

# A Geometric Framework for Evaluating Rare Variant Tests of Association

Keli Liu,<sup>1</sup> Shannon Fast,<sup>2</sup> Matthew Zawistowski,<sup>3,4†</sup> and Nathan L. Tintle<sup>5†\*</sup>

<sup>1</sup>Department of Statistics, Harvard University, Cambridge, Massachusetts; <sup>2</sup>Department of Mathematics, Statistics and Computer Science, St. Olaf College, Northfield, Minnesota; <sup>3</sup>Department of Biostatistics, University of Michigan, Ann Arbor, Michigan; <sup>4</sup>Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan; <sup>5</sup>Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, Iowa

Received 24 September 2012; Revised 12 February 2013; accepted revised manuscript 13 February 2013.  
Published online 21 March 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21722

**ABSTRACT:** The wave of next-generation sequencing data has arrived. However, many questions still remain about how to best analyze sequence data, particularly the contribution of rare genetic variants to human disease. Numerous statistical methods have been proposed to aggregate association signals across multiple rare variant sites in an effort to increase statistical power; however, the precise relation between the tests is often not well understood. We present a geometric representation for rare variant data in which rare allele counts in case and control samples are treated as vectors in Euclidean space. The geometric framework facilitates a rigorous classification of existing rare variant tests into two broad categories: tests for a difference in the lengths of the case and control vectors, and joint tests for a difference in either the lengths or angles of the two vectors. We demonstrate that genetic architecture of a trait, including the number and frequency of risk alleles, directly relates to the behavior of the length and joint tests. Hence, the geometric framework allows prediction of which tests will perform best under different disease models. Furthermore, the structure of the geometric framework immediately suggests additional classes and types of rare variant tests. We consider two general classes of tests which show robustness to noncausal and protective variants. The geometric framework introduces a novel and unique method to assess current rare variant methodology and provides guidelines for both applied and theoretical researchers.

Genet Epidemiol 37:345–357, 2013. © 2013 Wiley Periodicals, Inc.

**KEY WORDS:** rare variants; sequencing; burden tests

## Introduction

Several large sequencing efforts have established that an abundance of rare functional variation exists in the human population [1000 Genomes, 2010; Nelson et al., 2012; Tennessen et al., 2012]. The preponderance of such variants and their potential deleterious impact make them candidates for putative risk variants contributing to complex disease in humans. Thus, the development of powerful statistical methods to analyze these rare genetic variants observed in next-generation sequencing data is a critical area of current research in human genetics. Traditional single-marker tests used in genome-wide association studies lack sufficient power when applied to rare variants. Instead, many novel “gene-based” methods have been proposed with a common theme of combining association signals for multiple rare variants from the same gene into a single test of significance [Basu and Pan, 2011; Dai et al., 2012; Feng et al., 2011; Han and Pan, 2010; Ionita-Laza et al., 2011; Li and Leal, 2008; Li et al., 2011; Lin and Tang, 2011; Madsen and Browning 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini 2010; Neale et al., 2011; Pan and

Shen, 2011; Price et al., 2010; Sul et al., 2011; Wu et al., 2011; Zawistowski et al., 2010; Zhang et al., 2011; among others]. Several summaries and reviews of these methods are available [Asimit and Zeggini, 2010; Bansal et al., 2010; Cooper and Shendure 2011; Dering et al., 2011; Gibson, 2012].

As reflected by the number and variety of proposed gene-based methods, there is no clear-cut strategy to combine the information from multiple sites into a single test statistic. The existing methods differ not only in how individual variants are summarized and weighted before being combined but also the assumptions on the underlying disease model. Not surprisingly, the performance varies dramatically among proposed methods. The consensus among several simulation-based studies comparing performance between the gene-based tests is that there is no single best strategy for testing rare variants [e.g., Basu and Pan 2011; Ladouceur et al., 2012; Luedtke et al., 2011, Sun et al., 2011; Tintle et al., 2011]. Differences in the underlying disease models on which simulations are based, including frequency spectrum and effect sizes for risk variants, as well as analytic challenges such as inclusion of neutral variants all directly impact test performance. However, the severity of the effect differs among the proposed tests. Although the results of these simulation studies show that different tests are optimal for different scenarios, we often lack an intuitive understanding as to why certain methods perform the way that they do. For example,

<sup>†</sup>These authors contributed equally to this work.

\*Correspondence to: Nathan Tintle, Department of Mathematics, Statistics and Computer Science, Dordt College, 498 4th Avenue NE, Sioux Center, IA 51250. E-mail: nathan.tintle@dordt.edu

Basu and Pan [2011], after conducting an extremely comprehensive simulation study, conclude that while a particular variance components test appears to be the “best” across a variety of disease models, the result is “surprising and interesting.” Further complicating matters is that there have been relatively few published applications of gene-based rare variant tests on real data [e.g., Rivas et al., 2011, Torgerson et al., 2012], leaving us unsure about what types of rare variant genetic architectures exist in nature.

To address some of the gaps in our understanding of rare variant test performance, we introduce a novel framework for considering such tests. Specifically, we consider case-control sequence data as mathematical vectors in a geometric space and relate differences in rare variation between cases and controls to differences in the lengths and angles of vectors. Based on this geometric framework, we show that the null hypothesis of no association that is tested in gene-based rare variant tests can be decomposed into a compound geometric null hypothesis based on the lengths and angle between the vectors representing the case and control data. The geometric framework allows an intuitive classification of many existing tests into broad categories based on which portion of the compound null hypothesis is being tested. We show that within these categories, the general performance of individual tests is well predicted by the geometric properties of the case and control vectors. In turn, we describe how aspects of the underlying disease model and study conditions affect the geometry of the dataset, thus connecting test behavior to these more traditional study variables. We verify these analytic insights using simulation.

The main benefit of the geometric framework is that it provides a rigorous method with which to categorize existing rare variant tests. The classification helps to explain why certain tests perform more similarly than others and provides a means to evaluate new tests that are certain to be proposed in the future. In addition to a classification scheme, the geometric framework suggests additional tests of rare variant association unlike tests proposed to date. In particular, existing tests can be combined or modified to respond more optimally to differing distributions of neutral, risk, and protective variants.

## Methods

### The Geometric Framework

Assume a dataset consisting of sequence information for a gene of interest in  $N^+$  cases and  $N^-$  controls. We restrict attention to a subset of  $m$  variable sites in the dataset that are putative risk variants satisfying some predefined minor allele frequency (MAF) threshold and predicted functional annotation, for example,  $MAF < 1\%$  and nonsynonymous. Let  $c_j^+$  be the total number of rare alleles observed at site  $j = 1, \dots, m$  among the cases. Similarly, let  $c_j^-$  be the total number of rare alleles observed at site  $j = 1, \dots, m$  among the controls. Then we define  $\mathbf{f}^+ = (f_1^+, \dots, f_m^+)$  to be the vector of maximum likelihood estimates (MLEs) of allele frequencies

in cases at the  $m$  sites of interest, where  $f_j^+ = c_j^+/2N^+$ . Likewise, define  $\mathbf{f}^-$  in the same manner for controls.

Assuming no genotype-phenotype association at the  $j$ th site leads to the familiar single-marker null hypothesis of equal allele frequency in the populations of cases and controls, namely  $F_j^+ = F_j^-$ , where we let  $F_j$  be the population MAF at site  $j$ . Assuming the null hypothesis of no association holds for all  $m$  variable sites observed in the dataset, the null hypothesis that rare, putatively functional observed variation in the gene is not associated with disease risk can be formally stated as

$$H_0 : \mathbf{F}^+ = \mathbf{F}^- \quad (1)$$

Now, consider  $\mathbf{F}^+$  and  $\mathbf{F}^-$  as mathematical vectors in  $m$ -dimensional space. Two vectors are equivalent if and only if both their magnitudes (lengths) and directions (angles) are equal. Thus, the null hypothesis  $\mathbf{F}^+ = \mathbf{F}^-$  in Equation (1) is equivalent to the geometric compound null hypothesis,

$$H_0 : \|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p \text{ and } \theta = 0, \quad (2)$$

where  $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ ,  $p \geq 1$ , denotes the  $L^p$  norm of a vector  $x = (x_1, x_2, \dots, x_n)$  and  $\theta = \arccos(\frac{\mathbf{F}^+ \cdot \mathbf{F}^-}{\|\mathbf{F}^+\|_2 \cdot \|\mathbf{F}^-\|_2})$  is the angle between the vectors  $\mathbf{F}^+$  and  $\mathbf{F}^-$ . In order to show that the null hypothesis in Equation (2) does not hold, it is sufficient to show that either  $\|\mathbf{F}^+\|_p \neq \|\mathbf{F}^-\|_p$  or  $\theta \neq 0$ . Thus, the null hypotheses

$$H_{0, \text{Length}} : \|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p \quad (3)$$

and

$$H_{0, \text{Angle}} : \theta = 0 \quad (4)$$

are both necessary but not individually sufficient in order for Equation (1) to hold.

