-associated RVs.

Application to an Amyotrophic Lateral Sclerosis (ALS) Study

We applied DYNATE to an ALS study with samples that have either whole-exome sequencing (WES) or whole-genome sequencing (WGS). The goal of this study is to pinpoint the pathogenic RV coding sequence regions in individuals with estimated European ancestry.

We started from the diversified population with 3239 cases and 11,808 controls. Qualifying variants were defined as nonsynonymous coding or canonical splice variants that have a minor allele frequency (MAF) $\leq 0.1\%$ in cases and controls (internal MAF) and also a $\leq 0.1\%$ MAF imposed for each population represented in the ExAC Browser[18]. Gelfman et al.[11] filtered the variants based on their quality scores, quality by depth scores, genotype quality scores, read position rank sums, mapping quality scores, maximum Fisher's strand bias, alternative allele ratios, et al. Variants were annotated to Ensembl 73 using SnpEff[6].

Based on the filtered variants, we defined a European sub-cohort from the diversified population. We required the individuals 1) to self-declare as "White", 2) with an overall genotyping rate ≥ 0.87, and 3) with genotypes close to European, defined as the European ancestry probability ≥ 0.5. Here, the European ancestry probability was estimated through a multi-layer perceptron method used in Gelfman et al.[11].

The selected European sub-cohort contains $10,138$ samples: $2,634$ are cases ($2,590$ samples from WES and 44 from WGS) and $7,504$ are controls ($7,239$ samples from WES and 265 from WGS). There are $263,789$ qualified variants spanning $20,666$ qualified functional domains. The functional domains are defined as the genetic regions that aligned to the Conserved Domain Database (CDD)[24] and the unaligned genetic regions between CDD alignments[11]. A functional domain is qualified if it contains more than 5 RVs.

We applied DYNATE with FL (DYNATE-FL) and SS (DYNATE-SS) statistics. The SS statistics were obtained from fitting model (S1) with no non-genetic covariates ($\boldsymbol{\eta}_i = \mathbf{0}$). The node-FDR level is 5%. The default tuning parameters are $M = 7$ and $L = 5$.

We also compared the results from DYNATE with those from SCANG and DC. For SCANG, we used their default tunning parameter settings and considered three types of statistics: burden (SCANG-B), SKAT (SCANG-S), and SKAT-O (SCANG-O). Following Li et al.[21], the minimum and maximum numbers of variants in the sliding windows in SCANG were set as 1 and 103, respectively. For DC, we considered both FL (DC-FL) and SS (DC-SS) statistics. The desired FWER for SCANG and the domain-FDR for DC were set at 5%. A detected region was evaluated by its length and logOR. The region length is defined as the number of its containing RVs. The logOR is defined as

$$\text{logOR} = \log\left(\frac{\text{Odds of RV exposure among cases}}{\text{Odds of RV exposure among controls}}\right),$$
$$\text{Odds of RV exposure} = \frac{\text{number of subjects carrying RV mutations in the region}}{\text{number of subjects not carrying RV mutations in the region}}.$$

If logOR > 0, the region is likely to be pathogenic; if logOR < 0, the region is likely to be

protective.

Results

Simulation Studies: Type I Error, Power, and Robustness of DYNATE

Figure 1 depicts the average node-FDP and RV-level sensitivity of DYNATE with the default tuning parameter setting. The node-FDP measures type I error, and the RV-sensitivity measures power. We calculated the average node-FDR and RV-sensitivity at each layer up to the default layer number $L = 3$. In general, DYNATE controls node-FDR around the desired level 5%. Meanwhile, the RV-sensitivity increases as the layer number goes up. The results indicate that the multiple-layer testing approach can significantly improve the analysis power. DYNATE with different choices of statistics (DYNATE-FL and DYNATE-SS) have similar performance under various settings. Sometimes, DYNATE-SS has higher node-FDR, especially when the algorithm goes to layer 3, but DYNATE-SS also has higher RV-sensitivity than DYNATE-FL.

We further checked the robustness of DYNATE with different leaf sizes $M \in \{7, 9, 11, 13\}$ in Figure 2. For all the leaf sizes, DYNATE's average node-FDPs are between 4% and 10%. The desired level is 5%. The default tuning parameter setting ($M = 11$ and $L = 3$ based on **Tuning parameter selection**) yields reasonable average node-FDP (around 7%) and the highest RV-sensitivity. When $M = 13$, the algorithm stops at layer 2 because the remaining leaves might be insufficient to support another layer of analysis. The loss of layer 3 controls the average node-FDP better, but also affects the RV-Sensitivity.

Figure 2 also shows that DYNATE successfully detects many regions with enriched pathogenic RVs. The neutral RV proportions in the DPDs are $1 - \pi$ and in SPDs are close to 1. We checked the RV-FDPs in the identified region: they are consistently lower than $1 - \pi$. These results indicate that DYNATE successfully pinpointed those subregions with highly enriched pathogenic RVs.

