Association Analysis and Meta-Analysis of Multi-Allelic Variants for Large-Scale Sequence Data

Yu Jiang [1,†], Sai Chen [2,†], Xingyan Wang [1,†], Mengzhen Liu [3], William G. Iacono [4], John K. Hewitt [5], John E. Hokanson [6], Kenneth Krauter [5], Markku Laakso [7], Kevin W. Li [8], Sharon M. Lutz [9], Matthew McGue [3], Anita Pandit [8], Gregory J. M. Zajac [8], Michael Boehnke [8], Goncalo R. Abecasis [8], Scott I. Vrieze [3], Bibo Jiang [1,*], Xiaowei Zhan [10,*] and Dajiang J. Liu [1]

[1] Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA; ybj5037@psu.edu (Y.J.); xinyan.wang@psu.edu (X.W.); dajiang.liu@psu.edu (D.J.L.)

[2] Illumina Inc., 5200 Illuminay Way, San Diego, CA 92122, USA; saichen@umich.edu

[3] Department of Psychology, University of Minnesota, Minneapolis, MN 55454, USA; liu00282@umn.edu (M.L.); mcgue001@umn.edu (M.M.); vrieze@umn.edu (S.I.V.)

[4] Department of Psychiatry, University of Minnesota, Minneapolis, MN 55454, USA; wiacono@umn.edu

[5] Institute for Behavioral Genetics, University of Colorado Boulder, Aurora, CO 80045, USA; hewitt@ibg.colorado.edu (J.K.H.); krauter@colorado.edu (K.K.)

[6] Department of Epidemiology, School of Public Health, University of Colorado Denver, Aurora, CO 80045, USA; John.Hokanson@ucdenver.edu

[7] Department of Medicine, University of Eastern Finland and Kuopio University Hospital, 70211 Kuopio, Finland; Markku.Laakso@kuh.fi

[8] Center of Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; kevinwli@umich.edu (K.W.L.); anitapan@umich.edu (A.P.); gzajac@umich.edu (G.J.M.Z.); boehnke@umich.edu (M.B.); goncalo@umich.edu (G.R.A.)

[9] Department of Biostatistics and Informatics, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA; sharon.lutz@ucdenver.edu

[10] Department of Clinical Science, Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

* Correspondence: bjiang@phs.psu.edu (B.J.); xiaowei.zhan@utsouthwestern.edu (X.Z.)

† These authors contributed equally to this work.

Abstract: There is great interest in understanding the impact of rare variants in human diseases using large sequence datasets. In deep sequence datasets of >10,000 samples, ~10% of the variant sites are observed to be multi-allelic. Many of the multi-allelic variants have been shown to be functional and disease-relevant. Proper analysis of multi-allelic variants is critical to the success of a sequencing study, but existing methods do not properly handle multi-allelic variants and can produce highly misleading association results. We discuss practical issues and methods to encode multi-allelic sites, conduct single-variant and gene-level association analyses, and perform meta-analysis for multi-allelic variants. We evaluated these methods through extensive simulations and the study of a large meta-analysis of ~18,000 samples on the cigarettes-per-day phenotype. We showed that our joint modeling approach provided an unbiased estimate of genetic effects, greatly improved the power of single-variant association tests among methods that can properly estimate allele effects, and enhanced gene-level tests over existing approaches. Software packages implementing these methods are available online.

Keywords: multi-allelic variants; GWAS; meta-analysis; smoking

1. Introduction

Rare genetic variants are enriched with functional alleles that play an important role in a variety of complex human diseases, including hematological disorders [1], coronary artery disease [2–4], and others. The discovery of such rare-variant associations has contributed significantly to the generation of new mechanistic insights and the identification of novel therapeutic targets [4,5]. These discoveries are critical steps toward the successful implementation of precision medicine.

As the cost of sequencing continues to decrease, many sequence-based studies of rare variants have begun to emerge. Compared to array-based studies which only genotype variants at known sites, sequence-based studies unbiasedly reveal both known and novel variants across the frequency spectrum. The fraction of novel alleles/variants uncovered increases with increasing read depth and sample size. In addition to identifying novel variant sites, numerous novel alleles at known variant sites are being uncovered as well. As shown in the exome aggregation consortium (ExAC) [6], 8% of the variant sites in the human exome are multi-allelic and contain more than one alternative allele. A number of these multi-allelic variants are functional and have been shown to be disease relevant [6]. Despite the importance of multi-allelic variants, most of the methods developed so far for sequence-based association analysis consider only bi-allelic variants, and thus do not properly handle multi-allelic sites [7,8].

Interestingly, despite often being ignored in genome-wide association studies (GWAS), multi-allelic variants were well-studied prior to the GWAS era for microsatellite markers. However, the existing methods all have certain limitations, which make it challenging to analyze sequence data. Some methods focused on how to combine multiple alleles in the same position and perform an omnibus test [9]. Another method [10] made use of retrospective likelihood to model the joint distribution of multi-allelic variants at a single variant site, yet it is challenging to extend this model to multiple variant sites in linkage disequilibrium, and it is difficult to generalize this approach to analyze gene-level associations.

Moreover, it is unclear how to perform meta-analysis and combine samples across studies in the presence of multi-allelic variants. In addition, most of the identified rare-variant associations have small to moderate effect sizes [11]. There is growing recognition that large sample sizes are needed to attain sufficient power to uncover rare causal variants. Consortia efforts are underway to aggregate large sample sizes for the study of various complex human diseases. Meta-analysis plays a critical role in the vast majority of consortium efforts, where typically only summary-level information such as genetic effects and *p*-values are shared across different studies. Compared to sharing individual-level genotype and phenotype data from study participants, meta-analysis of summary statistics can be easier to implement, more protective of study participant privacy, and more robust against heterogeneity between studies [12]. It is therefore also necessary to extend existing meta-analysis methods and software to properly handle multi-allelic sites as well.

In this article, we discuss practical issues and proper methods to address the key analysis issues for multi-allelic variants, which represent 10% of the genomic variations. We developed novel methods to jointly model the effects of multiple alleles in single-variant association tests and facilitate convenient gene-level association analysis and meta-analysis. We evaluated these methods using extensive and realistic simulations and show that they consistently outperform existing naïve methods that either ignore multi-allelic sites or test each alternative allele separately. We also applied these methods to a large-scale meta-analysis of nicotine addiction phenotypes. We show that our method can uncover multi-allelic association in known loci of the cigarettes-per-day (CPD) phenotype. We have implemented these methods in RVTESTS [13] for association analysis and the generation of summary association statistics and RAREMETAL [14] for meta-analysis. Given the importance of the multi-allelic variants, we expect our methods and software to play a critical role in large-scale genetic discoveries with sequence data.