Alternatively, two vectors are equivalent if the difference vector that connects their endpoints is the zero vector, in this case  $\mathbf{F}^+ - \mathbf{F}^- = (f_1^+ - f_1^-, \dots, f_m^+ - f_m^-)$ . Thus, a further reformulation of Equation (1) is

$$H_0 : \mathbf{F}^+ - \mathbf{F}^- = \mathbf{0} \quad (5)$$

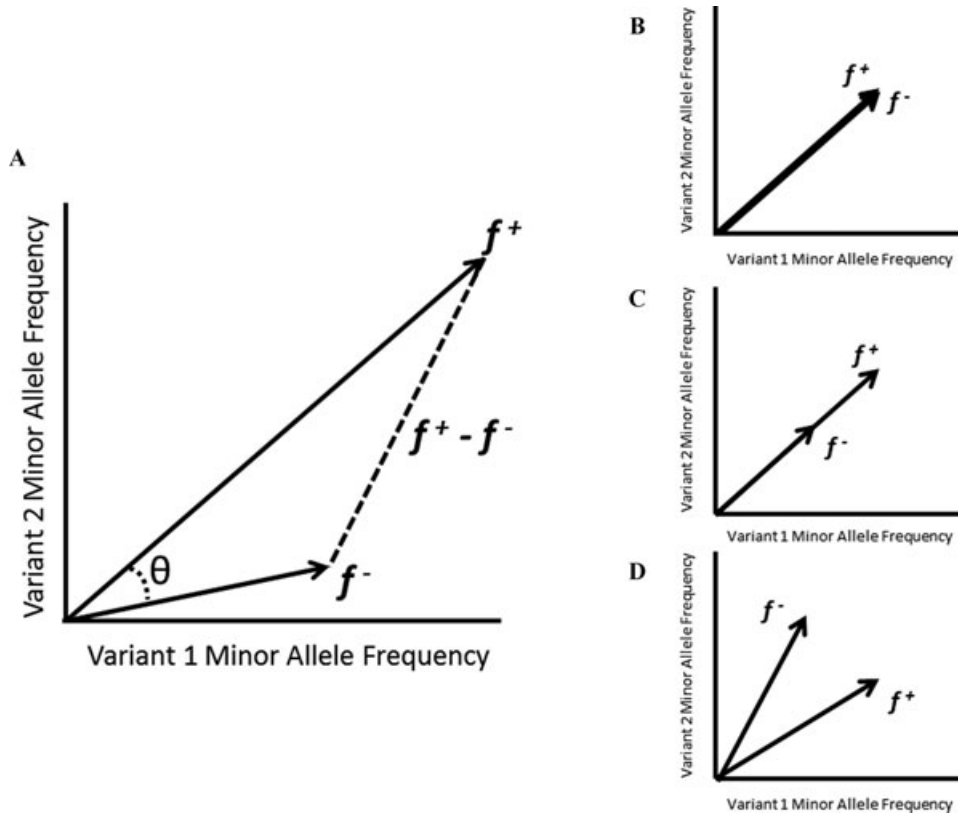
Since a vector  $\mathbf{x} = \mathbf{0}$  if and only if  $\|x\|_p = 0$ , it is sufficient to show that  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p \neq 0$  in order to show that Equation (3) does not hold. Therefore,

$$H_{0, \text{Joint}} : \|\mathbf{F}^+ - \mathbf{F}^-\|_p = 0 \quad (6)$$

is equivalent to Equation (1) and we refer to it as the joint null hypothesis since the difference vector  $\mathbf{F}^+ - \mathbf{F}^-$  jointly accounts for both the lengths and angle of the frequency vectors  $\mathbf{F}^+$  and  $\mathbf{F}^-$ . This is further illustrated by the fact that for  $L^2$

$$\begin{aligned} \|\mathbf{F}^+ - \mathbf{F}^-\|_2^2 &= (\|\mathbf{F}^+\|_2 - \|\mathbf{F}^-\|_2)^2 + 2 \cdot \|\mathbf{F}^+\|_2 \cdot \|\mathbf{F}^-\|_2 \\ &\quad \cdot (1 - \cos(\theta)). \end{aligned} \quad (7)$$

Essentially, Equation (7) combines a comparison of the differences in the lengths of the two vectors  $\mathbf{F}^+$  and  $\mathbf{F}^-$  with an evaluation of the size of the angle,  $\theta$ , observed between  $\mathbf{F}^+$  and  $\mathbf{F}^-$ . Figure 1 gives a graphical rendering of the geometric framework in order to provide visual intuition about the behavior of length, joint, and angle tests.



**Figure 1.** Two-dimensional rendering of the geometric framework for rare variant tests. The set of graphs shows simplified scenarios for two rare variant sites in a case-control dataset and are designed to provide intuition into the geometric interpretation of rare variant tests. (A) The vectors  $\mathbf{f}^+ = (f_1^+, f_2^+)$  and  $\mathbf{f}^- = (f_1^-, f_2^-)$  contain observed allele frequencies at two rare variant sites for cases and controls, respectively.  $\|\mathbf{f}^+\|_p$  and  $\|\mathbf{f}^-\|_p$  indicate the lengths of these frequency vectors with respect to the  $L^p$  norm,  $\theta$  is the measure of the angle between  $\mathbf{f}^+$  and  $\mathbf{f}^-$ , and  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p$  is the distance between the endpoints of  $\mathbf{f}^+$  and  $\mathbf{f}^-$ . The null hypothesis of no rare variant association ( $H_0: \mathbf{F}^+ = \mathbf{F}^-$ ) can be tested using any of the three following null hypotheses related to the geometry of the frequency vectors: (i)  $\|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p$ , the lengths of the vectors are equal, (ii)  $\theta = 0$ , the angle between the vectors is zero, or (iii)  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p = 0$ , the distance between the endpoints of the vectors is zero. We refer to tests of the three geometric null hypotheses as, respectively, length, angle, and joint tests. In the pictured scenario, the minor allele frequency is higher in cases for each variant ( $f_1^+ > f_1^-$  and  $f_2^+ > f_2^-$ ), indicating both as potential risk variants. (B) Under the null case of no association ( $\mathbf{F}^+ = \mathbf{F}^-$ ) each of the geometric null hypotheses hold: (i)  $\|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p$ , (ii)  $\theta = 0$ , and (iii)  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p = 0$ . (C) Both variants are causative with the case vector being a scalar multiple of the control vector ( $\mathbf{F}^+ = c\mathbf{F}^-$ ). This occurs if the case frequency and control frequency are the same across all variant sites. The result is that  $\|\mathbf{F}^+\|_p \neq \|\mathbf{F}^-\|_p$  and  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p \neq 0$ , but the null hypothesis of  $\theta = 0$  still holds. This scenario highlights the reason that angle tests are not powerful strategies and underscores why none have been proposed. (D) The scenario in which one rare variant is causative ( $f_1^+ > f_1^-$ ) and the other is protective ( $f_2^- > f_2^+$ ). In this case, it is possible that  $\|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p$  so that the signals from the two variants effectively cancel each other out, explaining reduced performance for length tests in the presence of a mix of risk and protective variants. Alternatively,  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p \neq 0$  and joint tests remain powerful.

In the following sections, we show that many rare variant tests of the null hypothesis  $H_0: \mathbf{F}^+ = \mathbf{F}^-$  can be formally classified according to which geometric null hypothesis (Equation (3), (4), or (6)) is being tested (see Table 1).

### Length Tests

We refer to rare variant tests of the null hypothesis  $H_{0, \text{Length}}: \|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p$  as length tests. Here we show two examples of published statistics that are length tests. The cumulative minor allele test [CMAT; Zawistowski et al., 2010] compares the total number of minor and major alleles in cases and controls across rare, functional variants within the same gene. Using our notation, the test

statistic for CMAT is  $\Sigma_{\text{CMAT}} = \frac{N}{2N^+N^-m} \frac{c^+(2N^- - c^-) - c^-(2N^+ - c^+) }{(c^+ + c^-)(2N^- - c^- + 2N^+ - c^+)}$ , where  $c^+ = \sum_{j=1}^m c_j^+$ , which can be simplified to  $\Sigma_{\text{CMAT}} = k(\|\mathbf{f}^+\|_1 - \|\mathbf{f}^-\|_1)$ , where  $k$  is a function of  $N^+$ ,  $N^-$ ,  $m$ , and  $c = c^+ + c^-$ . The CMAT statistic is significant (as determined by permutation) when  $\Sigma_{\text{CMAT}}$ , and therefore  $\|\mathbf{f}^+\|_1 - \|\mathbf{f}^-\|_1$ , becomes large.

Proportion regression [PR; Morris and Zeggini, 2010] uses a logistic regression framework to test for a rare variant association. Specifically, PR uses the model  $\text{logit}(\pi_i) = \alpha X_i + \beta \frac{r_i}{m}$  where  $X_i$  is a vector of  $p$  covariate values for the  $i$ th individual,  $\alpha$  is the vector of marginal effects of the  $p$  covariates on the disease phenotype and  $r_i = \sum_{j=1}^m c_{ij}$  is the total number of rare alleles possessed by the  $i$ th individual across the  $m$  variants. The score statistic  $S = (y - \hat{\pi}_0)' GG'(y - \hat{\pi}_0)$  is used

**Table 1. Classifying existing rare variant tests using the geometric framework**

	Length	Joint
Published examples	CAST [Morgenthaler and Thilly, 2007]	CMC [Li and Leal, 2008] <sup>a</sup>
	CMC [Li and Leal, 2008] <sup>a</sup>	C-alpha [Neale et al., 2011]
	WS [Madsen and Browning, 2009]	SKAT [Wu et al., 2011]
	PR [Morris and Zeggini, 2010]	SSU [Basu and Pan, 2011]
	CMAT [Zawistowski et al., 2010]	GFRV [Lin and Tang, 2011]
	aSUM [Han and Pan, 2010]	AT [Pan and Shen, 2011]
	Variable Threshold [Price et al., 2010]	PRVT [Ionita-Laza et al., 2011]
	ORWSS [Feng et al., 2011]	
	PWST [Zhang et al., 2011]	
	RWAS [Sul et al., 2011]	
	WHaT [Li et al., 2011]	
	WSCS [Dai et al., 2012]	

