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[1–5] share the following property: at a variant site, the imputed value $\widetilde{G}$ for a non-sequenced subject given the true genotype G and the set of genotypes observed on sequenced subjects, $\mathcal{G}_{\text{seq}} = \{G_1, \ldots, G_n\}$, has the expectation

$$\mathrm{E}(\widetilde{G}|G, \mathcal{G}_{\text{seq}}) = \overline{G}_{\text{seq}} + r^2(G - \overline{G}_{\text{seq}}), \tag{S1}$$

where $\overline{G}_{\text{seq}} = n^{-1}\sum_{i=1}^{n} G_i$. Note that $\widetilde{G}$ can be uniquely determined by G, $\mathcal{G}_{\text{seq}}$, and the observed genotypes at flanking GWAS SNPs for the non-sequenced subject and all sequenced subjects. Here, we are interested in $\mathrm{E}(\widetilde{G}|G, \mathcal{G}_{\text{seq}})$, 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 $\mathrm{E}(\widetilde{G}|G, \mathcal{G}_{\text{seq}}) = a + bG$, where a and b are determined as follows. First, by applying the linear relationship to the definition of the Pearson correlation coefficient, we obtain $\rho^2 = \{\mathrm{Cov}(\widetilde{G}, G)\}^2/\{\mathrm{Var}(\widetilde{G})\mathrm{Var}(G)\} = b^2\mathrm{Var}(G)/\mathrm{Var}(\widetilde{G})$. Since $\mathrm{Rsq} = \mathrm{Var}(\widetilde{G})/\mathrm{Var}(G)$, we conclude that $b = r^2$. It is reasonable to assume that the unconditional mean $\mathrm{E}(\widetilde{G})$ does not depend on r^2, so we can rewrite $a + r^2G$ as $a^* + r^2(G - \nu)$, where a^* does not involve r^2, and $\mathrm{E}(G - \nu) = 0$. There are two constraints: when the variant cannot be imputed with any accuracy (i.e., $r^2 = 0$), we have $\mathrm{E}(\widetilde{G}|G, \mathcal{G}_{\text{seq}}) = \overline{G}_{\text{seq}}$; when the variant is perfectly imputed (i.e., $r^2 = 1$), we have $\widetilde{G} = G$ and thus $\mathrm{E}(\widetilde{G}|G, \mathcal{G}_{\text{seq}}) = G$. 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 $\rho^2 = \mathrm{Rsq}$ 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 $\mathrm{E}(G|Z) = \widetilde{G}$ and $\mathrm{Cov}(\widetilde{G}, G) = \mathrm{E}\{\widetilde{G}\mathrm{E}(G|Z)\} - \mathrm{E}(\widetilde{G})\mathrm{E}\{\mathrm{E}(G|Z)\} = \mathrm{E}(\widetilde{G}^2) - \{\mathrm{E}(\widetilde{G})\}^2 = \mathrm{Var}(\widetilde{G})$. Setting $\mathrm{Cov}(\widetilde{G}, G)$ in the expression of ρ to $\mathrm{Var}(\widetilde{G})$, we obtain $\rho^2 = \mathrm{Rsq}$, such that $r^2 = \mathrm{Rsq}$.

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, $\mathrm{E}(\widetilde{G}_u) = (1 - r^2)\mathrm{E}(G_s) + r^2\mathrm{E}(G_u)$. 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 $\mathrm{E}(G_s) = \mathrm{E}(G_u)$. It follows that

$$\mathrm{E}(\widetilde{G}_s) = \mathrm{E}(\widetilde{G}_u), \tag{S2}$$

which is referred to as $\mathrm{E}(\widetilde{G})$.

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

$$\sum_{i=1}^{n} \left\{ Y_i - \overline{Y}_{\text{seq}} + r^2(\overline{Y}_{\text{seq}} - \overline{Y}) \right\}^2 (\widetilde{G}_i - \overline{G})^2 + \sum_{i=n+1}^{N} \left(Y_i - \overline{Y}_{\text{imp}} + \overline{Y}_{\text{imp}} - \overline{Y} \right)^2 (\widetilde{G}_i - \overline{G})^2,$$

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

$$n \left\{ \text{Var}(Y_s) + r^4(\mu_{\text{seq}} - \mu)^2 \right\} \text{Var}(\widetilde{G}_s) + (N-n) \left\{ \text{Var}(Y_u) + (\mu_{\text{imp}} - \mu)^2 \right\} \text{Var}(\widetilde{G}_u). \tag{S3}$$

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

$$N^{-1} \left[\sum_{i=1}^{n} \left\{ \left(Y_i - \overline{Y}_{\text{seq}} \right)^2 + \left(\overline{Y}_{\text{seq}} - \overline{Y} \right)^2 \right\} + \sum_{i=n+1}^{N} \left\{ \left(Y_i - \overline{Y}_{\text{imp}} \right)^2 + \left(\overline{Y}_{\text{imp}} - \overline{Y} \right)^2 \right\} \right] \times \left\{ \sum_{i=1}^{n} (\widetilde{G}_i - \overline{G})^2 + \sum_{i=n+1}^{N} (\widetilde{G}_i - \overline{G})^2 \right\},$$

which is approximately

$$N^{-1} \left[n \left\{ \text{Var}(Y_s) + (\mu_{\text{seq}} - \mu)^2 \right\} + (N-n) \left\{ \text{Var}(Y_u) + (\mu_{\text{imp}} - \mu)^2 \right\} \right] \times \left\{ n\text{Var}(\widetilde{G}_s) + (N-n)\text{Var}(\widetilde{G}_u) \right\}. \tag{S4}$$

Under perfect imputation (i.e., $r^2 = 1$ and $\text{Var}(\widetilde{G}_s) = \text{Var}(\widetilde{G}_u)$) or random sampling (i.e., $\mu_{\text{seq}} = \mu_{\text{imp}} = \mu$ 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 $\mu_{\text{seq}} = \mu_{\text{imp}} = \mu$ 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

$$n\text{Var}(Y_s)\text{Var}(\widetilde{G}_s) + (N-n)\text{Var}(Y_u)\text{Var}(\widetilde{G}_u) \tag{S5}$$

and

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

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

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

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

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

$$r^2_{\mathrm{min}} = (N-n)\left\{\mathrm{Var}(Y_s) + (\mu_{\mathrm{seq}} - \mu)^2 - \mathrm{Var}(Y_u) - (\mu_{\mathrm{imp}} - \mu)^2\right\} / \left\{2N(\mu_{\mathrm{seq}} - \mu)^2\right\}.$$

Because $D = 0$ at $r^2 = 1$, we conclude that $D > 0$ for $r^2 \in [0, 1)$ if and only if $r^2_{\mathrm{min}} \geq 1$. This condition simplifies to $\mathrm{Var}(Y_s) - \mathrm{Var}(Y_u) \geq \{N/(N-n)\}^2(\mu_{\mathrm{seq}} - \mu)^2$ because of the fact that $n\mu_{\mathrm{seq}} + (N-n)\mu_{\mathrm{imp}} = N\mu$. 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 $(\mu_{\mathrm{seq}} - \mu_{\mathrm{imp}})(1 - \mu_{\mathrm{seq}} - \mu_{\mathrm{imp}}) \geq (\mu_{\mathrm{seq}} - \mu_{\mathrm{imp}})^2$ because $\mathrm{Var}(Y_s) = \mu_{\mathrm{seq}}(1 - \mu_{\mathrm{seq}})$, $\mathrm{Var}(Y_u) = \mu_{\mathrm{imp}}(1 - \mu_{\mathrm{imp}})$, and $n\mu_{\mathrm{seq}} + (N-n)\mu_{\mathrm{imp}} = N\mu$. If $\mu_{\mathrm{seq}} > \mu_{\mathrm{imp}}$ or equivalently $\mu_{\mathrm{seq}} > \mu$, we obtain $\mu_{\mathrm{seq}} \leq 0.5$. If $\mu_{\mathrm{seq}} < \mu_{\mathrm{imp}}$ or equivalently $\mu_{\mathrm{seq}} < \mu$, we have $\mu_{\mathrm{seq}} \geq 0.5$. In other words, for the case-control sequencing study with an equal number of cases and controls (i.e., $\mu_{\mathrm{seq}} = 0.5$), V_{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 $G = (G_1, \ldots, G_m)^{\mathrm{T}}$, where G_j is the genotype of the jth variant. Write $\widetilde{G} = (\widetilde{G}_1, \ldots, \widetilde{G}_m)^{\mathrm{T}}$, where $\widetilde{G}_j$ 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:

