# An allelic series rare variant association test for candidate gene discovery

Zachary R McCaw[1,*], Hari Somineni[1], Michael Bereket[1], Christoph Klein[1],

Theofanis Karaletsos[1], Francesco Paolo Casale[2], Daphne Koller[1], Thomas W Soare[1,*]

December 22, 2022

**Abstract**

Allelic series are of candidate therapeutic interest due to the existence of a dose-response relationship between the functionality of a gene and the degree or severity of a phenotype. We define an allelic series as a gene in which increasingly deleterious mutations lead to increasingly large phenotypic effects, and develop a gene-based rare variant association test specifically targeted for the identification of allelic series. Building on the well-known burden and sequence kernel association tests (SKAT), we specify a variety of association models, covering different genetic architectures, and integrate these into a COding-variant Allelic Series Test (COAST). Through extensive simulations, we confirm that COAST maintains the type I error and improves power when the pattern of coding-variant effect sizes increases monotonically with mutational severity. We applied COAST to identify allelic series for 4 circulating lipid traits and 5 cell count traits among 145,735 subjects with available whole exome sequencing data from the UK Biobank. Compared with optimal SKAT (SKAT-O), COAST identified 29% more Bonferroni significant associations with circulating lipid traits, on average, and 82% more with cell count traits. All of the gene-trait associations identified by COAST have corroborating evidence either from rare-variant results in the full cohort (Genebass, $N = 400$K), or from common variant results in the GWAS catalog. In addition to detecting many gene-trait associations present in Genebass using only a fraction (36.9%) of the sample, COAST detects associations, such as *ANGPTL4* with triglycerides, that are absent from Genebass but which have clear common variant support.

**Keywords**: Allelic series, Candidate genes, Rare variant association testing, Whole exome sequencing

[1] Insitro, South San Francisco, CA, USA.

[2] Department of Informatics, Technical University of Munich, Munich, Germany.

[*] Correspondence: ZRM (zmccaw@insitro.com), TWS (tsoare@insitro.com).

# Introduction

The term *allelic series* refers to a collection of mutations or alleles in a gene or pathway that leads to a gradation of possible phenotypes [1]. Genes demonstrating allelic series are of candidate therapeutic interest due to the existence of a dose-response relationship between the functionality of the gene and the magnitude or severity of the phenotype [2]. Learning such dose-response relationships from 'experiments of nature' has confirmed the role of *PCSK9* in familial hypercholesterolaemia, LDL cholesterol, and coronary artery disease (as reviewed in [3]). More recently, Duan *et al* [4] described a naturally occurring allelic series in the *BHLHA9* locus associated with the clinical severity and penetrance of congenital limb malformations. Allelic series can also be generated experimentally, as was done recently by Mohajeri *et al* [5], who used CRISPR to engineer an allelic series of *MEF2C* mutations *in vitro*, providing a cellular model for human neurodevelopmental disorders. Motivated by these examples, we introduce a rare variant association test intended for identifying allelic series in large genotype-phenotype cohorts, such as the UK Biobank [6]. Hereafter, by allelic series, we refer specifically to an individual gene wherein increasingly deleterious mutations have increasingly large phenotypic effects.

Genome-wide association studies (GWAS) canonically seek to associate common genetic variants, typically those having a minor allele frequency (MAF) exceeding $1-5\%$, with complex traits and diseases [7]. The advent of whole genome and whole exome sequencing studies has enabled rare variant association analysis, wherein associations are sought for variants having lower MAFs [8, 9]. Because even large cohorts may have few examples of such subjects, specialized methodology is required to reliably associate phenotypes with rare variants [10]. For a given effect size, the power of single-variant association tests declines precipitously with the MAF [11]. To overcome this limitation, rare variant association tests typically aggregate signal within a biologically meaningful region, such as a gene.

Two major subdivisions of rare variant association tests are burden tests and variance component tests. Burden tests [12, 13, 14] associate the phenotype with a single genetic score formed by taking a weighted sum across the variants in a region. These tests differ with respect to the assignment of weights to variants but share in common the assumption that all rare variants in the region affect the phenotype in the same direction [10]. When this is

not the case, the power of burden testing is diminished. Variance component tests, notably the sequence kernel association test (SKAT) [15], relax the assumption of a common direction of effect by instead supposing the effect sizes are random, following a distribution with mean zero and finite variance. SKAT performs well when some of the rare variants in the region are non-causal or have conflicting directions of effect, but loses power when the genetic architecture is consonant with that assumed by the burden test. In practice the genetic architecture of a trait is seldom known. Optimal SKAT (SKAT-O) [16] adaptively combines the SKAT and burden tests to provide high power for rare variant association testing under the scenarios assumed by either the component tests.

There are several existing approaches that incorporate functional annotations to generally improve power for gene-based association testing [17, 18, 19]. For example, STAAR [19] uses annotation principle components to upweight the probability of a variant being causal based on the fractional rank of that variant's annotation score, relative to all sequenced variants, within functional categories such as epigenetics, conservation, and protein function. In a different vein, Regenie [20] allows for the definition of masks that specify which variants to include in an association test based on functional categories. However, no existing methods are specifically targeted to the identification of allelic series. We fill this gap by defining a rare variant association test that incorporates a set of adjustable weights to encourage rejection of the null hypothesis in cases where the expected effect of a variant on the phenotype increases monotonically with the severity of the mutation. Our approach focuses on the rare coding variants within a gene, as annotated by the Ensembl Variant Effect Predictor (VEP) [21]. Specifically, our test operates on three classes of variants: the benign missense variants (BMVs), deleterious missense variants (DMVs), and protein truncating variants (PTVs). Building on the burden and SKAT tests, we define a variety of gene-based association models targeting different genetic architectures. These tests are aggregated into a single, omnibus, COding-variant Allelic Series Test (COAST).

Through extensive simulation studies, we validate that COAST controls the type I error in the absence of a genotype-phenotype association, and benchmark its power relative to SKAT-O [10] under a variety of genetic architectures. We select SKAT-O as the comparator due to its widespread use, computational efficiency, and direct applicability using only the

data required by COAST. We note, however, that the set of genes targeted by COAST differs from that targeted by SKAT-O, and we are aware of no existing association test that specifically targets allelic series. Using whole exome sequencing data on up to 145,735 subjects and up to 17,225 genes from the UK Biobank, we apply COAST to identify allelic series for 4 circulating lipid traits and 5 cell count traits. Compared with SKAT-O, COAST identified 29% more Bonferroni significant associations with circulating lipids, on average, and 82% more with cell counts. Notably, all of the additional genes detected by COAST have some supporting evidence from published common or rare variant association studies, however not all of the rare variant associations have previously been reported by SKAT-O applied to the full UK Biobank (Genebass) [22].

# Material and Methods

## Setting

Let $Y$ denote a quantitative phenotype (e.g. circulating cholesterol or erythrocyte count), $X$ a set of covariates (e.g. age, sex, genetic principal components), and $G = (G_1, \cdots, G_J) \in \{0, 1, 2\}^J$ the additively coded genotype at the $J$ rare coding variants within a given gene. To each variant $j$, assign a categorical functional annotation $A_j \in \{1, \cdots, L\}$. For the present work, we specialize to the following $L = 3$ VEP categories:

$$A_j = \begin{cases} 1, & \text{benign missense variant (BMV)}, \\ 2, & \text{deleterious missense variant (DMV)}, \\ 3, & \text{protein truncating variant (PTV)}. \end{cases}$$

For a given subject, let $N_l$ count the total number of category $l$ alleles present in the gene:

$$N_l = \sum_{j=1}^{J} \mathbb{I}(A_j = l) G_j,$$

and let $N = (N_1, N_2, N_3)$ denote the vector of category-level allele counts.

