BinomiRare: A carriers-only test for association of rare genetic variants with a binary outcome for mixed models and any case-control proportion — Source link ↗

Tamar Sofer, Tamar Sofer, Jiwon Lee, Nuzulul Kurniansyah ...+37 more authors

Institutions: Harvard University, Brigham and Women's Hospital, University of Washington, Fred Hutchinson Cancer Research Center ...+12 more institutions

Related papers:

- A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies

- Utilizing population controls in rare-variant case-parent association tests.

- A Novel Statistic for Global Association Testing Based on Penalized Regression.

- Does accounting for gene-environment interactions help uncover association between rare variants and complex diseases?

- Meta-analysis of Complex Diseases at Gene Level with Generalized Functional Linear Models.

BinomiRare: A carriers-only test for association of rare genetic variants with a binary outcome for mixed models and any case-control proportion

Tamar Sofer[1,2,3,*], Jiwon Lee[3], Nuzulul Kurniansyah[3], Deepti Jain[4], Cecelia A. Laurie[4], Stephanie M. Gogarten[4], Matthew P. Conomos[4], Ben Heavner[4], Yao Hu[5], Charles Kooperberg[5], Jeffrey Haessler[5], Ramachandran S. Vasan[6,7], L. Adrienne Cupples[7,8], Brandon J. Coombes[9], Amanda Seyerle[10], Sina A. Gharib[11], Han Chen[12,13], Jeffrey R. O'Connell[14], Man Zhang[15], Daniel Gottlieb[3], Bruce M. Psaty[16,17], W.T. Longstreth Jr.[17], Jerome I. Rotter[19], Kent D. Taylor[19], Stephen S. Rich[20], Xiuqing Guo[19], Eric Boerwinkle[12,21], Alanna C. Morrison[12], James S. Pankow[22], Andrew D. Johnson[7,23], Nathan Pankratz[24], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Alex P. Reiner[5], Susan Redline[1,3], Nicholas L. Smith[18,25,26], Kenneth M. Rice[4], Elizabeth D. Schifano[27]

[1]Department of Medicine, Harvard Medical School, Boston, MA, USA

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[3]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA

[4]Department of Biostatistics, University of Washington, Seattle, WA, USA

[5]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[6]Departments of Medicine and Epidemiology, Boston University Schools of Medicine and Public Health, Boston, MA, USA

[7]Framingham Heart Study, Framingham, MA, USA

[8]Department of Biostatistics, Boston University, Boston, MA, USA

[9]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

[10]Division of Pharmaceutical Outcomes and Policy, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[11]Computational Medicine Core, Center for Lung Biology, Department of Medicine, University of Washington, Seattle, WA, USA

[12]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA

[13]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

[14]Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland School of Medicine, Baltimore, MD, USA

[15]Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

[16]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA

[17]Departments of Neurology and Epidemiology, University of Washington, Seattle, WA, USA

[18]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA

[19]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA

[20]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA.

[21]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

[22]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, USA

[23]Population Sciences Branch, Division Intramural Research, National Heart, Lung and Blood Institute, Framingham, MA, USA

[24]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA

[25]Department of Epidemiology, University of Washington, Seattle, WA USA

[26]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research & Development, Seattle WA USA.

[27]Department of Statistics, University of Connecticut, Storrs, CT, USA

*Correspondence: Tamar Sofer, tsofer@bwh.harvard.edu

Abstract

Whole genome and exome sequencing studies have become increasingly available and are being used to identify rare genetic variants associated with health and disease outcomes. Investigators routinely use mixed models to account for genetic relatedness or other clustering variables (e.g. family or household) when testing genetic associations. However, no existing tests of the association of a rare variant association with a binary outcome in the presence of correlated data controls the Type 1 error where there are (1) few carriers of the rare allele, (2) a small proportion of cases relative to controls, and (3) covariates to adjust for. Here, we address all three issues in developing the carriers-only test framework for testing rare variant association with a binary

trait. In this framework, we estimate outcome probabilities under the null hypothesis, and then use them, within the carriers, to test variant associations. We extend the BinomiRare test, which was previously proposed for independent observations, and develop the Conway-Maxwell-Poisson (CMP) test, and study their properties in simulations. We show that the BinomiRare test always controls the type 1 error, while the CMP test sometimes does not. We then use the BinomiRare test to test the association of rare genetic variants in target genes with small vessel disease stroke, short sleep, and venous thromboembolism, in whole-genome sequence data from the Trans-Omics for Precision Medicine program.

Introduction

Whole-genome and exome-sequencing studies are becoming increasingly available to public health researchers, for example, from NHLBI's Trans-Omics for Precision Medicine (TOPMed) program (1), NHGRI's Centers for Common Disease Genetics (CCDG), and the UK Biobank (2). As most variant in sequencing datasets are rare, researchers may be interested in using such datasets for detecting rare-variant associations, genome-wide or in a genomic region of interest. They may also seek to confirm suggested associations from other studies or populations, or to assess pathogenicity in large population-based studies of rare variant alleles reported from small family-based studies. For example, Amininejad et al. (3) studied the association of genetic variants within genes associated with monogenic immunodeficiency disorders with Crohn's Disease. Wright et al. (4) assessed the pathogenicity and penetrance of rare variants identified in clinical studies, in the population-based UK Biobank. Tuijnenburg et al. (5) studied rare genetic variants within *NGKB1* for association with primary immunodeficiency disease. Do et al. (6) studied risk of myocardial infarction in carriers of rare *LDLR* and *APOA5* alleles. Kendall et al.

(7) studied cognitive outcomes in carriers of rare copy number variants. These studies

demonstrate that there is an interest in testing single rare genetic variant associations with a wide

range of health outcomes, including binary outcomes such as disease or affection status.

Testing rare variant associations with binary traits is challenging. It was previously shown that

likelihood-based tests such as the Wald, Score, and likelihood ratio tests poorly control Type 1

error when testing for rare variant associations with a binary trait (8; 9). The Score test

performance depends on the case-control ratio, and for rare variants, even a small imbalance

causes "inflation" (i.e. too many false positive results). A few approaches have been used

previously to study rare variant associations in a set of unrelated individuals. Amininejad et al.

used a permutation approach to test for association of rare genetic variants with Crohn's Disease.