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

where β and γ are vectors of unknown regression parameters, and $a(.)$, $b(.)$, and $c(.,.)$ are specific functions. Define $b'(\theta) = db(\theta)/d\theta$ and $b''(\theta) = d^2b(\theta)/d\theta^2$. 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 $(Y_i, X_i, \widetilde{G}_i)$ $(i = 1, \ldots, N)$ is

$$U = \sum_{i=1}^{N} \left\{Y_i - b'(\widehat{\gamma}^{\mathrm{T}}X_i)\right\} \widetilde{G}_i,$$

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

$$V_{\mathrm{std}} = a(\widehat{\phi}) \sum_{i=1}^{N} b''(\widehat{\gamma}^{\mathrm{T}}X_i)\big(\widetilde{G}_i\widetilde{G}_i^{\mathrm{T}} - \widehat{\zeta}^{\mathrm{T}}X_i\widetilde{G}_i^{\mathrm{T}}\big),$$

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

We wish to prove that $\mathrm{E}(U) = 0$ 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 γ yields the asymptotic approximation

$$U = \sum_{i=1}^{N} \epsilon_i\big(\widetilde{G}_i - \zeta^{\mathrm{T}}X_i\big),$$

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

$$\zeta = \left[\mathrm{E}\left\{b''(\gamma^{\mathrm{T}}X)XX^{\mathrm{T}}\right\}\right]^{-1}\left[\frac{n}{N}\mathrm{E}\{b''(\gamma^{\mathrm{T}}X_s)X_s\widetilde{G}_s^{\mathrm{T}}\} + \frac{(N-n)}{N}\mathrm{E}\{b''(\gamma^{\mathrm{T}}X_u)X_u\widetilde{G}_u^{\mathrm{T}}\}\right].$$

We assume that G is independent of X, which implies that $\widetilde{G}$ is independent of X in both the sequenced and non-sequenced samples. Clearly,

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

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

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

such that

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

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

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

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

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 $\mathcal{G}_{\mathrm{seq}} = \{\widetilde{G}_{ji}; 1 \le j \le m, 1 \le i \le n\}$, which is redefined to include all m variants. We first decompose $\mathrm{Var}\left(U\right)$ as $\mathrm{E}\big\{\mathrm{Var}\big(U\big|\mathcal{Y},\mathcal{X}\big)\big\} + \mathrm{Var}\big\{\mathrm{E}\big(U\big|\mathcal{Y},\mathcal{X}\big)\big\}$, which is further expanded as

$$\mathrm{E}\big[\mathrm{E}\big\{\mathrm{Var}\big(U\big|\mathcal{Y},\mathcal{X},\mathcal{G}_{\mathrm{seq}}\big)\big|\mathcal{Y},\mathcal{X}\big\} + \mathrm{Var}\big\{\mathrm{E}\big(U\big|\mathcal{Y},\mathcal{X},\mathcal{G}_{\mathrm{seq}}\big)\big|\mathcal{Y},\mathcal{X}\big\}\big] + \mathrm{Var}\big\{\mathrm{E}\big(U\big|\mathcal{Y},\mathcal{X}\big)\big\}.$$

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

$$\sum_{i=n+1}^{N}\epsilon_i^2\mathrm{Var}\big(\widetilde{G}_i\big|\mathcal{Y},\mathcal{X},\mathcal{G}_{\mathrm{seq}}\big). \tag{S9}$$

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 jth variant, for a non-sequenced subject i,

$$\mathrm{E}(\widetilde{G}_{ji}|\mathcal{G}_{j,\mathrm{seq}}) = (1-r_j^2)\overline{G}_{j,\mathrm{seq}} + r_j^2\mathrm{E}(G_{ji}|\mathcal{G}_{j,\mathrm{seq}}), \tag{S10}$$

where $\widetilde{G}_{ji}$ and G_{ji} are the imputed and true genotypes for the jth variant of the ith subject, $\overline{G}_{j,\mathrm{seq}} = n^{-1}\sum_{i=1}^{n} G_{ji}$, $\mathcal{G}_{j,\mathrm{seq}} = \{\widetilde{G}_{ji}; 1 \le i \le n\}$, and r_j^2 is the value of r^2 for the jth variant. Note that $\mathrm{E}(G_{ji}|\mathcal{G}_{j,\mathrm{seq}}) = \mathrm{E}(G_{ji})$ 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 $\mathrm{E}(G_{ji})$ for a non-sequenced subject i by $\overline{G}_{j,\mathrm{seq}}$. Thus, we estimate (S10) by $\overline{G}_{j,\mathrm{seq}}$ and (S9) by $\sum_{i=n+1}^{N}\widehat{\epsilon}_i^2\big(\widetilde{G}_i - \overline{G}_{\mathrm{seq}}\big)\big(\widetilde{G}_i - \overline{G}_{\mathrm{seq}}\big)^{\mathrm{T}}$, where $\widehat{\epsilon}_i = Y_i - b'(\widehat{\gamma}^{\mathrm{T}}X_i)$, and $\overline{G}_{\mathrm{seq}} = (\overline{G}_{1,\mathrm{seq}}, \ldots, \overline{G}_{m,\mathrm{seq}})^{\mathrm{T}}$. The second variance term $\mathrm{Var}\big\{\mathrm{E}\big(U\big|\mathcal{Y},\mathcal{X},\mathcal{G}_{\mathrm{seq}}\big)\big|\mathcal{Y},\mathcal{X}\big\}$ pertains to the variability of observed genotypes. By equation (S10), this term becomes

$$\begin{aligned}
&\mathrm{Var}\left\{\sum_{i=1}^{n}\begin{bmatrix}\epsilon_i\widetilde{G}_{1i}\\ \vdots \\ \epsilon_i\widetilde{G}_{mi}\end{bmatrix} + \sum_{i=n+1}^{N}\begin{bmatrix}\epsilon_i(1-r_1^2)\overline{G}_{1,\mathrm{seq}}\\ \vdots \\ \epsilon_i(1-r_m^2)\overline{G}_{m,\mathrm{seq}}\end{bmatrix}\Bigg|\mathcal{Y},\mathcal{X}\right\} \\
&\qquad = \mathrm{Var}\left\{\sum_{i=1}^{n}\begin{bmatrix}\big\{\epsilon_i + (1-r_1^2)\big(n^{-1}\sum_{i'=n+1}^{N}\epsilon_{i'}\big)\big\}\widetilde{G}_{1i}\\ \vdots \\ \big\{\epsilon_i + (1-r_m^2)\big(n^{-1}\sum_{i'=n+1}^{N}\epsilon_{i'}\big)\big\}\widetilde{G}_{mi}\end{bmatrix}\Bigg|\mathcal{Y},\mathcal{X}\right\}.
\end{aligned} \tag{S11}$$

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

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

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

$$\sum_{i=1}^{n}S_i^*S_i^{*\mathrm{T}} + \sum_{i=n+1}^{N}\widehat{\epsilon}_i^2\big(\widetilde{G}_i - \overline{G}_{\mathrm{seq}}\big)\big(\widetilde{G}_i - \overline{G}_{\mathrm{seq}}\big)^{\mathrm{T}}. \tag{S12}$$