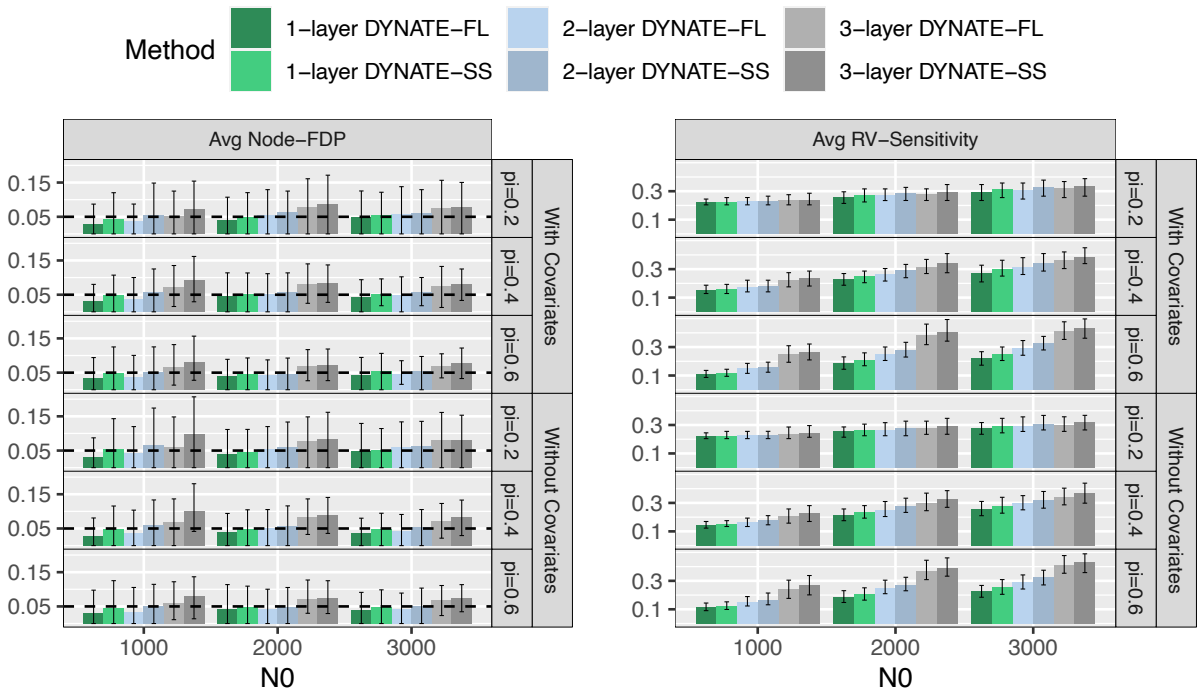

Figure 1: Performance of DYNATE after layer 1, layer 2, and layer 3. The number of cases $N_1 = 1000$; the number of controls $N_0 \in \{1000, 2000, 3000\}$. Different colors mark the DYNATE results with different statistics (FL and SS) and after different layers (1, 2, and 3). The main bars mark the average performance across 100 repetitions; the error bars mark the 90% confidence intervals – the 5% and 95% quantiles over the 100 simulations. The left panel shows the average node-FDP: the dashed horizontal lines mark the desired node-FDR level 5%. The right panel shows the average RV-sensitivity.

