Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations

Yi-Juan Hu, Yun Li, Paul L. Auer, and Dan-Yu Lin

*Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322; †Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420; ‡Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7264; and §Joseph J. Zilber School of Public Health, University of Wisconsin, Milwaukee, WI 53201-0413

Edited by Elizabeth A. Thompson, University of Washington, Seattle, WA, and approved December 9, 2014 (received for review April 3, 2014)

In the large cohorts that have been used for genome-wide association studies (GWAS), it is prohibitively expensive to sequence all cohort members. A cost-effective strategy is to sequence subjects with extreme values of quantitative traits or those with specific diseases. By imputing the sequencing data from the GWAS data for the cohort members who are not selected for sequencing, one can dramatically increase the number of subjects with information on rare variants. However, ignoring the uncertainties of imputed rare variants in downstream association analysis will inflate the type I error when sequenced subjects are not a random subset of the GWAS subjects. In this article, we provide a valid and efficient approach to combining observed and imputed data on rare variants. We consider commonly used gene-level association tests, all of which are constructed from the score statistic for assessing the effects of individual variants on the trait of interest. We show that the score statistic based on the observed genotypes for sequenced subjects and the imputed genotypes for nonsequenced subjects is unbiased. We derive a robust variance estimator that reflects the true variability of the score statistic regardless of the sampling scheme and imputation quality, such that the corresponding association tests always have correct type I error. We demonstrate through extensive simulation studies that the proposed tests are substantially more powerful than the use of accurately imputed variants only and the use of sequencing data alone. We provide an application to the Women’s Health Initiative. The relevant software is freely available.

R ecent technological advances have made it possible to conduct high-throughput DNA sequencing studies on rare variants, which have a stronger impact on complex diseases and traits than common variants (1). However, it is still economically infeasible to sequence all subjects in a large cohort, and, therefore, only a subset of cohort members can be selected for sequencing. A cost-effective sampling strategy is to preferentially select subjects in the extremes of a quantitative trait distribution or those with a specific disease (2, 3). For case–control studies, an equal number of cases and controls provides more power than other case–control ratios. For quantitative traits, the power increases as more extreme values are sampled (2).

Trait-dependent sampling has been adopted in many sequencing studies, including the National Heart, Lung, and Blood Institute (NHBLI) Exome Sequencing Project (ESP) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) resequencing project. The NHBLI ESP consists of three studies that sequenced subjects with the largest and smallest values of body mass index (BMI), low-density lipoprotein, and blood pressure, one case–control study on myocardial infarction, and one case-only study on stroke (2). The CHARGE resequencing project selected subjects with the highest values of 14 quantitative traits, as well as a random sample (4).

The cohorts from which subjects are drawn for sequencing often have collected genotyping array data on all or most cohort members through genome-wide association studies (GWAS). This is certainly the case with the cohorts used in the NHLBI ESP and CHARGE projects. If we impute the sequencing data from the GWAS data for the cohort members who are not selected for sequencing, we will dramatically increase the number of subjects with information on rare variants. Indeed, such imputation is being carried out in the Women’s Health Initiative (WHI) (5), which is a major component of the NHLBI ESP, and other cohorts (6). The algorithms for rare variant imputation include MaCH (Markov chain based haplotyper) (7), IMPUTE2 (8), BEAGLE (9), and minimac (10), all of which impute rare variants by leveraging the linkage disequilibrium (LD) between sequenced and GWAS variants.

The aforementioned imputation algorithms have been routinely used to impute untyped common single nucleotide polymorphisms (SNPs) (i.e., variants not present on the genotyping array) in GWAS (11, 12). Common SNPs can be imputed with high degrees of accuracy, and single-SNP association tests treating imputed genotype values as observed quantities have reasonable control of the type I error (13, 14).

There are major differences between imputing common SNPs in GWAS and imputing rare variants in sequencing studies. First, the former uses an external reference panel, such as the HapMap (15) or the 1,000 Genomes (16), whereas the latter relies primarily on an internal reference panel, i.e., the subset of GWAS subjects who are sequenced. Second, rare variants cannot be imputed very accurately because of the low minor allele count data integration | gene-level association tests | genotype imputation | linkage disequilibrium | whole-exome sequencing
and weak LD. Third, untyped common SNPs are unobserved and imputed for all subjects in GWAS whereas rare variants are observed in a subset of the study subjects (i.e., sequenced subjects) and imputed for the rest (i.e., nonsequenced subjects) in sequencing studies, creating within-study differential quality in genotype data. If the selection of subjects for sequencing depends on the trait values, then the variance of the trait for sequenced subjects is generally different from that of nonsequenced subjects. The combination of differential genotype quality and differential trait variability between the sequenced and nonsequenced samples will cause inflation of the type I error in association testing, as will be explained in Methods section.

To reduce within-study differential quality in genotype data, one may use a postimputation quality control (QC) procedure (6) to filter out variants that have been imputed with high degrees of uncertainty. As stated, the imputation accuracy for rare variants is typically low; therefore, the use of any reasonable QC filter will remove a large number of rare variants. The removal of variants results in loss of important information because the observed genotype data for sequenced subjects cannot be used if a variant is excluded due to low imputation quality for nonsequenced subjects and because the imputed genotypes, even for a variant that cannot be imputed accurately, may still contain valuable information about the association.

In this article, we show how to perform valid and efficient association tests when rare variants are imputed for nonsequenced subjects. Because single-variant tests and commonly used gene-level tests, such as burden (17, 18), variable threshold (VT) (19, 20), and sequence kernel association test (SKAT) (21), are all based on the score statistic for testing the associations between the genotypes of individual variants and the trait of interest, we investigate the properties of the score statistic based on the observed genotypes for sequenced subjects and the imputed genotypes for nonsequenced subjects. We find that the score statistic is unbiased (provided that there is no strong population stratification). We show that the standard variance estimator for the score statistic is valid if a random subset of the GWAS subjects is selected for sequencing but is invalid if the selection depends on the trait values and the imputation is inaccurate. In addition, we derive a robust variance estimator that reflects the true variability of the score statistic regardless of the sampling scheme and imputation quality, such that the corresponding association tests are guaranteed to have correct type I error. We show, in realistic simulation studies, that the proposed approach is substantially more powerful than the use of accurately imputed variants only and the use of sequencing data alone. We further demonstrate the advantages of the new methodology in an application to empirical data from the WHI.

Methods

We first consider single-variant analysis without covariates. Let G denote the genotype (i.e., number of minor alleles) at the variant site, and let Y denote the trait of interest. Suppose that a total of N unrelated cohort members are measured on Y and GWAS SNPs and that a subset of n subjects is selected for sequencing and thus measured on G. The selection may be completely random or trait dependent. For example, when Y is quantitative, one may select subjects with the largest and smallest values of Y, and when Y is binary, one may draw a case–control sample with an equal number of cases (i.e., Y = 1) and controls (i.e., Y = 0) (regardless of the proportion of cases in the entire cohort) or a case-only sample.

We infer the unknown values of G for the (N - n) nonsequenced subjects from their GWAS data by using the LD between the variant of interest and the GWAS SNPs among the sequenced subjects. We impute G by the expected count of the minor allele (i.e., dosage). The imputation is performed by any of the commonly used algorithms (7-10) without considering the trait information (e.g., by combining cases and controls in a case-control study). Let G denote the imputed value of G for the nonsequenced subject and the observed value of G for the sequenced subject.

We relate quantitative Y to G through the linear regression model

$$Y = \gamma_0 + \beta G + \epsilon,$$

and binary Y to G through the logistic regression model

$$\Pr(Y = 1) = \frac{\exp(\gamma_0 + \beta G)}{1 + \exp(\gamma_0 + \beta G)},$$

where $\gamma_0$ is the intercept, $\beta$ is the association parameter, and $\epsilon$ is normal with mean zero and variance $\sigma^2$. The score statistic for testing the null hypothesis $H_0: \beta = 0$ based on the data $(Y_i, G_i)$ ($i = 1, \ldots, N$) is

$$U = \sum_{i=1}^{N} (Y_i - \bar{Y}) (G_i - \bar{G}),$$

where $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$. The standard variance estimator for U based on the Fisher information is

$$V_{\text{std}} = N^{-1} \sum_{i=1}^{N} \left( Y_i - \bar{Y} \right)^2 \left( G_i - \bar{G} \right)^2,$$

where $\bar{G} = \frac{1}{N} \sum_{i=1}^{N} G_i$.