2. Materials and Methods

We describe methods and practical issues to encode multi-allelic variants, perform single-variant and gene-level analyses, and carry out meta-analysis. We advocate the estimation of effects of each allele relative to the reference allele in order to make the results interpretable and comparable across studies. The key idea is to jointly model the effects of multiple alternative alleles for multi-allelic variants in single-variant and gene-level tests. This joint modeling strategy gives a proper estimate of the alternative allele effect and facilitates the construction of gene-level tests from single-variant association test statistics of multi-allelic sites. This method improves power over methods that can unbiasedly estimate alternative allele effects, including the method that ignores multi-allelic variants and the method that models the effect of each alternative allele separately.

For a multi-allelic variant at site m with L alternative alleles, we can encode the genotype for individual i with an L-vector $G_{im} = \left(G_{im}^1, G_{im}^2, \ldots, G_{im}^L\right)$, where the lth entry is the number of the lth alternative allele. Assuming Hardy-Weinberg equilibrium, the counts $(2 - \sum_l G_{im}^l, G_{im}^1, \ldots, G_{im}^L)$ follow a multinomial distribution $Multinom(2, ((1 - \sum_l f_l), f_1, \ldots, f_L))$, where f_l is the alternative allele frequency for the lth alternative allele, and $1 - \sum_l f_l$ is the frequency for the reference allele. The counts for two different alternative alleles $G_{im}^l, G_{im}^{l'}$ are negatively correlated with covariance $(G_{im}^l, G_{im}^{l'}) = -2f_l f_{l'}$. The correlation between the two alternative alleles A_l and $A_{l'}$ can be large when they are common. We have illustrated this genotype coding with an example of a tri-allelic site in Table 1.

Table 1. Genotype coding for multi-allelic variant with two different alternative alleles.

Genotypes	Genotype Coding
A_0/A_0	(0,0)
A_0/A_1	(1,0)
A_1/A_1	(2,0)
A_1/A_2	(1,1)
A_2/A_2	(0,2)
A_0/A_2	(0,1)

When there are genotype uncertainties in the data, genotype dosages are often used instead of hard genotype calls for genetic association analyses [15–17]. Under our coding scheme, the calculation of the genotype dosages is similar to bi-allelic variants.

2.1. Joint Modeling Multi-Allelic Effects

We are interested in estimating and testing for the effect of each alternative allele A_l, $l = 1, \ldots, L$ relative to the reference allele. The effect of allele A_l measures the mean phenotype change for an additional copy of the A_l allele compared to the reference allele.

To properly analyze a multi-allelic variant, we propose a joint model that includes the genotypes for all alternative alleles in the model. Specifically, to estimate (or test for) the effect of the lth alternative allele, we perform the multiple regression $Y_i = \alpha + \beta_l G_{im}^l + \sum_{l' \neq l} \beta_{l'} G_{im}^{l'} + \epsilon_i$. The multiple regression coefficient β_l estimates $E\left(Y_i \middle| G_{im}^l = 1, G_{im}^{-l}\right) - E\left(Y \middle| G_{im}^l = 0, G_{im}^{-l}\right)$ where G_{im}^{-l} is the genotype vector at site m for the rest of the alleles $A_1, \ldots, A_{l-1}, A_{l+1}, \ldots, A_L$. The effect of the lth alternative allele can be unbiasedly estimated from multiple regression.

An alternative strategy, which we call single-allelic analysis, is to restrict our analysis to the set of individuals with genotypes $A_0/A_0, A_0/A_l, A_l/A_l$. As the analyzed samples are selected based on genotype only, the regression analysis is still valid and will give us an unbiased estimate of the effect

of A_l. However, depending on the frequency of other alternative alleles, the single-allelic analysis may discard a significant portion of the sample and the association analysis can be underpowered.

An additional advantage of joint multi-allelic analysis over single-allelic analysis is the convenience of constructing gene-level tests from single-variant association statistics. For single-allelic analysis, a different set of samples are analyzed for each different alternative allele. This makes it impossible to construct gene-level tests using single-variant association statistics calculated for different samples.

Finally, it is important to note that, when regressing Y over the count of only one allele (i.e., $Y_i \sim G_{im}^l \widetilde{\beta_l}$), the regression coefficient $\widetilde{\beta_l}$ measures the difference in phenotypic means between individuals carrying one copy of A_l and the baseline group of individuals that carry no copy of A_l. This is different from the estimates obtained by regressing Y over the multi-allele genotypes, i.e., $Y_i = \alpha + \beta_l G_{im}^l + \sum_{l' \neq l} \beta_{l'} G_{im}^{l'} + \epsilon_i$. For example, consider a tri-allelic site with reference allele A_0 and alternative alleles A_1, A_2. Regressing Y over the counts of allele A_1, the regression coefficient is equal to the mean phenotype difference between individuals that carry one copy of A_1 (i.e., with genotypes A_0A_1, A_1A_2), and the baseline group of individuals that carry no A_1 alleles (i.e., with genotypes A_0A_0, A_0A_2, A_2A_2). Similarly, regressing the phenotype over the allele count of A_2 will estimate the mean phenotype difference between individuals that carry one copy of A_2 (i.e., with genotypes A_0A_2, A_1A_2), and the baseline group of individuals with no A_2 alleles, (i.e., with genotypes A_0A_0, A_0A_1, A_1A_1). It would be difficult to interpret the effect $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$ or combine them in gene-level association tests, as the baseline group in two regressions differ. Also, as different studies may contain different sets of alleles (particularly when some alternative alleles are rare), estimating allele effect without a fixed baseline group will make it difficult to perform meta-analysis using estimated effects across studies. A numerical example is given in the Supplemental Methods and Figure S1 to illustrate the considerable bias for the approach that regresses phenotypes over the counts of each allele separately.

2.2. Meta-Analysis of Single-Variant Test in the Presence of Multi-Allelic Sites

We propose appropriate meta-analysis methods of single-variant and gene-level association tests in the presence of multi-allelic sites. We denote the sample genotype matrix at the multi-allelic site m as $\mathbf{G_m}$. We will calculate and share the marginal association statistic obtained from the regression analysis over the counts of each alternative allele, i.e., $Y_i = \alpha + \beta_l G_{im}^l + Z_i \gamma + \epsilon$, where Z_i is the vector of covariate for individual i.