Because of (S7), we replace $\overline{G}_{j,\text{seq}}$ and $\overline{G}_{\text{seq}}$ in (S12) by $\widehat{\zeta}_j^{\mathrm{T}} X_i$ and $\widehat{\zeta}^{\mathrm{T}} X_i$, respectively, to obtain

$$V_{\text{rob}} = \sum_{i=1}^{n} S_i S_i^{\mathrm{T}} + \sum_{i=n+1}^{N} \widehat{\epsilon}_i^2 \big(\widetilde{G}_i - \widehat{\zeta}^{\mathrm{T}} X_i\big)\big(\widetilde{G}_i - \widehat{\zeta}^{\mathrm{T}} X_i\big)^{\mathrm{T}}, \tag{S13}$$

where

$$S_i = \begin{bmatrix} \big\{\widehat{\epsilon}_i - (1-r_1^2)\overline{\epsilon}_{\text{seq}}\big\}\big(\widetilde{G}_{1i} - \widehat{\zeta}_1^{\mathrm{T}} X_i\big) \\ \vdots \\ \big\{\widehat{\epsilon}_i - (1-r_m^2)\overline{\epsilon}_{\text{seq}}\big\}\big(\widetilde{G}_{mi} - \widehat{\zeta}_m^{\mathrm{T}} X_i\big) \end{bmatrix},$$

and $\widehat{\zeta}_j$ is the jth column of $\widehat{\zeta}$. 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

$$\sum_{i=1}^{N} \widehat{\epsilon}_i^2 \big(\widetilde{G}_i - \widehat{\zeta}^{\mathrm{T}} X_i\big)\big(\widetilde{G}_i - \widehat{\zeta}^{\mathrm{T}} X_i\big)^{\mathrm{T}}.$$

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 $\mathcal{X}$ and the sampling scheme to obtain the appoximation $\sum_{i=n+1}^{N} \big[\big\{1 - b'(\widehat{\gamma}^{\mathrm{T}} X_i)\big\}^2 \mu_{\text{imp}} + \big\{b'(\widehat{\gamma}^{\mathrm{T}} X_i)\big\}^2 (1 - \mu_{\text{imp}})\big] \mathrm{Var}\big(\widetilde{G}_i \big| \mathcal{G}_{\text{seq}}\big)$, where μ_{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 $\mathcal{X}$ and the sampling scheme. Then we derive their estimators as before. We obtain a modified robust variance estimator as

$$\begin{aligned} V_{\text{rob}} = &\sum_{i=1}^{n} \big\{S_{i,1} S_{i,1}^{\mathrm{T}} \overline{Y}_{\text{seq}} + S_{i,0} S_{i,0}^{\mathrm{T}} (1 - \overline{Y}_{\text{seq}})\big\} \\ &+ \sum_{i=n+1}^{N} \big[\big\{1 - b'(\widehat{\gamma}^{\mathrm{T}} X_i)\big\}^2 \overline{Y}_{\text{imp}} + \big\{b'(\widehat{\gamma}^{\mathrm{T}} X_i)\big\}^2 (1 - \overline{Y}_{\text{imp}})\big] \big(\widetilde{G}_i - \widehat{\zeta}^{\mathrm{T}} X_i\big)\big(\widetilde{G}_i - \widehat{\zeta}^{\mathrm{T}} X_i\big)^{\mathrm{T}}, \end{aligned} \tag{S14}$$

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 $\overline{Y}_{\text{seq}}$ and $\overline{Y}_{\text{imp}}$ estimate the probabilities of being a case among sequenced and non-sequenced subjects,

respectively. The modified V_{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.[7] This meta-analysis is equivalent to the joint analysis of individual participant data of all studies and is thus statistically optimal.[8] The score statistic U should be normalized by $a(\widehat{\phi})$. 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_{rob} of each study by its own $a(\widehat{\phi})^{-1}$ and $a(\widehat{\phi})^{-2}$, respectively, before taking the summations; otherwise, the meta-analysis will no longer be optimal (unless ϕ is the same in all studies).

1. Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.

2. Howie, B., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

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

4. Liu, E. Y., Li, M., Wang, W., and Li, Y. (2013). MaCH-Admix: Genotype imputation for admixed populations. Genet. Epidemiol. *37*, 25–37.

5. Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.

6. Li, C. and Li, M. (2008). GWAsimulator: a rapid whole-genome simulation program. Bioinformatics *24*, 140–142.

7. Hu, Y. J., Berndt, S. I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K. E., Ingelsson, E., and Lin, D. Y. (2013). Meta-analysis of gene-level associations for rare variants based on single-variant statistics. Am. J. Hum. Genet. *93*, 236–248.

8. Lin, D. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika *97*, 321–332.

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

					p-value		
Gene	Accession	Chr	m	Rsq	T5-std	T5-rob	T5-seq
ODF2L	*NM_001007022*	1	2	0.941	1.5×10^{-4}	5.4×10^{-5}	9.4×10^{-2}
MRGPRX3	*NM_054031*	11	3	0.882	1.7×10^{-4}	3.1×10^{-4}	1.5×10^{-1}
TRDMT1	*NM_004412*	10	2	0.891	1.7×10^{-4}	2.2×10^{-4}	7.5×10^{-2}
ITSN1	*NM_003024*	21	3	0.943	2.8×10^{-4}	3.6×10^{-4}	7.0×10^{-1}
FANK1	*NM_145235*	10	2	0.792	2.8×10^{-4}	4.1×10^{-4}	4.5×10^{-1}
BDNF	*NM_001143805*	11	1	0.999	3.3×10^{-4}	2.8×10^{-4}	9.9×10^{-2}
CYP3A4	*NM_017460*	7	2	0.910	3.5×10^{-4}	1.9×10^{-4}	4.0×10^{-1}
FAM60A	*NM_001135811*	12	1	0.768	4.0×10^{-4}	2.3×10^{-4}	6.5×10^{-1}
MAP1B	*NM_005909*	5	6	0.815	4.5×10^{-4}	5.6×10^{-4}	7.1×10^{-2}
FXC1	*NM_012192*	11	2	0.911	5.1×10^{-4}	7.8×10^{-4}	4.0×10^{-1}

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.