Wright et al. used Fisher's exact test. While it is possible to adjust for covariates in the

permutation approach and when using Fisher's exact test to some extent through stratification

(10), they do not have the full flexibility of covariate adjustment of a generalized linear model;

i.e. they still require the identification of distinct groups in which no additional adjustment is

required. Further, permutation tests may also be computationally intensive if low p-values are

desired, because the number of required permutations may be large, although there are ways to

reduce this computational burden (10). Alternatively, Tuijnenburg et al. used a method called

BeviMed (11), implementing a Bayesian model to estimate posterior disease probabilities. The

BinomiRare test has also been proposed as a powerful method to test for rare variant associations

that can account for covariates (9). The BinomiRare test uses standard methods to compute the

disease probabilities in the entire dataset, under the null hypothesis of no association between a

specific genetic variant and the binary outcome. Then, for each specific genetic variant, it uses

the estimated probabilities in the variant carriers to test the hypothesis that the disease

probabilities under the null are the true outcome probabilities in the carriers. The null hypothesis

is rejected if the number of carriers with the outcome is inconsistent with their outcome

probabilities. However, the previously published version of this method assumed the sample

contains only unrelated individuals. Currently, there is no single-variant test that is generally

appropriate for testing rare variants when individuals are correlated, e.g., due to known or cryptic

genetic relatedness. Notably, the saddlepoint approximation to compute p-values (henceforth

SPA; (12)) was first developed to improve the calibration of the Score test when there is case-

control imbalance, and then extended in the SAIGE framework for the settings where related

individuals are used (13). However, it does not reliably control the Type I error rate when the

number of carriers of the rare variants is very small (i.e. tens of individuals; (14)). Therefore,

there is a need for a statistical test that is well-calibrated when the number of carriers is low,

individuals are potentially related, and there is case-control imbalance.

The previously published version of BinomiRare test (14) is useful in the presence of case-

control imbalance, allows for covariate adjustment, controls the Type I error rate for any number

of carriers, and can also be used when combining heterogeneous studies and here, we expand its

framework for testing rare variant associations when study individuals are correlated. We

developed two tests: first, we extended the BinomiRare test to the mixed models setting by

applying it on conditional probabilities computed with a mixed model, rather than on marginal

probabilities. Second, we developed the Conway-Maxwell-Poisson (CMP) test, which follows

the carriers-only framework by using estimated (conditional) disease probabilities like the

BinomiRare. For a given rare variant it uses the estimated disease probabilities in the variant

carriers to fit the parameters of the CMP distribution, under the null. It then tests whether the observed number of carriers with the outcome is consistent with this distribution. We study these tests using synthetic simulations with varying outcome probabilities, variant allele frequencies, and strengths of correlation between individuals due to genetic relatedness, and also in realistic simulation studies, using real phenotypes and WGS-based variant call set from the TOPMed program. We finally apply the BinomiRare test to test rare variant associations in known disease causing genes for specific disorders: the *NOTCH3* gene and small vessel disease (SVD)ischemic stroke; the *DEC2* (also known as *BHLHE41*) gene and "short sleep", and the *F5* gene and venous thromboembolism (VTE).

Methods

Statistical approach

Let D_i be an indicator of the disease, or another binary outcome, of participant i, with value 1 if the person is affected and 0 otherwise, where $i = 1, \dots, n$ and the n individuals may be correlated. Let $\boldsymbol{x}_i$ be a $p \times 1$ vector of covariate values for the ith participant, and g_i be their count of minor alleles for a specified genetic variant. Under the logistic disease model for correlated data:

$$\text{logit}(p_i) = \boldsymbol{x}_i^T \boldsymbol{\alpha} + g_i \beta + b_i, i = 1, \dots, n$$

with $p_i = \text{Pr}(D_i = 1 | \boldsymbol{x}_i, g_i, b_i)$ is the conditional outcome probability in the sample (regardless of the population probability), and b_i is the ith entry of the vector $\boldsymbol{b} = (b_1, \dots, b_n)^T \sim N(\boldsymbol{0}, \sum_{k=1}^K \sigma_k^2 V_k)$ of correlated random effects with possibly K variance components $\sigma_k^2, k = 1, \dots, K$ and V_k modelling the correlation structure corresponding to a particular source of correlation. While the methods proposed here can be applied for an arbitrary $K \geq 1$, we simplify

presentation by focusing on the scenario of a single correlation matrix modeling genetic relatedness, possibly cryptic, so that $b \sim N(0, \sigma_g^2 G)$ with G being any genetic relationship matrix (GRM), or possibly kinship matrix, and σ_g^2 is the corresponding variance component.

We assume that the genetic variant is rare, so that the minor allele frequency (MAF) is low and that carriers of the minor allele are overwhelmingly heterozygotes. While having homozygotes does not invalidate our approach, it also does not increase statistical power. Our carriers-only approach first estimates a disease probability for each individual in the sample under the null hypothesis of no association between the genetic variant and disease status, i.e. under the assumption that $\beta = 0$ by not including any variant of interest in the regression (step 1, demonstrated in Figure 1), and then considers carriers of the rare variant, testing whether the number of diseased carriers is consistent with their estimated disease probabilities (step 2, demonstrated in Figure 2).

Step 1: Estimating disease probabilities under the null hypothesis. At step 1, we fit a null model under the assumption $\beta = 0$, using the existing penalized quasi-likelihood algorithm for logistic mixed models (15). This approach is implemented in multiple software, including the GENESIS R package (16), GMMAT (17), and SAIGE (13). In both GENESIS and GMMAT, the vector of fixed effects α and the variance component σ_g^2 are estimated using an implementation of an AI-REML (Average Information Restricted Maximum Likelihood) algorithm on top of the penalized quasi-likelihood (PQL) approach (17), but the proposed tests do not depend on the specific algorithm used for estimating the outcome probabilities. From the fitted null model, we obtain estimates $\widehat{\alpha}, \widehat{b}$, and an estimated disease probability vector by plugging them in to obtain

$\hat{p}_i = \text{expit}(x_i^T \hat{\alpha} + \hat{b}_i), i = 1, \dots, n$, where *expit* is the inverse of the logit function. If the variance

component σ_g^2 is estimated as 0, so is $\hat{b} = 0$, and the analysis reverts to the independent

individual settings.

Step 2: Testing the association between a genetic variant and disease status.

Suppose that we obtained disease probability estimates $\hat{p}_i, i = 1, \dots, n$, under the null as

described above. Denote n_c as the number of carriers of the rare variant ("carriers" henceforth),

i.e. those with $g > 0$, so that $\sum_{i=1}^{n} 1(g_i > 0) = n_c$. Without loss of generality, assume that

participants $i = 1, \dots, n_c$ are the carriers. Let n_d be the number of diseased carriers

$$n_d = \sum_{i=1}^{n_c} 1(d_i = 1) = \sum_{i=1}^{n} 1(d_i = 1, g_i > 0).$$

Let $\widehat{p_{n_c}} = (\widehat{p_1}, \dots, \widehat{p_{n_c}})^T$ denote the vector of estimated disease probabilities for carriers of the

rare variant. Despite $\widehat{p_{n_c}}$ being estimated, we treat it as fixed. For testing, we assess the

goodness-of-fit of the estimated model to the observed disease status in the carriers, by testing

the null hypothesis:

$$H_0: p_{n_c} = \widehat{p_{n_c}},$$

where p_{n_c} is the true, unknown, vector of outcome probabilities among the carriers.