## Models

### Baseline Model

The **baseline allelic series model** takes the form:

$$\mathbb{E}(Y|N, X) = \alpha + \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 + \gamma' X. \tag{1}$$

(1) is similar to a standard burden model insofar as allele counts from distinct variants having a common functional annotation are aggregated into a single score [10]. However, unlike the standard burden test, aggregation is grouped by annotation category, and a separate coefficient is estimated for each. (1) is also a count-based model: the expected value of the phenotype increases linearly by $\beta_l$ for each additional category $l$ allele. An indicator-based model replaces the category $l$ allele count $N_l$ by an indicator that $N_l$ is non-zero:

$$\mathbb{E}(Y|N, X) = \alpha + \beta_1 \mathbb{I}(N_1 > 0) + \beta_2 \mathbb{I}(N_2 > 0) + \beta_3 \mathbb{I}(N_3 > 0) + \gamma' X.$$

The indicator-based model posits a genetic architecture in which the presence versus absence of category $l$ alleles has an effect on the phenotype, but beyond the first, the number of category $l$ alleles is immaterial. For example, the presence of any PTVs in a gene may abrogate its function, but additional PTVs beyond the first may have no effect. Under both the count and indicator models, the null hypothesis that the rare coding variants within the gene are unrelated to the phenotype is evaluated by testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ using a standard Wald test with 3 degrees of freedom [23].

### Allelic Series Sum Model

The baseline model provides a test for identifying genes associated with the phenotype and allows variants belonging to different annotation categories to have differing effects, however it is not yet tailored to the identification of allelic series. We next introduce a set of fixed weights $w = (w_1, \ldots, w_L)$, one for each variant category, in order to specify the *pattern* of effect sizes sought. Any monotone increasing pattern corresponds to an allelic series, but as a simple default, we adopt $w = (1, 2, 3)$. We define the **allelic series sum model** as a

specialization of (1) under the constraint that $\beta_l = \beta w_l$:

$$\mathbb{E}(Y|N, X) = \alpha + \beta(w_1 N_1 + w_2 N_2 + w_3 N_3) + \gamma' X. \tag{2}$$

(2) continues to allow different categories to have differing effects, but the constraint requires that the effect size vector $(\beta_1, \beta_2, \beta_3)$ is proportional to the allelic series weights:

$$(\beta_1, \beta_2, \beta_3) = \beta(w_1, w_2, w_3).$$

Compared with (1), (2) reduces the number of free parameters from 3 to 1. The null hypothesis of no association between the gene and the phenotype is evaluated by testing $H_0 : \beta = 0$. The baseline model may outperform the sum model in cases of allelic series whose pattern of effect size differs markedly from that sought by $w$, e.g. $(\beta_1, \beta_2, \beta_3) \propto (0, 0, 1)$. Conversely, the sum model will have a power advantage the more nearly the true pattern of effect sizes $(\beta_1, \beta_2, \beta_3)$ is proportional to the prespecified allelic series weights $(w_1, w_2, w_3)$.

### Allelic Series Max Model

Focusing on the parenthetical term in (2), we see that the allelic series sum model aggregates the per-category allele counts $(N_1, N_2, N_3)$ by taking a weighted sum. By using different methods of aggregation, we can construct additional allelic series tests. The **allelic series max model** aggregates via the weighted maximum:

$$\mathbb{E}(Y|N, X) = \alpha + \beta \max(w_1 N_1, w_2 N_2, w_3 N_3) + \gamma' X. \tag{3}$$

(3), particularly the indicator version, is motivated by a genetic architecture in which the expected phenotype is determined by the most deleterious variant present. For example, the presence of a BMV may have negligible phenotypic impact in the presence of a PTV. As for the sum model, the null hypothesis of no association is evaluated by testing $H_0 : \beta = 0$.

### Allelic Series SKAT Model

The preceding tests all make the burden-type assumption that all variants affect the phenotype in the same direction. The sequence kernel association test (SKAT) [15] was introduced to allow variants to have differing directions of effect. Briefly, SKAT posits that each variant $G_j$ has a separate effect $\beta_j$ and that these effects are drawn at random from an unspecified distribution with mean zero and variance $w_j^{\text{SKAT}}\tau^2$, where $\tau^2$ is a common variance component. The null hypothesis of no association is evaluated by testing $H_0 : \tau^2 = 0$. To develop an allelic series test that allows rare variants to have differing directions of effect, we build on the existing SKAT framework but modify the SKAT weights to incorporate the allelic series weights:

$$w_j^{\text{SKAT}} = \frac{\sum_{l=1}^{3} \mathbb{I}(A_j = l) w_l}{f_j(1 - f_j)}. \tag{4}$$

Under (4), the variance of the distribution from which $\beta_j$ is drawn is directly proportional to the weight of the annotation category to which $G_j$ belongs, and inversely proportional to the variance of $G_j$ ($f_j$ denoting the minor allele frequency of $G_j$). In the allelic series setting where $w_l$ increases monotonically from BMVs to DMVs to PTVs, the variance of the distribution from which $\beta_j$ is drawn is larger for more deleterious variants, allowing such variants to have larger effects on the phenotype.

### Allelic Series Omnibus Test

We have now defined several burden-type tests and a SKAT-based test. Which test is best-powered for a particular gene-phenotype pair will depend on which best approximates the true genetic architecture. Moreover, for a given phenotype, a single test may not be optimal for all genes, as the genetic architecture may vary gene by gene. To obviate the need for deciding among tests, we apply each of the preceding then combine their p-values via the Cauchy combination method [24, 25]. Given $M$ p-values $(p_1, \ldots, p_M)$, the **omnibus test statistic** is defined by:

$$T_{\text{omni}} = \sum_{m=1}^{M} w_m^{\text{omni}} \tan\{\pi(0.5 - p_m)\}, \tag{5}$$

where the omnibus weights $w_m^{\text{omni}}$ are selected to give burden-type and SKAT-type tests equal weight. Finally, the omnibus p-value is given by:

$$p_{\text{omni}} = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{T_{\text{omni}}}{\sum_{m=1}^{M} w_m^{\text{omni}}}\right).$$

We refer to the omnibus test and its associated p-value as the COding-variant Allelic Series Test (COAST).

## Simulation Methods

Although the sample size in our data application exceeds 100K, our simulations consider a sample size of only 1K to ensure the proposed method remains valid in small samples. As the sample size increases, the operating characteristics of any method that utilizes asymptotic approximations (including the burden and SKAT tests) will further improve. Rare variant genotypes were simulated having minor allele frequencies uniformly distributed over the range $[0.1\%, 1.0\%]$ (variants had a minimum minor allele count of 1, which corresponds to a MAF of 0.1%). Each variant was randomly assigned an annotation of BMV, DMV, or PTV. For each subject, covariates were simulated representing age, sex, and 3 genetic principal components (PCs); 10% of phenotypic variation was explained by age, 10% by sex, and 20% by the 3 genetic PCs collectively. For type I error simulations, the phenotype was independent of genotype. For power simulations, phenotypes were generated under multiple genetic architectures. Each simulated gene contained $10^2$ variants. Burden phenotypes were simulated from the {baseline, sum, max} models, varying the generative effect sizes $(\beta_1, \beta_2, \beta_3)$. The sum and max phenotypes were simulated by fixing $\beta = 1$ and varying the generative weights $(w_1, w_2, w_3)$. SKAT phenotypes were simulated by setting the effect size of each variant $\beta_j$ to the product of a random sign $r_j$ and a fixed effect size $\beta_l$ corresponding to the variant's annotation: $\beta_j = r_j \sum_{l=1}^{3} \mathbb{I}(A_j = l)\beta_l$. Note that the sum and max architectures are not compatible with a SKAT phenotype.