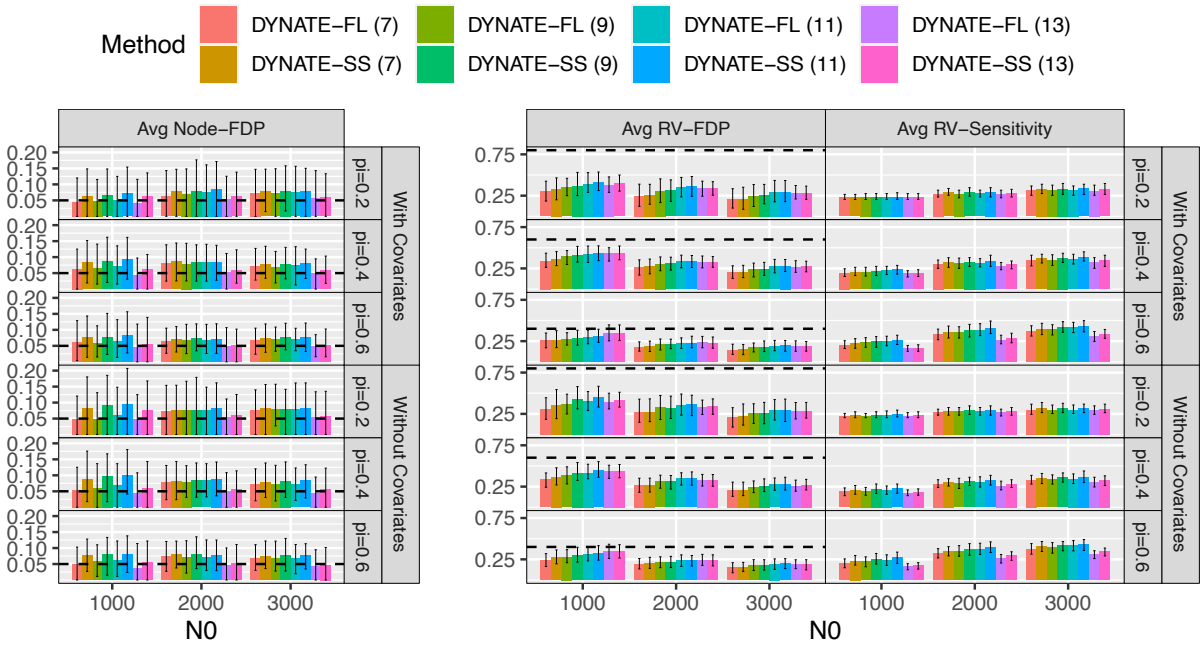

Figure 2: Performance of DYNATE across different leaf sizes. The number of cases $N_1 = 1000$; the number of controls $N_0 \in \{1000, 2000, 3000\}$. Different colors mark the DYNATE results with different statistics (FL and SS) and different leave sizes ($M \in \{7, 9, 11, 13\}$). The main bars mark the average performance across 100 repetitions; the error bars mark the 90% confidence intervals – the 5% and 95% quantiles over the 100 simulations. The left panel shows the average node-FDP: the dashed horizontal lines mark the desired node-FDR level 5%. The middle panel shows the average RV-FDP: the dashed horizontal lines mark the neutral RV proportion in DPDs $1 - \pi$. The right panel shows the average RV-sensitivity.

Simulation Studies: Comparing DYNATE with Competing Methods

We compared the performance of DYNATE with its competing methods: the domain-collapsing method (DC-FL and DC-SS) and the SCANG method (SCANG-B, SCANG-S, and SCANG-O).

Figure 3 compares these methods' domain-level accuracy. DYNATE and DC have similar average domain-FDP, while SCANG's average domain-FDP is higher, sometimes above 25%. In general, DYNATE and SCANG consistently have higher domain-sensitivity and domain-F1 score than DC. When $N_0 = 1,000$, DYNATE has higher domain-sensitivity and domian-F1 score than SCANG; When $N_0 \geq 2,000$, SCANG-O, SCANG-S, DYNATE-FL and DYNATE-SS have similar domain-sensitivity and domian-F1 score; all higher than SCANG-B, DC-FL and DC-SS. The results indicate that DYNATE and SCANG have better domain-level accuracy than DC. When the sample size is small ($N_0 = 1,000$), DYNATE has the best domain-level accuracy.

Figure 4 compares the methods' accuracy in pinpointing pathogenic RVs. Because DC cannot generate RV-level results, they were excluded from the comparison. SCANG, in general, has a higher RV-sensitivity than DYNATE. However, it also has extremely high average RV-FDPs, sometimes close to 1. Although the SCANG-identified regions capture more pathogenic RVs, these regions are so expansive that most of the RVs within them are neutral. This point is verified in Figure 6 for the ALS study. In contrast, DYNATE has a much lower average RV-FDP (lower than $1 - \pi$). The results suggest that DYNATE is much better in identifying pathogenic-RV-enriched regions than SCANG. After we summarized RV-FDP and RV-sensitivity into an overall accuracy score, RV-F1-score, we see that DYNATE has much higher overall RV-level accuracy than SCANG. When $\pi = 0.2$, DYNATE's average RV-level accuracy is at least four times higher than SGANG.