Let us order the data such that the first n subjects are the sequenced ones. By some simple algebra, $U = \sum_{i=1}^{n} (Y_i - \bar{Y}) (G_i - \bar{G}) + \sum_{i=n+1}^{N} (Y_i - \bar{Y}) (G_i - \bar{G})$. Under the null hypothesis $H_0$, Y is independent of G in both the sequenced and nonsequenced samples. By Eq. S2 of SI Appendix, SI Method A, the means of G are the same as the mean of G in both samples. Thus, the mean of U is zero regardless of the imputation quality. However, the standard variance estimator $V_{\text{std}}$ may not fully capture the variability of U with imputed rare variants, as explained below. First, the imputed values are less variable than the observed values when the imputation accuracy is not sufficiently high. Second, the variance of Y may be different between sequenced and nonsequenced subjects when the selection for sequencing depends on Y. Thus, the variance of Y may be related to the variance of G. Consequently, the standard variance estimator $V_{\text{std}}$ which treats the variance of Y and the variance of G as unrelated, tends to underestimate the true variance of U; a formal proof is provided in SI Appendix, SI Method C.

The following robust variance estimator fully captures the variability of U under any sampling scheme by properly adjusting for the imputation accuracy:

$$V_{\text{rob}} = n^{-1} \sum_{i=1}^{n} \left( Y_i - \bar{Y}_i \right)^2 \left( G_i - \bar{G}_i \right)^2 + \sum_{i=n+1}^{N} \left( Y_i - \bar{Y}_i \right)^2 \left( G_i - \bar{G}_i \right)^2,$$

where $\bar{Y}_i = \frac{1}{n} \sum_{j=1}^{n} Y_j$ and $\bar{G}_i = \frac{1}{n} \sum_{j=1}^{n} G_j$. $\bar{Y}_i$ and $\bar{G}_i$ are the means of $Y_{\text{seq}}$ and $G_{\text{seq}}$, respectively.

where $\bar{Y}_i$ and $\bar{G}_i$ are the means of $Y_{\text{seq}}$ and $G_{\text{seq}}$, respectively.

where $\bar{Y}_i$ and $\bar{G}_i$ are the means of $Y_{\text{seq}}$ and $G_{\text{seq}}$, respectively.

where $\bar{Y}_i$ and $\bar{G}_i$ are the means of $Y_{\text{seq}}$ and $G_{\text{seq}}$, respectively.

where $\bar{Y}_i$ and $\bar{G}_i$ are the means of $Y_{\text{seq}}$ and $G_{\text{seq}}$, respectively.

where $\bar{Y}_i$ and $\bar{G}_i$ are the means of $Y_{\text{seq}}$ and $G_{\text{seq}}$, respectively.
In SI Appendix, SI Method C, we extend our methodology in several directions. First, we consider multiple variants in a gene and derive a robust estimator for the variance–covariance matrix (of the vector-valued) score statistic, from which all commonly used gene-level tests (e.g., burden, VT, and SKAT) can be constructed (23). Second, we incorporate covariates (e.g., demographic variables and principal components for ancestry) into the regression model and show that the score statistic continues to be unbiased as long as covariates are independent of genotypes. (This implies that the score statistic is approximately unbiased when there is no strong population stratification.) We also provide the corresponding robust variance estimator in Eq. S13 in SI Appendix. Third, we accommodate other types of traits such as ordinal and count data by adopting the class of generalized linear models. Fourth, we modify the robust variance estimator for binary traits to improve numerical stability; see Eq. S14 in SI Appendix. Finally, we show how to perform meta-analysis of multiple studies.

Results

Simulation Studies. We conducted extensive simulation studies to evaluate the performance of the proposed methods in realistic settings. We chose one gene, NPHS2 (nephrosis 2, idiopathic, steroid-resistant) (accession NM_014625), on chromosome 1 and restricted our analysis to variants with MAFs ≤5% that are nonsense, missense, or splice site. For the 360 African Americans in the WHI who were sequenced by the NHLBI ESP, there are five variants in NPHS2, with MAFs of 0.002, 0.004, 0.001, 0.022, and 0.005 for snp.98398, snp.98400, snp.98401, snp.98418, and snp.98419, respectively. The number of variants and the total MAF (i.e., sum of MAFs over variant sites) for NPHS2 equal the median number of variants and the median total MAF for all genes, respectively, making this gene a good representation. We used GWA simulator (24) to generate genotype data for the variants identified by sequencing and for the flanking GWAS SNPs by mimicking the MAFs and LD patterns observed for the variants identified by sequencing and for the flanking GWAS SNPs.

We considered a cohort of 5,000 subjects. We generated quantitative traits by mimicking the MAFs and LD patterns observed for the variants identified by sequencing and for the flanking GWAS SNPs, respectively, making this gene a good representation. We constructed the burden, VT, and SKAT tests based on Vrob and Vstd. For the burden test, we adopted the MAF threshold of 5%, which corresponds to T5. For SKAT, we used the default weighted linear kernel function. We refer to these methods as T5-rob, T5-std, VT-rob, VT-std, SKAT-rob, and SKAT-std. As a benchmark, we included the tests based on sequenced subjects only, which are referred to as T5-seq, VT-seq, and SKAT-seq. We set the nominal significance level at 0.001.

We first considered the situation of no confounding (i.e., \( \xi = 0 \)) and set \( \gamma_1 = 0.2 \). The simulation results for the T5 tests under the null hypothesis are summarized in Table 1. As expected, the score statistic is virtually unbiased in every scenario, and Vrob accurately reflects the true variability of U. Consequently, T5-rob always has correct control of the type I error. Under random sampling (i.e., Q1 and B1), Vstd accurately and yields proper type I error. Under other sampling schemes, Vstd underestimates the true variability of U, and the type I error rate can be 50 times the nominal level. The results for VT and SKAT exhibit the same patterns as those of T5.

Table 1. Simulation results for the T5 tests under the null hypothesis

<table>
<thead>
<tr>
<th>Sampling scheme</th>
<th>Vrob</th>
<th>Vstd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: random sample of 500 subjects</td>
<td>0.000</td>
<td>0.120</td>
</tr>
<tr>
<td>Q2: 250 largest and 250 smallest values</td>
<td>0.001</td>
<td>0.181</td>
</tr>
<tr>
<td>Q3: 500 largest and 250 smallest values</td>
<td>0.006</td>
<td>0.192</td>
</tr>
<tr>
<td>Q4: 250 largest and 250 random sample of 250</td>
<td>0.006</td>
<td>0.134</td>
</tr>
<tr>
<td>Q5: 250 largest and 250 random sample of 1,000</td>
<td>0.005</td>
<td>0.171</td>
</tr>
<tr>
<td>B1: 50% disease rate</td>
<td>0.000</td>
<td>0.060</td>
</tr>
<tr>
<td>B2: 30% disease rate</td>
<td>0.001</td>
<td>0.057</td>
</tr>
<tr>
<td>B3: 20% disease rate</td>
<td>0.002</td>
<td>0.053</td>
</tr>
<tr>
<td>B4: 10% disease rate</td>
<td>0.003</td>
<td>0.047</td>
</tr>
<tr>
<td>B5: 5% disease rate</td>
<td>0.003</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Q1, Q2, Q3, Q4, and Q5 are five different sampling schemes for quantitative traits; B1, B2, B3, B4, and B5 are five different sampling schemes for binary traits; Vrob and Vstd are the robust and standard variance estimators, respectively. Bias and SE are, respectively, the bias and SE of the score statistic of T5. SEE is the mean of the SE estimator, and size is the type I error rate divided by the nominal significance level of 0.001. Each entry is based on 100,000 replicates.