## Preparation of UK Biobank Data

### Whole Exome Sequencing Data

Variants in the UKBB WES variant callset (N=200K) were filtered out if they were located > 100 bp away from the exome capture regions, located within the ENCODE black list [26], or within low-complexity regions [27]. We further filtered to autosomal, biallelic variants meeting the following criteria: allele balance fractions of $\leq 0.1$ (homozygous) or $\geq 0.25$ (heterozygous), call rate $\geq 0.95$, mean depth (DP) $12 \leq \text{DP} < 200$, mean genotype quality GQ $\geq 20$, Hardy-Weinberg equilibrium $p > 1 \times 10^{-6}$, and inbreeding coefficient $F > -0.3$. Variants were annotated with VEP v96 [21] using the following transcript consequences (for any transcript of a gene): protein-truncating variants (PTV): `splice acceptor variant`, `splice donor variant`, `stop gained`, or `frameshift variant`; missense: `missense variant`, `inframe deletion`, `inframe insertion`, `stop lost`, `start lost`, or `protein altering variant`. Missense variants were annotated to be damaging (DMV) if predicted to be `probably damaging` or `possibly damaging` by PolyPhen2 [28] and `deleterious` or `deleterious low confidence` by SIFT [29]. Missense variants were annotated to be benign (BMV) if predicted to be `benign` by PolyPhen2 and `tolerated` or `tolerated low confidence` by SIFT [30, 31]. PTVs were further filtered to remove low confidence loss-of-function by LOFTEE [32].

### Quality Control

Sample-level quality control (QC) was performed using 104,616 high-confidence variants passing the above filtering criteria, and had minor allele count of $\geq 5$, and were LD-pruned to $r^2 < 0.05$. Samples were first filtered to unrelated individuals without sex chromosome aneuploidy, within $\pm 7$ standard deviations of the mean of the first six genotype principal components [33], self-reported ethnic background of `White British`, `Irish`, or `White`, had call rate $\geq 0.95$, mean DP $\geq 19$, and mean GQ $\geq 47.5$. We regressed the top 20 genetic principal components from the following sample QC metrics and removed samples that were > 4 median absolute deviations (MAD) from the median for any metric: number of SNPs, number of heterozygous variants, number of homozygous alternate variants, number of insertions, number of deletions, number of transitions, number of transversions, the ratio between the number of transitions and transversions, the ratio between the number of heterozygous

and homozygous alternate variants, and the ratio between the number of insertions and deletions [32]. The sample size after filtering was 145,753. All data preprocessing was performed in hail 0.2.37 [34].

### Association Analysis

Processed WES data were restricted to BMVs, DMVs, and PTVs with a sample minor allele frequency $\leq 1\%$. Following [35], ultra-rare variants, those with a sample MAC $\leq 10$, were collapsed into a single pseudo marker separately for each gene $\times$ variant category. Genes were required to have at least 3 distinct rare variants (including the pseudo marker) for inclusion in the genome-wide screen; 17,225 genes passed this threshold. The circulating lipid phenotypes analyzed included low and high density lipoprotein (LDL and HDL; UKB IDs 30780 and 30760), triglycerides (30870), and total cholesterol (30690). The cell count phenotypes included erythrocytes (30010), leukoctyes (30000), lymphocytes (30120), neutrophils (30140), and thrombocytes (30080). In all cases, the phenotype was transformed to normality by applying the rank-based inverse normal transformation (INT) [36]. Covariates included age (to degree 3), genetic sex, age $\times$ sex interaction (to degree 3), 20 genetic principal components, and an indicator for being among the first 50K exomes sequenced.

## Overlap Analysis

For each trait of interest, common-variant associations were obtained from the NHGRI-EBI GWAS Portal [37, 38] and filtered to those having $p < 5 \times 10^{-8}$. Rare-variant associations from both putative loss of function (pLoF) variants and missense variants, including low-confidence pLoF variants and in-frame insertions or deletions, were obtained from the Genebass Portal [22, 39]. Associations were filtered to those Bonferroni significant based on the number of genes with non-missing p-values, and the union of significant associations from the pLoF and missense analyses was taken.

# Results

## Simulation Studies

Simulation studies were performed to evaluate the type I error (validity) and power of the proposed COding-variant Allelic Series Test (COAST). Under the null hypothesis of no association between the gene and the phenotype, the p-values from COAST were uniformly distributed (Panel A of **Figure 1**). The expected $\chi^2$ was 0.989 (95% CI, 0.986 to 0.992) and a genomic inflation factor of $\lambda = 0.881$, suggesting that COAST is slightly conservative in the absence of association. **Supplemental Figures 1-2** plot the expected $\chi^2$ statistics and p-value distributions for the individual components of COAST. The allelic series burden tests (base, max, sum) had expected $\chi^2$ statistics at the reference level, while the allelic series SKAT test was slightly conservative, consistent with the original publication [15]. Overall, these findings support the validity of COAST, and its components, for genome-wide analyses.

Panels B and C of **Figure 1** compare the power to detect allelic series between COAST and SKAT-O applied either to all variants (BMVs, DMVs, PTVs) or applied to PTVs only. The weights of COAST were fixed at $w = (1, 2, 3)$. The true pattern of effect sizes $\beta = (\beta_1, \beta_2, \beta_3)$ for BMVs, DMVs, and PTVs is indicated by the tuple beneath each genetic architecture. The first 3 genetic architectures correspond with models (1) through (3), while the remaining architectures are variations on (1) with different choices of $\beta$ (see the simulation methods for more detail). For the burden phenotype in Panel B, COAST was uniformly more powerful than SKAT-O, although the power advantage declined with the $L_1$ distance of the generative coefficient from that assumed by COAST (**Supplemental Figure 4**). Even in the case of uniform effect sizes across variant classes $\beta = (1, 1, 1)$, meaning that the variant annotation was uninformative, COAST was more powerful than SKAT-O. This is likely due to its ability to integrate evidence across multiple models via (5). For the SKAT phenotype in Panel C, COAST was most powerful when the allelic series weights matched the true pattern of effect sizes, but SKAT-O tended to acquire a slight power advantage when the weights did not match. To ensure that COAST remains powerful for SKAT-type phenotypes when its weights may be misspecified, we extended COAST to optionally incorporate SKAT-O applied to all variants and/or PTVs among the p-values combined by the omnibus tests (**Supplemental Figure 3**). Overall, these findings suggest that COAST is expected to provide the greatest

power benefit when the allelic series weights are aligned with the true pattern of effect sizes. Moreover, COAST is particularly advantageous for the burden architecture, meaning that the variants effect the phenotype in the same direction.