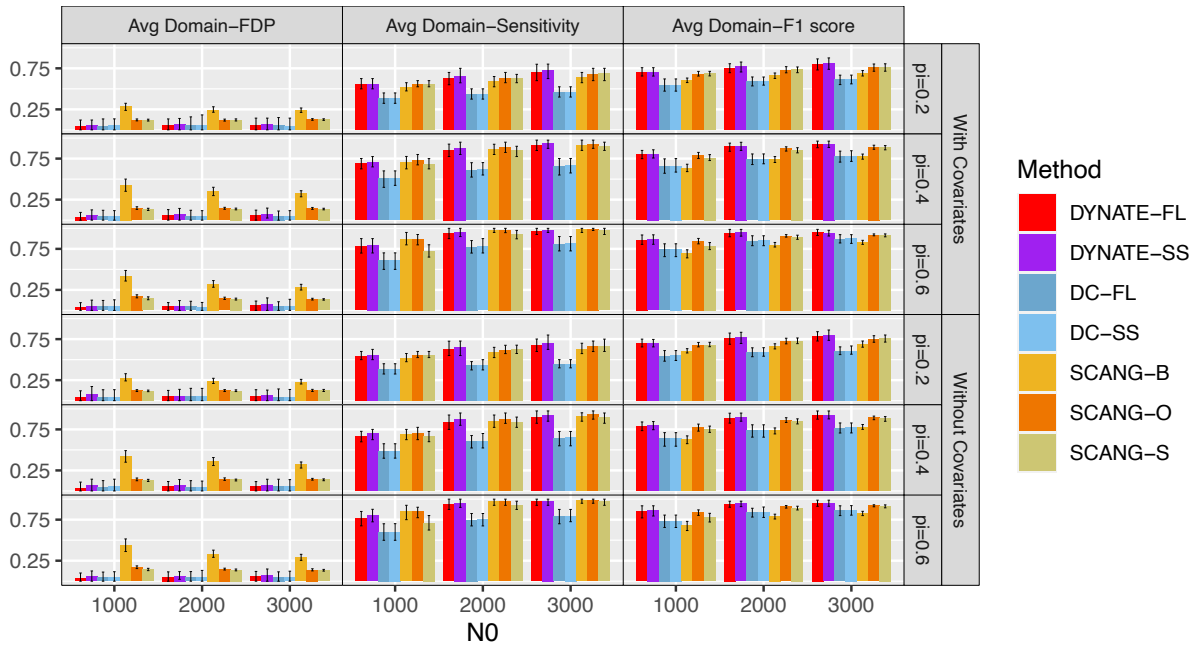

Figure 3: Domain-level performance comparison of the 3-layer DYNATE (DYNATE-FL, DYNATE-SS), DC (DS-FL, DC-SS) and SCANG (SCANG-B, SCANG-S, SCANG-O) under different simulation settings. The number of cases $N_1 = 1000$; the number of controls $N_0 \in \{1000, 2000, 3000\}$. Different colors mark different approaches of DYNATE, DC, and SCANG. The main bars mark the average performance across 100 repetitions; the error bars mark the 90% confidence intervals – the 5% and 95% quantiles over the 100 simulations. The left panel shows the average domain-FDP. The middle panel shows the average domain-sensitivity. The right panel shows the average domain-F1 score.