We examined the robustness of the proposed methods to confounding (i.e., \( \xi \neq 0 \)). We set \( \xi = 0.1 \) and 0.2, which correspond to Pearson correlation coefficients of ~0.03 and ~0.057, respectively. (The correlation coefficient of 0.03 is the third quartile of the correlation coefficients between the T5 burden scores and the percentage of African ancestry in the WHI.) In this case, the mean of the score statistic may not be zero. As shown in SI Appendix, Fig. S2, however, the type I error of T5-rob is still reasonable, especially for binary traits.
As mentioned previously, a common practice is to use a post-imputation QC procedure to filter out inaccurately imputed variants before association analysis. For example, Auer et al. (6) excluded variants if their Rsq values were less than the thresholds chosen such that within each MAF category, variants passing the threshold had an average Rsq of 0.8 or higher. In particular, they chose Rsq thresholds of 0.3, 0.6, 0.8, and 0.9 for variants with MAFs 3–5%, 1–3%, 0.5–1%, and 0.1–0.5%, respectively, and excluded all variants with MAFs <0.1%. If we had adopted such a QC procedure in our simulation studies, we would have excluded the gene NPHS2 entirely since the largest Rsq of the variants was merely 0.22. By contrast, our integrative analysis not only allowed association tests for this poorly imputed gene but also gained substantial power over the analysis of sequenced subjects only.

To further demonstrate the benefits of using inaccurately imputed variants, we considered a second gene, OR103 (olfactory receptor, family 10, subfamily J, member 3) (accession NM_001004467), on chromosome 1 that has seven variants, snp.88226, snp.88228, snp.88232, snp.88236, snp.88240, snp.88241, and snp.88244, with MAFs of 0.016, 0.01, 0.008, 0.002, 0.001, 0.008, and 0.002, respectively. The Rsq is 0.67 for snp.88226 and almost zero for the other variants. With the use of a QC procedure (6), only snp.88226 would be retained. As shown in Fig. S3, the T5-rob test based on pre-QC variants is substantially more powerful than the T5-std test based on post-QC variants.

The aforementioned simulation studies were based on two specific genes. We also considered all genes on chromosome 1 and generated a binary trait from the logistic regression model logit(Pr(Y = 1)) = βS + γX + γ0, where S is the total number of mutations the subject carries in five genes, and β = 0 and 1.2 under the null and alternative hypotheses, respectively, γ1 = 0.2, and γ0 was chosen to yield 10% disease rate. We simulated a cohort of 5,000 subjects, selected 250 cases and 250 controls for sequencing, and imputed the genotypes of the variants identified by sequencing for subjects that were not selected. As shown in SI Appendix, Fig. S4, the test based on Vrob correctly controls the type I error and identifies three out of five causal genes after Bonferroni correction. The test based on Vstd and pre-QC variants inflated type I error, while the one based on Vstd and post-QC variants identifies only one causal gene.

**WHI Data.** The WHI was established by the National Institutes of Health in 1991 to address major health issues causing morbidity and mortality among postmenopausal women (5). We focused on the BMI values for the African American participants of the WHI cohort. Among the 8,142 African American participants who were genotyped by the Affymetrix 6.0 arrays, 360 with BMI values >40 or <25 were selected for whole-exome sequencing in the NHLBI ESP (2). The distribution of the BMI values is displayed in SI Appendix, Fig. S5.

We used the 360 sequenced subjects as an internal reference panel to impute sequencing data from the GWAS data for the nonsequenced subjects. Specifically, we used MaCH (7) to construct a reference panel of 720 phased haplotypes consisting of both the variants discovered by exome sequencing and the SNPs on the GWAS arrays. We also prephased haplotypes at the GWAS SNPs for the remaining cohort members. We then used minimac (10) to impute genotypes at the variants discovered by exome sequencing for the nonsequenced subjects. We restricted our attention to nonsense, missense, and splice site variants with MAFs ≤5%. We ended up with a total of 19,135 genes containing at least one polymorphic site and a total of 143,273 variants in those genes. SI Appendix, Fig. S6 shows the Rsq values for the variants whose MAFs are lower than the 5% and

---

**Fig. 1.** Power of the T5 tests at the nominal significance level of 0.001 for the integrative analysis of sequencing and GWAS data based on the robust (T5-rob) and standard variance estimators (T5-std) and for the analysis of sequenced data only (T5-seq). The trait of interest is quantitative. The β values of 0.2, 0.3, and 0.4 correspond to 0.26%, 0.58%, and 1.0% of the trait variance explained by the causal variants. In Q2, Q4, and Q5, the critical values for T5-std were reset to achieve correct type I error. Each power estimate is based on 1,000 replicates.

**Fig. 2.** Power at the nominal significance level of 0.001 for the T5 test based on the robust variance estimator and pre-QC variants (T5-rob, pre-QC), the T5 test based on the standard variance estimator and pre-QC variants (T5-std, pre-QC), and the T5 test based on the standard variance estimator and post-QC variants (T5-std, post-QC) in the integrative analysis of sequencing and GWAS data on the gene OR103. The power of T5 for the analysis of pre-QC variants based on sequenced subjects only (T5-seq, pre-QC) is also included. The trait of interest is quantitative. In Q2, Q4, and Q5, the critical values for T5-std (pre-QC) were reset to achieve correct type I error. Each power estimate is based on 1,000 replicates.
1% thresholds. The variants with MAFs ≤ 1% account for a majority (83.3%) of all variants with MAFs ≤5% and are associated with lower Rsq values than those with MAFs >1%.

We used the log-transformed BMI value as the quantitative trait and included age and the proportion of African ancestry estimated by FRAPPE (frequentist approach for estimating individual ancestry proportion) (25) as covariates. The Pearson correlation coefficients between the ancestry variable and the burden scores of the variants with MAFs ≤5% have the first quartile, median, and third quartile of −0.013, 0.013, and 0.030, respectively. We refer to the burden tests with MAF thresholds of 1% and 5% as T1 and T5, respectively. We constructed the T1, T5, VT, and SKAT tests based on Vrob and Vstd. We also included the tests using only sequenced subjects.

The quantile–quantile plots are displayed in Fig. 3 and SI Appendix, Fig. S7. For all tests based on Vrob, the observed P values agree very well with the global null hypothesis of no association, except at the extreme right tails. By contrast, the observed P values for all tests associated with Vstd show early departures from the global null distribution, reflecting inflation of the type I error. All tests using only sequenced subjects yield less extreme P values than their counterparts in the integrative analysis based on Vrob.

The top 10 genes identified by T5-rob are listed in Table 2. The top gene ODF2L (accession NM_020729) is ranked the second by VT-rob and the fourth by SKAT-rob, and its imputation accuracy is quite high, with an average Rsq of 0.685. The common SNPs in the seventh gene BDNF were previously found common SNPs in the seventh gene BDNF were previously found to be associated with BMI in several GWAS (26, 27). This gene is not in the top 10 list by any test with Vstd.

For any type of test, the version based on Vstd prioritized the top genes differently from the one based on Vrob, as shown by the numerical marks in Fig. 3 and SI Appendix, Fig. S7. A gene identified by a test based on Vstd but not by its counterpart based on Vrob typically has very low average Rsq (i.e., <0.1). For example, the top gene TPSG1 (accession NM_012467) identified by T5-std is not among the top 10 genes by T5-rob and has an average Rsq of only 0.028. Because the tests based on Vstd tend to generate significant results when there are substantial differences in the genotype quality between the sequenced and nonsequenced subjects, the top genes identified by these tests are not reliable.