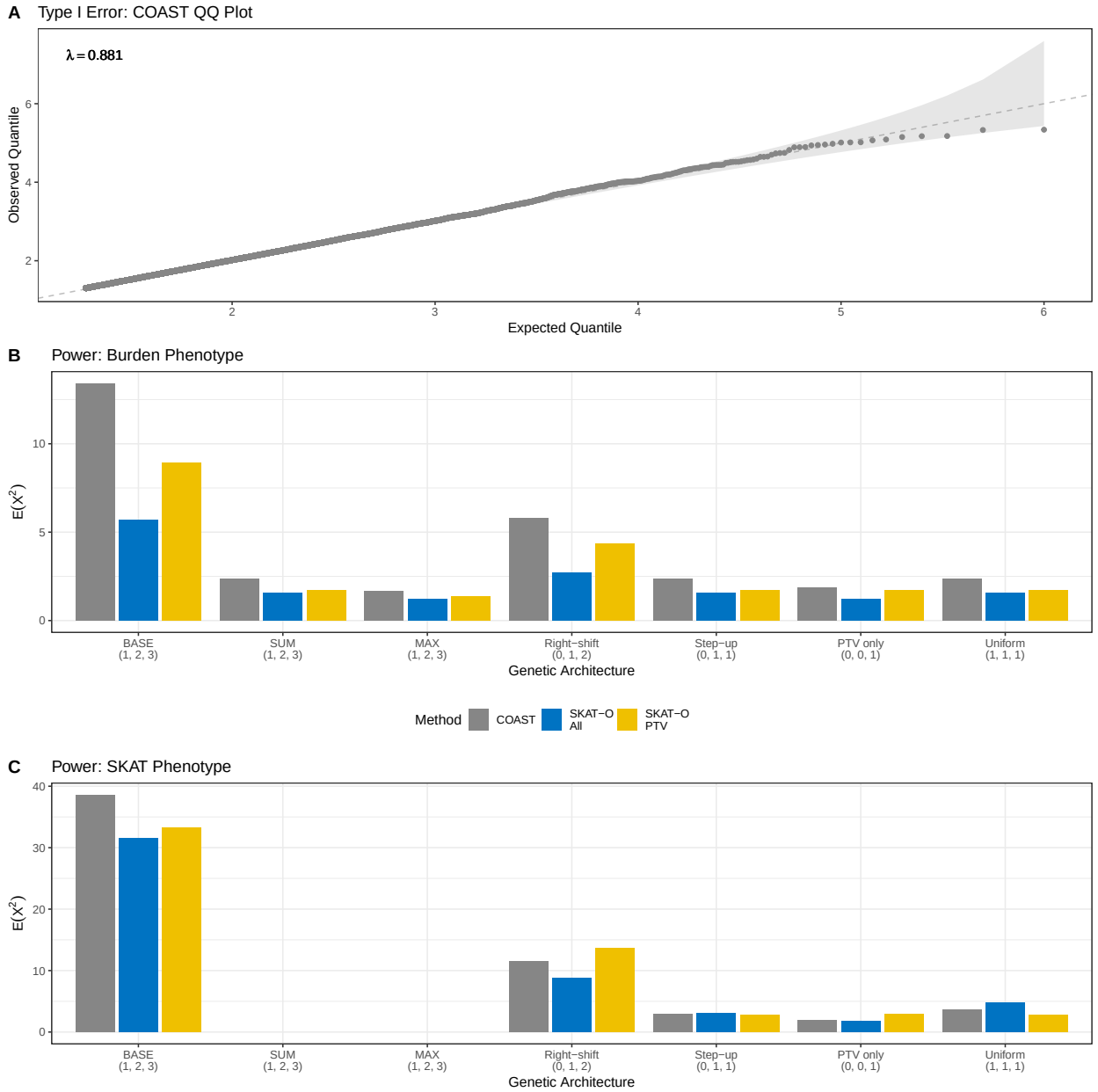


Figure 1: **Type I Error and Power of the Coding Allelic Series Test.** The sample size is $N = 10^3$, each gene contained $10^2$ variants, and results are aggregated across $10^6$ simulations in the case of type I error, and $10^4$ simulations in the case of power. **A:** Observed versus expected quantiles of $-\log_{10}(p)$ for COAST. Adherence to the dashed 45° line indicates that the observed p-values are uniformly distributed. $\lambda$ is the genomic inflation factor. **B & C:** Power across various genetic architectures. Each tick corresponds to a different genetic architecture, including the BASE (1), SUM (2), and MAX (3) architectures. The tuple beneath each tick denotes the relative effect sizes $(\beta_1, \beta_2, \beta_3)$ of benign missense variants, deleterious missense variants, and protein truncating variants. The allelic series weights remained fixed at $w = (1, 2, 3)$. For the phenotype, the effect sizes of all variants were in the same direction, whereas for the SKAT phenotype, the direction of effect was randomized. The SUM and MAX models are omitted for the SKAT phenotype because these architectures are incompatible with randomized signs.

## Application to Circulating Lipid Phenotypes

As our first case study, we applied COAST and SKAT-O to identify genes associated with the following circulating lipid phenotypes: cholesterol, HDL, LDL, and triglycerides. **Figure 2** presents the number of Bonferroni significant genes identified by COAST and SKAT-O. On average, COAST identified 29% more significant associations than SKAT-O applied to all variants (SKAT-O ALL), and 3.3-times as many associations as SKAT-O applied to PTVs only (SKAT-O PTV)(**Supplemental Table 1**). Across the union of genes declared significant by any association test, the average $\chi^2$ statistic of COAST was 24% higher than that of SKAT-O ALL and 2.3-times higher than that of SKAT-O PTV (**Supplemental Table 2**). Among 43 total associations, 9 (21%) were unique to COAST, 33 (77%) were in common, and 1 (2%) was unique to SKAT-O ALL (**Supplemental Figure 6**). The uniform quantile-quantile plots in **Supplemental Figure 5** evince no inflation of the type I error.

Taking cholesterol as an example, a down-sampling analysis was performed in which nested subsets of 25%, 50%, and 75% of the total sample were selected, and association analysis was performed on each (**Supplemental Figure 7**). At all sample sizes, COAST identified more Bonferroni significant associations than SKAT-O. Using only 50% of the sample, COAST recovered as many of the full-sample associations as SKAT-O ALL.

The mirrored Manhattan plots in **Figure 3** show that the signals identified by COAST and SKAT-O all are generally concordant, with several additional associations crossing the Bonferroni threshold under COAST. Manhattan plots for COAST stratified by trait are available in **Supplemental Figure 8**. **Supplemental Tables 3-4** assess the extent to which the gene-trait associations identified by COAST and SKAT-O have existing support from rare variants (Genebass) or common variants (GWAS Catalog). All gene-trait associations had support from at least one of Genebass or the GWAS Catalog, and most had support from both. Many well-known gene-trait associations appear as positive controls, for example *APOB* ($p = 3.8 \times 10^{-185}$) and *PCSK9* ($p = 5.5 \times 10^{-68}$) with LDL, *ABCA1* ($p = 1.3 \times 10^{-121}$) with HDL, and *ANGPTL3* ($1.0 \times 10^{-54}$) with triglycerides. It should be noted that while most of the associations reported here are already present in Genebass, the sample size of this analysis is only 36.9% that of Genebass ($145,735$ vs. $394,841$). The two genes, both associated with triglycerides, that were identified by COAST and were not

present in Genebass were *ANGPTL4* ($p = 6.7 \times 10^{-14}$) and *A1CF* ($p = 3.4 \times 10^{-7}$).

Figure 4 presents the pattern of effect sizes among genes significantly associated with circulating cholesterol by COAST. Analogous figures for the remaining lipid phenotypes are provided in **Supplemental Figure 9**. With few exceptions, the mean effect size magnitude increases monotonically from BMVs to DMVs to PTVs (e.g. *PCSK9*). The presence of genes with non-monotonic effect size patterns (e.g. *LDLR*) underscores that, although COAST is tailored for the identification of allelic series, it is not powerless against other genetic architectures. The two genes associated with triglycerides by COAST that were not present in Genebass (*ANGPTL4* and *A1CF*) both had a monotone increasing pattern of effect sizes (**Supplemental Figure 9**). The one gene (*APOE*) associated with LDL by SKAT-O but not COAST had an anti-allelic series pattern, meaning that the effect size magnitude decreased monotonically from BMVs to DMVs to PTVs.