The *p*-value for testing the null hypothesis of no variant-disease association is given by:

$$p\text{-value} = \Pr\{n_d \text{ diseased carriers or more extreme} | \widehat{p_{n_c}}\}. \tag{1}$$

This is a two-tailed p-value, because n_d can appear to be lower or higher than expected. When

only a single person carries the rare variant, i.e. $n_c = 1$, the calculation is trivial; Equation (1)

reduces to the single carrier's fitted probability $\hat{p}_i$ if they are a case, and $1 - \hat{p}_i$ if they are a

control. When $n_c > 1$, there are two special cases that are already developed. If $\hat{p}_i$ for all carriers

are equal, and outcomes for all carriers are independent, then $n_d \sim Binomial(n, \hat{p}_i)$, and the p-

value is the tail area (possibly two tails) of the standard Binomial distribution, i.e. a Binomial

exact test. If the $\hat{p}_i$ for the carriers differ but independence still holds, the distribution is the

Poisson-Binomial distribution, and the test is the previously-proposed BinomiRare test for

independent data (9). In the general case, an arbitrary sum of binomial variables, possibly

correlated, has the Conway-Maxwell-Poisson (CMP)-Binomial distribution, which can be

approximated by the CMP distribution (18; 19) when the number of carriers is "large enough"

(see appendix).

In addition to the p-value above, we also study the mid-p-value, which was previously shown to

improve properties of discrete tests (20) and to be less conservative. The mid-p-value is always

smaller than the p-value, because when summing the tail areas probabilities, it accounts for only

half of the probability of the observed event n_d, whereas the p-value uses it as it is, without

dividing in half.

BinomiRare and CMP tests using conditional probabilities

In the appendix, we show that the distribution of n_d in the general case can be approximated by

the CMP distribution and develop the CMP test. However, because approximations may not

work well in practice for low carrier count n_c, we also attempt a different approach. Note that for

two individuals i and j, we have that D_i and D_j were independent if the true conditional disease

probabilities were known. In other words, given conditional disease probabilities, knowing the

disease status of individual i does not inform of the disease status of individual j. Therefore, we

consider using the BinomiRare test which was developed for independent data – with the conditional probabilities. We note that this independence may not hold when probabilities are estimated, and therefore it is not trivially true that the BinomiRare is appropriate in this setting. Both the CMP and the BinomiRare tests for correlated data are available in the GENESIS R package for genetic association analysis (21).

Simulation study: testing rare variant associations using BinomiRare and CMP in a sample of trios

We carried out a simulation study to evaluate the performance of BinomiRare and CMP tests in samples of correlated individuals. In each simulation, we generated 3,000 individuals as 1,000 trios (two parents and one offspring), as follows. For 1,000 pairs of parents, and each of two chromosomal copies, we generated 20 independent "non-causal" genetic variants by first sampling minor allele frequencies (MAF) from a uniform $U[0.05, 0.5]$ distribution and setting MAF $\in \{0.05, 0.02, 0.01, 0.001\}$ for one "causal" variant, followed by sampling of genetic variants using a Binary distribution based on these MAF. For each parent, allele count was the sum of the two sampled alleles. For each variant independently, an offspring inherited one allele from each of the parents. The parental allele was sampled at random with equal probabilities from the two alleles. We used the 21 (1 causal and 20 non-causal) simulated genotypes to generate a variable mimicking a principal component (PC), as a weighted sum of all allele counts, with weights sampled from a standard Normal distribution $N(0,1)$. Next, we simulated probability of disease using a mixed logistic model:

$$logit[p(D_i = 1)] = \beta_0 + PC_i \times \beta_{pc} + g_i\beta_g + b_i, \ \ i = 1, \dots, n.$$

Here, $\exp(\beta_0) \in \{0.01, 0.05, 0.5\}$ is the probability of disease in non-carriers ($g_i = 0$) with genetic PC and b_i equal to zero; β_{pc} models the association of the PC with disease probability, β_g is the effect of the (causal) variant of interest, and $\boldsymbol{b} = (b_1, \ldots, b_n)^T$, representing the correlation across individuals, is sampled from a multivariate normal distribution $\boldsymbol{b} \sim MVT\text{-}N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{K})$, with the correlation matrix $\boldsymbol{K}$ being a block diagonal kinship matrix, having twice the kinship coefficient between a child and each of their parents, i.e. 0.5. We set $\sigma_g^2 \in \{0.06, 0.6\}$. In all simulations we had $\beta_{pc} = 0.1$. The variant effect was varied from zero when evaluating Type 1 error rate, to $\beta_g = \log(Odds\ Ratio) \in \{\log(2), \log(3), \log(4)\}$ when evaluating power. We then sampled disease status for each individual from a Binary distribution with the computed disease probability. Finally, we applied the BinomiRare and CMP tests and computed p-values and mid-p-values. We performed $1x10^7$ replicates to estimate Type 1 error rate and $1x10^5$ replicates to estimate power. We estimated Type 1 error rate and power for p-value threshold for declaring significance $\{1x10^{-2}, 1x10^{-3}, 1x10^{-4}\}$. For tests that did not, empirically, control the Type I error rate for a given p-value threshold (i.e. the proportion of simulations passing the threshold was higher than the threshold), we computed a "calibrated threshold", defined as a value for which the proportion of simulations with p-value less than this value was the desired threshold. We then used this calibrated threshold to estimate power, specifically power at an "honest alpha". Our main results are those focused in simulations in which the variance component had non-zero estimate, but we analyze all simulations.

The TOPMed whole genome sequencing Study

WGS was performed via TOPMed and the NHGRI's Centers for Common Disease Genetics (CCDG) programs. WGS was performed using DNA from blood at multiple sequencing centers

using Illumina X10 technology at an average sequencing depth of >30X. Studies and samples were sequenced in multiple phases. Periodically, the TOPMed Informatics Research Center (IRC) performed variant calling on the combined TOPMed and CCDG samples, resulting in multiple releases of data "freezes". Details regarding sequencing methods and quality control are provided elsewhere (22) and in the TOPMed website (https://www.nhlbiwgs.org/data-sets).