For sequence data, score statistics are often computed, and meta-analysis, which is evaluated under the null hypothesis, was used. Compared to methods that require maximizing the likelihood under the alternative hypothesis, score statistics are faster to compute and numerically more stable. The score statistic for the lth allele is equal to $U_{G_m^l} = \frac{1}{\hat{\sigma}^2} \sum_i G_{im}^l (Y_i - \hat{Y}_i)$, where $\hat{Y}_i = \hat{\alpha} + Z_i \hat{\gamma}$. The parameters $\hat{\alpha}$ and $\hat{\gamma}$ are estimates for α and γ and $\hat{\sigma}^2$ is the residual variance under the null hypothesis. The variance–covariance matrix between the score statistics for different alleles is given by:

$$\mathbf{V_{G_m G_m}} = 1/\hat{\sigma}^2 \left(\mathbf{G_m^T G_m} - \mathbf{G_m^T Z} \left(\mathbf{Z^T Z} \right)^{-1} \mathbf{Z^T G_m} \right) \tag{1}$$

To estimate and test for the effect of the lth alternative allele against the reference allele, we need to control for the effects of the rest of the $L - 1$ alternative alleles. Specifically, in a regression model that includes the counts of all alternative alleles, i.e.,

$$Y_i = \alpha + G_{im}^l \beta_l + \sum_{l' \neq l} G_{im}^{l'} \beta_{l'} + \epsilon_i,$$

the conditional score statistic for the *l*th alternative allele is equal to $\mathbf{U}_{G_m^l|G_m^{-l}} = \sum_i G_{im}^l (Y_i - \hat{Y}_i)$, where

$\hat{Y}_i = \sum_{l' \neq l} G_{im}^{l'} \hat{\beta}_{l'}$ and $\hat{\boldsymbol{\beta}}_{-l} = (\hat{\beta}_1, \ldots, \hat{\beta}_{l-1}, \hat{\beta}_{l+1}, \ldots, \hat{\beta}_L) = \mathbf{V}_{G_m^{-l}G_m^{-l}}^{-1} \mathbf{U}_{G_m^{-l}}$.

Throughout the manuscript, the matrix inverse is implemented with the generalized inverse. The alternative allele counts for each multi-allelic site are almost surely non-colinear, as the different alternative alleles are inherited independently from the parental generation. In the very rare cases where the co-linearity occurs by chance, the generalized inverse can still be calculated, and the test statistic would have very large variance, yield insignificant results, and would not result in false positive signals.

The conditional score statistic can be calculated using marginal association statistics:

$$\mathbf{U}_{G_m^l|G_m^{-l}} = \mathbf{U}_{G_m^l} - \mathbf{V}_{G_m^l G_m^{-l}} \mathbf{V}_{G_m^{-l}G_m^{-l}}^{-1} \mathbf{U}_{G_m^{-l}}. \tag{2}$$

The variance of the conditional score statistic is equal to:

$$\mathbf{V}_{G_m^l|G_m^{-l}} = \left(\mathbf{V}_{G_m^l G_m^l} - \mathbf{V}_{G_m^l G_m^{-l}} \mathbf{V}_{G_m^{-l}G_m^{-l}}^{-1} \mathbf{V}_{G_m^{-l}G_m^l} \right) \hat{\sigma}_{Y|G_m^{-l}}^2$$

In meta-analysis, the score statistics will be combined using the Mantel–Haenszel method. This well approximates the commonly used inverse variance meta-analysis method based upon maximum likelihood estimates of the genetic effects when the genetic effects are small [18]. Specifically, given the score statistics at site *m* (i.e., $U_{1,G_m^l|G_m^{-l}}, \ldots, U_{K,G_m^l|G_m^{-l}}$) and their variances in *K* studies (i.e., $\mathbf{V}_{1,G_m^l|G_m^{-l}}, \ldots, \mathbf{V}_{K,G_m^l|G_m^{-l}}$), the meta-analysis score statistic can be calculated using $\mathbf{U}_{META,G_m^l|G_m^{-l}} = \sum_k \mathbf{U}_{k,G_m^l|G_m^{-l}}$, and $\mathbf{V}_{META,G_m^l|G_m^{-l}} = \sum_k \mathbf{V}_{k,G_m^l|G_m^{-l}}$.

The standardized score statistic is equal to $T_{META,G_m^l} = \frac{\mathbf{U}_{META,G_m^l|G_m^{-l}}^2}{\mathbf{V}_{META,G_m^l|G_m^{-l}}}$, which follows a chi-square distribution with 1 degree of freedom.

2.3. Meta-Analysis of Gene-Level Association Test in the Presence of Multi-Allelic Sites

As we showed for single-variant analysis, it is necessary to jointly model the effects of all alternative alleles in the same site in order to attain unbiased association analysis of each allele. Most commonly used gene-level association tests, such as the burden test, SKAT, and VT, can be constructed using single-variant association statistics and their covariance matrices [19]. When the gene region contains rare alternative alleles from multi-allelic sites, the score statistic from joint multi-allelic analysis (i.e., $\mathbf{U}_{G_m^l|G_m^{-l}}$) needs to be used to construct a gene-level test. As in single-variant analysis, using the marginal score statistic $U_{G_m^l}$ without adjusting the effects of other alternative alleles leads to estimates that are difficult to interpret.

Below, we describe an extension of gene-level tests to scenarios where the gene region contains multi-allelic sites. The calculation of gene-level tests requires score statistics from variant sites that contain rare alleles, including the score statistics from bi-allelic sites and the score statistic from joint multi-allelic analysis, as well as the covariance matrix between them. Single-variant association statistics from bi-allelic and multi-allelic sites have been described in the above section. We next derive the variance–covariance matrix between these score statistics and then discuss how to use them to construct commonly used rare-variant tests.

For notational convenience, we denote the genotype matrices for common alternative allele genotypes from multi-allelic sites as $\mathbf{G_C}$, the rare allele genotypes from the multi-allelic sites as $\mathbf{G_R}$, and the rare alleles from bi-allelic sites as $\mathbf{G_B}$. We denote the vector of score statistics for all rare alleles as $\mathbf{U_{GENE}} = \left(\mathbf{U}_{G_B|G_C}, \mathbf{U}_{G_R|G_C} \right)$, which includes the score statistics from bi-allelic sites (conditional on the common alternative alleles from multi-allelic sites) and the score statistics from joint multi-allelic analysis.

Below, we illustrate how to calculate the covariance matrix between score statistics. The covariance matrix between score statistics of rare alleles at multi-allelic sites is equal to:

$$\mathbf{V_{G_R G_R}} = \frac{1}{\hat{\sigma}^2}\left[\mathbf{G_R^T G_R} - \mathbf{G_R^T G_C}\left(\mathbf{G_C^T G_C}\right)^{-1}\mathbf{G_R^T G_C}\right]$$

The covariance between rare bi-allelic variants is equal to:

$$\mathbf{V_{G_B G_B}} = \frac{1}{\hat{\sigma}^2}\left[\mathbf{G_B^T G_B} - \mathbf{G_B^T G_C}\left(\mathbf{G_C^T G_C}\right)^{-1}\mathbf{G_C^T G_B}\right]$$

The covariance matrix between rare bi-allelic variants and rare multi-allelic variants is equal to:

$$\mathbf{V_{G_B G_R}} = \frac{1}{\hat{\sigma}^2}\left[\mathbf{G_B^T G_R} - \mathbf{G_B^T G_C}\left(\mathbf{G_C^T G_C}\right)^{-1}\mathbf{G_C^T G_R}\right]$$

When non-genetic covariates $\mathbf{Z}$ are present, we just need to replace $\mathbf{G_C}$ with $\tilde{\mathbf{G}}_{\mathbf{C}} = (\mathbf{G_C}, \mathbf{Z})$, and the calculation of covariance matrix remains the same.