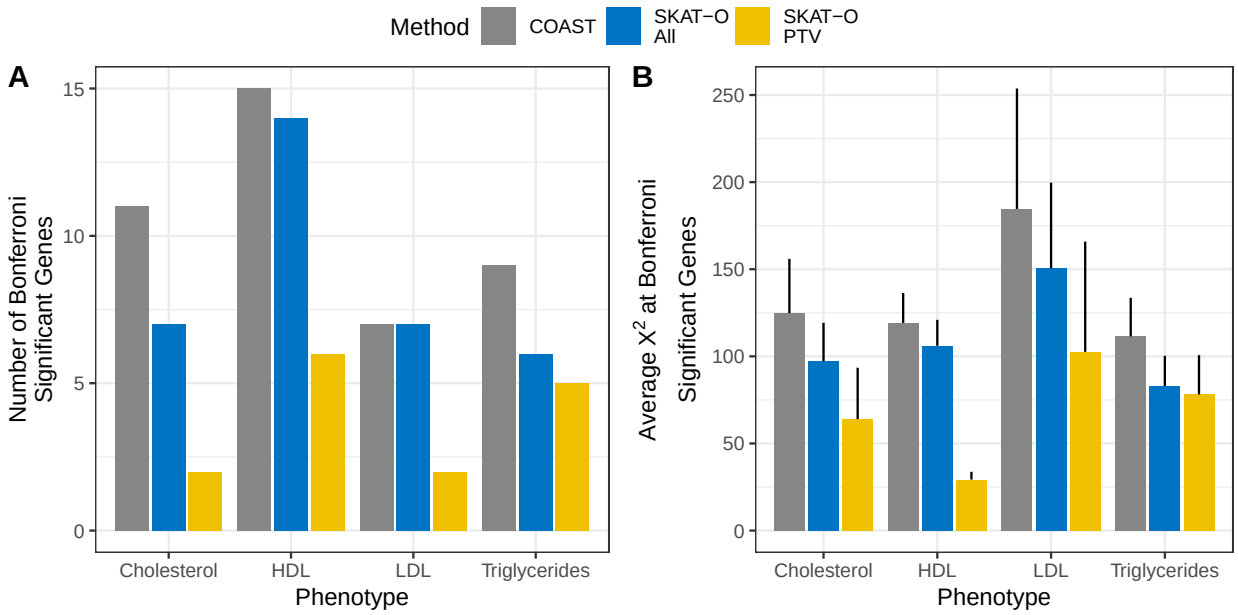


Figure 2: **Number and significance of genes significantly associated with circulating lipid phenotypes. A.** Number of significantly associated genes by phenotype and association test. **B.** Average $\chi^2$ across the union of genes considered Bonferroni significant by any association test. Error bars are 95% confidence intervals.

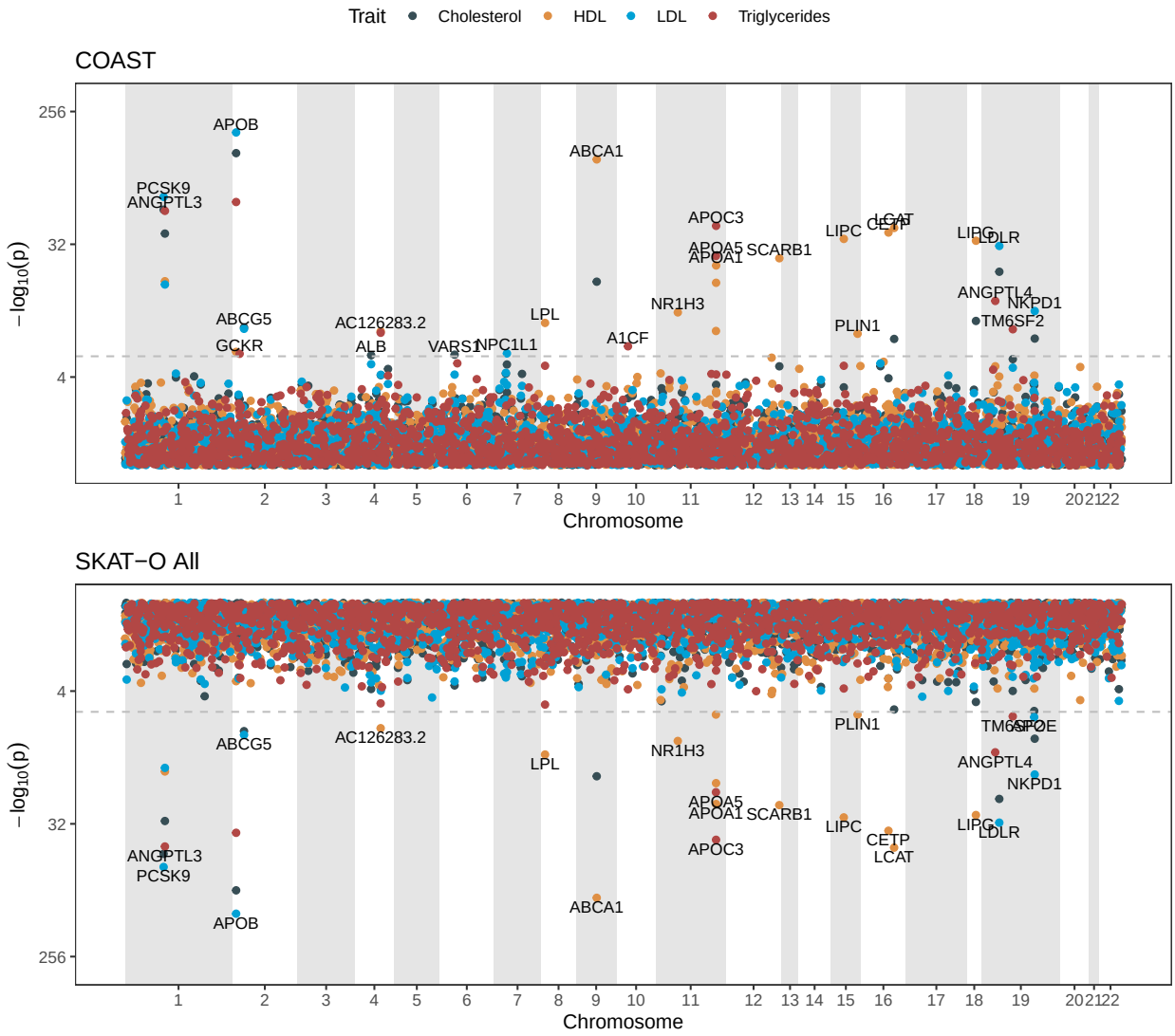


Figure 3: **Mirrored Manhattan plots for lipid phenotypes.** Upper panel is the coding allelic series test (COAST). Lower panel is SKAT-O applied to all rare coding variants.