Applying the postimputation QC procedure of Auer et al. (6) to the WHI data, we found that only 17.1% of variants and 47.9% of genes passed the QC criteria. To be specific, 92.8% of variants with MAFs 3–5% passed QC, 69.6% with MAFs 1–3%, 29.5% with MAFs 0.5–1%, and only 3.5% with MAFs 0.1–0.5%. We repeated our association analysis for the post-QC variants and genes. As shown in SI Appendix, Figs. S8 and S9, the tests with Vstd no longer exhibit inflation of the type I error. However, their top P values are less extreme than those of Vrob without

Table 2. Top 10 genes for BMI identified by T5-rob in the analysis of the WHI data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession</th>
<th>Chr</th>
<th>m</th>
<th>Rsq</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODF2L</td>
<td>NM_020729</td>
<td>1</td>
<td>11</td>
<td>0.685</td>
<td>3.1 × 10⁻⁵</td>
</tr>
<tr>
<td>ITSN1</td>
<td>NM_003024</td>
<td>21</td>
<td>7</td>
<td>0.609</td>
<td>1.7 × 10⁻⁵</td>
</tr>
<tr>
<td>KDM6B</td>
<td>NM_001080424</td>
<td>17</td>
<td>30</td>
<td>0.266</td>
<td>5.0 × 10⁻⁵</td>
</tr>
<tr>
<td>SOCS1</td>
<td>NM_003745</td>
<td>16</td>
<td>2</td>
<td>0.348</td>
<td>5.8 × 10⁻⁵</td>
</tr>
<tr>
<td>ODF2L</td>
<td>NM_001007022</td>
<td>1</td>
<td>9</td>
<td>0.689</td>
<td>7.8 × 10⁻⁵</td>
</tr>
<tr>
<td>ACADVL</td>
<td>NM_000018</td>
<td>17</td>
<td>15</td>
<td>0.189</td>
<td>1.1 × 10⁻⁴</td>
</tr>
<tr>
<td>BDNF</td>
<td>NM_170734</td>
<td>11</td>
<td>2</td>
<td>0.628</td>
<td>1.6 × 10⁻⁴</td>
</tr>
<tr>
<td>TRDMT1</td>
<td>NM_004412</td>
<td>10</td>
<td>3</td>
<td>0.718</td>
<td>2.1 × 10⁻⁶</td>
</tr>
<tr>
<td>FAM60A</td>
<td>NM_001135811</td>
<td>12</td>
<td>2</td>
<td>0.768</td>
<td>2.3 × 10⁻⁴</td>
</tr>
<tr>
<td>PDGFRA</td>
<td>NM_006206</td>
<td>4</td>
<td>12</td>
<td>0.563</td>
<td>2.4 × 10⁻⁴</td>
</tr>
</tbody>
</table>

Chr is the chromosome number, m is the number of variants in the gene, and Rsq is the Rsq value averaged over the variant sites in the gene. T5-rob and T5-std are the T5 tests in the integrative analysis of sequencing and GWAS data with the robust and standard variance estimators, respectively, and T5-seq is the T5 test using only sequenced subjects.

Fig. 3. Quantile–quantile plots of −log₁₀(P values) for the T1 and T5 tests in the analysis of the BMI data in the WHI. (Left) The top five genes identified by T1-rob are marked as 1–5. (Right) The top five genes identified by T5-rob are marked as 1–5.
QC. We list the top 10 genes identified by T5-std in SI Appendix, Table S1, which can be compared with Table 2. We see that the association signal for BDNF is weaker after QC. None of the other genes in SI Appendix, Table S1 has previously been found to be associated with BMI.

Discussion
This article provides a valid and efficient approach to integrative analysis of sequencing and GWAS data. The approach is very general in several aspects: (i) It can handle any type of trait and any sampling scheme; (ii) it encompasses all commonly used gene-level tests for rare variants; (iii) it includes single-variant tests as a special case; and (iv) it allows for covariates. The computation is the same as the usual gene-level tests except for the replacement of the standard variance estimator with the robust one. The proposed methods have been implemented in the software program SEQGWAS (integrative analysis of SEQuencing and GWAS data), which is freely available at web1.sph.emory.edu/users/yha30/software.html. It took ~2 h on an IBM HS22 machine to analyze the WHI data.

For binary traits, the GWAS subjects may come from a cohort or case-control study. The standard variance estimator is valid if the case–control ratios are the same between the sequencing study and the parent study. This condition typically holds when the parent study is a case–control study but likely fails when the parent study is a cohort study. It is also desirable to integrate other types of genetic data. For example, a subset of the WHI African Americans was genotyped on Metabochip (28) and used as the reference panel to impute Metabochip variants for the remaining African Americans with GWAS data (29). In addition, the GWAS genotyping arrays can be replaced by the Metabochip (28) or the Exomechip (30), which can then be imputed against sequencing data (28). Our approach can be readily used to analyze such mixtures of observed and imputed data.

Our approach is built upon existing imputation algorithms and focused on properly analyzing a mixture of observed and imputed genotype data. Thus, we can take full advantage of newly developed imputation programs to improve imputation accuracy and computational efficiency. Although imputed data may pertain to the genotype dosage, the most likely genotype, or the genotype probabilities, our approach is tailored to the dosage data. In practice, the dosage is the most commonly used because, unlike the most likely genotype, it partially accounts for the uncertainty in the imputed value and because the dosage is computationally more tractable than the genotype probabilities while still retaining most information.

The postimputation QC process can alleviate the inflation of the type I error caused by the use of $V_{out}$. However, this strategy tends to remove many rare variants. The removal of variants results in loss of important information because the observed genotype data for sequenced subjects cannot be used if a variant is excluded due to low imputation quality and because the imputed genotypes, even for a variant with low Rsq, may still contain valuable information about the association. Our approach does not exclude any variants with low Rsq. The imputed values for those variants have small variances and do not add much noise. When Rsq is close to zero, the integrative analysis reduces to the analysis of sequenced subjects only and thus can at least use the information in sequenced subjects. Our approach can certainly be applied to data after QC, but the QC criteria can be much less stringent.

ACKNOWLEDGMENTS. The authors thank the WHI investigators and staff for their dedication and the study participants for making the program possible. The authors thank two referees for their helpful comments.

Data from the Exome Sequencing Project were supported through NHLBI RC2 HL-102924, RC2 HL-102925, and RC2 HL-102926. This work was supported by NIH Grants R01CA082659, P01CA142538, R37GM047845, R37GM073052, and R01HG006703. The WHI program is funded by the NHLBI (HHSN268201100004C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C).

Supporting Information

Hu et al.

SI Methods

A. Properties of Imputed Values

Assume that the imputation is performed independently of the trait and covariates. The commonly used imputation algorithms\(^1\text{-}^5\) share the following property: at a variant site, the imputed value \(\tilde{G}\) for a non-sequenced subject given the true genotype \(G\) and the set of genotypes observed on sequenced subjects, \(G_{\text{seq}} = \{G_1, \ldots, G_n\}\), has the expectation

\[
E(\tilde{G}|G, G_{\text{seq}}) = \overline{G}_{\text{seq}} + r^2 (G - \overline{G}_{\text{seq}}),
\]

(S1)

where \(\overline{G}_{\text{seq}} = n^{-1} \sum_{i=1}^{n} G_i\). Note that \(\tilde{G}\) can be uniquely determined by \(G, G_{\text{seq}}\), and the observed genotypes at flanking GWAS SNPs for the non-sequenced subject and all sequenced subjects. Here, we are interested in \(E(\tilde{G}|G, G_{\text{seq}})\), which is the imputed value averaged over the genotypes of the flanking SNPs. To verify equation (S1), we use the fact that the mean of the imputed genotype is linearly related to the true genotype given the sequenced genotypes, as illustrated in Figure S10. We write the linear relationship as

\[
E(\tilde{G}|G, G_{\text{seq}}) = a + bG,
\]

where \(a\) and \(b\) are determined as follows. First, by applying the linear relationship to the definition of the Pearson correlation coefficient, we obtain

\[
\rho^2 = \frac{\text{Cov}(\tilde{G}, G)^2}{\text{Var}(\tilde{G})\text{Var}(G)} = b^2 \frac{\text{Var}(G)}{\text{Var}(\tilde{G})}.\]

Since \(\text{Rsq} = \frac{\text{Var}(\tilde{G})}{\text{Var}(G)}\), we conclude that \(b = r^2\). It is reasonable to assume that the unconditional mean \(E(\tilde{G})\) does not depend on \(r^2\), so we can rewrite \(a + r^2 G\) as \(a^* + r^2 (G - \nu)\), where \(a^*\) does not involve \(r^2\), and \(E(G - \nu) = 0\). There are two constraints: when the variant cannot be imputed with any accuracy (i.e., \(r^2 = 0\)), we have \(E(\tilde{G}|G, G_{\text{seq}}) = \overline{G}_{\text{seq}}\); when the variant is perfectly imputed (i.e., \(r^2 = 1\)), we have \(\tilde{G} = G\) and thus \(E(\tilde{G}|G, G_{\text{seq}}) = G\). These two constraints imply that \(a^* = \nu = \overline{G}_{\text{seq}}\), so that equation (S1) holds.