<sup>a</sup> CMC is a length test when all variants are below an arbitrarily defined MAF threshold; when some variants are above the threshold it acts as a combined length/joint test. See Appendices A and B for details.

to test  $H_0 : \beta = 0$ , where  $y$  is a length  $N$  vector ( $N =$  total number of individuals in the study) of 0s and 1s, where 1 = the individual is a case, 0 otherwise,  $\hat{\pi}_0$  is a vector of predicted disease probabilities estimated under the null logistic model, and  $G$  is a vector containing  $r_i/m$  for each of the  $N$  individuals in the study. Under the null hypothesis of no association (Equation (1)),  $S$  is distributed as a one degree of freedom chi-squared random variable. For simplicity, consider the case of no covariates and an equal number of cases and controls, which yields  $\hat{\pi}_0 = \frac{1}{2}$  for all individuals. Then the PR score statistic can be written in terms of the vector length as follows:

$$\begin{aligned} S &= (y - \hat{\pi}_0)' G G' (y - \hat{\pi}_0) \propto 4 (y - \hat{\pi}_0)' G G' (y - \hat{\pi}_0) \\ &= (2y - 1)' G G' (2y - 1) \\ &= (G' y - G' (1 - y))' (G' y - G' (1 - y)) \\ &= \frac{N}{m^2} (\|f^+ \|_1 - \|f^- \|_1)^2. \end{aligned}$$

The null hypothesis  $H_0 : \beta = 0$  will be rejected for large  $S$ , which occurs when  $\|f^+ \|_1 \neq \|f^- \|_1$ .

We identified a total of 12 recently proposed rare variant tests as length tests based on the fact that significance of the test statistics was equivalent to testing  $\|f^+ \|_p \neq \|f^- \|_p$  (Appendix A). Many length tests have been referred to in the literature as “burden” tests or “collapsing” tests [e.g., Dering et al., 2011]. In general, length tests measure how rare an individual is based on some index of the individual’s cumulative rare allele profile across all variants in the set. This can be viewed as measuring the overall disease “burden” or as “collapsing” all variants in the set. The aggregate level of burden within the cases is then compared to the aggregate burden in the controls to test for association.

### Angle Tests

Tests of the null hypothesis  $H_{0, \text{Angle}} : \theta = 0$  are referred to as angle tests. To the best of our knowledge, no rare variant tests of this specific null hypothesis have been proposed. This

is not necessarily a surprising observation because, as shown in Figure 1, if the case allele frequency vector,  $F^+$ , is a scalar multiple of the control allele frequency vector,  $F^-$ , as is the case when there is a consistent level of increased risk (the scalar multiple) across the set of variants, there will be no angle between the two vectors.

### Joint Tests

We refer to rare variant tests of the null hypothesis  $H_{0, \text{Joint}} : \|F^+ - F^-\|_p = 0$  as joint tests since they jointly consider both the lengths and angles of the observed frequency vectors  $f^+$  and  $f^-$ . A common example of a joint test is the Sequence Kernel Association Tests [SKAT, Wu et al., 2011]. Using a scaled version of SKAT with a linear unweighted kernel, we see that the test statistic is

$$\begin{aligned} &(y - \hat{\pi}_0 1)' A A' (y - \hat{\pi}_0 1) \\ &\propto \left( \frac{N}{(N^+)(N^-)} \right)^2 (y - \hat{\pi}_0 1)' A A' (y - \hat{\pi}_0 1), \end{aligned}$$

where  $A$  is an  $N \times m$  genotype matrix containing the rare allele count of each individual at each site,  $y$  is an  $N \times 1$  vector indicating disease status (1 or 0), and  $\hat{\pi}_0$  is the fraction of cases in the sample. We note that

$$\begin{aligned} &\left( \frac{N}{(N^+)(N^-)} \right)^2 (y - \hat{\pi}_0 1)' A A' (y - \hat{\pi}_0 1) \\ &= \left( \frac{1}{N^+} y - \frac{1}{N^-} (1 - y) \right)' A A' \left( \frac{1}{N^+} y - \frac{1}{N^-} (1 - y) \right) \\ &= (f^+ - f^-)' (f^+ - f^-) = \|f^+ - f^-\|_2^2. \end{aligned}$$

Using similar rationale, we classified six additional rare variant tests as joint tests (Appendix B). Appendix B also illustrates how SKAT can be considered as a joint test for many, but not all, kernel choices.

Unlike length tests, joint tests do not “collapse” or measure the overall individual disease burden in cases as compared to controls. Instead joint tests, by evaluating the length of the difference between the case and control vectors, consider differences in case-control allele frequency on a variant-by-variant basis, and then combine the variant-by-variant differences to obtain a statistic.

### Novel Tests Suggested by the Geometric Framework

In addition to providing clarity on existing tests of association, the geometric framework also suggests alternative rare variant tests of association. In the following two sections, we describe two alternative, generalized classes of rare variant tests of association that are direct implications of the geometric framework.

*Alternative choice of norm.* As shown in the previous sections, many length tests use  $p = 1$ , while most joint tests use  $p = 2$ . While these choices are natural due to asymptotic theory of resulting test statistics and typical conceptualizations



of geometric spaces, the geometric framework suggests that the choice of norm,  $p$ , in  $\|x\|_p$ , does not necessarily need to take values of 1 or 2. In particular, we can select any positive value for  $p$ , including infinity,  $\infty$ , where we define  $\|x\|_\infty = \arg \max_i(x_i)$ . Thus, we define the following generalized length ( $L_p$ ) and joint ( $J_p$ ) test statistics, with arbitrary choice of norm,  $p$ , as  $L_p = \|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p$  and  $J_p = \|\mathbf{f}^+ - \mathbf{f}^-\|_p$ . Appendix C provides an overview of how to modify  $L_p$  and  $J_p$  to handle covariates. Statistical significance is assessed via disease permutation. Note that  $L_1$  is approximately equivalent to CMAT and  $J_2$  is approximately equivalent to SKAT.

**Weighted length-angle test.** Another way, the geometric framework can be used to generate new rare variant tests is by recognizing that length and angle tests represent two “extremes” in rare variant testing strategies. As noted earlier, joint tests of the form  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p$  weight the “length” and “angle” portions of the test statistic approximately equally. This is a reasonable, though not necessary, choice. We define the following generalized joint test statistic,  $W$ , which gives the ability to increase or decrease the length or angle portion of the test by modifying  $w_1$ , as well as giving control over the choice of norm used through choice of  $p$  and  $q$ .

$$W(p, q, w_1) = w_1 (\|f^+\|_p - \|f^-\|_p)^2 + (1 - w_1) \cdot 2 \cdot \|f^+\|_q \cdot \|f^-\|_q \cdot (1 - \cos(\theta)).$$

Again, statistical significance is assessed via disease permutation. Appendix C provides a covariate adjusted version of  $W$ .

## Simulation

We use two main simulation studies to validate select findings from our analysis. In the first simulation study, we simulate 1,500 cases and 1,500 controls according to the following disease architecture. There are two rare variants in the set, each with an MAF of 1% in the controls in the populations from which the samples are selected. The choice of two variants was made to optimally illustrate the behavior of length and joint tests. As we will demonstrate later (Results), results easily generalize to any number of variants,  $m$ . Variant one has a fixed relative risk of 1.25, and we change the relative risk of variant two to take values of 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.5, and 3.0, for a total of 12 separate settings. At each setting, we simulated 1,000 sets of data. Empirical power estimates are calculated as the percentage of the 1,000 simulated sets yielding a  $P$ -value less than 0.05. Three length tests (combined multivariate and collapsing [CMC], CMAT, and PR) and two joint tests (SKAT [linear kernel] and C-alpha) were applied to the simulated data. Significance was determined by asymptotic distributions or 1,000 permutations, depending upon the availability of an asymptotic distribution for the test. A smaller follow-up study considered a select set of these relative risks (0.2, 0.6, 1.0, 1.4, and 2.0) and applied the weighted length/angle test ( $W(p, q, w_1)$ ).

In the second simulation study, we simulated 1,000 cases and 1,000 controls, according to the following disease model. We simulated eight causal variants: two with an MAF of 1%, and six with an MAF of 0.1% in the populations from which the samples are selected. All eight causal variants have a relative risk of 2.0. We then added increasing numbers of noncausal variants (in sets of eight, two at 1%/six at 0.1%). Ultimately, we considered 10 different simulation settings representing 0, 8, 16, 24, 32, 40, 48, 56, 64, and 72 noncausal variants in the set with the eight causal variants. Similar to the first simulation, we simulated 1,000 sets of data at each setting to estimate empirical power. We considered the generalized length test statistic ( $L_p$ ), with  $p = 1, 2, 4$ , or  $\infty$ , and the generalized joint test statistic ( $J_p$ ), with  $p = 1, 2, 4$ , or  $\infty$ . Significance of these eight test statistics was assessed using 1,000 permutations of case-control status.