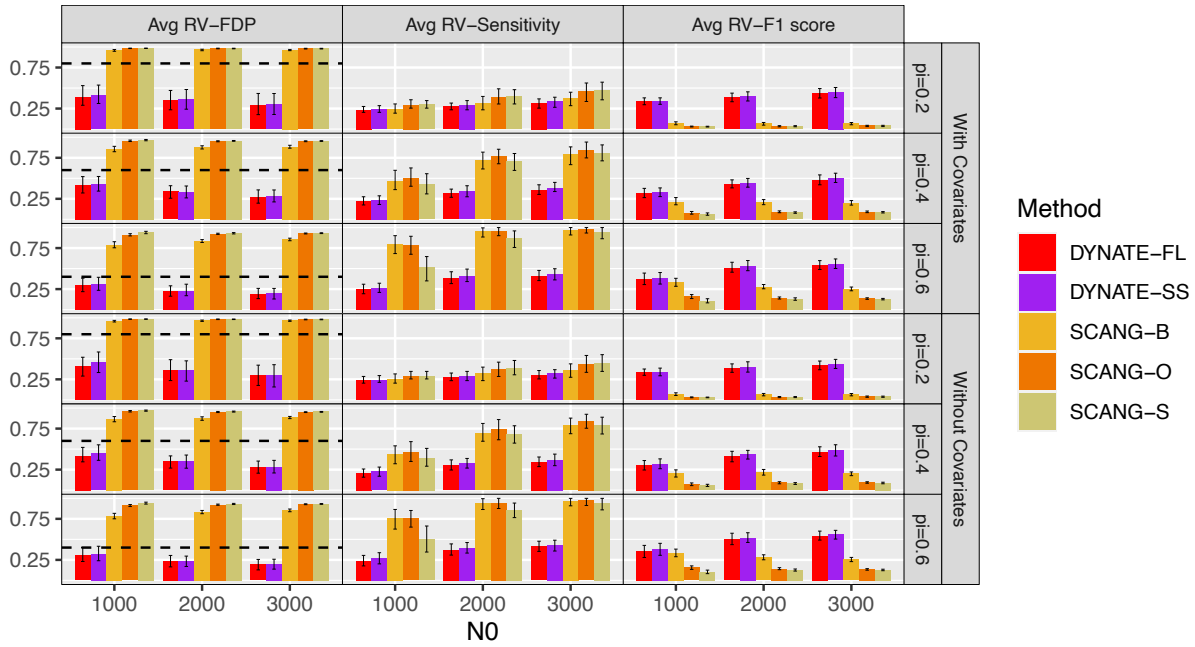

Figure 4: RV-level performance comparison of the 3-layer DYNATE (DYNATE-FL, DYNATE-SS) and SCANG (SCANG-B, SCANG-S, SCANG-O) under different simulation settings. The number of cases $N_1 = 1000$; the number of controls $N_0 \in \{1000, 2000, 3000\}$. Different colors mark different approaches of DYNATE and SCANG. The main bars mark the average performance across 100 repetitions; the error bars mark the 90% confidence intervals – the 5% and 95% quantiles over the 100 simulations. The left panel shows the average RV-FDP. The middle panel shows the average RV-sensitivity. The right panel shows the average RV-F1 score.

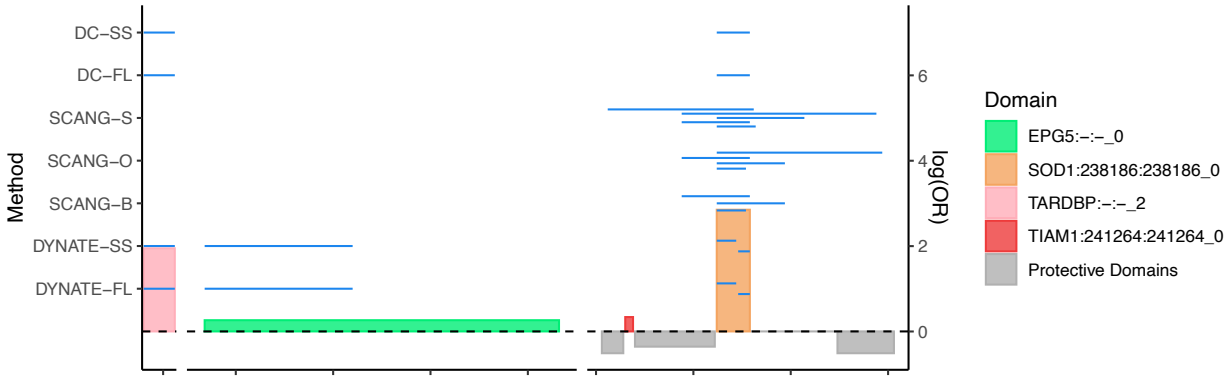

Figure 5: Genetic landscape of the regions identified by DYNATE, SCANG and DC. The blue solid lines mark the identified region. The regions are located in chromosome 1 (chr1), chromosome 18 (chr18) and chromosome 21 (chr21). The first row of x-axis labels the RV id in the identified regions. The RV id is ordered based on their location on each chromosome. The left y-axis labels the method that identified the region. The right y-axis labels the domain-level logOR. Each rectangle bar represents a domain: the bar width represents the number of RVs in the domain; the bar height represents the domain-level logOR. Here, protective domains are those whose logOR < 0.