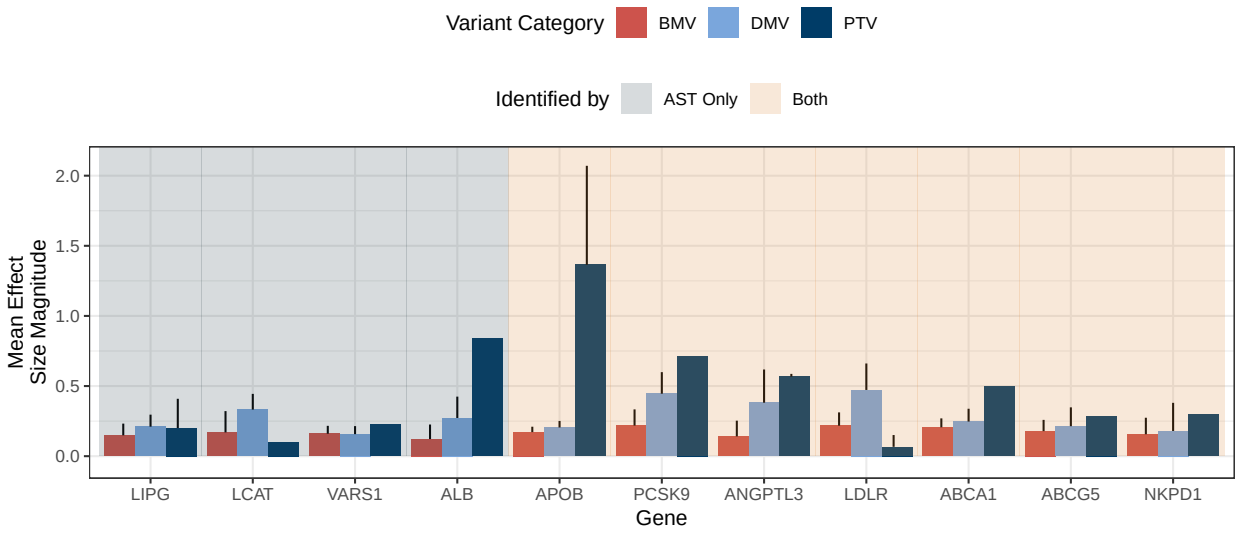


Figure 4: **Effect size patterns among genes significantly associated with circulating cholesterol by the allelic series test.** Effect sizes were estimated by standard linear regression. Each bar represents the mean effect size for variants within a gene and variant category. Error bars are 95% confidence intervals. The absence of an error bar indicates insufficiently many variants to estimate a standard error.

## Application to Cell Count Phenotypes

As a second case study, we applied COAST and SKAT-O to a collection of cell count phenotypes: erythrocytes, leukoctyes, lymphocytes, neutrophils, and thrombocytes (platelets). The uniform quantile-quantile plots in **Supplemental Figure 10** again demonstrate control of the type I error. For all cell count traits, COAST identified more Bonferroni significant associations than SKAT-O (**Figure 5**): 82% more than SKAT-O ALL and 6.3-times as many as SKAT-O PTV (**Supplemental Table 5**). Across the union of genes declared significant by any association test, the average $\chi^2$ statistic of COAST was 37% higher than that of SKAT-O ALL and 4.8-times higher than that of SKAT-O PTV (**Supplemental Table 6**). Among 61 total associations, 25 (41%) were unique to COAST, 33 (54%) were in common, and 3 (5%) were unique to SKAT-O ALL (**Supplemental Figure 11**). All associations except 1 (*TNXB* with thrombocytes, identified only by SKAT-O) either had common variant support, from the GWAS catalog, rare variant support, from Genebass, or both (**Supplemental Tables 7-8**). One gene, *DOK2* ($p = 4.8 \times 10^{-7}$), associated with lymphocytes by COAST, did not previously reach significance in Genebass, although its association with missense variants was suggestive ($p = 7.6 \times 10^{-6}$). Many well-known gene-trait associations were recapitulated, including *HBB* ($p = 1.2 \times 10^{-8}$) for erythrocytes, *CXCR2* ($2.6 \times 10^{-41}$) for neutrophils, and *JAK2* ($p = 2.2 \times 10^{-86}$) for thrombocytes (**Figure 6, Supplemental Figure 12**). For cell count traits, the pattern of effect sizes among genes identified by COAST was not always monotone increasing, although there were many such examples: *SH2B3* with erythrocytes, *S1PR2* with leukoctyes, *SBNO2* with lymphocotyes, and *JAK2* with thrombocytes (**Supplemental Figure 13**). Among genes identified as significant by SKAT-O ALL but not COAST, none had an allelic series pattern of effect sizes.

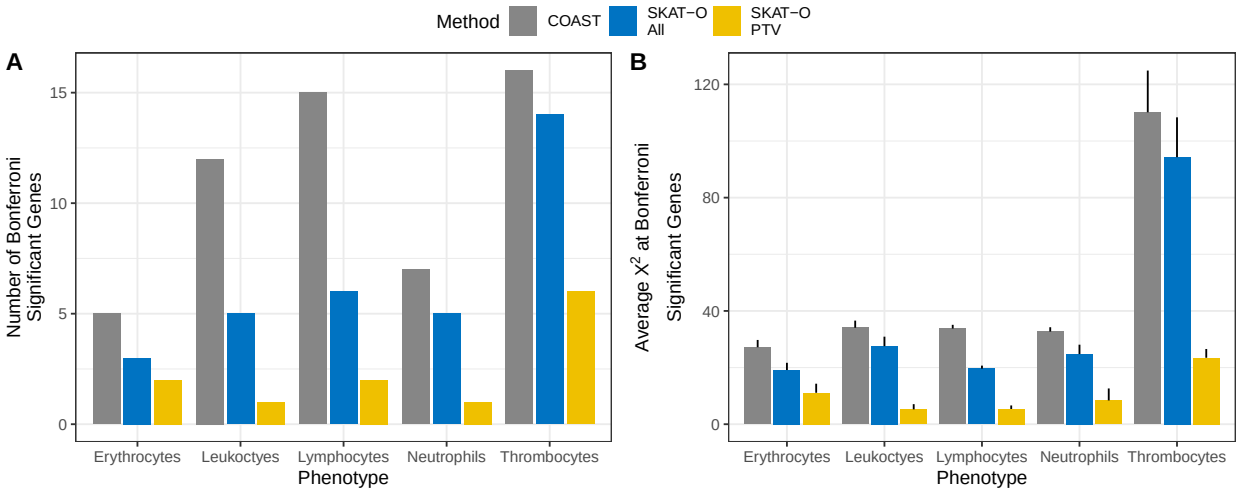


Figure 5: **Number and significance of genes significantly associated with cell count phenotypes. A.** Number of significantly associated genes by phenotype and association test. **B.** Average $\chi^2$ across the union of genes considered Bonferroni significant by any association test.