## Results

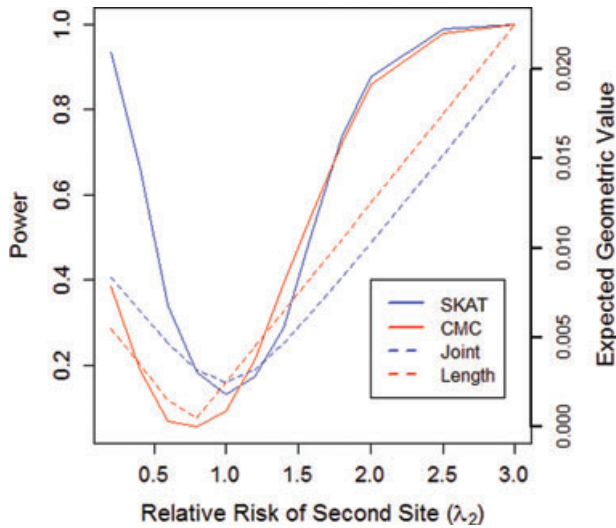
We have classified the majority of rare variant tests into one of two types: length or joint tests. In the following sections, we (a) present simulation results illustrating how the behavior of length and joint tests follows patterns suggested by the geometric framework, (b) provided a detailed analysis of test behavior in light of genetic architecture, and (c) demonstrate two specific ways that the geometric framework can be used to suggest modifications to existing rare variant tests of association.

### Simulation Results

Figure 1 and the overview of the geometric framework in the Methods section suggested that length tests focus solely on testing the null hypothesis that there is no difference in the lengths of the two allele frequency vectors, namely  $\|\mathbf{F}^+\|_p = \|\mathbf{F}^-\|_p$ , while joint tests focus on testing the null hypothesis that the length of the difference between the two allele frequency vectors  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p$  is zero. We used simulation to confirm that the behavior of these values directly corresponds to the power of both length and joint tests (Fig. 2).

### Analytic Insights into Test Behavior

Having provided an overview of the geometric framework and suggested intuition behind the behavior of different classes (Fig. 1), as well as simulation results confirming this intuition (Fig. 2), in the following sections, we will explicitly demonstrate how genetic architecture affects the behavior of length, angle, and joint tests. In particular, we will explore how the three types of tests behave in relationship to three components of the genetic architecture of disease: (1) The relative risk of disease  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ , where  $\lambda_j \approx F_j^+/F_j^-$ , (2) the number of variants,  $m$ , in the gene, and (3) the population minor allele frequencies at the  $m$  variant sites,  $F$ .



**Figure 2.** Power of length and joint tests corresponds to the behavior predicted by geometric framework. This graph illustrates the power of length and joint tests in relation to the expected value of the difference in lengths or length of difference of  $F^+$  and  $F^-$ . In particular, we consider the simplified scenario of a gene containing two rare variants, both with an allele frequency of 1%. The sample consists of 1,500 cases and 1,500 controls. The relative risk at the first variant site ( $\lambda_1$ ) is fixed at 1.25, while the relative risk at the second site ( $\lambda_2$ ) varies along the x-axis. We plot the power of a length test (CMC) alongside the expected value of  $\|\mathbf{f}^+\|_2 - \|\mathbf{f}^-\|_2$ . The expected value of the length test statistic will be minimized (taking a value of zero) when the relative risk at site two is 0.75. As the relative risk at site two ( $\lambda_2$ ) moves away from the value 0.75, the expected value of the length test statistic increases linearly. A similar pattern of behavior is observed for other length tests (e.g., CMAT and PR) though not shown here. We also plot the expected value of  $\|\mathbf{f}^+ - \mathbf{f}^-\|_2$  along with the power of a joint test (SKAT). Unlike length tests, joint tests attain their minimum value when  $\lambda_2 = 1$ , with the value of the expected value of both statistics increasing symmetrically as  $\lambda_2$  moves away from 1. A similar pattern of behavior is observed for other joint tests (e.g., C-alpha), though not shown here.

### Changes in the Relative Risk Distribution

For low prevalence diseases,  $F_j^- \approx F_j$ , and  $F_j^+ \approx \lambda_j F_j$ , where  $\lambda_j$  is the relative risk of site  $j$ , and so  $\lambda_j > 1$  denotes a site where the rare allele increases disease risk,  $\lambda_j < 1$  denotes a site which reduces disease risk, and  $\lambda_j = 1$  denotes a site that does not impact disease risk. We will use the terms risk, protective, and noncausal site to denote these three cases, respectively.

**Length tests.** Length tests typically evaluate the difference between  $\|\mathbf{f}^+\|_p$  and  $\|\mathbf{f}^-\|_p$ , which is equivalent to comparing  $(\sum_{j=1}^m |\hat{\lambda}_j f_j|^p)^{\frac{1}{p}}$  to  $(\sum_{j=1}^m |f_j|^p)^{\frac{1}{p}}$ , where  $\hat{\lambda}_j$  is the MLE of  $\lambda_j$ . If  $\mathbf{F}^+ = \mathbf{F}^-$ , then,  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m) = (1, 1, \dots, 1) = \mathbf{1}$ , and so,

$$\left( \sum_{j=1}^m |\lambda_j F_j|^p \right)^{\frac{1}{p}} = \left( \sum_{j=1}^m |F_j|^p \right)^{\frac{1}{p}}. \quad (8)$$

The limitation of length tests described earlier is demonstrated explicitly here, since  $\lambda = \mathbf{1}$  is not a necessary condition for Equation (8) to be true. In fact, there are an infinite number of values of  $\lambda$  which, for any given  $\mathbf{F}$ , make Equation (8) true, since Equation (8) is an underdetermined equation (i.e., Equation (8) is a single equation with  $m$  unknowns, namely  $\lambda_1, \lambda_2, \dots, \lambda_m$  for  $m > 1$ ). For example, consider the case where  $F_j = k$  for all  $j$ , and  $p = 1$ . Then for any  $\lambda$  where  $\sum_{j=1}^m \lambda_j = m$ , Equation (8) will be approximately true. In particular, if  $m = 2$ ,  $\lambda_1 = 1.25$  and  $\lambda_2 = 0.75$ , then  $\sum_{j=1}^m \lambda_j = 2$  and, so, Equation (8) is true. This example demonstrates that, within a gene with multiple variants at similar MAF, there will be little difference in  $\|\mathbf{f}^+\|_p$  and  $\|\mathbf{f}^-\|_p$  if the relative risks cancel out because some variants are risk and other variants are protective. Figure 2 illustrated this result.

Generally, a robust rare variant test of association will have the characteristic that the value of its test statistic will move farther from the null hypothesis value as  $|\lambda_j - 1|$  increases. However, for length tests, the behavior of  $\|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p$  as  $|\lambda_j - 1|$  increases is variable. In situations where all variants are risk ( $\lambda_j > 1$  for all  $j$ ) or all variants are protective ( $\lambda_j < 1$  for all  $j$ ), the difference in the lengths of  $\|\mathbf{f}^+\|_p$  and  $\|\mathbf{f}^-\|_p$  will increase as  $|\lambda_j - 1|$  increases. However, in cases where there is a mix of protective and risk variants (some  $\lambda_j$  are  $> 1$  and others are  $< 1$ ), the behavior of  $\|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p$  may or may not increase. In particular, if  $\|\mathbf{f}^+\|_p > \|\mathbf{f}^-\|_p$  then increasing any  $\lambda_j$  will increase the difference in lengths of  $\|\mathbf{f}^+\|_p$  and  $\|\mathbf{f}^-\|_p$ , whereas if  $\|\mathbf{f}^+\|_p < \|\mathbf{f}^-\|_p$  decreasing any  $\lambda_j$  will increase the difference in lengths of  $\|\mathbf{f}^+\|_p$  and  $\|\mathbf{f}^-\|_p$ . Thus, length tests demonstrate a lack of robustness in the presence of mixes of risk and protective variants.

**Joint tests.** While length tests illustrate a lack of robustness in the presence of mixes of risk and protective variants, joint tests, in testing  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p = 0$ , do not have this same limitation. In general,  $0 = \|\mathbf{F}^+ - \mathbf{F}^-\|_p = (\sum_{j=1}^m |\lambda_j F_j - F_j|^p)^{1/p} = (\sum_{j=1}^m F_j^p |\lambda_j - 1|^p)^{1/p}$ , which is true if and only if,

$$|\lambda_j - 1|^p = |k_j|^p = 0 \quad \text{for all } j = 1, \dots, m, \quad (9)$$

where  $\lambda_j = 1 + k_j$  and  $k_j$  can be interpreted as the “risk deviation.” Equation (9) is only true if and only if  $\lambda_j = 1$  for all  $j$ , alternatively  $k_j = 0$  for all  $j$ , which is true if and only if  $\mathbf{F}^+ = \mathbf{F}^-$ . Thus, a joint test which considers  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p$ , equals 0 (the null hypothesis value) exactly when the null hypothesis is true. Alternatively, joint tests can be interpreted as finding the length of the allele frequency weighted risk deviations since  $\|\mathbf{F}^+ - \mathbf{F}^-\|_p = \|\mathbf{F}k\|_p$ . Implicitly, joint tests, unlike length tests, do not allow risk deviations to cancel out. Furthermore, it is clear from Equation (9) that joint tests are robust to mixes of risk and protective variants since the value of  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p$  will increase as  $|\lambda_j - 1|$  increases. Figure 2 provides an example of this behavior.

**Angle tests.** Our results thus far suggest that handling mixes of protective/risk variants may be problematic for length tests, while joint tests more appropriately handle mixes of protective/risk variants. Thus, since joint tests are testing a

combined hypothesis which considers both length and angle differences in the vectors (see Methods), this suggests that angle tests might be robust to mixes of protective and risk variants.