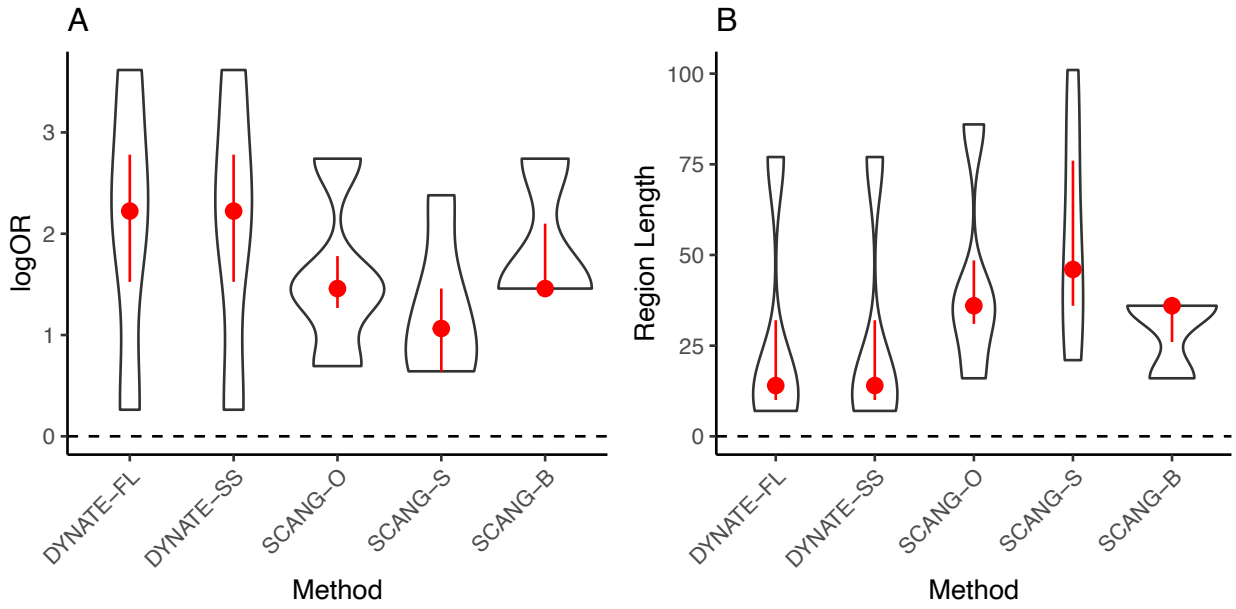

Figure 6: comparison of the DYNATE-identified regions and the SCANG-identfied regions. Violin plot of the distribution of detected regions' length and logOR.The red dot indicates the median and the red vertical lines go from the 25th to the 75th percentile.

Application to an ALS Study

We applied DYNATE, DC, and SCANG to the ALS European sub-cohort, and compared their identified regions in Figure 5. All three methods identify regions overlapping with the domain *SOD1*:238186:238186_0 in *SOD1* (logOR $= 2.85$). *SOD1* is a well-known ALS gene[30,34,35]. While SCANG also identify the regions overlapping with the *SOD1* domain, its regions are too expansive, stretching to the neighboring domains: some of them are in different genes. Also, some SCANG-identified regions extend to both pathogenic (logOR > 0) and protective (logOR < 0) domains, causing confusion in interpretation.

To further compare the identified regions, we compared the lengths of the DYNATE- and SCANG-identified regions (Figure 6B). Here the region length is defined as the number of RVs in the region. In general, the SCANG-identified regions are much longer than the DYNATE-identified regions: the median region lengths of DYNATE-FL and DYNATE-SS are 14; in contrast, the median region lengths of three SCANG approaches range from 36 to 46 (Supplementary Figure S2). We also compared their region-level logORs (Figure 6A). The DYNATE-identified regions have higher logORs: the median region logOR of both DYNATE-FL and DYNATE-SS are 2.22; in contrast, the median region-level logOR of three SCANG approaches range from 1.06 to 1.46. These results suggest that the DYNATE-identified regions are potentially more enriched with the pathogenic RVs than the SCANG-identified regions.

Besides the regions in *SOD1*, DC and DYNATE identified the domain *TARDBP*:–:–_2. *TARDBP* has been reported to associate with ALS in other studies[11,31]. In contrast, SCANG did not identify this domain. This domain has a slightly lower OR than *SOD1*:238186:238186_0; yet it is still high (logOR $= 1.95$).

In addition to the regions in *SOD1* and *TARDBP*, DYNATE identified a new region located in the *EPG5*:–:–_0. In contrast, SCANG and DC failed in identifying any regions overlapping with this domain. *EPG5* is a well-known pathogenic gene of Vici Syndrome[7,27], another neurodegenerative disease. One previous study found that mice with the *EPG5* deficiency would

exhibit key characteristics of neurodegeneration[38]. However, to the best of the authors' knowledge, no literature has reported the association between *EPG5* and ALS on human. Thus, the domain *EPG5*:–:–_0 is an interesting finding from DYNATE and merits further investigation.

Computation Time

We compared the computation time of DYNATE with DC and SCANG. DC only conducts domain level analysis; thus, unsurprisingly, it is computationally efficient. Therefore, we focused on the comparison between DYNATE and SCANG. This comparison is reasonable as both dynamically scan short and long regions. We used the ALS dataset as the benchmark dataset and evaluated the computation time of both methods on a 2.10 GHz Intel Xeon Gold 6252 processor with 16 Gb of memory. DYNATE-FL took 1.7 minutes, and DYNATE-SS took 14.0 minutes. DYNATE-SS took longer because computing the score statistics with the saddle-point-approximation is computationally demanding. Meanwhile, the SCANG package took 32.5 minutes to generate the results of SCANG-B, SCANG-S, and SCANG-O simultaneously. Most of their computation time was used to dynamically scan regions with different lengths. On average, each SCANG approach took 10.8 minutes, while each DYNATE approach took less time on average, 7.8 minutes.

Although both SCANG and DYNATE are varying-window methods, DYNATE is faster because SCANG uses an exhaustive search strategy to search all regions with lengths between the minimal and maximal constraints; in contrast, DYNATE uses a hierarchical aggregation strategy that significantly reduces the number of candidate regions (*i.e.*, the test burden).

Discussion

In this paper, we developed a novel method, DYNATE, to pinpoint disease-associated regions harboring RVs. DYNATE adopts a dynamic aggregation and testing strategy to test for disease

associations of regions with varying lengths. DYNATE is a flexible testing framework and can easily be adapted to handle studies with various disease outcomes, demographic and clinical covariates, statistical models, and test statistics. We name a few examples below.