We can use the arguments of Browning and Browning\(^5\) to show that \(\rho^2 = \text{Rsq}\) under accurate calibration of the imputed posterior probabilities. Specifically, we define \(Z = (Z_0, Z_1, Z_2)\)
with \( Z_g \) representing the posterior probability of genotype \( g \) \((g = 0, 1, 2)\) and assume that the posterior probabilities are accurately calibrated such that \( \Pr(G = g | Z) = Z_g \). It follows that \( \text{E}(G|Z) = \tilde{G} \) and \( \text{Cov}(\tilde{G}, G) = \text{E}\{\tilde{G}\text{E}(G|Z)\} - \text{E}(\tilde{G})\text{E}(\text{E}(G|Z)) = \text{E}(\tilde{G}^2) - \{\text{E}(\tilde{G})\}^2 = \text{Var}(\tilde{G}) \). Setting \( \text{Cov}(\tilde{G}, G) \) in the expression of \( \rho \) to \( \text{Var}(\tilde{G}) \), we obtain \( \rho^2 = \text{Rsq} \), such that \( r^2 = \text{Rsq} \).

We numerically verified equation (S1) for MaCH.\(^1\) We first generated genotype data for 1,000 subjects for variants discovered by exome sequencing on chromosome 21, as well as flanking SNPs from the GWAS arrays. We adopted GWASimulator\(^6\) and used the 720 phased haplotypes from the sequenced WHI subjects as the templates. We then masked the genotypes of sequenced variants for 500 subjects and imputed them by MaCH\(^1\) using the other 500 subjects as the reference panel. Figure S10 compares the imputed genotypes with the true genotypes for ten variants with a wide range of Rsq values, and Figure S11 compares the imputed genotypes with the expected genotypes given by equation (S1) (with \( r^2 \) replaced by the sample Rsq) for the same set of variants. Clearly, the means of the imputed values agree very well with the expected values. Indeed, the linear regression fit for each variant coincides with the diagonal line except for the top left plot, indicating that equation (S1) is well satisfied at modest to large values of Rsq. When Rsq is extremely low, as in the top left plot, equation (S1) automatically holds. Figure S12 provides an overview of imputed variants on chromosome 21 binned into ten Rsq intervals. The linear regression fit in each bin coincides with the diagonal line.

Equation (S1) implies that, for a sequenced subject \( s \) and a non-sequenced subject \( u \), \( \text{E}(\tilde{G}_u) = (1 - r^2)\text{E}(G_s) + r^2\text{E}(G_u) \). Although sequenced subjects may not be a random sample of the GWAS cohort, the distributions of the true genotypes are the same between sequenced and non-sequenced subjects under \( H_0 \) and independence of \( G \) and \( X \), such that \( \text{E}(G_s) = \text{E}(G_u) \). It follows that

\[
\text{E}(\tilde{G}_s) = \text{E}(\tilde{G}_u),
\]

which is referred to as \( \text{E}(\tilde{G}) \).

**B. Properties of the Standard Variance Estimator**
The robust variance estimator $V_{rob}$ in equation (1), which characterizes the true variability of $U$, can be written as

$$\sum_{i=1}^{n} \{Y_i - \bar{Y}_{seq} + r^2(\bar{Y}_{seq} - \bar{Y})\}^2 (\tilde{G}_i - \bar{G})^2 + \sum_{i=n+1}^{N} \{Y_i - \bar{Y}_{imp} + \bar{Y}_{imp} - \bar{Y}\}^2 (\tilde{G}_i - \bar{G})^2,$$

where $Y_{imp} = (N - n)^{-1} \sum_{i=n+1}^{N} Y_i$. Let $\mu$, $\mu_{seq}$ and $\mu_{imp}$ be the limits of $Y$, $Y_{seq}$ and $Y_{imp}$, respectively. Then $V_{rob}$ is approximately

$$n \{\text{Var}(Y_s) + r^4(\mu_{seq} - \mu)^2\} \text{Var}(\tilde{G}_s) + (N - n) \{\text{Var}(Y_u) + (\mu_{imp} - \mu)^2\} \text{Var}(\tilde{G}_u).$$

(S3)

The standard variance estimator $V_{std}$ can be written as

$$N^{-1} \left[ \sum_{i=1}^{n} \{Y_i - \bar{Y}_{seq}\}^2 + (\bar{Y}_{seq} - \bar{Y})^2 \right] + \sum_{i=n+1}^{N} \{Y_i - \bar{Y}_{imp}\}^2 + (\bar{Y}_{imp} - \bar{Y})^2 \right]$$

$$\times \left\{ \sum_{i=1}^{n} (\tilde{G}_i - \bar{G})^2 + \sum_{i=n+1}^{N} (\tilde{G}_i - \bar{G})^2 \right\},$$

which is approximately

$$N^{-1} \left[ n \{\text{Var}(Y_s) + (\mu_{seq} - \mu)^2\} + (N - n) \{\text{Var}(Y_u) + (\mu_{imp} - \mu)^2\} \right]$$

$$\times \left\{ n\text{Var}(\tilde{G}_s) + (N - n)\text{Var}(\tilde{G}_u) \right\}. \quad \text{(S4)}$$

Under perfect imputation (i.e., $r^2 = 1$ and $\text{Var}(\tilde{G}_s) = \text{Var}(\tilde{G}_u)$) or random sampling (i.e., $\mu_{seq} = \mu_{imp} = \mu$ and $\text{Var}(Y_s) = \text{Var}(Y_u)$), expression (S4) is the same as expression (S3). We show below that (S4) is generally smaller than (S3) in other scenarios.

We first consider the case where $\mu_{seq} = \mu_{imp} = \mu$ and $\text{Var}(Y_s) > \text{Var}(Y_u)$. For quantitative traits with extreme-trait sampling, this means that the sampling is balanced between the two extremes. In this case, (S3) and (S4) reduce to

$$n\text{Var}(Y_s)\text{Var}(\tilde{G}_s) + (N - n)\text{Var}(Y_u)\text{Var}(\tilde{G}_u)$$

(S5)

and

$$N^{-1} \{n\text{Var}(Y_s) + (N - n)\text{Var}(Y_u)\} \left\{ n\text{Var}(\tilde{G}_s) + (N - n)\text{Var}(\tilde{G}_u) \right\}, \quad \text{(S6)}$$

3
respectively. Suppose that \( \text{Var}(\tilde{G}_s) > \text{Var}(\tilde{G}_u) \) (i.e., under imperfect imputation). It then follows from Chebyshev’s sum inequality that (S6) is smaller than (S5), so \( V_{\text{std}} \) underestimates the variance of \( U \).

We now consider the case where \( \mu_{\text{seq}} \neq \mu \). We assume that the posterior probabilities are accurately calibrated such that \( r^2 = \text{Rsq} \) and \( \text{Var}(\tilde{G}_u) = r^2 \text{Var}(\tilde{G}_s) \). Then the subtraction of (S4) from (S3) yields

\[
D = \left[ n(\mu_{\text{seq}} - \mu)^2 r^4 - nN^{-1}(N - n) \left\{ \text{Var}(Y_s) + (\mu_{\text{seq}} - \mu)^2 - \text{Var}(Y_u) - (\mu_{\text{imp}} - \mu)^2 \right\} r^2 
+ nN^{-1}(N - n) \left\{ \text{Var}(Y_s) + (\mu_{\text{seq}} - \mu)^2 - \text{Var}(Y_u) - (\mu_{\text{imp}} - \mu)^2 \right\} - n(\mu_{\text{seq}} - \mu)^2 \right] \text{Var}(\tilde{G}_s).
\]

The term inside the square brackets is a quadratic function of \( r^2 \), whose global minimum is reached at

\[
r_{\text{min}}^2 = (N - n) \left\{ \text{Var}(Y_s) + (\mu_{\text{seq}} - \mu)^2 - \text{Var}(Y_u) - (\mu_{\text{imp}} - \mu)^2 \right\} / \left\{ 2N(\mu_{\text{seq}} - \mu)^2 \right\}.
\]