We used three TOPMed multi-ethnic data sets: a data set of small vessel disease stroke (SVD stroke) in the Women Health Initiative (WHI), a study of short-sleep, and a study of venous thromboembolism (VTE), with the latter two comprised of individuals from multiple TOPMed cohorts. We performed data analysis to demonstrate the BinomiRare test. The approaches for data analysis were similar. GRMs were constructed based on the analytic datasets of each of the analyses, using all genetic variants with minor allele frequency≥0.001. Logistic mixed models under the null were fit and adjusted for age, sex, and self-reported race/ethnic group, and for short-sleep, also for parent study/cohort. SVD stroke and short sleep analyses used TOPMed freeze 5b release, while the VTE analysis used TOPMed freeze 8 genotype release. All participants provided written informed consent at their recruitment centers.

The TOPMed WHI stroke dataset

The Women's Health Initiative (WHI) is a long-term health study following postmenopausal women aged 50-79 years who were recruited from 1993 through 1998 from 40 clinical centers throughout the U.S. (23). In the present analysis, we focus on a subset of 5,358 WHI participants who were sequenced through TOPMed with data available via freeze 5b, and had SVD stroke case-control classification, according to the following methodology: stroke diagnosis requiring

and/or occurring during hospitalization was based on the rapid onset of a neurological deficit
attributable to an obstruction or rupture of an arterial vessel system. Hospitalized incident stroke
events were identified by semiannual questionnaires and adjudicated following medical record
review, which occurred both locally (at individual study sites) and centrally. Ischemic strokes
were further classified by the central neurologist adjudicators into cardio-embolic stroke, larger
artery stroke, and SVD stroke according to the Trial of Org 10172 Acute Stroke Trial (TOAST)
criteria (24). The TOAST classification focuses on the presumed underlying stroke mechanism
and requires detailed investigations (such as brain computed tomography, magnetic resonance
imaging, angiography, carotid ultrasound, and echocardiography). Baseline stroke cases were
excluded from the analysis and VTE cases were excluded from the control samples. Further,
participants who had non-SVD stroke were excluded.

The TOPMed short sleep dataset

We used sleep duration data from multiple TOPMed cohorts, as described in the Supplementary
Information detailing phenotype harmonization for short sleep analysis. Short sleep was defined
as self-reported sleep duration during weekday, or usual sleep (if sleep duration during the
weekdays was not available) being 5 hours or less. Otherwise, if self-reported sleep duration was
6 hours or longer and less than 9 hours, sleep was "normal". Individuals with self-reported sleep
duration longer than 5 hours and shorter than 6 hours were excluded to minimize risk of
misclassification. Because of a well-known "U-shaped" relationship between sleep duration and
cardiovascular disease (25), suggesting that potential non-linearity in genetic associations may
exist as well, we also excluded "long sleepers" reporting 9 or more hours of usual sleep,

The TOPMed VTE dataset

The TOPMed VTE dataset includes TOPMed participants from six studies, combining prospective cohort and case-only studies. Individuals were matched across groups defined to be homogeneous with respect to race/ethnicity and sex, and strata defined by age at event (determined according to cases). The matching strategy resulted in a sample set mimicking a case-control study, with 11,627 individuals of which 3,793 are cases, and 7,834 are controls.

Association testing of rare coding variants within known disease causing genes

For each of the SVD stroke, short sleep, and VTE datasets, we considered a known gene associated with the disorder. For stroke, we focused on the *NOTCH3* gene, in which mutations may cause Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), which causes ischemic stroke (26). For short sleep, we focused on the gene *DEC2* (also known as *BHLHE41*), a transcription inhibitor of orexin, a neuropeptide that regulates wakefulness (27; 28). For VTE, we focused on the coagulation factor V gene, *F5* (29; 30), which has a known common variant highly associated with VTE, factor V Leiden (rs6025). We performed single variant analysis within the candidate genes, as follows. We selected a subset of rare variants within the genes based on functional annotations, with the goal of increasing power by focusing on variants that are more likely to be functional compared to others. In detail, the filter based on functional annotation included the selection of variants that were: (a) high confidence loss of function variants according to the Ensembl Variant Effect Predictor (31), (b) missense variants if they are predicted deleterious by either SIFT 4G (32), Polyphen2-HDIV (33), Polyphen2-HVAR (33), or LRT-pred (34), (c) inframe indels with FATHMM-XF coding score > 0.5 (35), or (d) variants that are synonymous according to the

Ensemble Variant Effect Predictor and have FATHMM-XF coding score > 0.5. The annotation

based variant filtering was performed using the Annotation explorer application on the NHLBI's

BioData Catalyst (36). We further filtered variants to those that passed the TOPMed QC filter

(22), had at least 3 carriers of the rare allele, and had no more than 300 carriers. This upper

threshold was defined because we were interested specifically in rare variants and because it was

previously shown that properties of statistical tests of rare variant associations depend on carrier

count, rather than on allele frequency (14). Finally, we further restricted the set of variants to

those that had reasonable statistical power according to a power analysis performed as follows.

We arbitrarily assumed an odds ratio (OR) of 2 for a causal variant, and for each variant we

computed power based on a function developed for the BinomiRare test (provided here

https://github.com/tamartsi/Binary_combine/blob/master/compute_power.R) . The function uses

the estimated outcome probabilities in the sample, an OR, the number of variant carriers, and p-

value threshold, to compute power. To increase accuracy, for each variant we specifically used

the estimated disease probabilities among the variant carriers.

Results

Simulations studies

We studied the performance of the tests in simulations of 1,000 trios. In the setting where

$\sigma_g^2 = 0.06$, about half of the simulations estimated the variance component to be zero. When

$\sigma_g^2 = 0.6$, this happened in about a third of the simulations. The number of carriers of the

simulated rare variant allele was in the range [0, 27] when MAF = 0.001, [17, 119] when MAF =

0.01, [58, 203] when MAF = 0.02, and [195, 401] when MAF = 0.05. Table 1 provides estimated

Type 1 error rates in the simulations, restricted to those simulations in which the estimated variance component was $\hat{\sigma}_g^2 > 0$. For BinomiRare, we only provide results for the mid-p-value, because in our simulations it always controlled the Type 1 error rate, while the usual p-value controlled it as well while being more conservative. For CMP, we only provide results for the usual p-value, because it sometimes did not control the type 1 error and the lack of control was worse with the mid-p-value. The CMP test usually did not control the type 1 error when the variant was very rare (MAF =0.001), and when the case proportion was low $(\exp(\beta_0) = 0.05)$. Its performance improved as the MAF increased. In the Supplementary Information, Table S3-S5, we provide complete simulation results, including both mid-p-value and the usual p-value for both the CMP and BinomiRare tests, and results computed over all simulations, and computed over the simulations in which $\hat{\sigma}_g^2 = 0$

In power analysis, after appropriately calibrating the p-value threshold for the CMP test, CMP was either equally powerful as BinomiRare or more powerful (see Figures S1-S3 in the Supplementary Information). The patterns were similar across p-value thresholds used, and across the two variance component parameters used in the simulations. Notably, when the disease was common $(\exp(\beta_0) = 0.5)$ the power was lower when the variance component was high $(\sigma_g^2 = 0.6)$ compared to when it was low $\sigma_g^2 = 0.06$. When the disease was rare $(\exp(\beta_0) = 0.05)$, the power was essentially the same with both values of variance components.