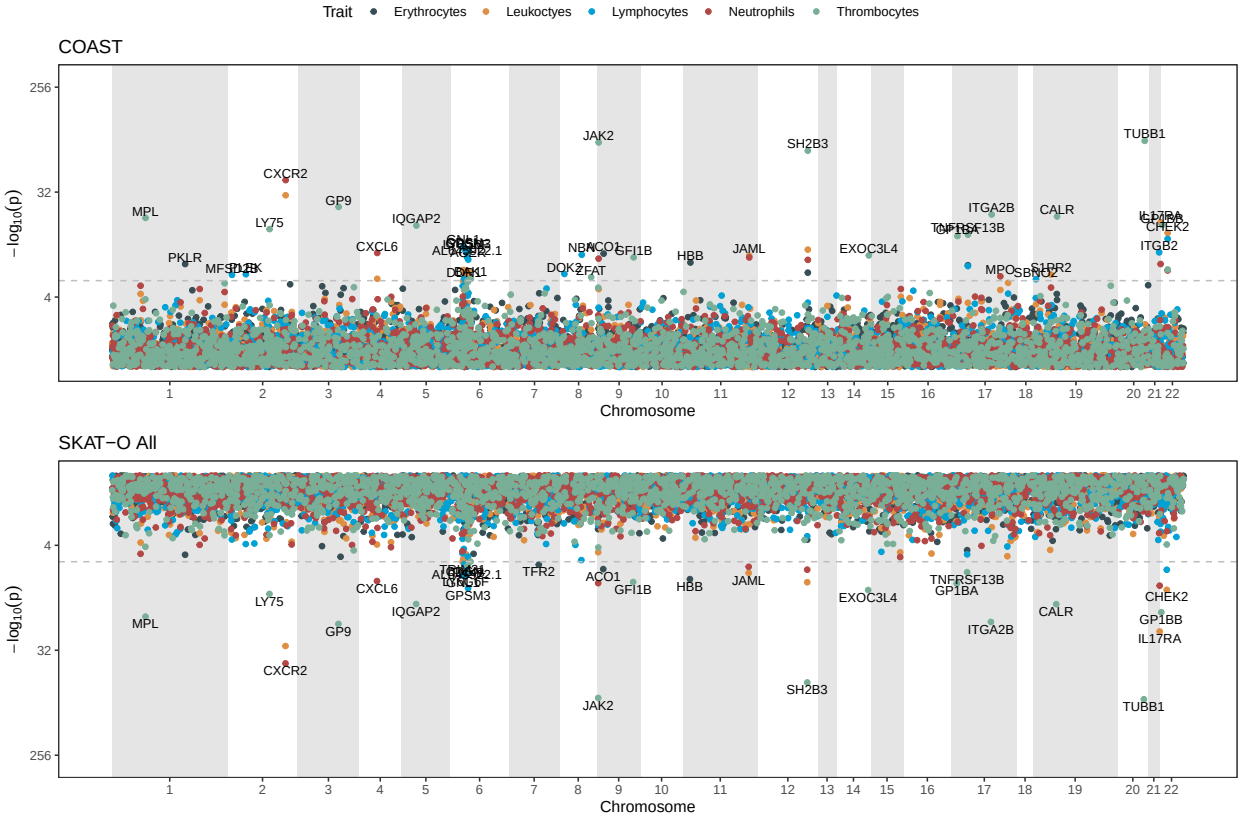


Figure 6: **Mirrored Manhattan plots for cell count phenotypes.** Upper panel is the coding-variant allelic series test (COAST). Lower panel is SKAT-O applied to all rare coding variants.

# Discussion

We have developed a gene-based rare variant association test targeted to the identification of allelic series: genes where increasingly deleterious mutations have increasingly large phenotypic effects. In simulation studies, COAST controlled the type I error and tended to improve power when the pattern of effect sizes, from BMVs to DMVs to PTVs, was monotone increasing. In applications to circulating lipid and cell count phenotypes from the UK Biobank, COAST consistently identified more Bonferroni significant associations than SKAT-O, and did so at a higher level of significance. In all cases, the additional genes detected by COAST had supporting evidence from at least one of Genebass or the GWAS catalog, mitigating the risk that these associations are spurious. Empirically, it was shown that in many but not all cases the pattern of effect sizes for genes identified by COAST formed an allelic series. Open-source code for implementing COAST is being released with this article.

Although in simulations SKAT-O PTV achieved greater power than SKAT-O ALL, empirically SKAT-O ALL outperformed SKAT-O PTV. We posit that this relates to the frequency of PTVs in real populations subject to selective pressure. Whereas in simulations BMVs, DMVs, and PTVs occurred with equal frequency, in reality more damaging variants are expected to occur less frequently. This observation does not diminish the utility of the simulations for demonstrating that COAST is most powerful when the allelic series weights align with the generative pattern of effect sizes.

COAST is not intended as a replacement for existing rare-variant associations tests, such as SKAT-O [16], ACAT [25], or STAAR [19]. These methods seek to identify the existence of a gene-trait association in general, placing no prior on what form that relationship should take. In contrast, COAST is tailored to the identification of genes that harbor a dose-response relationship between mutational severity and phenotypic impact. As such, the collection of genes targeted by COAST is only a subset of that targeted by existing methods. Empirically, the effect sizes of genes identified by COAST were not always monotone increasing in mutational severity. In part this is likely due to uncertainty in estimating the effect sizes of individual rare variants, however it is also due in part to COAST retaining power for detecting gene-trait associations even when the effect sizes do not follow a pattern of monotonic increase. In practice, when the goal of an analysis is to identify allelic series, the initial

results of COAST may require *post hoc* filtering to remove those genes whose effect size patterns are not monotone increasing.

COAST identified several gene-trait associations that are not present in Genebass but which garner support from the GWAS catalog: *ANGPTL4* (triglycerides), *A1CF* (triglycerides), and *DOK2* (lymphocytes). A low-frequency coding variant (rs116843064) in *ANGPTL4* was associated with HDL and triglycerides among 42K individuals of European ancestry (outside of UK Biobank) [40], and a meta-analysis of 50K individuals associated this variant with fasting triglyceride levels [41]. A candidate gene study associated rare inactivating mutations in *ANGPTL4* with reduced levels of fasting triglycerides and reduced odds of coronary artery disease [42]. Consistent with that work, 25 of the 27 rare variants in *ANGPTL4* considered in the present study, including all PTVs, were marginally associated with decreased triglyceride levels. A low-frequency coding variant (rs41274050) in *A1CF* was associated with triglycerides and total cholesterol in over 300K individuals of various ancestries, and the association was validated experimentally in a mouse model [43]. A large trans-ethnic analysis of 750K individuals, including UK Biobank, identified a low-frequency coding variant (rs56094005) in *DOK2* as associated with lymphocyte count, in addition to monocytes, neutrophils, and platelets [44, 45].

COAST has several limitations and opportunities for extension. The current implementation only incorporates rare coding variants annotated as a BMV, DMV, or PTV. While a clear ordering exists among the expected phenotypic impacts of these variant categories, in other cases the effect of a variant on gene function is complex and difficult to predict [46]. One possible extension of COAST is to assign a continuous measure of mutational severity, such as the CADD score [47, 48], to all variants within a gene, including non-coding variants. The continuous metric could then be cut into a finite number of intervals, and the variants categorized accordingly. These mutational severity intervals would replace the current categorization into BMV, DMV, and PTV. Open questions include what metric to use for measuring mutational severity and how to discretize it. Alternatively, the idea of categorizing could be discarded, and models could be defined that allow the expected value of the phenotype to vary continuous with, for example, the total value of the mutational severity metric aggregated across all rare variants within the gene.

We view the allelic series weights as hyperparameters that specify the pattern of effect sizes sought. The default values of $(w_1, w_2, w_3) = (1, 2, 3)$ for BMVs, DMVs, and PTVs specify a monotone increasing pattern. However, any choice of weights such that $w_1 \leq w_2 \leq w_3$ suffices to tailor COAST for detecting allelic series. In general the optimal weighting scheme will depend on the genetic architecture of a given gene-trait pair. Nevertheless, a weighting scheme that is informed by the expected distribution of effect sizes for a particular phenotype would likely improve power. For example, data from a single large biobank might be split into training and evaluation sets. COAST with default weights could be run on the training data, and a collection of candidate allelic series identified. Among the candidate allelic series, the average effect sizes of BMVs, DMVs, and PTVs could be calculated. These average effect sizes would then serve as the weights $(w_1, w_2, w_3)$ when applying COAST to the evaluation data.

Although COAST does not consider the effects of common variants, many prominent examples of allelic series have supporting evidence from both rare and common variants. For example, cholesterol levels are associated with common non-coding variants at the *ANGPTL3* locus [49], and with rare loss-of-function variants [50, 3]. Additionally, both a common missense variant and rare PTVs in *SLC30A8* have independently been associated with type 2 diabetes [51]. Extending COAST to include common variant effects would increase power to detect allelic series that have both rare and common variant support.