Because \( D = 0 \) at \( r^2 = 1 \), we conclude that \( D > 0 \) for \( r^2 \in [0, 1) \) if and only if \( r_{\text{min}}^2 \geq 1 \). This condition simplifies to \( \text{Var}(Y_s) - \text{Var}(Y_u) \geq \{N/(N - n)\}^2 (\mu_{\text{seq}} - \mu)^2 \) because of the fact that \( n\mu_{\text{seq}} + (N - n)\mu_{\text{imp}} = N\mu \). For quantitative traits, this condition is satisfied if the sequenced sample has a sufficiently larger trait variance than the non-sequenced sample. For binary traits, this condition becomes \( (\mu_{\text{seq}} - \mu_{\text{imp}})(1 - \mu_{\text{seq}} - \mu_{\text{imp}}) \geq (\mu_{\text{seq}} - \mu_{\text{imp}})^2 \) because \( \text{Var}(Y_s) = \mu_{\text{seq}}(1 - \mu_{\text{seq}}) \), \( \text{Var}(Y_u) = \mu_{\text{imp}}(1 - \mu_{\text{imp}}) \), and \( n\mu_{\text{seq}} + (N - n)\mu_{\text{imp}} = N\mu \). If \( \mu_{\text{seq}} > \mu_{\text{imp}} \) or equivalently \( \mu_{\text{seq}} > \mu \), we obtain \( \mu_{\text{seq}} \leq 0.5 \). If \( \mu_{\text{seq}} < \mu_{\text{imp}} \) or equivalently \( \mu_{\text{seq}} < \mu \), we have \( \mu_{\text{seq}} \geq 0.5 \). In other words, for the case-control sequencing study with an equal number of cases and controls (i.e., \( \mu_{\text{seq}} = 0.5 \)), \( V_{\text{std}} \) will underestimate the variance of \( U \) unless the disease rate in the cohort happens to be 50%.

**C. Details of the Proposed Methods**

We now extend our methodology to an arbitrary trait with multiple variants in a gene and with covariates. Suppose that we are interested in \( m \) rare variants within a gene. Their
genotypes are represented by \( G = (G_1, \ldots, G_m)^T \), where \( G_j \) is the genotype of the \( j \)th variant. Write \( \tilde{G} = (\tilde{G}_1, \ldots, \tilde{G}_m)^T \), where \( \tilde{G}_j \) is equal to \( G_j \) if it is observed and to the imputed value otherwise. Let \( X \) denote a set of covariates, including the unit component. The selection of subjects for sequencing may depend on \( Y \) and \( X \). We specify the conditional density of \( Y \) given \( G \) and \( X \) through the generalized linear model:

\[
\exp \left\{ \frac{Y \left( \beta^T G + \gamma^T X \right) - b(\beta^T G + \gamma^T X) \theta}{a(\phi)} + c(Y, \phi) \right\},
\]

where \( \beta \) and \( \gamma \) are vectors of unknown regression parameters, and \( a(.) \), \( b(.) \), and \( c(., .) \) are specific functions. Define \( b'(\theta) = db(\theta)/d\theta \) and \( b''(\theta) = d^2 b(\theta)/d\theta^2 \). For the linear model, \( a(\phi) = \sigma^2 \), \( b(\theta) = \theta^2/2 \), \( b'(\theta) = \theta \), and \( b''(\theta) = 1 \). For the logistic regression model, \( a(\phi) = 1 \), \( b(\theta) = \log(1 + e^\theta) \), \( b'(\theta) = e^\theta/(1 + e^\theta) \), and \( b''(\theta) = e^\theta/(1 + e^\theta)^2 \).

The score vector for testing the null hypothesis \( H_0 : \beta = 0 \) based on the data \((Y_i, X_i, \tilde{G}_i)\) \((i = 1, \ldots, N)\) is

\[
U = \sum_{i=1}^{N} \{Y_i - b'(\tilde{\gamma}^T X_i)\} \tilde{G}_i,
\]

where \( \tilde{G}_i \) is the value of \( \tilde{G} \) for the \( i \)th subject, and \( \tilde{\gamma} \) is the restricted maximum likelihood estimation (MLE) of \( \gamma \) under \( H_0 \). The standard variance estimator for \( U \) based on the Fisher information matrix is

\[
V_{\text{std}} = a(\hat{\phi}) \sum_{i=1}^{N} b''(\tilde{\gamma}^T X_i) \left( \tilde{G}_i \tilde{G}_i^T - \tilde{\zeta}^T X_i \tilde{G}_i^T \right),
\]

where \( \hat{\phi} \) is the restricted MLE of \( \phi \), and \( \tilde{\zeta} = \left\{ \sum_{i=1}^{N} b''(\tilde{\gamma}^T X_i) X_i X_i^T \right\}^{-1} \left\{ \sum_{i=1}^{N} b''(\tilde{\gamma}^T X_i) X_i \tilde{G}_i^T \right\} \).

We wish to prove that \( \text{E}(U) = 0 \) under \( H_0 \). Note that the expectation is conditional on the sampling scheme that the first \( n \) subjects are selected for sequencing and the other \((N-n)\) subjects are not. The Taylor series expansion of \( U \) at \( \gamma \) yields the asymptotic approximation

\[
U = \sum_{i=1}^{N} \epsilon_i (\tilde{G}_i - \zeta^T X_i),
\]

where \( \epsilon_i = Y_i - b'(\gamma^T X_i) \), and

\[
\zeta = \left[ \text{E} \left\{ b''(\gamma^T X) XX^T \right\} \right]^{-1} \left[ \frac{n}{N} \text{E} \left\{ b''(\gamma^T s) X_s \tilde{G}_s^T \right\} + \frac{(N-n)}{N} \text{E} \left\{ b''(\gamma^T u) X_u \tilde{G}_u^T \right\} \right].
\]
We assume that $G$ is independent of $X$, which implies that $\tilde{G}$ is independent of $X$ in both the sequenced and non-sequenced samples. Clearly,

$$
\frac{n}{N} E\{b''(\gamma^T X_u) X_u\} + \frac{(N-n)}{N} E\{b''(\gamma^T X_u) X_u\} = E\{b''(\gamma^T X)\}.
$$

Write $X = (1, \tilde{X}^T)^T$. Then

$$
\zeta^T X = E(\tilde{G}) E\left\{b''(\gamma^T X) \begin{pmatrix} 1 & \tilde{X}^T \end{pmatrix}\right\} \left[E\left\{b''(\gamma^T X) \begin{pmatrix} 1 & \tilde{X}^T \end{pmatrix}\right\}\right]^{-1} \begin{pmatrix} 1 \\ \tilde{X} \end{pmatrix} = E(\tilde{G}) \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{X} \end{pmatrix},
$$

such that

$$
\zeta^T X = E(\tilde{G}). \tag{S7}
$$

Under $H_0$ and independence of $G$ and $X$, $\tilde{G}$ is independent of $X$ and $Y$ in both the sequenced and non-sequenced samples, such that $E(\tilde{G}_s|\mathcal{Y}, \mathcal{X}) = E(\tilde{G})$ and $E(\tilde{G}_u|\mathcal{Y}, \mathcal{X}) = E(\tilde{G})$, where $\mathcal{Y}$ and $\mathcal{X}$ represent the data sets $\{Y_1, \ldots, Y_N\}$ and $\{X_1, \ldots, X_N\}$, respectively. Consequently,

$$
E(U|\mathcal{Y}, \mathcal{X}) = \sum_{i=1}^{N} \epsilon_i \{E(\tilde{G}) - E(\tilde{G})\} = 0, \tag{S8}
$$

which holds under any sampling scheme and is a stronger result than $E(U) = 0$.