First, DYNATE can extend beyond case-control studies with binary outcomes. DYNATE is based on leaf P-values. When the outcome is continuous, ordinal, categorical, or censored, etc., we can use an outcome-appropriate statistical model to generate P-values. For example, we can use linear regression for continuous outcomes or Cox regression for time-to-event outcomes. Thus, DYNATE works for studies with different forms of outcomes. Second, DYNATE uses a collapsing approach to summarize the leaf P-values. The collapsing approach assumes that the RVs in the leaves are associated with the phenotype in the same direction. However, DYNATE can also adopt other statistics like SKAT and SKAT-O. These statistics do not restrict the effect direction of the RVs in the same leaf.

Finally, in this paper, we focused on coding regions and partitioned the genome based on homology-based protein domains. For whole-genome studies, we can consider more general partitions, including those based on maps of regulatory elements and their potential connections to target genes in disease-relevant cell types. Further, nodes could be aggregated based on their functional annotation similarities or regulatory relationship. Together, this forms an attractive strategy for biologically informed and interpretable WGS analysis. We plan to explore these approaches in future studies.

Declaration of Interests

The authors declare no competing interests.

Acknowledgment

Andrew Allen, Korous Owzar, Mathew Harms, and Jichun Xie's research was partially supported by NIH 1R01HG012555. Xuechan Li and John Pura's research was funded by Duke University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data and Code Availability

The code to reproduce the experiments is available at `https://github.com/jichunxie/DYNATE_manu`. We also provide an R package to implement the proposed method, available at `https://github.com/jichunxie/DYNATE`. The leaf P-values in the European-based ALS corhort are available at Mendeley (DOI: 10.17632/gvykg8pwyw.1). The aggregated genotypes of case and control cohorts of the mixed population are available for download as version 3.0 of the ALS database `http://www.alsdb.org`.

References

[1] Asimit, J.L., Day-Williams, A.G., Morris, A.P., Zeggini, E., 2012. Ariel and amelia: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73, 84–94. doi:10.1159/000336982.

[2] Bansal, V., Libiger, O., Torkamani, A., Schork, N.J., 2010. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11, 773–785.

[3] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300.

[4] Bhatia, G., Bansal, V., Harismendy, O., Schork, N.J., Topol, E.J., Frazer, K., Bafna, V., 2010. A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol 6, e1000954. doi:10.1371/journal.pcbi.1000954.

[5] Biddle, D.A., Morris, S.B., 2011. Using lancaster's mid-p correction to the fisher's exact test for adverse impact analyses. J Appl Psychol 96, 956–65. doi:10.1037/a0024223.

[6] Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92. doi:10.4161/fly.19695.

[7] Cullup, T., Kho, A.L., Dionisi-Vici, C., Brandmeier, B., Smith, F., Urry, Z., Simpson, M.A., Yau, S., Bertini, E., McClelland, V., et al., 2013. Recessive mutations in epg5 cause vici syndrome, a multisystem disorder with defective autophagy. Nature genetics 45, 83–87.

[8] Daniels, H.E., 1954. Saddlepoint approximations in statistics. The Annals of Mathematical Statistics 25, 631–650. URL: http://dx.doi.org/10.1214/aoms/1177728652, doi:10.1214/aoms/1177728652.

[9] Dey, R., Schmidt, E.M., Abecasis, G.R., Lee, S., 2017. A fast and accurate algorithm to test for binary phenotypes and its application to phewas. Am J Hum Genet 101, 37–49. doi:10.1016/j.ajhg.2017.05.014.

[10] Fisher, R., 1925. Statistical method for research workers. Oliver and Boyd, Edinburgh ;London.

[11] Gelfman, S., Dugger, S., Moreno, C.d.A.M., Ren, Z., Wolock, C.J., Shneider, N.A., Phatnani, H., Cirulli, E.T., Lasseigne, B.N., Harris, T., et al., 2019. A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. Genome research 29, 809–818.

[12] Kanoungi, G., Nothnagel, M., Becker, T., Drichel, D., 2020. The exhaustive genomic scan approach, with an application to rare-variant association analysis. Eur J Hum Genet 28, 1283–1291. doi:10.1038/s41431-020-0639-3.

[13] Katsumata, Y., Fardo, D.W., 2020. Quantitative phenotype scan statistic (qpss) reveals rare variant associations with alzheimer's disease endophenotypes. BMC Med Genet 21, 106. doi:10.1186/s12881-020-01046-6.

[14] Lancaster, H.O., 1961. Significance tests in discrete distributions. Journal of the American Statistical Association 56, 223–234. URL: http://www.jstor.org/stable/2282247.

[15] Lawrence, C.B., Goldman, D.A., 1988. Definition and identification of homology domains. Comput Appl Biosci 4, 25–33. doi:10.1093/bioinformatics/4.1.25.