To test this claim, note that when the null hypothesis ( $\mathbf{F}^+ = \mathbf{F}^-$ ) is true, there is no angle between the two vectors ( $\theta = 0$ ). When  $\theta = 0$ , then  $\mathbf{F}^+ \cdot \mathbf{F}^- = \|\mathbf{F}^+\|_2 \cdot \|\mathbf{F}^-\|_2$ , or alternatively,  $\sum_{j=1}^m \lambda_j F_j^2 = \sqrt{\sum_{j=1}^m (\lambda_j F_j)^2} \sqrt{\sum_{j=1}^m (F_j)^2}$ . But, we note that if  $\lambda = l$  ( $l > 0$ ; a constant value for all  $\lambda_j$ ), then  $\sqrt{\sum_{j=1}^m (\lambda_j F_j)^2} \sqrt{\sum_{j=1}^m (F_j)^2} = \sqrt{\sum_{i=1}^m \sum_{j=1}^m (\lambda_i F_i)^2 (F_j)^2} = l \sqrt{\sum_{i=1}^m \sum_{j=1}^m (F_i)^2 (F_j)^2} = l \sqrt{(\sum_{j=1}^m F_j^2)^2} = l(\sum_{j=1}^m F_j^2)$ . Thus, there is no angle between vectors  $\mathbf{F}^+$  and  $\mathbf{F}^-$  whenever  $\lambda = l$ , not only when the null hypothesis is true ( $l = 1$ ). In general, larger angles will occur as the values of  $\lambda_j$  become increasingly different (spread out), as will occur when there are mixes of risk and protective variants. However, when values of  $\lambda_j$  are not that different, there will be little difference in the angle between the two vectors, even if the values of  $\lambda_j$  are different than 1. Thus, importantly, sets of variants showing consistent levels of risk or consistent levels of protection will yield minimal angles between the case and control allele frequency vectors. For example, if all variants in the set had a relative risk of 2 ( $\lambda = 2$ ), there would be no angle between the two vectors; refer also to Figure 1C. Hence, angle tests, like length, show lack of robustness to a plausible scenario for the values of  $\lambda$ .

### Changes in MAF

In the previous section, we described how changes to the relative risk distribution impact length, angle, and joint tests. With these results in mind, we now turn our attention to how changes in the MAF vector,  $\mathbf{F}$  impact the general behavior of length and joint tests.

Increasing the MAF,  $F_j$ , at a particular site  $j$ , will have different effects on the difference in lengths of the case and control allele frequency vectors,  $\|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p$ , depending upon the value of  $\lambda_j$  and the starting value of  $\|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p$ . In particular, increasing  $F_j$  for a risk variant ( $\lambda_j > 1$ ) will increase the value of  $\|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p$  if  $\|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p > 0$  before the change. Increasing  $F_j$  for a protective variant ( $\lambda_j < 1$ ) will decrease the difference in lengths, and when  $\lambda_j = 1$ , changing  $F_j$  does not change the difference in lengths. In contrast, for joint tests, the impact of MAF,  $\mathbf{F}$ , on  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p$  is straightforward. Simply stated, since  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p = (\sum_{j=1}^m F_j^p |\lambda_j - 1|^p)^{1/p}$ , increases to the minor allele frequencies  $F_j$  will always increase the value of  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p$ , regardless of  $\lambda$ .

### Number of Variants

For length tests, the impact of increasing the number of variants,  $m$ , in the set is straightforward. Simply stated, if the length of the case vector ( $\|\mathbf{f}^+\|_p$ ) is greater than the length of the control vector ( $\|\mathbf{f}^-\|_p$ ) before adding the additional variant, then if the additional variant has  $f_j^+ > f_j^-$ , the dif-

ference in lengths will increase whereas if  $f_j^+ < f_j^-$  the difference in lengths will decrease. The addition of noncausal variants will, on average, increase both the case and control vectors a similar amount and, thus, have no impact on the difference in lengths. However, for joint tests, the value of  $\|\mathbf{f}^+ - \mathbf{f}^-\|_p = (\sum_{j=1}^m f_j^p |\lambda_j - 1|^p)^{1/p}$  will increase with the addition of each causal variant ( $\lambda_j \neq 1$ ) and, on average, remain the same for noncausal variants ( $\lambda_j = 1$ ) that are added to the set being tested.

## Examples of Further Implications of the Geometric Framework

In the previous sections, we have described how the geometric framework provides direct insight into the behavior of length and joint tests. The geometric framework also suggests alternative rare variant tests of association, with predictable behavior. Earlier, we proposed two generalized tests suggested by the geometric framework. In the following two sections, we describe the behavior of these tests on simulated data.

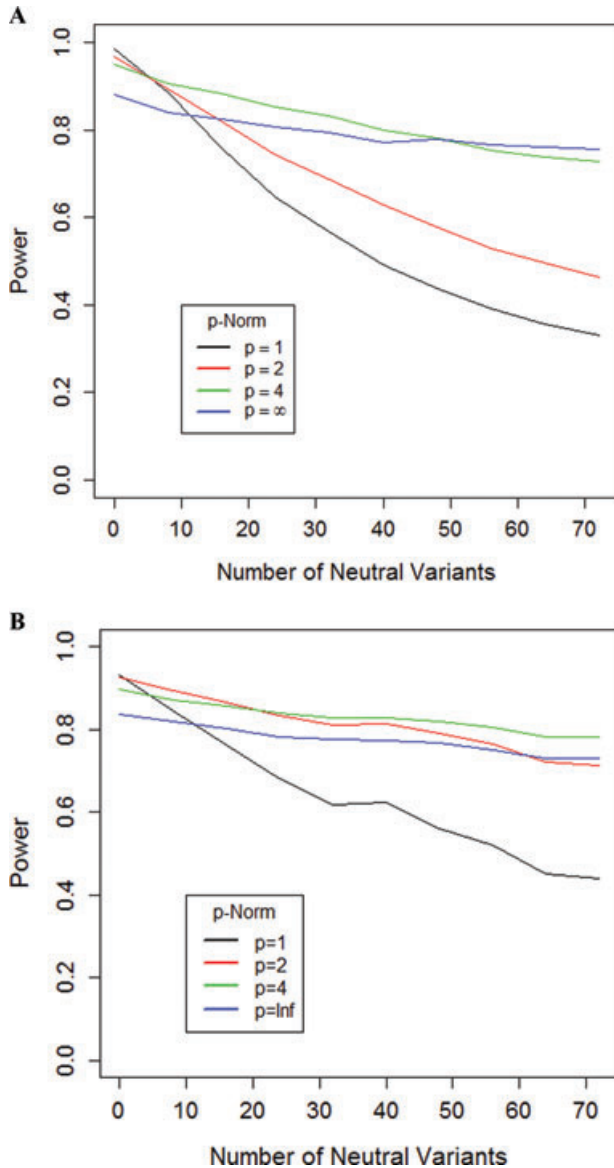
### The Impact of the Choice of Norm, $p$

Earlier, we proposed generalized length and joint test statistics,  $L_p$  and  $J_p$ , which allowed researchers to select any positive value,  $p > 0$  or  $\infty$ , for  $p$ . We now explore the behavior of these test statistics as a function of the choice of norm,  $p$ .

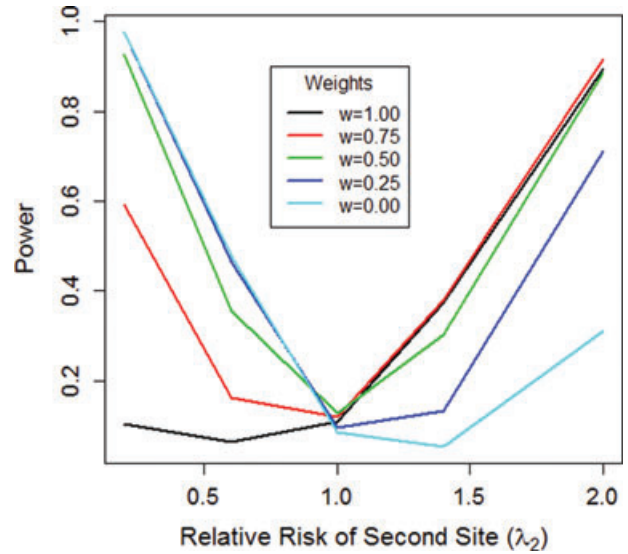
In most realistic situations, a fraction of the  $m$  variants will have  $\lambda_j = 1$  since it is difficult to identify, a priori, which of the variants in a set are causal ( $\lambda_j \neq 1$ ). Thus, in the observed sample, many of the estimated relative risks will be different from 1 only by chance—not reflecting a population relative risk different than 1. Intuitively, by increasing  $p$ , we are able to mitigate the effect of noncausal sites on the test statistic, by, in essence, up-weighting larger observed effects. Ultimately, this translates into a mitigation of decreased power from the addition of noncausal sites, as seen in Figure 3A and B for  $L_p$  and  $J_p$ , respectively. Importantly, we see a complete reversal in the ordering of most powerful tests as we move from 0 to 72 noncausal sites for the length test, and a near complete reversal for the joint test.