Table 1: Estimated Type 1 error rates of BinomiRare and CMP tests in simulations with related individuals. Bolded numbers highlight settings in which the Type 1 error was not controlled, defined according to Type 1 error rate being larger than the highest value in a 95% confidence intervals around the expected Type 1 error rate, based on Binomial distribution with parameters being the p-value threshold and number of simulations used.

MAF	$\exp(\beta_0)$	Estimated type 1 error by p-value threshold					
		$\sigma_g^2 = 0.06$			$\sigma_g^2 = 0.6$		
		10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
		BinomiRare (mid-p-value)					
0.001	0.01	3.38E-03	2.08E-04	1.25E-05	3.78E-03	2.55E-04	1.88E-05
0.001	0.05	5.61E-03	4.45E-04	2.76E-05	5.63E-03	4.4E-04	3.24E-05
0.001	0.5	5.78E-03	3.22E-04	1.57E-05	5.44E-03	2.94E-04	1.26E-05
0.01	0.01	6.22E-03	4.62E-04	3.25E-05	6.36E-03	4.64E-04	3.35E-05
0.01	0.05	7.70E-03	6.58E-04	4.42E-05	7.83E-03	6.58E-04	5.1E-05
0.01	0.5	8.68E-03	8.28E-04	6.67E-05	8.21E-03	7.3E-04	6.56E-05
0.02	0.01	6.17E-03	4.27E-04	2.73E-05	6.41E-03	4.71E-04	3.38E-05
0.02	0.05	7.53E-03	6.27E-04	5.34E-05	7.52E-03	6.15E-04	5.27E-05
0.02	0.5	7.90E-03	6.87E-04	6.51E-05	7.56E-03	6.34E-04	5.55E-05
0.05	0.01	4.99E-03	2.88E-04	1.84E-05	5.21E-03	2.97E-04	1.48E-05
0.05	0.05	5.94E-03	4.49E-04	3.10E-05	5.93E-03	4.26E-04	2.73E-05
0.05	0.5	6.02E-03	4.69E-04	3.91E-05	5.67E-03	4.16E-04	3.34E-05
		CMP (usual p-value)					
0.001	0.01	5.75E-02	6.59E-03	4.18E-04	6.01E-02	6.54E-03	4.41E-04
0.001	0.05	**4.50E-02**	**4.44E-03**	**2.83E-04**	**4.22E-02**	**3.89E-03**	**2.26E-04**
0.001	0.5	**3.44E-02**	9.29E-04	6.28E-06	**3.34E-02**	7.78E-04	8.21E-06
0.01	0.01	2.34E-02	2.19E-03	1.70E-04	2.25E-02	2.09E-03	1.68E-04
0.01	0.05	**1.62E-02**	**1.55E-03**	**1.37E-04**	**1.53E-02**	**1.44E-03**	**1.20E-04**
0.01	0.5	9.30E-03	7.53E-04	4.41E-05	8.71E-03	6.74E-04	4.65E-05
0.02	0.01	1.76E-02	1.50E-03	1.11E-04	1.69E-02	1.44E-03	1.16E-04
0.02	0.05	**1.11E-02**	**1.10E-03**	1.02E-04	1.02E-02	9.77E-04	8.58E-05
0.02	0.5	7.86E-03	6.30E-04	5.49E-05	7.45E-03	5.76E-04	4.36E-05
0.05	0.01	1.06E-02	7.29E-04	4.14E-05	1.01E-02	7.19E-04	4.04E-05
0.05	0.05	6.53E-03	4.94E-04	3.44E-05	6.38E-03	4.52E-04	3.38E-05
0.05	0.5	5.80E-03	4.37E-04	3.30E-05	5.46E-03	3.83E-04	2.99E-05

Data analysis: TOPMed data sets

For each of the three TOPMed datasets that we considered, Table 2 provides the sample sizes, gene of interest, and number of variants according to sequential filtering: the number of available (non-monomorphic) variants in the sample that passed the functional filters described in the Methods section, number of variants after applying quality filters, and after restricting to those with at least 3 carriers of the rare alleles and less than 300 carriers, and the number of variants with at least 50% power to reject the null hypothesis at the 0.05 level under the assumption of odds ratio =2. There were 3 such variants in the *NOTCH3*-SVD stroke analysis, 1 variant in the *DEC2*-short sleep analysis, and 4 variants in the *F5*-VTE analysis.

Table 2: Characteristics of the TOPMed datasets and variants considered for association testing.

	SVD stroke	Short sleep	VTE
# individuals in the analysis	5,358	20,021	11,627
# cases	692 (12.9%)	2,408 (12%)	3,793 (32.6%)
# controls	4,666 (87.1%)	17,613 (88%)	7,834 (67.4%)
Gene of interest	*NOTCH3*	*DEC2/ BHLHE41*	*F5*
# potentially functional non-monomorphic variants identified	122	58	142
# variants further passing TOPMed quality filters	117	49	132
# variants further having 2<carriers <300	20	9	25
# variants with estimated power >0.5 at the 0.05 α level	3	1	4

Table 3 provides the results from testing each of the variants passing this estimated power filter. Of the three tested *NOTCH3* variants, rs115582213 had p-value=0.03. For short sleep, only a single *DC2* variant was tested, it had BinomiRare mid-p-value=0.03, suggesting association with short sleep. For the *F5* gene and VTE, none of the four tested variants showed evidence of association.