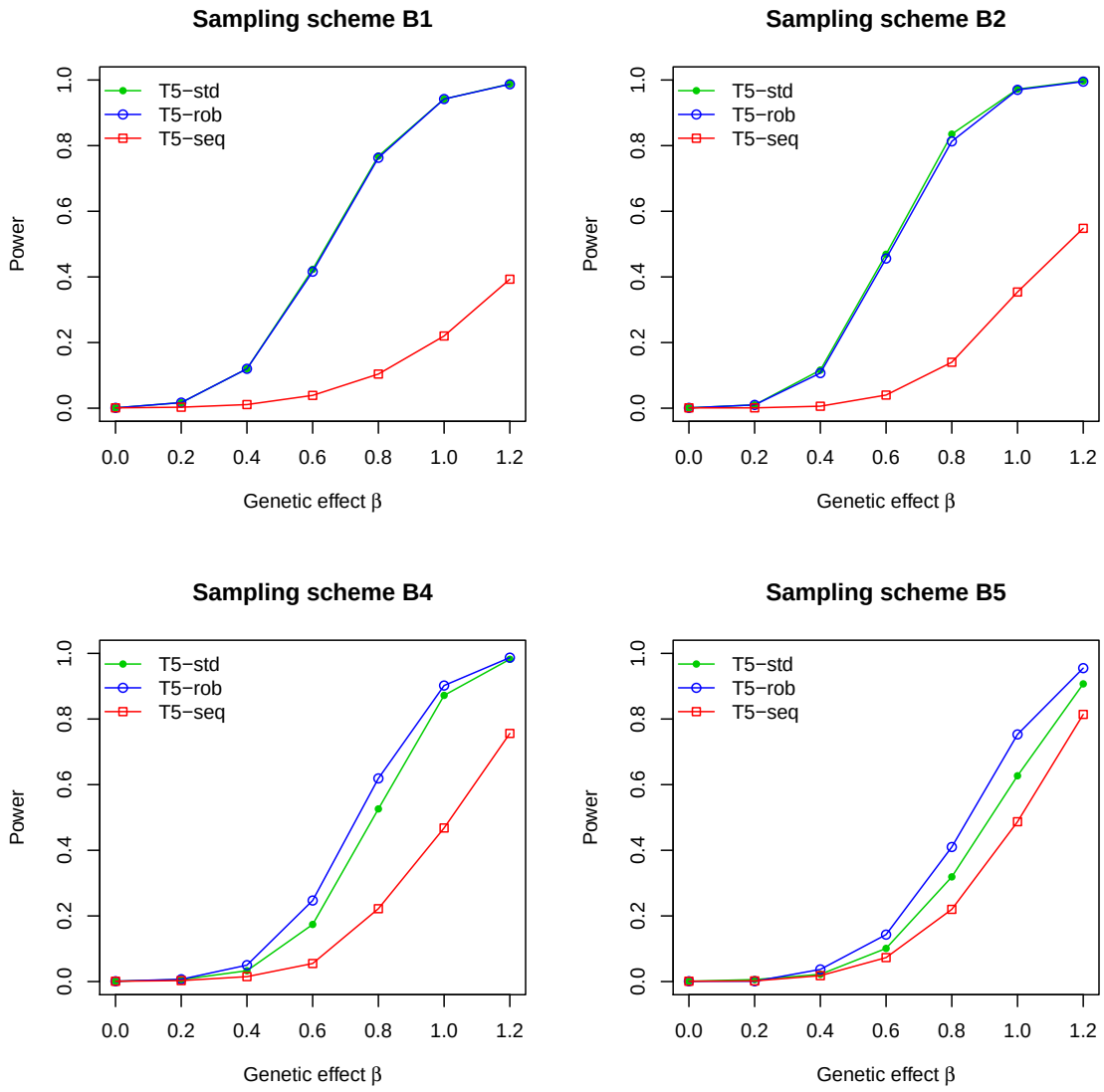

Figure S1. Power of the T5 tests at the nominal significance level of 0.001 for the integrative analysis of sequencing and GWAS data based on the robust (T5-rob) and standard variance estimators (T5-std) and for the analysis of sequenced data only (T5-seq). The trait of interest is binary. In B2, B4 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.

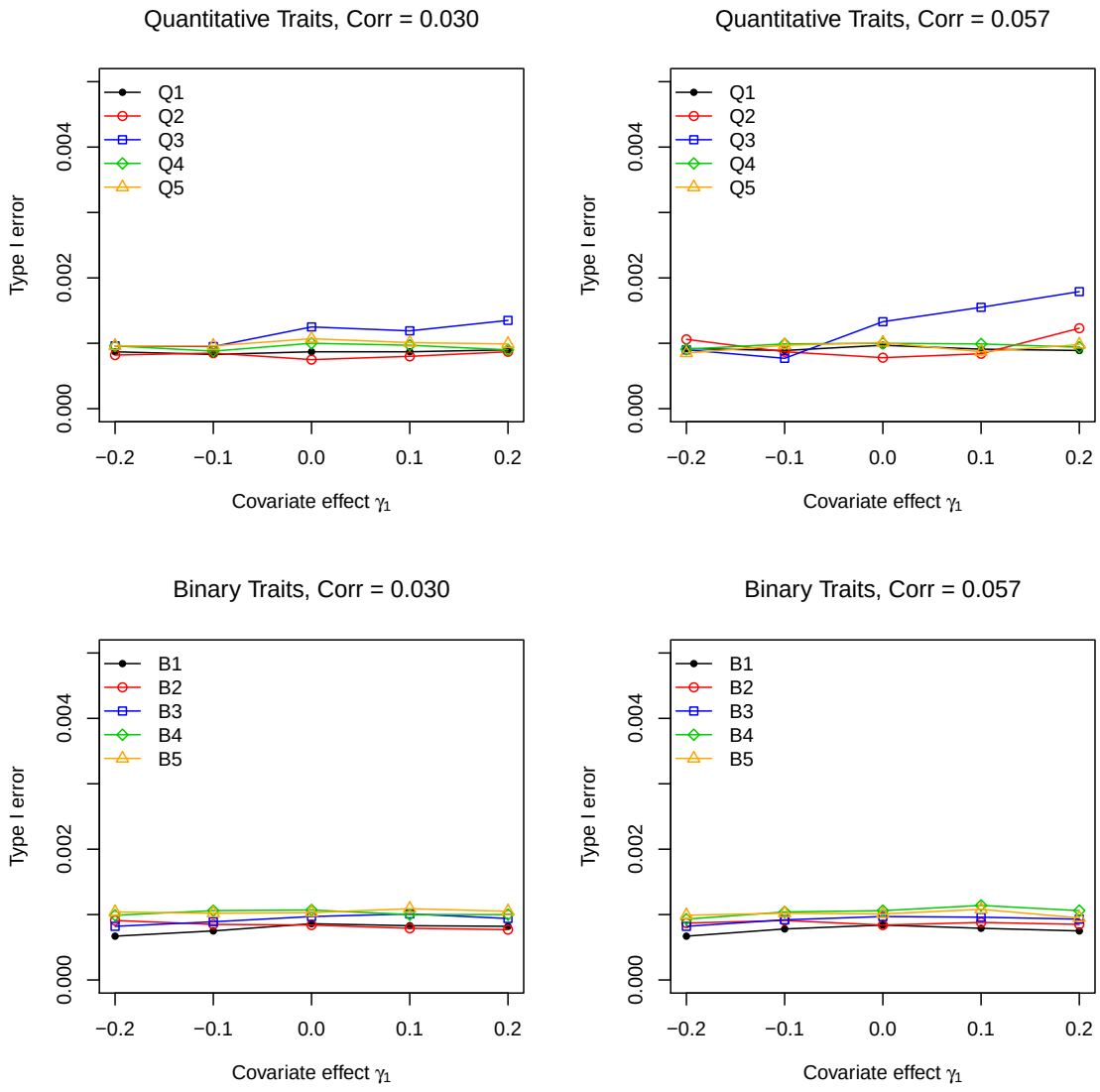

Figure S2. Type I error of the T5 test with the robust variance estimator at the nominal significance level of 0.001 for quantitative traits (upper panel) and binary traits (lower panel) when the burden score and covariate are correlated. Corr is the Pearson correlation coefficient between the burden score and covariate. Each type I error estimate is based on 100,000 replicates.