The covariance matrix for the score statistic is denoted by $\mathbf{V_{GENE}} = \begin{pmatrix} \mathbf{V_{G_R G_R}} & \mathbf{V_{G_R G_B}} \\ \mathbf{V_{G_B G_R}} & \mathbf{V_{G_B G_B}} \end{pmatrix}$. In the implementation of RAREMETAL [14,20] and RVTESTS [13], the covariance matrix is generated and shared between studies. When the covariance matrix is not shared, they can be approximated using a reference panel [21,22].

The burden test statistic [23] and its variance are equal to $U_{BURDEN} = \mathbf{w^T U_{GENE}}$ and $V_{BURDEN} = \mathbf{w^T V_{GENE} w}$, where $\mathbf{w}$ is the weight assigned to each variant. The standardized burden statistic satisfies $T_{BURDEN} = \frac{U_{BURDEN}^2}{V_{BURDEN}} \sim \chi^2_{df=1}$. The SKAT statistic [24] is equal to $Q_{SKAT} = \mathbf{U_{GENE}^T \Omega\, U_{GENE}}$, where $\mathbf{\Omega}$ is a diagonal matrix with the diagonal entries being the weights assigned to each variant site. The SKAT statistic follows a mixture chi-square distribution with mixture proportions being the eigenvalues for $\mathbf{V_{GENE}^{1/2} \Omega V_{GENE}^{1/2}}$. The VT statistic calculates a burden test statistic for each minor allele frequency threshold and corrects for the multiple comparison issue using the minimal *p*-value method [25,26]. The *p*-values can be calculated using the distribution function for a multivariate normal distribution.

2.4. Design of Simulation Evaluation

We conducted extensive simulations to evaluate the proposed methods. To generate genetic data with realistic patterns of multi-allelic sites, we used the allele frequency spectra estimated from large-scale exome sequencing projects. We downloaded data from the ExAC project (version 0.3.1), which consists of summary information for coding variants from 60,706 exomes.

To benchmark single-variant association tests, in each replicate, we randomly picked one variant site from 219,680 sites that contain multiple alternative alleles. To illustrate the advantage of joint multi-allelic analysis, we separately considered the power for detecting associations with the primary alternative allele (i.e., the most frequent alternative allele) and the secondary alternative allele (i.e., the less frequent allele(s)). We simulated the genotype (i.e., the reference and alternative allele counts) for each sample based on a multinomial distribution: $multinom(2, (1 - \sum_j f_j, f_1, \ldots, f_L))$. For each variant site, we randomly chose one alternative allele as causal with effects being 0.1, 0.25, or 0.5 sd. The power for detecting associations with the primary (or secondary) alternative allele was assessed by the fraction of the replicates with significant *p*-values ($<5 \times 10^{-8}$) among the replicates where the primary (or secondary) alternative allele is causal.

We also assessed the power for single allelic and joint-multi-allelic analysis as omnibus tests to identify associated variant sites (instead of identifying associated alleles). We compared it with the off-the-shelf method of collapsing the multiple alternative alleles.

In order to evaluate the gene-level association test under the most realistic patterns of linkage disequilibrium and multi-allelic variant allele frequency spectrums, we made use of real genotype data

from eight cohorts, including the Minnesota Center for Twin and Family Research (MCTFR), SardiNIA, METabolic Syndrome In Men (METSIM), Genes for Good, COPDGene with samples of European ancestry, COPDGene with samples of African American ancestry, and the Center for Antisocial Drug Dependence (CADD). Based upon the real genotype, we simulated phenotypes: for each replicate, we randomly chose one gene with at least one multi-allelic variant. We chose a fraction (20% or 50%) of the genetic variants as causal, with effects simulated from $N(0, 0.2^2)$. We considered three commonly used tests, including the simple burden test, SKAT, and VT. The type I error and power for the meta-analyses were evaluated under $\alpha = 2.5 \times 10^{-6}$.

2.5. Analysis of Cigarettes-Per-Day Phenotype (CPD)

In order to benchmark our method and its implementation, we applied our method to perform a meta-analysis on large genetic datasets from eight cohorts for the CPD phenotype. The eight cohorts included the Minnesota Center for Twin and Family Research (MCTFR), SardiNIA, METabolic Syndrome In Men (METSIM), Genes for Good, COPDGene with samples of European ancestry, COPDGene with samples of African American ancestry, and the Center for Antisocial Drug Dependence (CADD). Summary association statistics from the eight cohorts were generated using RVTESTS [13], and meta-analysis was performed centrally using RAREMETAL [14]. Detailed descriptions of the cohorts are available in Supplemental Methods Section 2, including information on the methods for association analyses and the adjusted covariates.

In order to ensure the validity of our association analysis results, we conducted extensive quality control for the imputed genotype data. We filtered out variant sites with the imputation quality metric $R^2 < 0.7$, and removed variant sites that showed large differences in allele frequencies from the reference panel. We performed single-variant tests using joint multi-allelic analysis and single-allelic analysis. We also performed gene-level tests using the burden test, SKAT, and VT under two different allele frequency cutoffs, 1% and 5%. As a comparison, we analyzed the data using the method that discards the multi-allelic sites as well.

3. Results

3.1. Type I Error and Power Evaluation for Single-Variant Association Test

Simulations indicated that jointly modeling the allelic effects of multiple alternative alleles leads to more powerful single-variant association tests (Table 2). The power for joint multi-allelic analysis is consistently higher than single allelic analysis. We separately considered the power for the analysis of the primary and secondary alternative alleles. For the analysis of secondary alternative alleles, the single allelic analysis did not consider samples that carry the primary alleles. The power for single allelic analysis was much lower than multi-allelic analysis. For example, in the scenario where the causal allele effect is 0.25, the power for single allelic analysis is 0.36, whereas the power for multi-allelic analysis is 0.43. On the other hand, the power of multi-allelic analysis for detecting associations with primary alternative allele has a smaller advantage.

We also compared the power for the single allelic and multi-allelic analyses as omnibus tests for identifying associated variant sites (Table S1). Testing each allele separately may slightly increase the burden for multiple testing. In a deep sequencing study, 10% of the variant site can be multi-allelic. Using single allelic or multi-allelic analysis as omnibus test, a variant site is deemed to be associated if at least one alternative allele has p-values $< 5 \times 10^{-8}/1.1$, a threshold that corrects for the increased load of multiple testing. The power for the collapsing method was evaluated under the threshold of 5×10^{-8}. We considered models where (1) only the primary alternative allele is causal, (2) only the secondary alternative is causal, and (3) the model where all alternative alleles are causal. The power for single allelic and multi-allelic analysis is higher than the method that collapses multiple alleles under nearly all scenarios. When all alternative alleles are causal with effects in the same direction, the collapsing method is only slightly more powerful. When only the secondary alternative is causal,

the presence of a non-causal primary alternative allele can severely weaken the association signal and substantially reduce the power for the collapsing method.