To derive a robust variance estimator for $U$ under $H_0$, we calculate the variance of $U$ conditional on $(\mathcal{Y}, \mathcal{X})$ so as to account for possible trait-dependent sampling, treat sequenced and non-sequenced subjects separately, and account for the dependence of imputed genotypes on observed genotypes $G_{seq} = \{\tilde{G}_{ji}; 1 \leq j \leq m, 1 \leq i \leq n\}$, which is redefined to include all $m$ variants. We first decompose $\text{Var}(U)$ as $E\{\text{Var}(U|\mathcal{Y}, \mathcal{X})\} + \text{Var}\{E(U|\mathcal{Y}, \mathcal{X})\}$, which is further expanded as

$$
E[\text{Var}(U|\mathcal{Y}, \mathcal{X}, G_{seq})|\mathcal{Y}, \mathcal{X}] + \text{Var}\{E(U|\mathcal{Y}, \mathcal{X}, G_{seq})|\mathcal{Y}, \mathcal{X}\} + \text{Var}\{E(U|\mathcal{Y}, \mathcal{X})\}.
$$

The first variance term $\text{Var}(U|\mathcal{Y}, \mathcal{X}, G_{seq})$ corresponds to the variation induced by imputation. In light of equation (S7), this term becomes

$$
\sum_{i=n+1}^{N} \epsilon_i^2 \text{Var}(\tilde{G}_i|\mathcal{Y}, \mathcal{X}, G_{seq}). \tag{S9}
$$
Under $H_0$ and independence of $G$ and $X$, $\widetilde{G}$ is independent of $X$ and $Y$ in the non-sequenced sample, such that $\mathcal{Y}$ and $\mathcal{X}$ can be omitted. Applying equation (S1) to the $j$th variant, for a non-sequenced subject $i$,

$$E(\tilde{G}_{ji}|G_{j, \text{seq}}) = (1 - r_j^2)\overline{G}_{j, \text{seq}} + r_j^2E(G_{ji}|G_{j, \text{seq}}),$$  \hspace{1cm} (S10)$$

where $\tilde{G}_{ji}$ and $G_{ji}$ are the imputed and true genotypes for the $j$th variant of the $i$th subject, $\overline{G}_{j, \text{seq}} = n^{-1}\sum_{i=1}^{n}G_{ji}$, $G_{j, \text{seq}} = \{G_{ji}; 1 \leq i \leq n\}$, and $r_j^2$ is the value of $r^2$ for the $j$th variant. Note that $E(G_{ji}|G_{j, \text{seq}}) = E(G_{ji})$ due to the independence of sequenced and non-sequenced subjects. Because the distributions of $G$ are the same between sequenced and non-sequenced subjects under $H_0$ and independence of $G$ and $X$, we estimate $E(G_{ji})$ for a non-sequenced subject $i$ by $\overline{G}_{j, \text{seq}}$. Thus, we estimate (S10) by $\overline{G}_{j, \text{seq}}$ and (S9) by $\sum_{i=n+1}^{N}\tilde{e}_i^2(\tilde{G}_i - \overline{G}_{\text{seq}})(\tilde{G}_i - \overline{G}_{\text{seq}})^T$, where $\tilde{e}_i = Y_i - b(\hat{\gamma}^TX_i)$, and $\overline{G}_{\text{seq}} = (\overline{G}_{1, \text{seq}}, \ldots, \overline{G}_{m, \text{seq}})^T$. The second variance term $\text{Var}\{E(U|\mathcal{Y}, \mathcal{X}, G_{\text{seq}})|\mathcal{Y}, \mathcal{X}\}$ pertains to the variability of observed genotypes. By equation (S10), this term becomes

$$\text{Var} \left\{ \sum_{i=1}^{n} \left[ \tilde{e}_i \tilde{G}_{1i} \right] + \sum_{i=n+1}^{N} \left[ \tilde{e}_i(1 - r_i^2)\overline{G}_{1, \text{seq}} \right] \right\} \text{Var} \{\mathcal{Y}, \mathcal{X}\} \hspace{1cm} (S11)$$

Because $X$ includes the unit component, $\sum_{i=n+1}^{N}\tilde{e}_i = -\sum_{i=1}^{n}\tilde{e}_i$. Let

$$S_i^* = \begin{bmatrix} \{\hat{\epsilon}_i - (1 - r_1^2)\overline{G}_{\text{seq}}\}(\tilde{G}_{1i} - \overline{G}_{1, \text{seq}}) \\ \vdots \\ \{\hat{\epsilon}_i - (1 - r_m^2)\overline{G}_{\text{seq}}\}(\tilde{G}_{mi} - \overline{G}_{m, \text{seq}}) \end{bmatrix},$$

where $\overline{G}_{\text{seq}} = n^{-1}\sum_{i=1}^{n}\hat{e}_i$. Then (S11) can be estimated by $\sum_{i=1}^{n}S_i^*S_i^{*T}$. The third variance term $\text{Var}\{E(U|\mathcal{Y}, \mathcal{X})\}$ is zero by equation (S8). In summary, $\text{Var}(U)$ can be estimated by

$$\sum_{i=1}^{n}S_i^*S_i^{*T} + \sum_{i=n+1}^{N}\tilde{e}_i^2(\tilde{G}_i - \overline{G}_{\text{seq}})(\tilde{G}_i - \overline{G}_{\text{seq}})^T. \hspace{1cm} (S12)$$
Because of (S7), we replace $G_{j,seq}$ and $G_{seq}$ in (S12) by $\hat{\zeta}_j^T X_i$ and $\hat{\zeta}^T X_i$, respectively, to obtain

$$V_{rob} = \sum_{i=1}^{n} S_i S_i^T + \sum_{i=n+1}^{N} \hat{\epsilon}_i^2 (\hat{G}_i - \hat{\zeta}_j^T X_i) (\hat{G}_i - \hat{\zeta}^T X_i)^T,$$

(S13)

where

$$S_i = \begin{bmatrix}
\{ \hat{\epsilon}_i - (1 - r_j^2) \bar{e}_{seq} \} (\hat{G}_{1i} - \hat{\zeta}_1^T X_i) \\
\vdots \\
\{ \hat{\epsilon}_i - (1 - r_m^2) \bar{e}_{seq} \} (\hat{G}_{mi} - \hat{\zeta}_m^T X_i)
\end{bmatrix},$$

and $\hat{\zeta}_j$ is the $j$th column of $\hat{\zeta}$. Equation (S13) is preferable to equation (S12) because the former coincides with the empirical variance estimator under certain circumstances. In particular, if $r_j^2 = 1$ for all $j$, then $V_{rob}$ in equation (S13) reduces to the empirical variance estimator

$$\sum_{i=1}^{N} \hat{\epsilon}_i^2 (\hat{G}_i - \hat{\zeta}_j^T X_i) (\hat{G}_i - \hat{\zeta}^T X_i)^T.$$

Because the empirical variance estimator is valid in this case, the robust variance estimator is valid even when $G$ and $X$ are strongly correlated.

For binary traits, equation (S13) is unstable for low disease rates because the small number of subjects who have the disease tend to dominate the summation in (S13). To fix this problem, we take the expectation of (S9) with respect to $\{Y_{n+1}, \ldots, Y_N\}$ conditional on $X$ and the sampling scheme to obtain the approximation $\sum_{i=n+1}^{N} \left[ \{1 - b'(\hat{\gamma}^T X_i)\}^2 \mu_{imp} + \{b'(\hat{\gamma}^T X_i)\}^2 (1 - \mu_{imp}) \right] \text{Var}(\tilde{G}_i|G_{seq})$, where $\mu_{imp}$ now means the disease rate among non-sequenced subjects. Likewise, we take the expectation of (S11) with respect to $\{Y_1, \ldots, Y_n\}$ conditional on $X$ and the sampling scheme. Then we derive their estimators as before. We obtain a modified robust variance estimator as

$$V_{rob} = \sum_{i=1}^{n} \{ S_{i,1} S_{i,1}^T \bar{Y}_{seq} + S_{i,0} S_{i,0}^T (1 - \bar{Y}_{seq}) \}$$

$$+ \sum_{i=n+1}^{N} \left[ \{1 - b'(\hat{\gamma}^T X_i)\}^2 \bar{Y}_{imp} + \{b'(\hat{\gamma}^T X_i)\}^2 (1 - \bar{Y}_{imp}) \right] (\tilde{G}_i - \hat{\zeta}_j^T X_i) (\tilde{G}_i - \hat{\zeta}^T X_i)^T,$$

(S14)

where $S_{i,1}$ and $S_{i,0}$ are the values of $S_i$ at $Y_i = 1$ and $Y_i = 0$, respectively. Note that $\bar{Y}_{seq}$ and $\bar{Y}_{imp}$ estimate the probabilities of being a case among sequenced and non-sequenced subjects,
respectively. The modified $V_{\text{rob}}$ in equation (S14) is asymptotically equivalent to (S13) but is more stable in small samples.