Table 3: Results from association analysis of rare genetic variants within monogenic disease genes of interest. Genetic variants presented are those that passed functional annotation and statistical power filters. For each variant we provide its BinomiRare p-value and mid-p-value, the number of carriers of the rare allele n_c, the number of carriers with the outcome n_d, the estimated power computed while assuming effect size OR=2 and p-value threshold=0.05, pathogenicity interpretation from ClinVar, CADD score, and FATHMM-XF coding score.

rsID	Variant	BinomiRare pval	BinomiRare midp	n_c	n_d	Estimated Power (OR=2)	ClinVar Interpretation	CADD PHRED	FATHMM-XF coding
SVD Stroke: *NOTCH* 3 gene									
rs115582213	chr-19-15162524-C-T	0.04	0.03	87	17	0.7	Benign/ Likely benign	25.4	0.66
rs112197217	chr-19-15179425-G-T	0.53	0.49	166	23	0.91	Benign/ Likely benign	21	0.42
rs11670799	chr-19-15188240-G-A	0.81	0.77	180	23	0.94	Benign/ Likely benign	28.8	0.68
Short sleep: *DEC2* gene									
rs121912617	chr-12-26122364-G-T	0.04	0.03	127	38	0.98	Not available	27.5	0.66
VTE: *F5* gene									
rs6026	chr-1-169528054-C-T	0.37	0.34	115	31	0.94	Benign/ Likely benign	25.7	0.75
rs6034	chr-1-169529782-G-C	1.00	0.94	46	16	0.57	Conflicting interpretations	21.3	0.56
rs78958618	chr-1-169542985-G-A	0.67	0.63	130	32	0.94	Benign	15.18	0.11
rs9332485	chr-1-169586344-C-T	0.37	0.34	222	55	1	Benign/ Likely benign	22.5	0.23

Discussion

We extended the BinomiRare test and proposed a CMP test for testing the association of a rare genetic variant with a binary outcome in the mixed model framework. These tests were specifically developed to handle variants with very low minor allele counts (tens of carriers), because it was previously shown that other tests that allow for covariates adjustment such as the naïve Score test and the SPA test do not always control the type 1 error in the very low count settings (14). Both carriers-only tests first estimate the outcome probabilities for each person in a

data set, while accounting for covariates and for genetic relatedness (and possibly other covariance matrices) via a mixed model, and then use the estimated conditional disease probabilities. For a single variant, the carriers of the rare alleles are identified, and based on their disease probabilities and the observed number of "cases", a p-value is computed, as the probability of observing the given number of cases or more extreme given the estimated outcome probabilities. The BinomiRare test with conditional probabilities performed well, while, surprisingly, the CMP test did not control the Type 1 error rate for settings with low carrier counts. This was likely because the approximations on which it relies are asymptotic in its non-centrality parameter λ, which is related to the number of carriers.

We demonstrated the application of the BinomiRare test using three TOPMed studies: of SVD stroke, short sleep, and VTE. Due to the low power for testing low-count variants, we filtered variants according to functional annotation, and according to computed statistical power. The limitation of this approach is that (1) The deleteriousness predicting annotations used and the filters applied to them may not have captured the true functional variant set; (2) the power analysis was based on an arbitrarily selected OR parameters. In this study, we chose OR=2 and only considered the handful of variants that had estimated power > 0.5 for testing while requiring p-value (α level)<0.05. We recognize that many rare variants have larger effect sizes. However, if we specified a larger OR parameter, and thus included more variants in our analysis, a more stringent α level would be needed. Thus, the resulting list of variants to test may have been similar. More work is needed developing strategies for identifying single rare variant associations.

For each of the phenotypes, SVD stroke, VTE, and short sleep, we searched for rare variants within genes with known trait associations. For SVD stroke, we considered *NOTCH3*, because some *NOTCH3* variants have been in CADASIL patients, which poses a risk for stroke. Most *NOTCH3* mutations reported as associated with CADASIL are those involving loss or gain of a cysteine residue, leading to unpaired cysteine (37). Single nucleotide variants in *NOTCH3* have not yet consistently identified as associated with SVD stroke in population-based studies. Here, we identified the rare variant rs115582213 (BinomiRare mid-p-value=0.03). This variant was rare, with 87 variant carriers out of 5,358 individuals in the dataset. Of these, 17 individuals had SVD stroke.

For VTE, we considered the *F5* gene. The *F5* gene harbors the strongest known, relatively common, genetic risk factor for VTE, the rs6025 variant (38; 39). This motivated the search for rare variants in this gene. We did not identify any variant associated with VTE at the p-value<0.05 level. We did not consider rs6025 as part of our testing strategy because it was common with MAF=0.04, and had 839 carriers of the rare allele, a setting in which other tests such as the SPA should be able to control the type 1 error well and also be more powerful. Still, as a positive control we tested its association with VTE using BinomiRare, and the p-value was 1.5×10^{-14}.

Short sleep has been consistently associated with cardiovascular and cardiometabolic disease (40; 41). Genetic determinant of short sleep may help elucidate this connection (42). We considered the *DEC2/BHLHE4* gene, which a mutation with a know familial aggregation associated with short sleep. Our filtering strategy resulted in a single variant considered for

testing: rs121912617, the known short sleep mutation (27). In our data, it was associated with short sleep with BinomiRare mid-p-value=0.03. Rs121912617 is substantially more common (yet is still rare) in African Americans compared to European Americans (0.01 MAF in African Americans from the TOPMed short sleep datasets, compared to MAF < 0.001 in European Americans from the same dataset), allowing for observing this association in a population-based, rather than a family-based, study.

Here, we demonstrated the BinomiRare test for testing single-variant associations in data with known or cryptic relatedness. It can also be used to test sets of rare variants, by defining a carrier as an individual with at least one rare allele in the variant set. It is a topic of future research to extend this framework to use the counts of the rare variant allele and increase power.

Acknowledgements

contract HHSN2682018000011). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Author contributions

References

1.   Taliun DaH, Daniel N and Kessler, Michael D and Carlson, Jedidiah and Szpiech, Zachary A and Torres, Raul and Taliun, Sarah A Gagliano and Corvelo, André and Gogarten, Stephanie M and Kang, Hyun Min. 2019. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*:563866

2.   Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, et al. 2020. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586:749-56

3.   Amininejad L, Charloteaux B, Theatre E, Liefferinckx C, Dmitrieva J, et al. 2018. Analysis of Genes Associated With Monogenic Primary Immunodeficiency Identifies Rare Variants in XIAP in Patients With Crohn's Disease. *Gastroenterology* 154:2165-77

4.   Wright CF, West B, Tuke M, Jones SE, Patel K, et al. 2019. Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am J Hum Genet* 104:275-86

5.   Tuijnenburg P, Lango Allen H, Burns SO, Greene D, Jansen MH, et al. 2018. Loss-of-function nuclear factor kappaB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *J Allergy Clin Immunol* 142:1285-96

6.   Do R, Stitziel NO, Won HH, Jorgensen AB, Duga S, et al. 2015. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 518:102-6

7.   Kendall KM, Bracher-Smith M, Fitzpatrick H, Lynham A, Rees E, et al. 2019. Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br J Psychiatry* 214:297-304

8.   Ma C, Blackwell T, Boehnke M, Scott LJ, Go TDi. 2013. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* 37:539-50

9.   Sofer T. 2017. BinomiRare: A robust test of the association of a rare variant with a disease for pooled analysis and meta-analysis, with application to the HCHS/SOL. *Genet Epidemiol* 41:388-95