Table 2. The power for single-variant association analysis. We compared the power of single allelic analysis and joint multi-allelic analysis for detecting associations with each alternative allele. The power was evaluated under the threshold of $\alpha = 5 \times 10^{-8}$, adjusting for the increased multiple testing burden for analyzing multiple alleles.

Sample Size	Genetic Effects	Single Allelic Analysis	Multi-Allelic Analysis
Type I Error/Power for the Analysis of the Primary (i.e., more frequent) Alt Alleles			
	0	4.7×10^{-8}	4.2×10^{-8}
10,000	0.1	0.24	0.25
	0.25	0.57	0.57
	0.5	0.75	0.76
	0	4.6×10^{-8}	4.2×10^{-8}
20,000	0.1	0.36	0.37
	0.25	0.67	0.68
	0.5	0.82	0.82
Type I Error/Power for the Analysis of Secondary (i.e., less frequency) Alt Alleles			
	0	4.1×10^{-8}	4.8×10^{-8}
10,000	0.1	0.037	0.056
	0.25	0.24	0.3
	0.5	0.48	0.55
	0	4.9×10^{-8}	4.3×10^{-8}
20,000	0.1	0.087	0.12
	0.25	0.36	0.43
	0.5	0.6	0.66

3.2. Type I Error and Power Evaluation for Gene-Level Association Test

We evaluated the power of two different analysis strategies for gene-level tests in the presence of multi-allelic sites: (1) the joint modeling approach that simultaneously considers the effects of multi-allelic and bi-allelic sites, and (2) the approach that discards multi-allelic sites from the gene-level analysis.

We also evaluated the power under a variety of scenarios with different combinations of sample sizes, genetic effect distributions, and proportions of causal variants. Causal variant effects were sampled from a normal distribution $N\left(0, \sigma_\beta^2\right)$, with $\sigma_\beta = 0.25$. Under each scenario, three commonly used gene-level tests were considered: the simple burden test, SKAT, and VT, analyzing rare variants with minor allele frequency (MAF) < 1% or 5%.

Type I errors were well-controlled across all scenarios. The power for gene-level tests was consistently higher when we jointly modeled the effects of all alternative alleles for multi-allelic sites (Table 3). The strategy that discards multi-allelic sites could lead to ~20% decrease in power, particularly when the effects of causal alleles are in the same directions. For example, when the MAF cutoff of 0.01 and 20% of the variants were causal, power for the burden/SKAT/VT tests were respectively 50%/39%/68%, which were substantially higher than the power for the three tests analyzing only bi-allelic variants (42%/35%/61%, respectively). This is consistent with the benchmark of rare-variant association methods [23,27], where the erroneous exclusion of causal variants drastically reduces power.

Table 3. The Type I Error and Power for Gene-level Association Tests. We compared the power for simple burden, SKAT, and VT tests for the joint multi-allelic analysis and the analysis that discards multi-allelic sites. The power and type I error were assessed under a threshold of $\alpha = 2.5 \times 10^{-6}$ using 1,000,000 replicates.

MAF Cutoff	Pct of Causal Variants	Power	
		Burden/SKAT/VT	Burden/SKAT/VT
		Joint Multi-allelic Analysis	**Discard Multi-allelic Sites**
		Type I Error	
0.01	0%	$2.6 \times 10^{-6}/2.1 \times 10^{-6}/3.0 \times 10^{-6}$	$2.5 \times 10^{-6}/3.1 \times 10^{-6}/2.6 \times 10^{-6}$
0.05	0%	$2.5 \times 10^{-6}/2.3 \times 10^{-6}/2.3 \times 10^{-6}$	$3.0 \times 10^{-6}/2.1 \times 10^{-6}/2.7 \times 10^{-6}$
		Power—Causal Variants Have Uni-directional Effects	
0.01	20%	0.50/0.39/0.68	0.42/0.35/0.61
	50%	0.93/0.79/0.99	0.90/0.77/0.98
0.05	20%	0.42/0.39/0.71	0.37/0.36/0.64
	50%	0.88/0.80/0.99	0.87/0.79/0.99
		Power—Causal Variants Have Bi-directional Effects	
0.01	20%	0.06/0.16/0.13	0.05/0.13/0.11
	50%	0.14/0.44/0.31	0.14/0.40/0.29
0.05	20%	0.05/0.17/0.14	0.05/0.15/0.12
	50%	0.12/0.42/0.30	0.12/0.40/0.30

When variants have bi-directional effects, SKAT is the most powerful test. The power of SKAT based upon multi-allelic analysis was considerably higher than the method that discards multi-allelic sites. Simple burden and VT tests were underpowered in this scenario. However, the tests based upon multi-allelic analysis were still consistently more powerful.

3.3. Analysis of Cigarettes-Per-Day Phenotype

We analyzed the genetic and phenotype data from the eight cohorts for the study of the CPD trait. The eight cohorts were genotyped with GWAS arrays and imputed to the 1000Genome reference panel with the Michigan Imputation Server [28,29]. After quality control, a total of 29,124,949 variants sites were segregating in at least one cohort. Among them, 289,809 (1%) contained multiple alternative alleles. The fraction of multi-allelic sites was lower than what was discovered in sequence-based studies, due to the sample size of the reference panel, the exclusion of the variant sites with low alternative allele counts from the reference panel, and the removal of rare imputed alleles due to low imputation quality. We performed careful quality control to ensure the quality of the results. Specifically, we carefully examined the quality of the data through additional metrics, including Hardy Weinberg Equilibrium (HWE) *p*-values. We excluded variants with HWE *p*-values $< 1 \times 10^{-8}$ (we checked and noted that all reported significant variants or variants in significant genes have HWE *p*-values > 0.05). Additionally, the histogram verifies that HWE *p*-values for genome-wide variants are predominantly insignificant (Figure S2). The allele frequency spectrum is also as expected with peaks for low-frequency variants, showing that a majority of the variants are low-frequency (Figure S3). Most of the multi-allelic variants were in the intergenic region, and only 105,727 belonged to the genic region. Among the 19,321 genes that were analyzed, 2475 contained coding multi-allelic variants (nonsynonymous, stop or splice), and 2319/2417 contained rare alternative alleles with MAF $< 1\%/5\%$ at their multi-allelic sites.