To perform meta-analysis, we sum the score vectors from the participating studies to obtain the overall score vector and sum the corresponding robust covariance matrices to obtain the overall covariance matrix. We then construct gene-level tests in the same manner as before. This meta-analysis is equivalent to the joint analysis of individual participant data of all studies and is thus statistically optimal. The score statistic $U$ should be normalized by $a(\hat{\phi})$. For analysis of a single study, this factor is omitted because it cancels between the numerator and denominator of the test statistic. For meta-analysis, we multiply $U$ and $V_{\text{rob}}$ of each study by its own $a(\hat{\phi})^{-1}$ and $a(\hat{\phi})^{-2}$, respectively, before taking the summations; otherwise, the meta-analysis will no longer be optimal (unless $\phi$ is the same in all studies).


Table S1. Top ten genes for BMI identified by T5-std in the analysis of the WHI data after post-imputation QC

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession</th>
<th>Chr</th>
<th>m</th>
<th>Rsq</th>
<th>p-value</th>
<th>T5-std</th>
<th>T5-rob</th>
<th>T5-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODF2L</td>
<td>NM_001007022</td>
<td>1</td>
<td>2</td>
<td>0.941</td>
<td>9.4×10⁻²</td>
<td>1.5×10⁻⁴</td>
<td>5.4×10⁻⁵</td>
<td>9.4×10⁻²</td>
</tr>
<tr>
<td>MRGPRX3</td>
<td>NM_054031</td>
<td>11</td>
<td>3</td>
<td>0.882</td>
<td>3.1×10⁻⁴</td>
<td>1.7×10⁻⁴</td>
<td>1.5×10⁻¹</td>
<td>1.7×10⁻⁴</td>
</tr>
<tr>
<td>TRDMT1</td>
<td>NM_004412</td>
<td>10</td>
<td>2</td>
<td>0.891</td>
<td>7.5×10⁻²</td>
<td>2.8×10⁻⁴</td>
<td>3.6×10⁻⁴</td>
<td>7.0×10⁻¹</td>
</tr>
<tr>
<td>ITSN1</td>
<td>NM_003024</td>
<td>21</td>
<td>3</td>
<td>0.943</td>
<td>7.0×10⁻¹</td>
<td>2.8×10⁻⁴</td>
<td>3.6×10⁻⁴</td>
<td>7.0×10⁻¹</td>
</tr>
<tr>
<td>FANK1</td>
<td>NM_145235</td>
<td>10</td>
<td>2</td>
<td>0.792</td>
<td>4.5×10⁻¹</td>
<td>8.0×10⁻⁴</td>
<td>1.9×10⁻⁴</td>
<td>4.5×10⁻¹</td>
</tr>
<tr>
<td>BDNF</td>
<td>NM_001143805</td>
<td>11</td>
<td>1</td>
<td>0.999</td>
<td>9.9×10⁻²</td>
<td>3.3×10⁻⁴</td>
<td>2.8×10⁻⁴</td>
<td>9.9×10⁻²</td>
</tr>
<tr>
<td>CYP3A4</td>
<td>NM_017460</td>
<td>7</td>
<td>2</td>
<td>0.910</td>
<td>4.0×10⁻¹</td>
<td>3.5×10⁻⁴</td>
<td>1.9×10⁻⁴</td>
<td>4.0×10⁻¹</td>
</tr>
<tr>
<td>FAM60A</td>
<td>NM_001135811</td>
<td>12</td>
<td>1</td>
<td>0.768</td>
<td>6.5×10⁻¹</td>
<td>4.0×10⁻⁴</td>
<td>2.3×10⁻⁴</td>
<td>6.5×10⁻¹</td>
</tr>
<tr>
<td>MAP1B</td>
<td>NM_005909</td>
<td>5</td>
<td>6</td>
<td>0.815</td>
<td>7.1×10⁻²</td>
<td>4.5×10⁻⁴</td>
<td>5.6×10⁻⁴</td>
<td>7.1×10⁻²</td>
</tr>
<tr>
<td>FXC1</td>
<td>NM_012192</td>
<td>11</td>
<td>2</td>
<td>0.911</td>
<td>4.0×10⁻¹</td>
<td>5.1×10⁻⁴</td>
<td>7.8×10⁻⁴</td>
<td>4.0×10⁻¹</td>
</tr>
</tbody>
</table>

Chr is the chromosome number, m is the number of variants in the gene, and Rsq is the Rsq value averaged over the variants in the gene. T5-std and T5-rob are the T5 tests with the standard and robust variance estimators, respectively, and T5-seq is the T5 test using only sequenced subjects.
Figure S1. Power of the T5 tests at the nominal significance level of 0.001 for the integrative analysis of sequencing and GWAS data based on the robust (T5-rob) and standard variance estimators (T5-std) and for the analysis of sequenced data only (T5-seq). The trait of interest is binary. In B2, B4 and B5, the critical values for T5-std were reset to achieve correct type I error. Each power estimate is based on 1,000 replicates.
Figure S2. Type I error of the T5 test with the robust variance estimator at the nominal significance level of 0.001 for quantitative traits (upper panel) and binary traits (lower panel) when the burden score and covariate are correlated. Corr is the Pearson correlation coefficient between the burden score and covariate. Each type I error estimate is based on 100,000 replicates.
Figure S3. Power at the nominal significance level of 0.001 for the T5 test based on the robust variance estimator and pre-QC variants (T5-rob, pre-QC), the T5 test based on the standard variance estimator and pre-QC variants (T5-std, pre-QC), and the T5 test based on the standard variance estimator and post-QC variants (T5-std, post-QC) in the integrative analysis of sequencing and GWAS data on the gene OR10J3. The power of T5 for the analysis of pre-QC variants based on sequenced subjects only (T5-seq, pre-QC) is also included. The trait of interest is binary. In B2, B4 and B5, the critical values for T5-std (pre-QC) were reset to achieve correct type I error. Each power estimate is based on 1,000 replicates.
Figure S4. Quantile-quantile plots of $-\log_{10}(p$-values) for the T5 tests of all genes on chromosome 1 when the disease status was simulated with a disease rate of 10%. The left and right sides pertain to the null hypothesis and alternative hypothesis (with five causal genes), respectively. On the right side, the five causal genes are marked as 1–5. Only one causal gene appears in the bottom right plot because the others were filtered out by the QC procedure. All $p$-values smaller than $1.0 \times 10^{-5}$ are truncated.
Figure S5. Distributions of original BMI values (upper panel) and log-transformed BMI values (lower panel) for sequenced subjects (left side) and all subjects (right side) in the WHI African Americans.
Figure S6. Distributions of Rsq values for variants with MAFs ≤ 5% and ≤ 1% in the WHI data.
Figure S7. Quantile-quantile plots of $-\log_{10}(p$-values) for the VT and SKAT tests in the analysis of the BMI data in the WHI. On the left side, the top five genes identified by VT-rob are marked as 1–5. On the right side, the top five genes identified by SKAT-rob are marked as 1–5.
Figure S8. Quantile-quantile plots of $-\log_{10}(p$-values) for the T1 and T5 tests in the analysis of the BMI data in the WHI after post-imputation QC. On the left side, the top five genes identified by T1-rob are marked as 1–5. On the right side, the top five genes identified by T5-rob are marked as 1–5.
Figure S9. Quantile-quantile plots of $-\log_{10}(p$-values) for the VT and SKAT tests in the analysis of the BMI data in the WHI after post-imputation QC. On the left side, the top five genes identified by VT-rob are marked as 1–5. On the right side, the top five genes identified by SKAT-rob are marked as 1–5.
Figure S10. Scatter plots of imputed genotypes against the true values for ten variants in the simulation studies based on the WHI data. The blue line is the linear regression fit to the data for each variant.
Figure S11. Scatter plots of imputed genotypes against the expected values given by equation (S5) for ten variants in the simulation studies based on the WHI data. The diagonal line is shown in red. The blue line is the linear regression fit to the data for each variant, with the formula shown on the bottom right.
**Figure S12.** Scatter plots of imputed genotypes against the expected values given by equation (S5) for all variants binned into one of ten Rsq intervals in the simulation studies based on the WHI data. The diagonal line is shown in red. The blue line is the linear regression fit to all variants in each bin, with the formula shown on the bottom right.