The intuition in Figure 3 can be confirmed by looking more carefully at the formulation of the test statistics. For  $J_p$ , in all cases, increasing  $p$ , means that larger allele frequency weighted risk deviations,  $f_j^p |\lambda_j - 1|^p$ , are counted proportionally more in  $J_p = \|\mathbf{f}^+ - \mathbf{f}^-\|_p = (\sum_{j=1}^m f_j^p |\lambda_j - 1|^p)^{1/p}$ . For  $L$ , a similar result also holds. Consider a case where the allele frequency at all variant sites is constant,  $f_j = f$  for all  $j$ . Then,

$$\begin{aligned} L_p &= \|\mathbf{f}^+\|_p - \|\mathbf{f}^-\|_p \approx \left( f^p \sum_{j=1}^m |\hat{\lambda}_j|^p \right)^{\frac{1}{p}} - (mf^p)^{\frac{1}{p}} \\ &= f \left( \sum_{j=1}^m |\hat{\lambda}_j|^p \right)^{\frac{1}{p}} - f(m)^{\frac{1}{p}} = f \left( \left( \sum_{j=1}^m |\hat{\lambda}_j|^p \right)^{\frac{1}{p}} - (m)^{\frac{1}{p}} \right). \end{aligned}$$



**Figure 3.** Power of length and joint tests corresponds to the behavior predicted by geometric framework. The two graphs illustrate the power of length ( $L_p = \|f^+ \|_p - \|f^- \|_p$ ) and joint tests ( $J_p = \|f^+ - f^- \|_p$ ) with different norms ( $p = 1, 2, 4$ , and  $\infty$ ). In each case, the test statistic is computed and significance is assessed via permutation of case-control status. We consider a scenario where a gene contains eight causal risk variants, all with a relative risk of 2.0. Two of the risk variants have MAF = 1%, the other six have MAF = 0.1%. We simulated a sample of 1,000 cases and 1,000 controls for this setting. We then considered nine additional settings where we added 8, 16, 24, 32, 40, 48, 56, 64, and 72 additional noncausal variants (relative risk = 1), always maintaining 3:1 ratio of low MAF (0.1%) to high MAF (1%) variants in the set. (A) For  $L_p$ , as we move from no noncausal variants to 72 noncausal variants, the order of most powerful tests completely reverses, suggesting that higher norms are more optimal in situations with large numbers of noncausal variants. (B) For  $J_p$ , as we move from no noncausal variants to 72 noncausal variants, the order of most powerful tests nearly reverses, suggesting that, once again, higher norms are more optimal in situations with large numbers of noncausal variants.



**Figure 4.** Arbitrary combining of length and angle tests. We conducted a simulation analysis and used  $W(w_1, p, q)$  with  $p = q = 2$ , and let  $w_1 = 1$  (length only),  $w_1 = 0.75$ ,  $w_1 = 0.5$  (typical joint test),  $w_1 = 0.25$ , and  $w_1 = 0$  (angle only test). This figure illustrates the power of these five tests across different values of  $\lambda_2$ . Power curves are as predicted. In particular, the length only ( $w_1 = 1$ ) and angle only ( $w_1 = 0$ ) tests show the least robustness, while the ( $w_1 = 0.5$ ) test is quite robust. As expected, the ( $w_1 = 0.75$ ) weighted test outperforms the ( $w_1 = 0.5$ ) test when both variants are risk-inducing, while providing more power than the length only test when there is a mix of risk-inducing and protective variants. The reverse is true for the ( $w_1 = 0.25$ ) test.

Thus, if all  $\hat{\lambda}_j > 1$ , increasing  $p$  means that variants with larger estimated risk will be counted proportionally more when finding the difference in lengths. In situations where  $f$  is not constant, the same general result holds, but may be mitigated or exacerbated based on the allele frequencies of the noncausal and causal variants. Thus, higher choices of norms exhibit more robustness to the inclusion of neutral variants by upweighting larger effects and downweighting weaker effects observed in the sample data.

#### Fine-Tuned Combinations of Length and Joint Tests

Earlier, we defined  $W(w_1, p, q) = w_1(\|f^+ \|_p - \|f^- \|_p)^2 + (1 - w_1) \cdot 2 \cdot \|f^+ \|_q \cdot \|f^- \|_q \cdot (1 - \cos(\theta))$ . Thus,  $W(w_1, p, q)$  gives more fine-tuned control over the contribution of length and angle differences to the test statistic. Most joint tests implicitly let  $w_1 = 0.5$ . Figure 4 illustrates that more powerful tests can be obtained for particular relative risks by modifying  $w_1$ .

#### Discussion

Observing all variation in large genetic datasets is fast becoming a reality due to high throughput sequencing technologies. However, an open question is how to best use these datasets in order to test for association between rare variants



and phenotypes of interest. The current strategy is to simultaneously analyze multiple rare variants in the same gene in a single statistic, with at least 18 such methods having been proposed to date. While the underlying intuition for some tests provides a means for determining which are similar, in other cases, it is less obvious how they are related. Here, we have derived a formal geometric framework that provides a rigorous method for comparing many existing rare variant tests. Further, we have identified strategies to increase statistical power both by adjusting features of existing tests and by combining existing tests with different properties.

When placed in context of this geometric framework, we find that the major distinguishing characteristic of rare variant tests is how they handle variants with opposing directions of effect. The geometric framework differentiates between tests that are robust to a mix of risk and protective variants (joint tests) and those that are most powerful when all causal variants have the same direction of effect (length tests). While this distinction was previously known to exist, we are now able to attribute the difference in performance to a theoretical difference in the underlying null hypotheses of the respective tests. By decomposing the compound null hypothesis of no phenotype association within a set of multiple rare variants into simple null hypotheses, it becomes clear that a “rare variant association” can be interpreted in different fashions. Namely, a rare variant association can be either a difference between cases and controls in the cumulative frequency of all rare alleles as assumed by the length tests, or a pattern of frequency differences for individual variants as assumed by the joint tests. Both definitions of rare variant association are entirely reasonable and there are likely to be traits and susceptibility genes that satisfy each definition.

We classified many existing rare variant tests according to the geometric framework, including the most common and cited methods; most could be categorized as either a length or a joint test. Thus, despite the number and seeming diversity of rare variant tests proposed to date, they are actually quite similar in philosophy. However, we did observe some tests which did not obviously fit the geometric framework; for example, Bayesian approaches [Quintana et al., 2011; Yi and Zhi, 2011], extensions of single-marker methods [Li et al., 2010], and other approaches [Wu et al., 2011 with a nonadditive kernel]. We do not view the existence of such methods as a limitation of the geometric framework, but rather a dramatic shift in methodology from what has largely been developed to date. If a test satisfies the definition of neither a length nor a joint test, it could benefit the field as it would introduce alternative interpretations of the rare variant problem and potentially improve our ability to discover rare variant associations.

The geometric framework provides several practical applications to rare variant analysis. With performance known to differ between tests, however subtly, investigators are likely to apply multiple gene-based tests to the same dataset. The geometric classification method can be a useful tool in planning an analysis that includes multiple tests on the same set of rare variants. The investigator can choose a set of tests that provide

complementary pieces of information, for example, ensuring both length- and joint-style tests are applied. If some prior knowledge of the underlying genetic architecture is available, it can inform the set of tests that are likely to perform best. The geometric framework can also aid in the interpretation of the results from multiple rare variant tests. Given that we can predict which tests should produce similar *P*-values based on their classification (length vs. joint), differences in significance can be attributed to additional factors such as variant weighting schemes, which in turn can be valuable for interpreting a significant association signal. For example, frequency-based weights that increase significance of a test may provide some information on the underlying frequency spectrum of causal variants. Furthermore, tests within each category are likely to perform similarly with respect to artifacts in the data such as genotyping errors [Mayer-Jochimsen et al., 2012; Powers et al., 2011] and population stratification [Zawistowski et al., 2012]. Thus, methodological research in these areas may not need to consider each of the many rare variant tests individually, but rather consider implications within each broad class of tests.

Our exploration of the effect of norms on test power also has implications for rare variant analysis designs. Combining multiple rare variants into a single test statistic leads to an inevitable signal-to-noise problem due to the inclusion of noncausal variants. In our analysis, we observed that the value of the norm used to compute distance between case and control frequency vectors in a rare variant statistic contributes to the robustness of that statistic to the inclusion of noncausal variants in the analysis. We showed that in some cases, increasing the norm of the statistic can reduce power loss for inclusion of noncausal variants. However, only norms of  $p = 1$  and  $p = 2$  have been used to date, indicating that this is a relatively unexplored aspect of handling the signal-to-noise problem in rare variant testing.

This is noteworthy because the common current strategy for reducing the inclusion of noncausal variants is to combine only nonsynonymous and nonsense coding variants within a gene since these are assumed to be most likely to be causal. This strategy ignores the potential deleterious effect of noncoding and synonymous variants, both of which have been shown in some cases to contribute to disease risk. In particular, the ENCODE database highlights the critical role that the noncoding portion of the genome plays in gene expression [ENCODE, 2012]. Efficiently incorporating noncoding variation into rare variant tests while controlling for noncausal variants may be a powerfully strategy for detecting novel associations. The increased-norm approach provides an alternative to the nonsynonymous-only approach of controlling power loss due to inclusion of noncausal variants with the advantage that it allows a more complete set of rare variation in the gene, particularly noncoding variants, to be investigated.

The categorization of tests based on assumptions of risk and protective variants helps to systematically explain some of the differences in performance between rare variant tests. However, there are likely several additional factors that can

impact performance, most notably variant weighting strategies. Weighting rare variants based on some criteria of evidence for association is a common strategy to attempt to increase power and includes allele frequency-based weights, quantitative predictions of the damaging impact of a variant or a measure of conservation across species. Though we have not directly considered it here, the geometric framework can provide a context for evaluating the effect of various weighting strategies. The weighted variants can be viewed as transformations of the original genotype frequency vectors, and the merits of a weighting strategy could be determined by comparing the change in the original and transformed vectors. For example, if we considered weighting strategies for a length test, a set of weights would be beneficial if the weighted case and control vectors have a larger difference in normed length than did the original unweighted vectors.

The geometric framework presented here provides a rather unique perspective on the current field of rare variant association tests. In addition to improving our understanding of why certain tests perform as they do and which are methodologically most similar, we have also uncovered areas in which future modifications can be made to existing rare variant tests to improve power to detect associations.