We first performed single-variant association tests. The analysis results are well-behaved. We examined the genomic control values for all variants in different frequency bins (0, 0.001], (0.001, 0.01], and (0.01, 0.5]. All genomic control inflation factors were < 1.03. We also separately

examined the genomic control inflation factor for multi-allelic variants only and ensured that the tests all had well-calibrated type I errors (Figures S4 and S5).

In single-variant association analysis, we recovered a well-known locus associated with CPD. In fact, all top hits in the meta-analysis came from the *CHRNA5-CHRNA3-CHRNB4* locus. The top variant was 15:78886947_G/A (rs4887067), which is a variant from the untranslated region in the gene *CHRNA5*. No other loci, or novel loci, were uncovered in this study with genome-wide significance.

The top association signals for multi-allelic variants also lay in the *CHRNA5-CHRNA3-CHRNB4* locus (Table 4, Figures S4 and S5). Most of the top association signals appeared as insertion–deletion polymorphisms (indels). The most significant variant 15:78915370(rs34573245) had a *p*-value of 1.6×10^{-11} and is located in the intergenic regions. There were five other significantly associated multi-allelic variants in the genes *CHRNA5*, *CHRNA3*, and *IREB2*.

We compared the association results using the new method and the method that relies on single-allelic analysis. Single-allelic analysis identified only 5 significantly associated multi-allelic variants in the locus, while the joint multi-allelic analysis identified 6 variants. The *p*-values for multi-allelic analysis were consistently smaller. The mean chi-square statistic at known loci could be used as an estimate for the non-centrality parameter and then used as a metric to empirically assess the power for an association statistic [30]. The mean chi-square statistics for the multi-allelic and single allelic analyses were 34.5 and 33.3, respectively, with multi-allelic variants 4% higher. This is consistent with the observations in our simulation studies. We plotted the $-\log_{10}(P)$ for the two methods (Figure S6). We observed a higher concordance between multi-allelic and single-allelic association analysis for lower-frequency variants (MAF < 1%) than for common variants (MAF > 1%), with rank correlations for common and rare variants at 98% and 90%, respectively.

In addition, we also implemented the method that collapses multiple alternative alleles. The collapsing method is an omnibus test, which can be used to identify associated variant sites, instead of associated alleles. Given that all the top association signals are driven by the common primary alternative allele, all collapsing *p*-values were less significant than multi-allelic analysis *p*-values. No additional significant variant sites were identified. Among the 6 top variant sites identified using joint multi-allelic analysis, only two variant sites, 15:78915370 and 15:78859605, had genome-wide significant collapsing *p*-values ($p < 5 \times 10^{-8}$).

We also performed gene-level association tests analyzing variants with MAF <1% and 5%. Type I errors were well-controlled for all gene-level tests (Figures S7 and S8). For genes with rare multi-allelic variants, no significant associations were found (Table 5). Only one gene, *SHCBP1L*, was identified as significant under the Bonferroni threshold $\alpha = 2.5 \times 10^{-6}$ for testing 20,000 genes (Table S2). The gene is a testis-specific spindle-associated factor that plays a role in spermatogenesis [31,32]. It was shown that smoking is detrimental to human reproduction [33]. Here, the direction of effect for the rare-variant burden is consistent with the previous observation, and increased burden is associated with increased tobacco use and reduces the sperm production. Single variant association results within identified genes are shown in (Table S3).

Finally, we compared the gene-level test *p*-values for the analysis that discards multi-allelic sites, and the analysis that only analyzes rare alternative alleles without controlling for the genotypes of the common alternative allele in the same site. Considerable discrepancies were observed in the scatterplots for different analysis strategies (Figure S9), which shows that the naïve method does not provide a useful approximation for the principled methods.

Our software implementation scaled well with this large-scale analysis. The generation of single-variant association statistics took 15.1 CPU hours. The computing time scales linearly with the sample size and the number of genetic variants. It required 2.1 CPU hours for single-variant meta-analysis and 6.2 CPU hours for three gene-level association tests conducted under two different MAF cutoffs.

Table 4. Top Single-Variant Association Signals for the Cigarettes-Per-Day Phenotype Using Multi-Allelic Analysis. Results are shown for variants with *p*-values less than 5×10^{-8}. We report the *p*-values and genetic effect estimates for each alternative allele at multi-allelic sites. As a comparison, we also report the *p*-values and test statistics from single-allelic analysis, as well as the omnibus test that collapses multiple alleles.

Position	Ref Allele	Alt Allele	Alt Allele Freq	*p*-Value	β	β SD	N	Direction of Effects*	Anno	Stat Single-Allelic Analysis	*p*-Value Single-Allelic Analysis	*p*-Value Collapsing Multi-Allelic Sites
15:78915370	CT	C	0.41	1.6×10^{-11}	0.078	0.012	17,512	-++++++	Intergenic	44.67	2.3×10^{-11}	1.0×10^{-10}
15:78915370	CT	CTTT	0.019	0.61	0.022	0.044	17,512	-++++-+-	Intergenic	0.34	0.55	
15:78859605	AAAAAG	A	0.33	2.3×10^{-11}	0.079	0.012	17,512	-++++++	Deletion; CHRNA5	43.74	3.8×10^{-11}	5.1×10^{-10}
15:78859605	A	G	0.00077	0.38	0.29	0.33	17,512	+++-+-+-	Intron; CHRNA5	0.091	0.76	
15:78913353	CGCGGGCGG	C	0.47	2.4×10^{-9}	0.072	0.012	17,512	-++++++	Deletion; CHRNA3	33.31	7.8×10^{-9}	2.8×10^{-7}
15:78913353	CGCGGGCGG	CGCGGGCGGGCGG	0.033	0.10	−0.057	0.035	17,512	+-+—-	Insertion; CHRNA3	0.75	0.38	
15:78785944	AT	ATT	0.29	7.7×10^{-9}	0.079	0.014	17,512	-++++++	Insertion; IREB2	31.81	1.7×10^{-8}	1.5×10^{-4}
15:78785944	AT	A	0.18	0.71	0.0056	0.016	17,512	+——+	Deletion; IREB2	0.00057	0.98	
15:78871382	CT	CTT	0.40	1.4×10^{-8}	0.080	0.014	13,723	XX++++++	Insertion; CHRNA5	27.81	1.3×10^{-7}	6.2×10^{-7}
15:78871382	CT	C	0.070	0.37	0.022	0.025	17,512	-+—+	Deletion; CHRNA5	0.29	0.58	
15:78751667	G	GTTTTTTGTTTGTTTGT	0.29	1.6×10^{-8}	0.071	0.013	17,512	-++++++	Insertion; IREB2	22.25	2.4×10^{-6}	1.1×10^{-7}
15:78751667	G	GTTTTTTTGTTTGTTTG	0.0019	0.97	0.0048	0.14	17,512	++-+—	Insertion; IREB2	0.065	0.80	

*: "+" and "-" denote the direction of relationship between alternative alleles and Cigarettes-Per-Day phenotype.