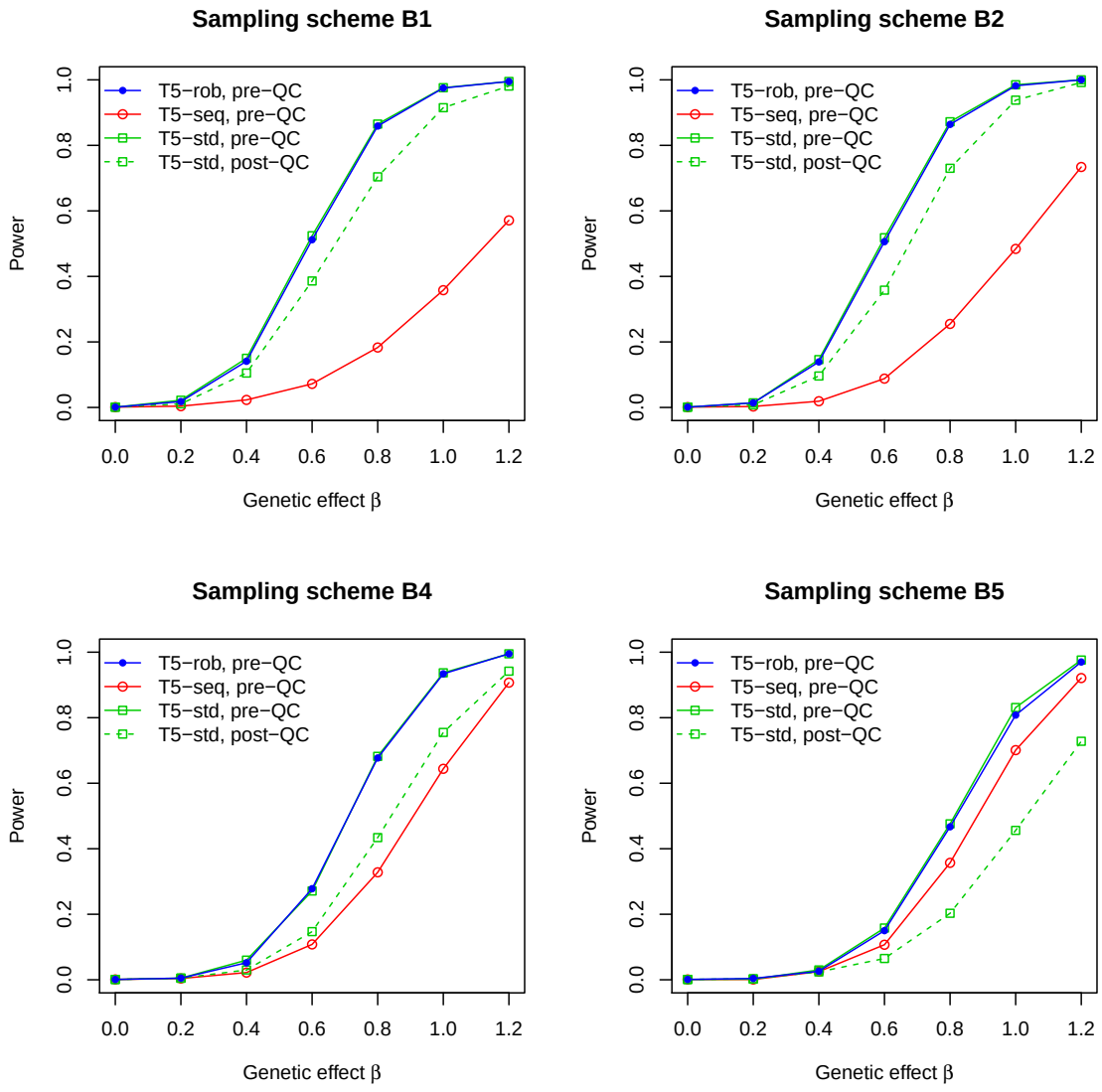

Figure S3. Power at the nominal significance level of 0.001 for the T5 test based on the robust variance estimator and pre-QC variants (T5-rob, pre-QC), the T5 test based on the standard variance estimator and pre-QC variants (T5-std, pre-QC), and the T5 test based on the standard variance estimator 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.

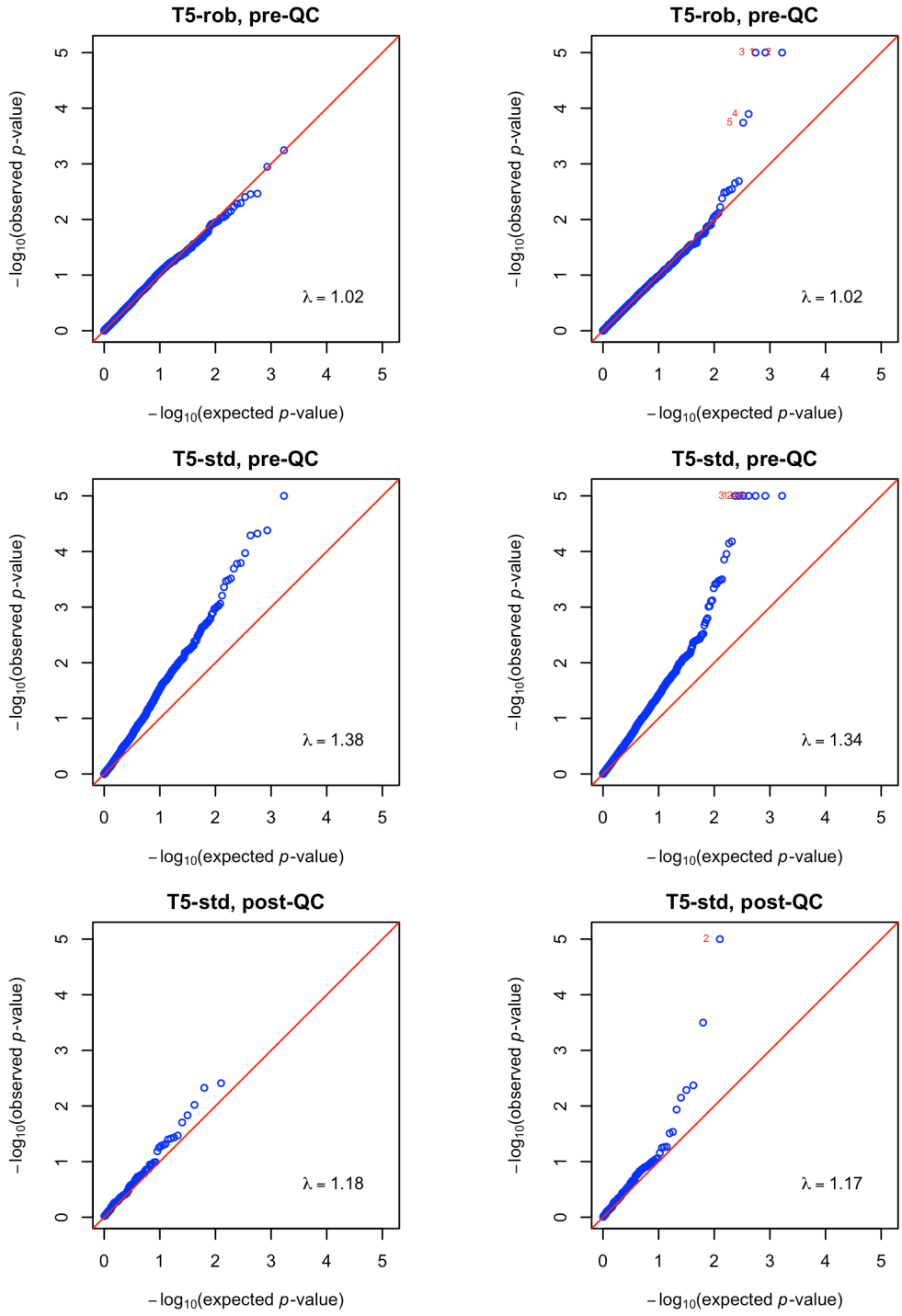

Figure S4. Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the T5 tests of all genes on chromosome 1 when the disease status was simulated with a disease rate of 10%. The left and right sides pertain to the null hypothesis and alternative hypothesis (with five causal genes), respectively. 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 1.0×10^{-5} are truncated.