## Acknowledgments

This work was funded by the National Human Genome Research Institute (R15HG004543; R15HG006915) with M.Z. supported by HG005855. We acknowledge the use of the Hope College parallel computing cluster for assistance in data simulation and analysis and for helpful discussions about this work with Mark Reppell, Michael Boehnke, Goncalo Abecasis, and Sebastian Zollner.

## References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

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

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. 11:773–785.

Basu S, Pan W. 2011. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 35:606–619.

Cooper GM, Shendure J. 2011. Needles in stacks of needs: finding disease causal variants in a wealth of genomic data. *Nat Genet* 12:628–640.

Dai Y, Jiang R, Dong J. 2012. Weighted selective collapsing strategy for detecting rare and common variants in genetic association study. *BMC Genet* 13:7.

Dering C, Pugh E, Ziegler A. 2011. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol* 35:S12–S17.

Feng T, Elston RC, Zhu X. 2011. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet Epidemiol* 35:398–409.

Gibson G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet* 13:135–145.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70:42–54.

Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. 2011. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 7:e1001289.

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB. 2012. The empirical power of rare variant association methods: results from sanger sequencing in 1998 individuals. *PLoS Genet* 8:e1002496.

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

Li Q, Zhang H, Yu K. 2010. Approaches for evaluating rare polymorphisms in genetic association studies. *Hum Hered* 69:219–228.

Li Y, Byrnes AE, Li M. 2011. To identify associations with rare variants, just WHaT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 87:728–735.

Lin D, Tang Z. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89:354–367.

Luedtke A, Powers S, Petersen A, Sitarik A, Bekmetjev A, Tintle NL. 2011. Evaluating methods for the analysis of rare variants in sequence data. *BMC Proc* 5:S119.

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

Mayer-Jochimsen M, Fast S, Tintle NL. 2012. Assessing the impact of differential genotyping errors on rare variant tests of association. *PLoS One* 8:e56626.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res Fund Mol Mech Mut* 615:28–56.

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

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e1001322.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, Shen J, Tang Z, Bacanu S, Fraser D and others. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337:100–104.

Pan W, Shen X. 2011. Adaptive tests for association analysis of rare variants. *Genet Epidemiol* 35:381–388.

Powers S, Gopalakrishnan S, Tintle NL. 2011. Assessing the impact of non-differential genotyping errors on rare variant tests of association. *Hum Hered* 72:152–159.

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86:832–838.

Quintana MA, Berstein JL, Thomas DC, Conti DV. 2011. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genet Epidemiol* 35:638–649.

Rivas MA, Beaudooin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burt N and others. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 43:1066–1075.

Sul J, Han B, He D, Eskin E. 2011. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genet* 188:181–188.

Sun YV, Sung YJ, Tintle NL, Ziegler A. 2011. Identification of genetic association of multiple rare variants using collapsing methods. *Genet Epidemiol* 35:S101–S106.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G and others. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Sci. Science* 3337:64–69.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Tintle NL, Aschard H, Hu I, Nock N, Wang H, Pugh E. 2011. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from group 7 at genetic analysis workshop 17. *Genet Epidemiol* 35:S56–S60.

Torgerson DG, Capurso D, Mathias RA, Graves PE, Hernandez RD, Beaty TH, Bleecker ER, Raby BA, Meyers DA, Barnes KC and others. 2012. Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am J Hum Genet* 90:273–281.

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

Yi N, Zhi D. 2011. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 35:57–69.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 87:604–617.

Zawistowski M, Reppell M, Wegmann D, St. Jean PL, Ehm MG, Nelson MR, Novembre J, Zollner S. 2012. Differential stratification in rare variant tests observed using an analytic model of joint site frequency spectra. Unpublished manuscript.

Zhang Q., Irvin MR, Arnett DK, Province MA, Borecki I. 2011. A data-driven method for identifying rare variants with heterogeneous trait effects. *Genet Epidemiol* 35:679–685.

## Appendix A: Length Tests

Detailed justification is provided for 10 of the 12 length tests (cohort allelic sums test [CAST], CMC, weighted sum [WS], data-adaptive Sum test [aSUM], Variable Threshold, odds ratio weighted sum statistic [ORWSS], *P*-value weighted sum test [PWST], rare variant weighted aggregate

statistic [RWAS], weighted haplotype and imputation-based tests [WHAIT], and weighted selective collapsing strategy [WSCS]; see Table 1) in the following sections of Appendix A. Justification for CMAT and PR is provided in the main text of the manuscript and not repeated here.

### CAST [Morgenthaler and Thilly, 2007]

The CAST [Morgenthaler and Thilly, 2007] compares the proportion of cases with at least one rare allele across the  $m$  variant sites of interest,  $\frac{\sum_{i=1}^{N^+} d_i}{N^+}$ , to the proportion of controls with at least one rare allele,  $\frac{\sum_{i=1}^{N^-} d_i}{N^-}$ , where  $d_i = \begin{cases} 1 & \text{if } \sum_{j=1}^m c_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$ . Note that  $\|f^+\|_1 = \sum_{j=1}^m f_j^+ = \frac{\sum_{i=1}^{N^+} \sum_{j=1}^m c_{ij}}{2N^+} \approx 2 \frac{\sum_{i=1}^{N^+} d_i}{N^+}$  since most individuals will only have a single rare variant. Thus, CAST is a length test because the quantities it compares are a function of  $\|f^+\|_1$  and  $\|f^-\|_1$ .

### CMC [Li and Leal, 2008]

The CMC test [Li and Leal, 2008] is similar to CAST in that, for all variants in a subset of  $m$  (typically, all the variants below some MAF threshold), an indicator variable,  $d_i$ , is created indicating whether an individual has at least one rare variant in the set. When all variants are in the same set, the test is equivalent to CAST. When there are two or more subsets of the  $m$  variants, then CMC becomes a joint test, as described in Appendix B.

### Weighted-Sum [Madsen and Browning, 2009]

The WS statistic [Madsen and Browning 2009], like other length tests, first collapses across variants to obtain a measure of rarity,  $r_i$  for each individual  $i$ . That is  $r_i = \sum_{j=1}^m \frac{c_{ij}}{\hat{w}_j}$ , where  $\hat{w}_j$  is the weight for the  $j$ th variant calculated using the MAF in the control sample. Information is then combined across individuals using ranks, where  $l^+ = \sum_{\text{cases}} \text{rank}(r_i)$  and  $l^- = \sum_{\text{controls}} \text{rank}(r_i)$ , and the statistic of interest is the size of the difference between  $l^+$  and  $l^-$ . If we consider the essence of WS to be a comparison of rarity indices ( $l^+$  and  $l^-$ ) in cases, then WS is a length test as follows:

$$\begin{aligned} (l^+ - l^-)^2 &\sim \left( \sum_{\text{cases}} r_i - \sum_{\text{controls}} r_i \right)^2 = \left( \sum_{\text{cases}} \sum_{j=1}^m \frac{I_{ij}}{\hat{w}_j} - \sum_{\text{controls}} \sum_{j=1}^m \frac{I_{ij}}{\hat{w}_j} \right)^2 \\ &= \left( \sum_{j=1}^m \sum_{\text{cases}} \frac{I_{ij}}{\hat{w}_j} - \sum_{j=1}^m \sum_{\text{controls}} \frac{I_{ij}}{\hat{w}_j} \right)^2 \\ &= (\|f^+\|_1 - \|f^-\|_1)^2, \end{aligned}$$

where the first equivalence is not formally true because of the nonlinear nature of the rank operation, but it suffices to show how WS is a test of the length variety.

### aSUM [Han and Pan, 2010]

Han and Pan recognize that collapsing variants is equivalent to assuming that all variants have the same odds ratio (OR). To get around this assumption, they propose using data-adaptive weights, where variants with larger ORs are given more weight, and variants with ORs  $< 1$  (protective variants) are given negative weight so that risk and protective variants do not cancel each other out. Marker weights,  $w_j$ , are obtained through single-marker logistic regression. All markers are then collapsed into a “super variant” as  $G_i^* = \sum_{j=1}^m w_j G_{ij}$ , where  $G_{ij}$  is the genotype of the  $i$ th individual at variant  $j$ . The super variant is then incorporated into a logistic regression model in the manner of PR. Thus, the arguments used for PR hold here, and aSUM is a length test.

### Variable Threshold Approach [Price et al., 2010]

Similar to aSUM, Price et al. [2010] attempt to use the data to determine marker weights,  $w_j$ . In the Price et al. [2010] approach, they consider a simple (0,1) weighting strategy whereby each marker is included or excluded based on whether its MAF is less than a threshold,  $T$ . The approach is called a Variable Threshold approach, because  $T$  is varied across a range of values and each time  $G_i^* = \sum_{j=1}^m w_j G_{ij}$  is regressed against the phenotype. Because, at its root, the Variable Threshold approach is similar to PR, for any particular value of  $T$ , the Variable Threshold approach is a length test.

### ORWSS [Feng et al., 2011]

As mentioned by the authors, the ORWSS approach is similar to the WS statistic. The main difference is in the choice of weights. Briefly, Feng et al. [2011] suggest a weight that assigns more weight to a variant which appears to be an “outlier” (large effect size) relative the majority of other variants in the set. Each individual is assigned a genetic score which is the WS of their genotypes. The ranks of the genetic scores in the cases are compared to the controls. The similarity to WS mean that ORWSS is a length test.