[16] Lee, S., Abecasis, G.R., Boehnke, M., Lin, X., 2014. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics 95, 5–23.

[17] Lee, S., Wu, M.C., Lin, X., 2012. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762–75. doi:10.1093/biostatistics/kxs014.

[18] Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Tukiainen, T., Birnbaum, D.P., Kosmicki, J.A., Duncan, L.E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D.N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M.I., Moonshine, A.L., Natarajan, P., Orozco, L., Peloso, G.M., Poplin, R., Rivas, M.A., Ruano-Rubio, V., Rose, S.A., Ruderfer, D.M., Shakir, K., Stenson, P.D., Stevens, C., Thomas, B.P., Tiao, G., Tusie-Luna, M.T., Weisburd, B., Won, H.H., Yu, D., Altshuler, D.M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J.C., Gabriel, S.B., Getz, G., Glatt, S.J., Hultman, C.M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Neale, B.M., Palotie, A., Purcell, S.M., Saleheen, D., Scharf, J.M.,

Sklar, P., Sullivan, P.F., Tuomilehto, J., Tsuang, M.T., Watkins, H.C., Wilson, J.G., Daly, M.J., MacArthur, D.G., Exome Aggregation Consortium, 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–91. doi:10.1038/nature19057.

[19] Li, B., Leal, S.M., 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83, 311–21. doi:10.1016/j.ajhg.2008.06.024.

[20] Li, X., Sung, A., Xie, J., 2021. Distance assisted recursive testing. arXiv preprint arXiv:2103.11085 .

[21] Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A.C., Boerwinkle, E., Lin, X., 2019. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. The American Journal of Human Genetics 104, 802–814.

[22] Liu, W., et al., 2013. Gaussian graphical model estimation with false discovery rate control. The Annals of Statistics 41, 2948–2978.

[23] Liu, Y., Xie, J., 2020. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. J Am Stat Assoc 115, 393–402. doi:10.1080/01621459.2018.1554485.

[24] Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., et al., 2012. Cdd: conserved domains and protein three-dimensional structure. Nucleic acids research 41, D348–D352.

[25] Morgenthaler, S., Thilly, W.G., 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutat Res 615, 28–56. doi:10.1016/j.mrfmmm.2006.09.003.

[26] Morris, A.P., Zeggini, E., 2010. An evaluation of statistical approaches to rare variant

analysis in genetic association studies. Genet Epidemiol 34, 188–93. doi:10.1002/gepi.20450.

[27] Nam, S.E., Cheung, Y.W.S., Nguyen, T.N., Gong, M., Chan, S., Lazarou, M., Yip, C.K., 2021. Insights on autophagosome–lysosome tethering from structural and biochemical characterization of human autophagy factor epg5. Communications biology 4, 1–14.

[28] Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., Daly, M.J., 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7, e1001322. doi:10.1371/journal.pgen.1001322.

[29] Pan, W., 2009. Asymptotic tests of association with multiple snps in linkage disequilibrium. Genet Epidemiol 33, 497–507. doi:10.1002/gepi.20402.

[30] Pansarasa, O., Bordoni, M., Diamanti, L., Sproviero, D., Gagliardi, S., Cereda, C., 2018. Sod1 in amyotrophic lateral sclerosis:"ambivalent" behavior connected to the disease. International journal of molecular sciences 19, 1345.

[31] Pesiridis, G.S., Lee, V.M.Y., Trojanowski, J.Q., 2009. Mutations in tdp-43 link glycine-rich domain functions to amyotrophic lateral sclerosis. Human molecular genetics 18, R156–R162.

[32] Petrovski, S., Todd, J.L., Durheim, M.T., Wang, Q., Chien, J.W., Kelly, F.L., Frankel, C., Mebane, C.M., Ren, Z., Bridgers, J., et al., 2017. An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. American journal of respiratory and critical care medicine 196, 82–93.

[33] Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A.S., Goldstein, D.B., 2019. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nature Reviews Genetics 20, 747–759.

[34] Renton, A.E., Chiò, A., Traynor, B.J., 2014. State of play in amyotrophic lateral sclerosis genetics. Nature neuroscience 17, 17–23.

[35] Rosen, D.R., Siddique, T., Patterson, D., Figlewicz, D.A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J.P., Deng, H.X., et al., 1993. Mutations in cu/zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362, 59–62.

[36] Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics 89, 82–93.

[37] Xie, J., Li, R., 2018. False discovery rate control for high dimensional networks of quantile associations conditioning on covariates. J R Stat Soc Series B Stat Methodol 80, 1015–1034. doi:10.1111/rssb.12288.

[38] Zhao, Y.G., Zhao, H., Sun, H., Zhang, H., 2013. Role of epg5 in selective neurodegeneration and vici syndrome. Autophagy 9, 1258–1262.