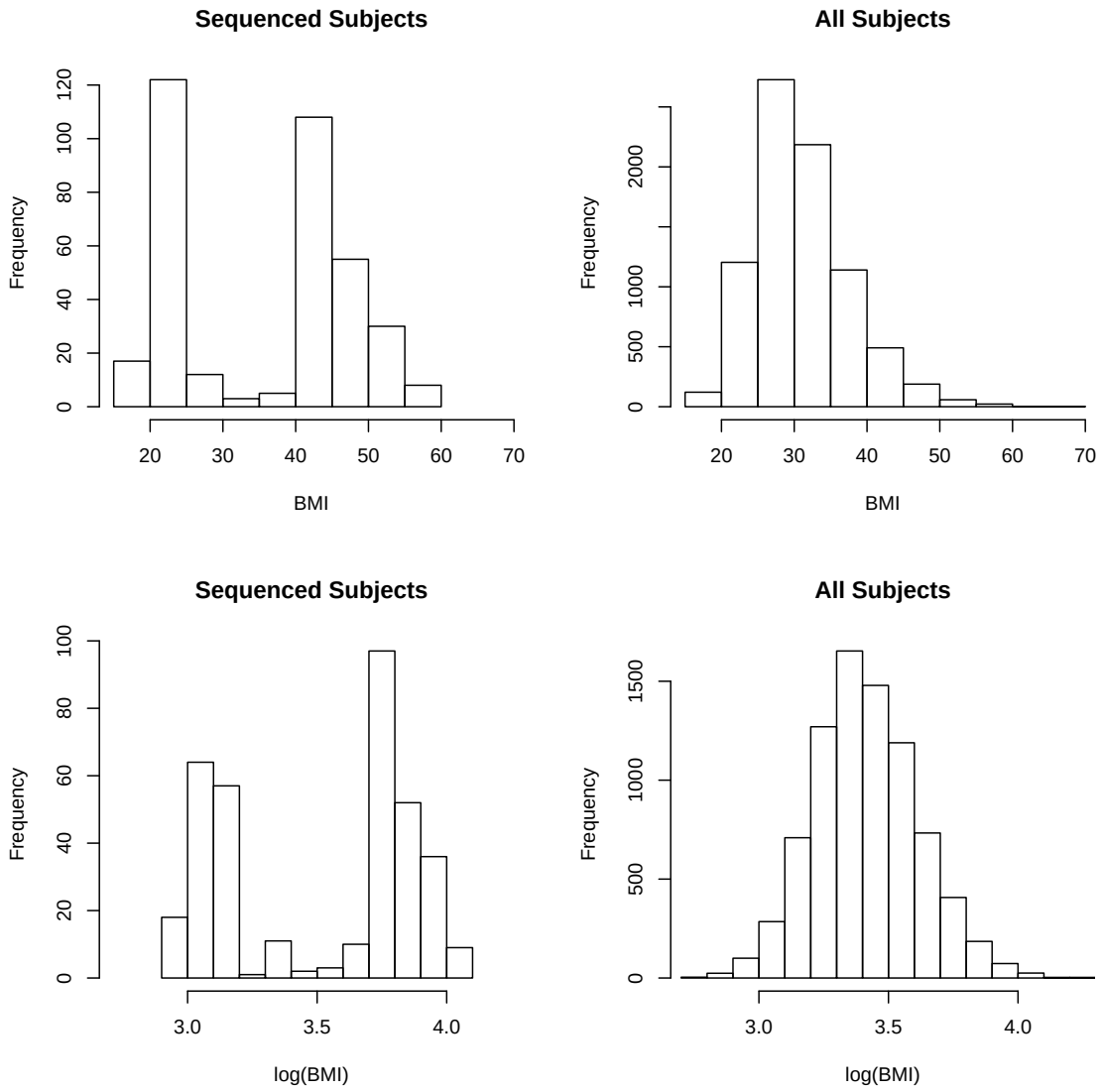

Figure S5. Distributions of original BMI values (upper panel) and log-transformed BMI values (lower panel) for sequenced subjects (left side) and all subjects (right side) in the WHI African Americans.

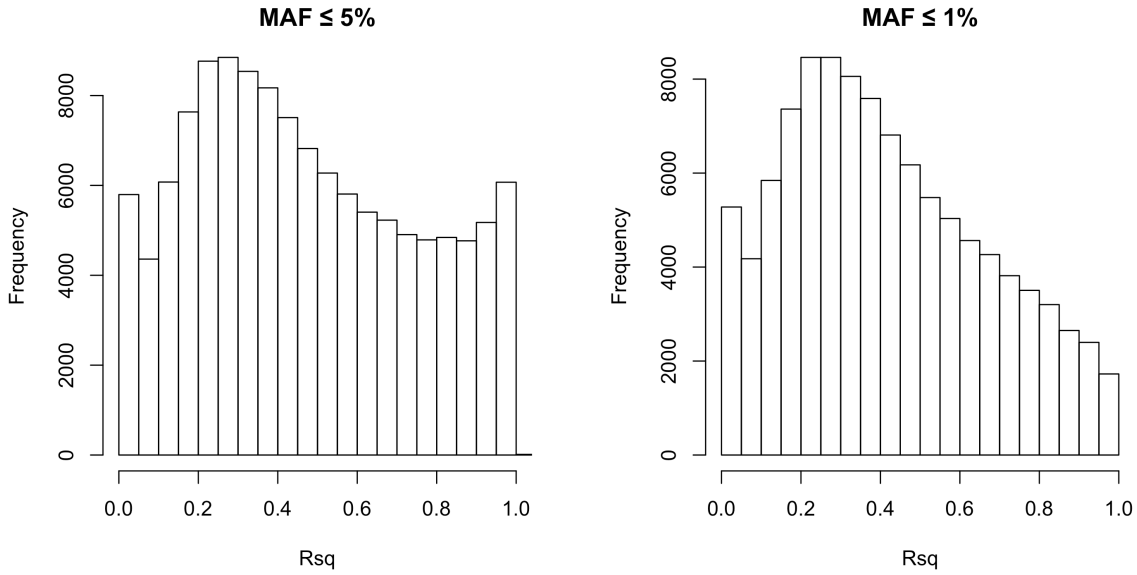

Figure S6. Distributions of Rsq values for variants with MAFs $\leq 5\%$ and $\leq 1\%$ in the WHI data.

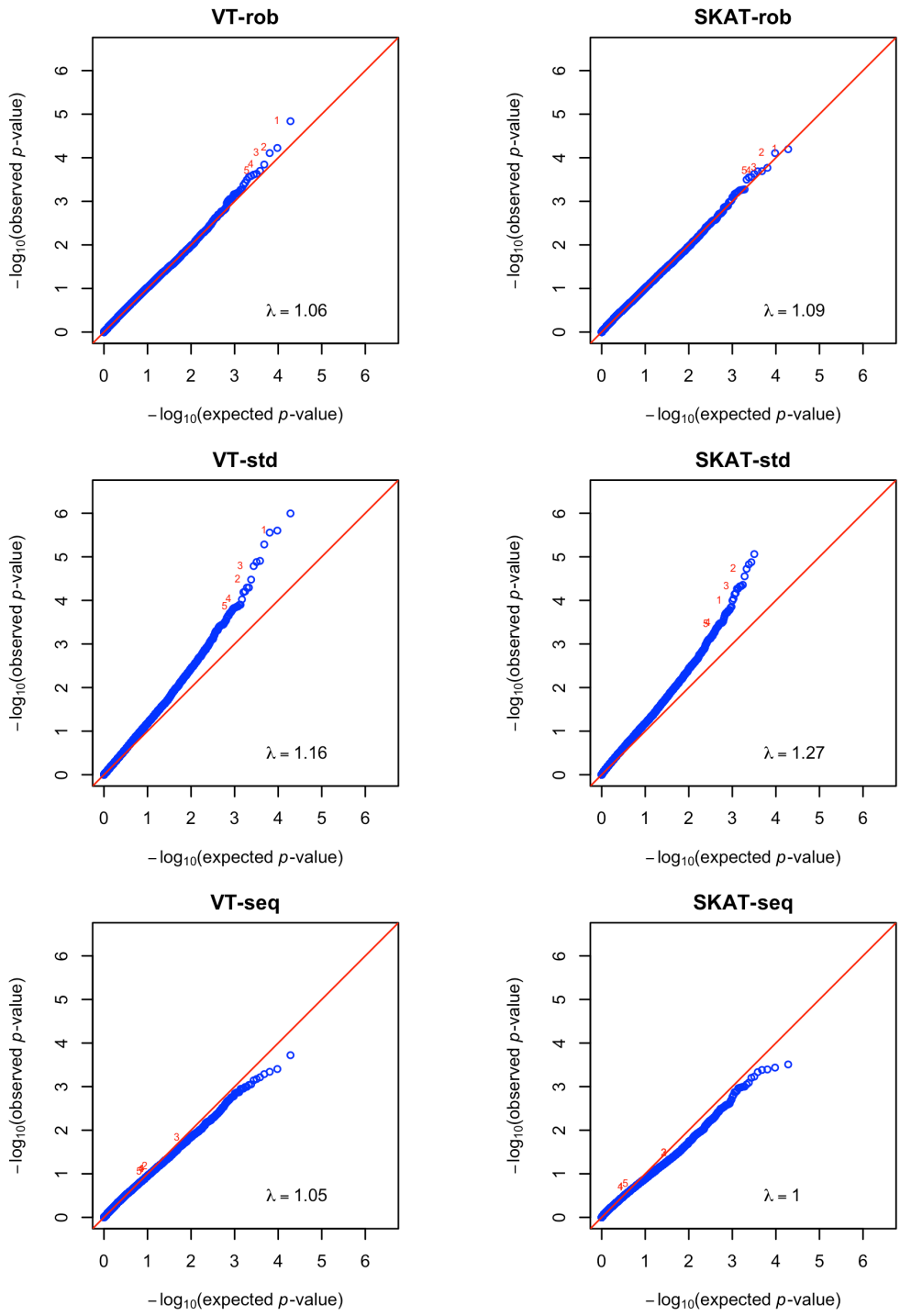

Figure S7. Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the VT and SKAT tests in the analysis of the BMI data in the WHI. On the left side, the top five genes identified by VT-rob are marked as 1–5. On the right side, the top five genes identified by SKAT-rob are marked as 1–5.