10. Sondhi A, Rice KM. 2018. Fast permutation tests and related methods, for association between rare variants and binary outcomes. *Ann Hum Genet* 82:93-101

11. Greene D, BioResource N, Richardson S, Turro E. 2017. A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases. *Am J Hum Genet* 101:104-14

12. Dey R, Schmidt EM, Abecasis GR, Lee S. 2017. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* 101:37-49

13. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, et al. 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 50:1335-41

14. Sofer T, Guo N. 2020. Rare variants association testing for a binary outcome when pooling individual level data from heterogeneous studies. *Genetic Epidemiology*:1–13

15. Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, et al. 2016. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am J Hum Genet* 98:165-84

16. Conomos MP, Gogarten SM, Brown L, Chen CH, Rice K, et al. 2018. GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. p. R package

17. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, et al. 2016. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* 98:653-66

18. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54:127-42

19. Kadane JB. 2016. Sums of Possibly Associated Bernoulli Variables: The Conway-Maxwell-Binomial Distribution. *Bayesian Anal.* 11:403--20

20. Graffelman J, Moreno V. 2013. The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat Appl Genet Mol Biol* 12:433-48

21. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, et al. 2019. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*

22. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, et al. 2019. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *In Revision to Nature*

23. Hays J, Hunt JR, Hubbell FA, Anderson GL, Limacher M, et al. 2003. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol* 13:S18-77

24. Adams HP, Bendixen BH, Kappelle LJ, Biller J, Love BB, et al. 1993. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* 24:35-41

25. Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. 2011. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *European Heart Journal* 32:1484-92

26. Ross OA, Soto-Ortolaza AI, Heckman MG, Verbeeck C, Serie DJ, et al. 2013. NOTCH3 variants and risk of ischemic stroke. *PLoS One* 8:e75035

27. He Y, Jones CR, Fujiki N, Xu Y, Guo B, et al. 2009. The transcriptional repressor DEC2 regulates sleep length in mammals. *Science* 325:866-70

28. Hirano A, Hsu PK, Zhang L, Xing L, McMahon T, et al. 2018. DEC2 modulates orexin expression and regulates sleep. *Proc Natl Acad Sci U S A* 115:3434-9

29. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, et al. 1994. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 369:64-7

30. Dahlback B, Hildebrand B. 1994. Inherited resistance to activated protein C is corrected by anticoagulant cofactor activity found to be a property of factor V. *Proc Natl Acad Sci U S A* 91:1396-400

31. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, et al. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* 17:122

32. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc* 11:1-9

33. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7:248-9

34. Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. *Genome Res* 19:1553-61

35. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. 2018. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* 34:511-3

36. 'National Heart L, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services'. 2020. The NHLBI BioData Catalyst

37. Coupland K, Lendahl U, Karlström H. 2018. Role of NOTCH3 Mutations in the Cerebral Small Vessel Disease Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy. *Stroke* 49:2793-800

38. HEIT JA, ARMASU SM, ASMANN YW, CUNNINGHAM JM, MATSUMOTO ME, et al. 2012. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *Journal of Thrombosis and Haemostasis* 10:1521-31

39. Lindstrom S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, et al. 2019. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood* 134:1645-57

40. Hoevenaar-Blom MP, Spijkerman AMW, Kromhout D, van den Berg JF, Verschuren WMM. 2011. Sleep Duration and Sleep Quality in Relation to 12-Year Cardiovascular Disease Incidence: The MORGEN Study. *Sleep* 34:1487-92

41. Sabanayagam C, Shankar A. 2010. Sleep Duration and Cardiovascular Disease: Results from the National Health Interview Survey. *Sleep* 33:1037-42

42. Dashti HS, Jones SE, Wood AR, Lane JM, van Hees VT, et al. 2019. Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nature Communications* 10:1100

43. Daly F, Gaunt RE. 2016. The Conway-Maxwell-Poisson distribution: Distributional theory and approximation. *ALEA* 13:635—58

44. Sellers K, Lotze T, Raim A. 2018. COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression. R package version 0.6.1. https://CRAN.R-project.org/package=COMPoissonReg

Appendix

The CMP test

Let $W = \sum_{i=1}^{m} D_i$, for $D_i \sim Binom(p_i, 1)$ be a random variable with the CMP-Binomial

probability function. When m increases, this distribution is approximated by the CMP

distribution (Theorem 4.1. in Daly and Gaunt (43)) so that $W \sim CMP(\lambda, \nu)$. Consider

proposition 2 in Kadane (19) stating:

Proposition 2 (Kadane, 2016): *Suppose $D_1, \ldots, D_m$ take values on $\{0,1\}$. Let $P(W = k) = \widetilde{p_k} \geq$ 0, where $\sum_{k=0}^{m} \widetilde{p_k} = 1$. Then there exists a unique distribution on $D_1, \ldots, D_m$ such that $D_1, \ldots, D_m$ are exchangeable of order m, and $\sum_{i=1}^{m} D_i$ has the same distribution as W.*

According to this proposition, an arbitrary sum of binary variables is distributed as a sum of exchangeable binary variables, where the exchangeable variables are such that there is a unique combination of probability parameter $p = \Pr(D_1 = 1) = \cdots = \Pr(D_m = 1)$ and a parameter ρ modeling the dependency between each pair $D_i, D_j, i \neq j$. Therefore, two parameters suffice to characterize the distribution of an arbitrary sum of binary variables. Specifically, for a given set of carriers of a rare genetic variant, the sum of their disease statuses

$$W_{n_c} = \sum_{i=1}^{n_c} D_i$$

is distributed like a unique sum of exchangeable binary variables. Based on the estimated disease probabilities, we estimate the two parameters (different than the probability and dependency parameters p and ρ above) of the CMP distribution to obtain an estimated probability function in a variation of a method-of-moment approach that is based on estimated probabilities, rather than on the observed data. Daly and Gaunt (43) provided an approximation to the CMP distribution:

Proposition 2.3. (Daly and Gaunt): *Let $W \sim CMP(\lambda, \nu)$. Then, for $k \in \mathbb{N}$,*

$$E[W^k] \sim \lambda^{\frac{k}{\nu}}\left[1 + O\left(\lambda^{-\frac{1}{\nu}}\right)\right],$$

as $\lambda \to \infty$.