### RWAS [Sul et al., 2011]

It is straightforward to see that the RWAS [Sul et al., 2011] is a length test. The RWAS statistic is simply  $S_{RWAS} = \frac{\sum_{j=1}^m w_j z_j}{\sqrt{\sum_{j=1}^m w_j^2}}$ , where  $w_j$  is a variant weight based on the markers allele frequency and  $z_j = \frac{f_j^+ - f_j^-}{k_j}$  and  $k_j$  is another weight based on the observed MAF at site  $j$ . Thus,  $S_{RWAS} \propto \sum_{j=1}^m w_j f_j^+ - \sum_{j=1}^m w_j f_j^- \propto \|f^+\|_1 - \|f^-\|_1$ .

### PWST [Zhang et al., 2011]

The PWST [Zhang et al., 2011] extends the aSUM approach of Han and Pan [2010], using a different, data-adaptive, weighting strategy. Thus, PWST is a length test because aSUM is a length test.

## WHaIT [Li et al., 2011]

The WHaIT approach [Li et al., 2011] is related to both WS and ORWSS, and, thus, is a length test. Briefly, they compute a weighted haplotype score for each individual (akin to genetic score for ORWSS and WS), and then use rank test to assess significance. The weighted haplotype score consists of weights computed in a manner similar to ORWSS, except they use a training set of data to help determine optimal weights.

## WSCS [Dai et al., 2012]

Similar to aSUM, the authors attempt to find data-dependent weights,  $w_j$ , with which to collapse a group of variants into a super variant  $G_i^* = \sum_{j=1}^m w_j G_{ij}$ . The authors consider numerous approaches to determining the weights, which we do not describe in detail here. Ultimately, the test is a length test because a logistic regression model (see main text section *Length tests in section Methods*) is used to determine whether the set of variants is associated with the disease.

## Appendix B: Joint Tests

Detailed justification is provided for six of the seven joint tests (CMC, C-alpha, SSU, PWST; Lin and Tang, 2011) in the following sections of Appendix B. Justification for SKAT with a linear kernel is provided in the main text of the manuscript and not repeated here; however, we do discuss SKAT with nonlinear kernels.

## CMC [Li and Leal, 2008]

As noted in Appendix A and Table 1, CMC can act as both a length and/or joint test. When all variants in the set are in the same group, CMC is equivalent to CAST and acts as a length test. However, when there are multiple groups of variants, CMC is a joint test like SKAT. In particular, CMC is a joint test because it tests  $(y - \bar{y})' A^* W A^{*'} (y - \bar{y})$ , where  $A^*$  is a matrix of genotypes similar to  $A$ , but where some variants may be collapsed. Thus, because the formulation of the CMC test statistic is equivalent to SKAT acting on groups of variants, CMC can act as a joint test.

## C-Alpha [Neale et al., 2011]

As argued by Wu et al. [2011], C-alpha is a special case of SKAT when there are no covariates, the phenotype of interest is dichotomous and all variants are weighted equally. Thus, C-alpha is a joint test.

## SKAT [Wu et al., 2011]

Wu et al. [2011] showed that the SKAT statistic may be written as a WS  $\sum_{j=1}^m w_j S_j^2$ , where  $S_j$  is a score for the  $j$ th variant. When the linear kernel is used, the score can be written as

$S_j^2 = \sum_i \sum_{i'} (y_i - \hat{\pi}_{0i}) A_{ij} A_{i'j} (y_{i'} - \hat{\pi}_{0i'})$ . By changing the kernel, we essentially change the way that  $S_j^2$  is computed. As long as the kernel function  $K(A_i, A_{i'})$  can be linearized as  $K(A_i, A_{i'}) = \sum_{j=1}^m K_j(A_{ij}, A_{i'j})$ , the kernel function will be additively combining information across variants, and so the test will be a joint test. There are, however, kernel choices which could lead to nonadditive combination of information across variants (e.g., quadratic kernel).

## SSU [Basu and Pan, 2011]

As noted by Basu and Pan [2011], the unweighted SSU statistic is equivalent to SKAT using a linear kernel. As shown earlier, SKAT is a joint test and, thus, SSU is a joint test. Basu and Pan [2011] go on to propose a variety of different variant weights that can be used with SSU; however, variant weights do not fundamentally change the classification of the test.

## GFRV [Lin and Tang, 2011]

Lin and Tang [2011] propose a general framework for rare variant (GFRV) testing. The focus of their work is on choosing weights related to biological context. Each variant is given a weight, and then the SSU test is applied. Thus, GFRV is a joint test since SSU is a joint test.

## AT [Pan and Shen, 2011]

Pan and Shen [2011] propose adaptive tests (AT) for rare variant testing. In short, Pan and Shen [2011] propose an adaptive approach to attempt to eliminate variants a priori since they argue that power loss can be substantial from the inclusion of no or weak effects variants.

## PRVT [Ionita-Laza et al., 2011]

The Poisson rare variant test (PRVT) [Ionita-Laza et al., 2011] attempts to leverage the fact that the frequency of rare variants is approximately Poisson. The PRVT is a joint test because it evaluates the difference in allele frequency between cases and controls at each marker, and then sums over the markers in the set. The unique aspect of PVRT is that rather than simply take the difference in the case and control allele frequencies, the PVRT assesses the likelihood of observing at least  $c_j^+$  rare alleles in the cases at site  $j$  and at most  $c_j^-$  rare alleles in the controls, given a Poisson distribution (with parameter estimated from the data). In particular, at each variant site  $j$ , the distance,  $D_j^*$ , between the case and control counts is defined as

$$D_j^*(c_j^+, c_j^-) = \begin{cases} -\log(p(c_j^+, c_j^-)), & c_j^+ > c_j^- \\ -\log(p(c_j^-, c_j^+)), & c_j^- > c_j^+ \end{cases},$$

where  $p(c_j^+, c_j^-) = ppois(c_j^-, \hat{\lambda}_j)(1 - ppois(c_j^+ - 1, \hat{\lambda}_j))$ ,  $ppois(\cdot)$  represents the Poisson cdf and  $\hat{\lambda}_j = \frac{c_j^+ + c_j^-}{2}$ . For a generic joint



test (see main text), instead of the distance metric shown above, we use  $D(c_j^+, c_j^-) = |\frac{c_j^+}{2N^+} - \frac{c_j^-}{2N^-}|^p$ . Since the alternative distance metric,  $D^*$ , used by PRVT is monotonic increasing in terms of  $D$ , PVRT is a joint test.

### Appendix C: Handling Covariates in $L_p$ and $J_p$

In the main text, we show how the geometric framework suggests alternative rare variant test statistics  $L_p$  and  $J_p$ . Our presentation of these statistics in the manuscript does not consider the case of covariates. We consider covariates briefly here.

The presence of meaningful covariates suggests that the contribution of each individual toward the value of the statistic should not be equal, but rather weighted based on their “background” covariates which increase or decrease an individual’s disease risk independent of the presence of particular genetic variants. Thus, we propose that in order to modify  $L_p$  and  $J_p$  to account for covariates, we up- or down-weight each individual according to their disease risk given covariates under the null model.

To obtain intuition for the correct weighting scheme, consider the case of no covariates and equal numbers of cases and controls. In this situation, the estimated null probability of disease is  $\hat{\pi}_0 = 0.5$ . We note that

$c_j^+ - c_j^- = A_j' y - A_j'(1 - y) = A_j'(y - 0.5) - A_j'((1 - y) - 0.5) = A_j'(y - \hat{\pi}_0) - A_j'((1 - y) - (1 - \hat{\pi}_0)) \propto (\sum_{\text{cases}} A_{ij}(1 - \hat{\pi}_{0i})) - (\sum_{\text{controls}} A_{ij}\hat{\pi}_{0i})$ . This suggests that the correct weight for each individual,  $i$ , in the cases is  $1 - \hat{\pi}_{0i}$  and is  $\hat{\pi}_{0i}$  in the controls. This motivates the following adjusted genotype counts:  $\tilde{c}_j^+ = \sum_{\text{cases}} A_{ij}(1 - \hat{\pi}_{0i})$  and  $\tilde{c}_j^- = \sum_{\text{controls}} A_{ij}\hat{\pi}_{0i}$ , where for a set of covariates,  $\mathbf{X}$ , we estimate  $\hat{\pi}_{0i}$  under the null model:

$$\text{logit}(P(Y_i = 1)) = \alpha \mathbf{X}_i.$$

Thus, we can write covariate adjusted versions of  $L_p$  and  $J_p$  as

$$\tilde{L}_p = \|\tilde{\mathbf{f}}^+\|_p - \|\tilde{\mathbf{f}}^-\|_p$$

$$\tilde{J}_p = \|\tilde{\mathbf{f}}^+\| - \|\tilde{\mathbf{f}}^-\|.$$

Where,  $\tilde{f}_j^+ = \tilde{c}_j^+$  and  $\tilde{f}_j^- = \tilde{c}_j^-$ . We note that  $\tilde{L}_1$  is equivalent to PR and  $\tilde{J}_p$  is equivalent to SKAT with a linear kernel and equal weights.

Similarly,  $\tilde{W}(p, q, w_1) = w_1(\|\tilde{\mathbf{f}}^+\|_p - \|\tilde{\mathbf{f}}^-\|_p)^2 + (1 - w_1) \cdot 2 \cdot \|\tilde{\mathbf{f}}^+\|_q \cdot \|\tilde{\mathbf{f}}^-\|_q \cdot (1 - \cos(\hat{\theta}))$ , where  $\hat{\theta}$  is the angle between  $\tilde{\mathbf{f}}^+$  and  $\tilde{\mathbf{f}}^-$ .