Finally, a single gene may have a dose-response relationship with multiple related traits. For example, *APOB* was significantly associated with all of cholesterol, HDL, LDL, and triglycerides via COAST. In such cases, a multivariate outcome model, such as a linear mixed effects model, can improve power by simultaneously estimating the effects of rare variants on multiple correlated traits. Alternatively, COAST can be extended to non-quantitative phenotypes, such as binary, count, or time-to-event phenotypes, by generalizing the constituent models to generalized linear models or Cox proportional hazards models. These extensions are under active development.

## Data and code availability

This work uses genotypes and phenotypes from the UK Biobank (https://www.ukbiobank.ac.uk/) accessed pursuant to approved application number **51766**. The coding-variant allelic series test has been implemented as an `R` [52] package, which is available on GitHub at https://github.com/insitro/AllelicSeries.

## Acknowledgements

The authors would like to acknowledge Baris Ungun for contributions to curating the UK Biobank data, and to thank the participants of the UK Biobank, whose data were used with permission.

## Author Contributions

TWS, DK, and FPC conceived of the project. FPC curated the phenotypic data. TWS curated the genetic data and performed the proof-of-principle analysis. ZRM developed the method and performed the final analysis. All authors provided scientific input. CK prepared the software for production. ZRM and TWS wrote the first draft of the manuscript. All authors contributed to critical revision of the final manuscript.

## References

[1] B McClintock. The relation of homozygous deficiencies to mutations and allelic series in maize. *Genetics*, 29(5):478 – 502, 1944.

[2] PM Plenge, EM Scolnick, and D Altshuler. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*, 12(8):581 – 594, 2013.

[3] K Musunuru and S Kathiresan. Genetics of common, complex coronary artery disease. *Cell*, 177(1):132 – 145, 2019.

[4] R Duan, H Hijazi, EY Gulec, et al. Developmental genomics of limb malformations:

Allelic series in association with gene dosage effects contribute to the clinical variability. *HGG Adv*, 3(4):100132, 2022.

[5] K Mohajeri, R Yadav, E D'haene, et al. Transcriptional and functional consequences of alterations to mef2c and its topological organization in neuronal models. *Am J Hum Genet*, 109(11):2049 – 2067, 2022.

[6] C Sudlow, J Gallacher, N Allen, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12(3):e1001779, 2015.

[7] PM Visscher, NR Wray, Q Zhang, P Sklar, MI McCarthy, MA Brown, and J Yang. 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5 – 22, 2017.

[8] PL Auer and G Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome Med*, 7(1):16, 2015.

[9] JA Kosmicki, CL Churchhouse, MA Rivas, and BN Neale. Discovery of rare variants for complex phenotypes. *Hum Genet*, 135(6):625 – 634, 2016.

[10] S Lee, GA Abecasis, M Boehnke, and X Lin. Rare-variant association analysis: Study designs and statistical tests. *Am J Hum Genet*, 95(1):5–23, 2014.

[11] J Asimit and E Zeggini. Rare variant association analysis methods for complex traits. *Annu Rev Genet*, 44:293–308, 2010.

[12] BE Madsen and SR Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5:e1000384, 2009.

[13] AP Morris and E Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*, 34(2):188 – 193, 2010.

[14] S Morgenthaler and WG Thilly. 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, 2014.

[15] MC Wu, S Lee, T Cai, Y Li, M Boehnke, and X Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82 – 93, 2011.

[16] Seunggeun Lee, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani, Mark M. Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

[17] Z He, B Xu, S Lee, and I Ionita-Laza. Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am J Hum Genet*, 101(3):340 – 352, 2017.

[18] Y Ma and P Wei. Funspu: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLoS Genet*, 15(4):e1008081, 2019.

[19] X Li, Z Li, H Zhou, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*, 52(9):969 – 983, 2020.

[20] J Mbatchou, L Barnard, J Backman, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*, 53(7):1097 – 1103, 2021.

[21] W McLaren, L Gil, SE Hunt, et al. The ensembl variant effect predictor. *Genome Biol*, 17(1):122, 2016.

[22] KJ Karczewski, M Solomonson, KR Chao, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 uk biobank exomes. *Cell Genomics*, 2(9):100168, 2022.

[23] G Seber. *The Linear Model and Hypothesis*. Springer, 2015.

[24] Y Liu and J Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[25] Y Liu, S Chen, Z Li, AC Morrison, E Boerwinkle, and X Lin. Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am J Hum Genet*, 104(3):410–421, 2019.

[26] HM Amemiya, A Kundaje, and AP Boyle. The encode blacklist: Identification of problematic regions of the genome. *Sci Reports*, 9:9354, 2019.

[27] H Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843 – 2851, 2014.

[28] Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248 – 249, 2010.

[29] Ng PC and Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*, 11(5):863 – 874, 2001.

[30] ET Cirulli, S White, RW Read, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun*, 11:542, 2020.

[31] CV Van Hout, I Tachmazidou, JD Backman, et al. Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature*, 586:749 – 756, 2020.

[32] KJ Karczewski, LC Francioli, G Tiao, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434 – 443, 2020.

[33] C Bycroft, C Freeman, D Petkova, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203 – 209, 2018.

[34] Hail team. hail 0.2. `https://github.com/hail-is/hail`.

[35] W Zhou, W Bi, Z Zhao, et al. Saige-gene+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet*, 54(10):1466 – 1469, 2022.

[36] ZR McCaw, JM Lane, R Saxena, S Redline, and X Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262 – 1272, 2020.

[37] A Buniello, JAL MacArthur, M Cerezo, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2019.

[38] Gwas catalog: The NHGRI-EBI catalog of human genome-wide association studies. `https://www.ebi.ac.uk/gwas/`. Accessed: 2022-11-23.

[39] Genebass: gene-based association summary statistics. `https://app.genebass.org/`. Accessed: 2022-11-23.

[40] GM Peloso, PL Auer, JC Bis, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet*, 94(2):223 – 232, 2014.

[41] EM van Leeuwen, A Sabo, JC Bis, et al. Meta-analysis of 49,549 individuals imputed with the 1000 genomes project reveals an exonic damaging variant in angptl4 determining fasting tg levels. *J Med Genet*, 53(7):441 – 449, 2016.

[42] FE Dewey, V Gusarova, C O'Dushlaine, et al. Inactivating variants in angptl4 and risk of coronary artery disease. *N Engl J Med*, 374(12):1123 – 33, 2016.

[43] DJ Liu, GM Peloso, H Yu, et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat Genet*, 49(12):1758 – 1766, 2017.

[44] MH Chen, LM Raffield, A Mousas, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell*, 182(5):1198 – 1213, 2020.

[45] D Vuckovic, EL Bao, P Akbari, et al. The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182(5):1214 – 1231, 2020.

[46] T Lappalainen and DG MacArthur. From variant to function in human disease genetics. *Science*, 24(373):1464 – 1468, 2021.

[47] M Kircher, DM Witten, P Jain, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310 – 315, 2014.

[48] P Rentzsch, DM Witten, GM Cooper, J Shendure, and M Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886 – D894, 2019.

[49] T Teslovich, K Musunuru, A Smith, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466:707 – 713, 2010.

[50] FE Dewey, V Gusarova, RL Dunbar, et al. Genetic and pharmacologic inactivation of angptl3 and cardiovascular disease. *N Engl J Med*, 377(3):211 – 221, 2017.

[51] J Flannick, JM Mercader, C Fuchsberger, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*, 570:71 – 76, 2019.

[52] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.