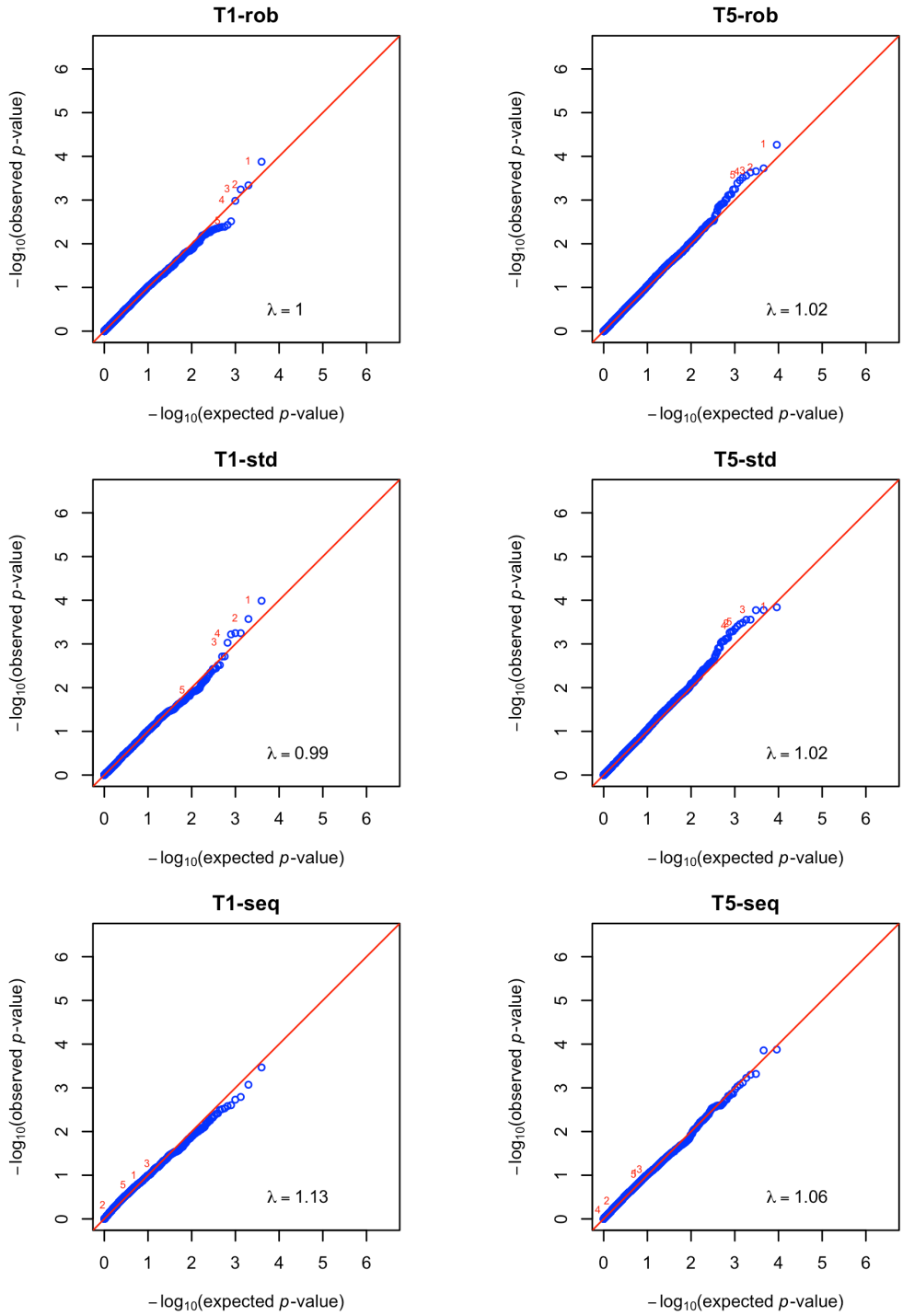

Figure S8. Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the T1 and T5 tests in the analysis of the BMI data in the WHI after post-imputation QC. On the left side, the top five genes identified by T1-rob are marked as 1–5. On the right side, the top five genes identified by T5-rob are marked as 1–5.

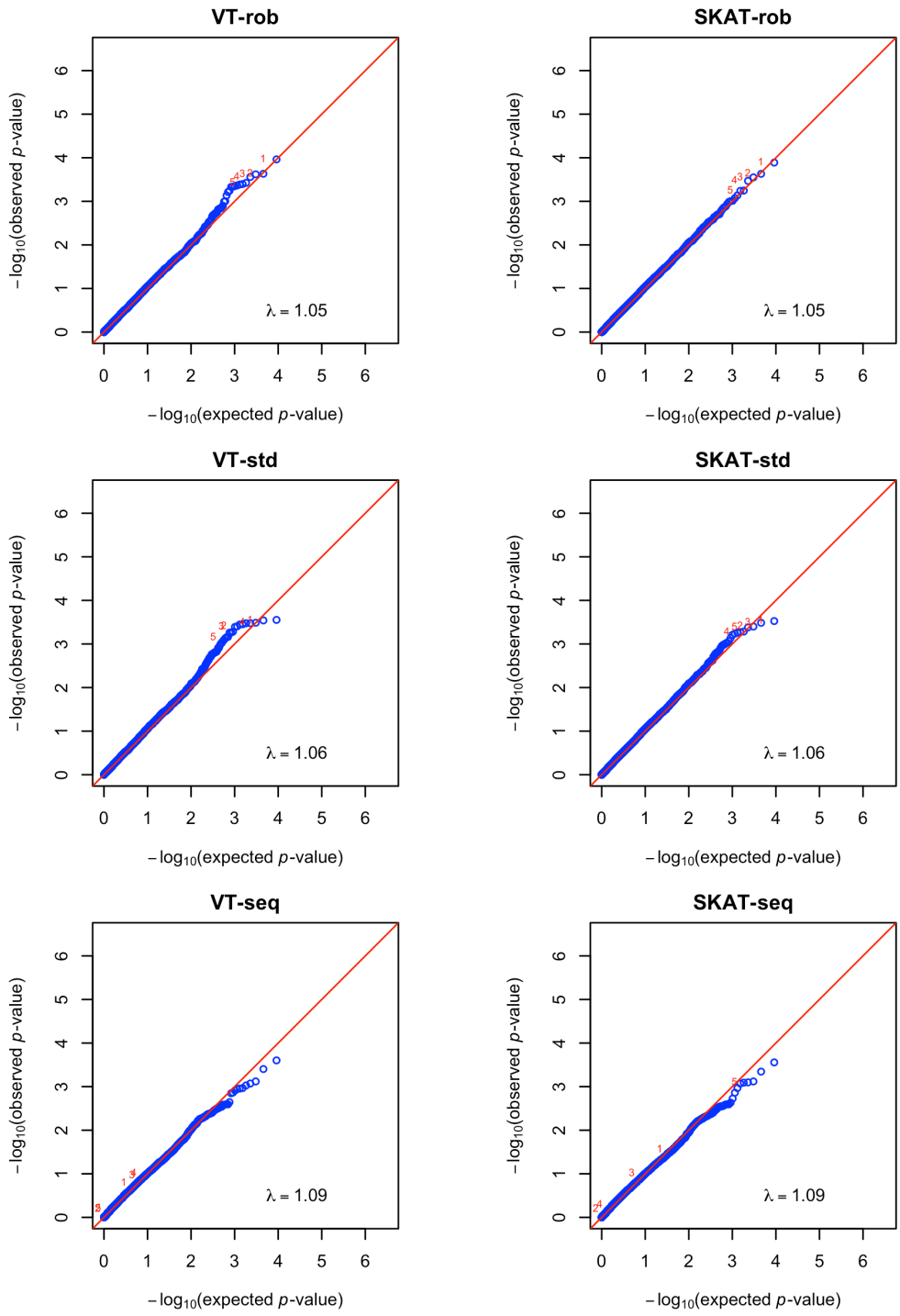

Figure S9. Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the VT and SKAT tests in the analysis of the BMI data in the WHI after post-imputation QC. On the left side, the top five genes identified by VT-rob are marked as 1–5. On the right side, the top five genes identified by SKAT-rob are marked as 1–5.