Assuming that $\lambda^{-1/\nu}$ is small (which, as we shall see, is true when λ is very large, because ν tends to be well bounded), we get that, approximately:

$$E[W] \approx \lambda^{1/\nu}. \tag{2}$$

Daly and Gaunt also showed, in their equation 2.4 and based on the result in Shmueli et al (18),

that:

$$\mathrm{Var}(W) \approx \frac{1}{\nu}\lambda^{\frac{1}{\nu}} + O(1), \ \text{as} \ \lambda \to \infty. \tag{3}$$

Therefore, noting that $\mathrm{Var}(W) = E[W^2] - (E[W])^2$, once we estimate $E[W]$ and $E[W^2]$, we

use (4) and (5) to obtain estimators of λ and ν by:

$$\hat{\nu} = \frac{\widehat{E[W]}}{\widehat{E[W^2]} - \left(\widehat{E[W]}\right)^2} = \frac{\widehat{E[W]}}{\widehat{Var}(W)} \tag{4}$$

$$\hat{\lambda} = \left(\widehat{E[W]}\right)^{\hat{\nu}} \tag{5}$$

Estimating parameters of the CMP distribution from estimated diseased probabilities. We

consider two approaches to estimate components of $\hat{\lambda}$ and $\hat{\nu}$, i.e. $E[W]$, $E[W^2]$, and $\mathrm{Var}(W)$: an

analytic approach, and a sampling-based approach. In the analytic approach, we compute

$\widehat{E[W]} = \sum_{i=1}^{n_c} \hat{p}_i$, and $\widehat{Var}[W] = \sum_{i=1}^{n_c} \hat{p}_i (1 - \hat{p}_i)$. In the sampling-based approach, we generate

random variables $\widetilde{W}$ with the same distribution as W_{n_c} (the sum of disease statuses among the n_c

carriers of a genetic variant), and treat them as observed data to estimate the desired quantities.

More specifically, let $\widetilde{D}_{i,s} \sim Binom(\hat{p}_i)$ be the sampled disease status of the ith individual in the

$s = 1, \dots, S$ sample. Then:

$$\widetilde{W}_s = \sum_{i=1}^{n_c} \widetilde{D}_{i,s} \tag{6}$$

and we estimate:

$$\widehat{E[W]} = \frac{1}{S}\sum_{s=1}^{S} \widetilde{W}_s$$

$$E\widehat{[W^2]} = \frac{1}{S}\sum_{s=1}^{S}\widetilde{W}_s^2$$

To summarize, to calculate the p-value and the mid-p-value, formally given by:

$$p\text{-value} = \widehat{\Pr}(W = n_d) + \sum_{k=1}^{n_c}\widehat{\Pr}(W = k) \times 1[\widehat{\Pr}(W = k) < \widehat{\Pr}(W = n_d)] \quad (9)$$

$$\text{mid-}p\text{-value} = \frac{\widehat{\Pr}(W = n_d)}{2}$$

$$+ \sum_{k=1}^{n_c}\widehat{\Pr}(W = k) \times 1[\widehat{\Pr}(W = k) < \widehat{\Pr}(W = n_d)] \quad (10)$$

we estimate probabilities for each potential number of disease carriers, in the following process:

1. Obtain individual disease probability estimates $\widehat{p_1}, \dots, \widehat{p_{n_c}}$ via standard approaches (e.g. logistic mixed model).

2. Compute estimates $\widehat{E[W]}$ and $\widehat{Var[W]}$ in the analytic approach, or compute $\widehat{E[W]}$ and $\widehat{E[W^2]}$ in the sampling approach.

3. Compute estimates $\hat{\lambda}, \hat{v}$ using (6) and (7).

4. Compute $\widehat{\Pr}(W = k)$ for $k = 1, \dots, n_c$ using the R package COMPoissonReg (44).

Figures

Figure 1: Step 1 of testing genetic association using the carriers-only tests framework. A "null model" of association between the binary outcomes and covariates of interest is fitted, accounting for genetic relationship. Then, estimated conditional outcome probabilities are extracted to be used in the testing step.
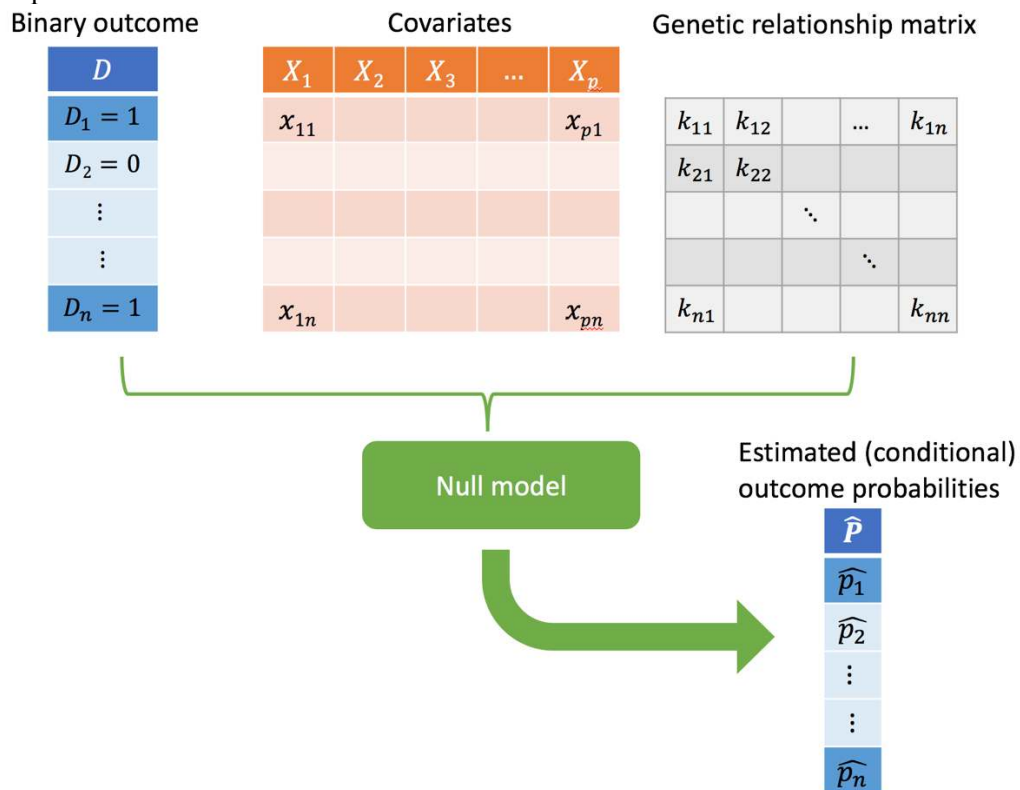
Figure 2: Step 2 of testing genetic associations using the carriers-only tests framework. Based on estimated outcome probabilities, variants are inspected one at a time. For a given variant, carriers of the rare allele are identified, and a test of the null hypothesis $H_0: P_{n_c} = \widehat{P_{n_c}}$ is performed testing whether n_d is consistent with the outcome probabilities within the carriers, based on the null model.