Table 5. Top Gene-level Association Signals for Genes with Multi-allelic Sites. We performed simple burden, SKAT, and VT tests under the two different minor allele frequency cutoffs, 0.01 and 0.05. No results were significant under the threshold $\alpha = 2.5 \times 10^{-6}$. For each rare-variant test performed, we show the test statistics, p-values, the number of variant sites, and the number of multi-allelic variant sites for the top 3 signals.

Gene	Statistic	p-Value	Number of Variant Site	Number of Multi-allelic Site	Number of Multi-allelic Site with Rare Variant	Gene	Statistic	p-Value	Number of Variant Site	Number of Multi-allelic Site	Number of Multi-allelic Site with Rare Variant
Burden Test with MAF < 1%						Burden Test with MAF < 5%					
MLKL	16.81	4.1×10^{-5}	16	8	4	PTPN22	13.49	0.00024	19	10	3
DMBX1	13.11	0.00029	4	2	2	CROCC	11.14	0.00085	80	11	8
BRD3	10.73	0.0011	5	6	4	HLA-DQA1	10.25	0.0014	6	15	0
SKAT Test with MAF < 1%						SKAT Test with MAF < 5%					
ABTB1	1,654,137.32	5.5×10^{-5}	12	6	5	ABTB1	1,834,395.76	0.00015	12	6	5
SEMA7A	1,056,004.78	0.00032	5	2	1	DTNBP1	4,098,263.16	0.00049	7	6	5
METTL8	1,075,541.99	0.00036	6	15	8	NRBF2	1,454,883.06	0.00074	3	4	2
VT Test with MAF < 1%						VT Test with MAF < 5%					
TTC15	21.98	1.9×10^{-5}	17	23	9	TTC15	21.98	2.46×10^{-5}	17	23	9
MLKL	16.81	0.00031	16	8	4	WNK1	16.11	0.00049	41	25	8
ARHGEF40	15.08	0.00078	26	4	2	MLKL	15.71	0.00068	16	8	4

4. Discussion

Multi-allelic variants represent a highly important class of genetic variation in large-scale sequencing studies. Multi-allelic variants have been largely ignored in the GWAS era due to the extensive use of common bi-allelic single-nucleotide polymorphisms (SNPs) as markers to tag regions that harbor causal variants. As deep sequence data become increasingly available on larger sample sizes, many more low-frequency and rare multi-allelic variants are expected to be discovered. Many of the novel variants will be identified at previously monomorphic sites, while others will appear as novel alleles at known sites. For example, in a sample with ~65,000 exomes, 8% of the known variant sites were found to be multi-allelic. It will be important to be able to properly analyze such multi-allelic variants for disease association and assess their functional impact.

Here, we developed and evaluated a new method for analyzing multi-allelic sites in sequence-based association studies and meta-analysis. The method proceeds by jointly modeling the effects of different alternative alleles for multi-allelic variants. It allows unbiased estimates of multi-allelic effects relative to the reference allele and leads to more powerful gene-level tests than the method that discards multi-allelic variants.

Most of the multi-allelic variants from imputation-based GWAS contain no more than 2 different alternative alleles. For single nucleotide variants, there can be at most 3 different alternative alleles. We focus on testing the effect of each alternative allele for association, while jointly modeling the effects of other alternative alleles. This strategy appears to be the most biologically relevant. We are interested in knowing if a given base pair change is associated with the phenotype, so that we can follow up with precise functional experiment to validate these discoveries.

For the analysis of other types of variants such as indels or copy number variations, there may be more alternative alleles. It may be of interest to perform an omnibus test (e.g., by collapsing multiple alternative alleles) and examine if at least one allele at the site is associated with the disease outcome. One possibility is to consider multivariate tests, such as the multivariate score test [34], the method by collapsing multiple alternative alleles, or the variance component score-based test [35]. It is well known that these multivariate tests can be calculated using the shared summary association statistics and their covariance matrices. Thus, our framework can be easily adapted to omnibus tests for multi-allelic variants. It should be noted that the omnibus test can be extremely underpowered if only the secondary alternative alleles (the alternative allele with lower frequencies) are causal. As most of the novel alternative allele identified by a sequencing study is rare, the utility of omnibus test in the association of multi-allelic variants remain to be understood in the upcoming large-scale sequencing studies.

As an application, we applied our method to a large-scale meta-analysis of the cigarettes-per-day phenotype using the 1000 Genomes Project-based imputations. The analysis type I error rates were well-behaved and sufficiently powerful, confirming the known association for the *CHRNA5-CHRNA3-CHRNB4* locus. However, no new loci were uncovered from our meta-analysis using ~18,000 samples. This may be because our dataset is smaller than some of the largest studies on tobacco addiction [36]. It is clear that larger sample sizes may be necessary to uncover novel nicotine addiction-related loci.

As multi-allelic variants are often ignored by existing GWAS and sequence-based association analysis software packages, the representation of multi-allelic variants is still not unified. The output from the popular imputation software and imputation servers [29] represent multiple alleles from the same site in separate lines, with the genotypes in each line representing the number of the corresponding alternative alleles. For instance, the variant at chromosome 19 and position 55178198 has reference allele C, and two possible alternative alleles G and T. The VCF file contains two lines for this variant site: one line with reference/alternative alleles being C/G and the other line with reference/alternative alleles C/T. To represent individual genotypes, one must combine information across the two lines in the VCF. For example, genotypes of G/T (i.e., heterozygous for both alternative alleles) would be represented with a genotype coding of 0/1, 0/1 in the two lines of VCF file. Similarly, genotypes of C/G would be encoded as 1/0 and 0/0. In other VCFs, such as the VCF files released

by the ExAC project [6], the multi-allelic variant may be represented in one single line. For instance, the same variant 19:55178198, 0,1, and 2 may be used to represent the reference allele C and the first and second alternative alleles G and T, respectively. In this case, the genotype G/T is coded as 1/2. Software packages are available to recode the genotypes of multi-allelic site in separate lines. Our implementation of the method supports both representations. In the future, it will be helpful to standardize the representation of the multi-allelic sites and streamline the support in software packages and libraries.

In conclusion, we discussed several methods and practical issues for multi-allelic association analysis and meta-analysis, which provide unbiased effect estimates for multi-allelic variants and improve power over currently available approaches. As large-scale sequencing studies become more prevalent, multi-allelic variants will become an even more important class of genetic variation. We envision that our methods and software tools will be highly applicable for understanding the functional impact and disease associations of multi-allelic variants in large-scale sequencing studies.

References

1. Auer, P.L.; Teumer, A.; Schick, U.; O'Shaughnessy, A.; Lo, K.S.; Chami, N.; Carlson, C.; de Denus, S.; Dube, M.P.; Haessler, J.; et al. Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat. Genet.* **2014**, *46*, 629–634. [CrossRef] [PubMed]
2. Do, R.; Stitziel, N.O.; Won, H.H.; Jorgensen, A.B.; Duga, S.; Angelica Merlini, P.; Kiezun, A.; Farrall, M.; Goel, A.; Zuk, O.; et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **2015**, *518*, 102–106. [CrossRef] [PubMed]
3. Myocardial Infarction, G.; Investigators, C.A.E.C.; Stitziel, N.O.; Stirrups, K.E.; Masca, N.G.; Erdmann, J.; Ferrario, P.G.; Konig, I.R.; Weeke, P.E.; Webb, T.R. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *N. Engl. J. Med* **2016**, *374*, 1134–1144. [CrossRef]
4. TG and HDL Working Group of the Exome Sequencing Project; National Heart, Lung, and Blood Institute; Crosby, J.; Peloso, G.M.; Auer, P.L.; Crosslin, D.R.; Stitziel, N.O.; Lange, L.A.; Lu, Y.; Tang, Z.; et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* **2014**, *371*, 22–31. [CrossRef]

5. Cohen, J.C.; Boerwinkle, E.; Mosley, T.H., Jr.; Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **2006**, *354*, 1264–1272. [CrossRef] [PubMed]

6. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B.; et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285–291. [CrossRef]

7. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, 7. [CrossRef]

8. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]

9. El Galta, R.; Hsu, L.; Houwing-Duistermaat, J.J. Methods to test for association between a disease and a multi-allelic marker applied to a candidate region. *BMC Genet.* **2005**, *6*, S101. [CrossRef]

10. Terwilliger, J.D. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* **1995**, *56*, 777–787.

11. Zuk, O.; Schaffner, S.F.; Samocha, K.; Do, R.; Hechter, E.; Kathiresan, S.; Daly, M.J.; Neale, B.M.; Sunyaev, S.R.; Lander, E.S. Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E455–E464. [CrossRef] [PubMed]

12. Evangelou, E.; Ioannidis, J.P. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **2013**, *14*, 379–389. [CrossRef]

13. Zhan, X.; Hu, Y.; Li, B.; Abecasis, G.R.; Liu, D.J. RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **2016**, *32*, 1423–1426. [CrossRef] [PubMed]

14. Feng, S.; Liu, D.; Zhan, X.; Wing, M.K.; Abecasis, G.R. RAREMETAL: Fast and powerful meta-analysis for rare variants. *Bioinformatics* **2014**, *30*, 2828–2829. [CrossRef] [PubMed]

15. Li, Y.; Sidore, C.; Kang, H.M.; Boehnke, M.; Abecasis, G.R. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* **2011**, *21*, 940–951. [CrossRef] [PubMed]

16. Howie, B.; Fuchsberger, C.; Stephens, M.; Marchini, J.; Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **2012**, *44*, 955–959. [CrossRef]

17. Howie, B.N.; Donnelly, P.; Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **2009**, *5*, e1000529. [CrossRef]

18. Tang, Z.Z.; Lin, D.Y. Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs. *Am. J. Hum. Genet.* **2015**, *97*, 35–53. [CrossRef]

19. Lee, S.; Teslovich, T.M.; Boehnke, M.; Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **2013**, *93*, 42–53. [CrossRef]

20. Liu, D.J.; Peloso, G.M.; Zhan, X.; Holmen, O.L.; Zawistowski, M.; Feng, S.; Nikpay, M.; Auer, P.L.; Goel, A.; Zhang, H.; et al. Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **2014**, *46*, 200–204. [CrossRef]

21. Jiang, Y.; Chen, S.; McGuire, D.; Chen, F.; Liu, M.; Iacono, W.G.; Hewitt, J.K.; Hokanson, J.E.; Krauter, K.; Laakso, M.; et al. Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLoS Genet.* **2018**, *14*, e1007452. [CrossRef] [PubMed]

22. Yang, J.; Ferreira, T.; Morris, A.P.; Medland, S.E.; Genetic Investigation of, A.T.C.; Replication, D.I.G.; Meta-analysis, C.; Madden, P.A.; Heath, A.C.; Martin, N.G.; et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **2012**, *44*, 369–375. [CrossRef] [PubMed]

23. Li, B.; Leal, S.M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **2008**, *83*, 311–321. [CrossRef] [PubMed]

24. Wu, M.C.; Lee, S.; Cai, T.; Li, Y.; Boehnke, M.; Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **2011**, *89*, 82–93. [CrossRef]

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

26. Price, A.L.; Kryukov, G.V.; de Bakker, P.I.; Purcell, S.M.; Staples, J.; Wei, L.J.; Sunyaev, S.R. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **2010**, *86*, 832–838. [CrossRef]

27. Liu, D.J.; Leal, S.M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* **2010**, *6*, e1001156. [CrossRef]

28. McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A.R.; Teumer, A.; Kang, H.M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **2016**, *48*, 1279–1283. [CrossRef]

29. Das, S.; Forer, L.; Schonherr, S.; Sidore, C.; Locke, A.E.; Kwong, A.; Vrieze, S.I.; Chew, E.Y.; Levy, S.; McGue, M.; et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **2016**, *48*, 1284–1287. [CrossRef]

30. Zaitlen, N.; Pasaniuc, B.; Patterson, N.; Pollack, S.; Voight, B.; Groop, L.; Altshuler, D.; Henderson, B.E.; Kolonel, L.N.; Le Marchand, L.; et al. Analysis of case-control association studies with known risk variants. *Bioinformatics* **2012**, *28*, 1729–1737. [CrossRef]

31. Sood, R.; Bonner, T.I.; Makalowska, I.; Stephan, D.A.; Robbins, C.M.; Connors, T.D.; Morgenbesser, S.D.; Su, K.; Faruque, M.U.; Pinkett, H.; et al. Cloning and characterization of 13 novel transcripts and the human RGS8 gene from the 1q25 region encompassing the hereditary prostate cancer (HPC1) locus. *Genomics* **2001**, *73*, 211–222. [CrossRef] [PubMed]

32. Liu, M.; Shi, X.; Bi, Y.; Qi, L.; Guo, X.; Wang, L.; Zhou, Z.; Sha, J. SHCBP1L, a conserved protein in mammals, is predominantly expressed in male germ cells and maintains spindle stability during meiosis in testis. *Mol. Hum. Reprod.* **2014**, *20*, 463–475. [CrossRef] [PubMed]

33. Dai, J.B.; Wang, Z.X.; Qiao, Z.D. The hazardous effects of tobacco smoking on male fertility. *Asian J. Androl.* **2015**, *17*, 954–960. [CrossRef] [PubMed]

34. Hotelling, H. The Generalization of Student's Ratio. *Ann. Math. Statist.* **1931**, *3*, 360–378. [CrossRef]

35. Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **1997**, *84*, 309–326. [CrossRef]

36. The Tobacco; Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **2010**, *42*, 441–447. [CrossRef]