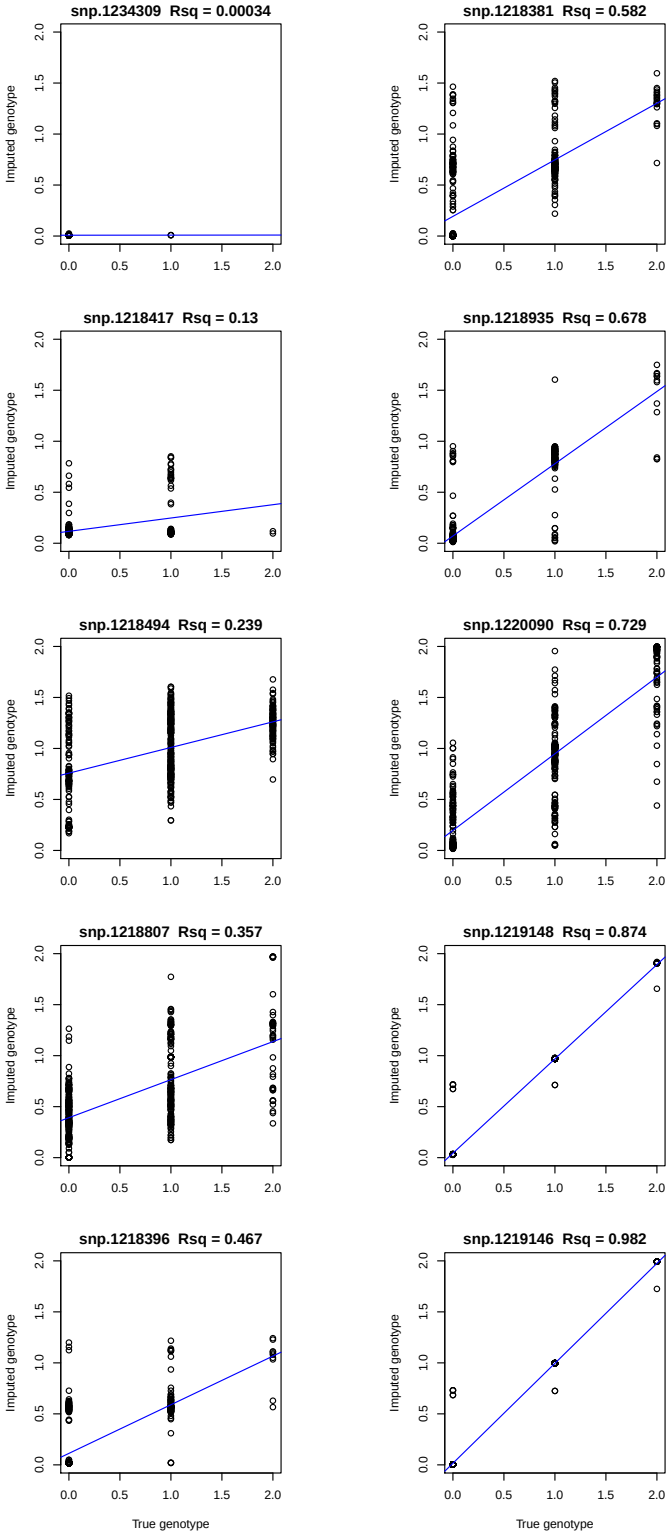

Figure S10. Scatter plots of imputed genotypes against the true values for ten variants in the simulation studies based on the WHI data. The blue line is the linear regression fit to the data for each variant.

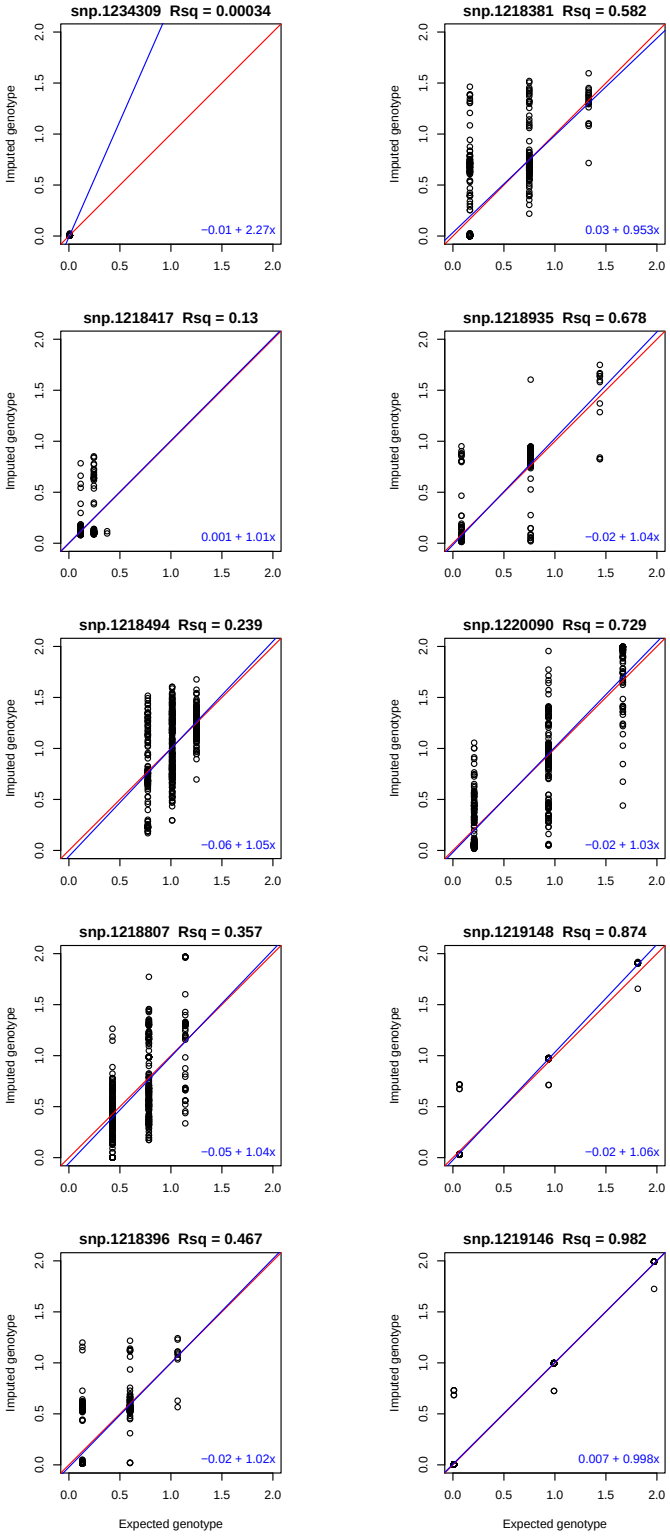

Figure S11. Scatter plots of imputed genotypes against the expected values given by equation (S5) for ten variants in the simulation studies based on the WHI data. The diagonal line is shown in red. The blue line is the linear regression fit to the data for each variant, with the formula shown on the bottom right.

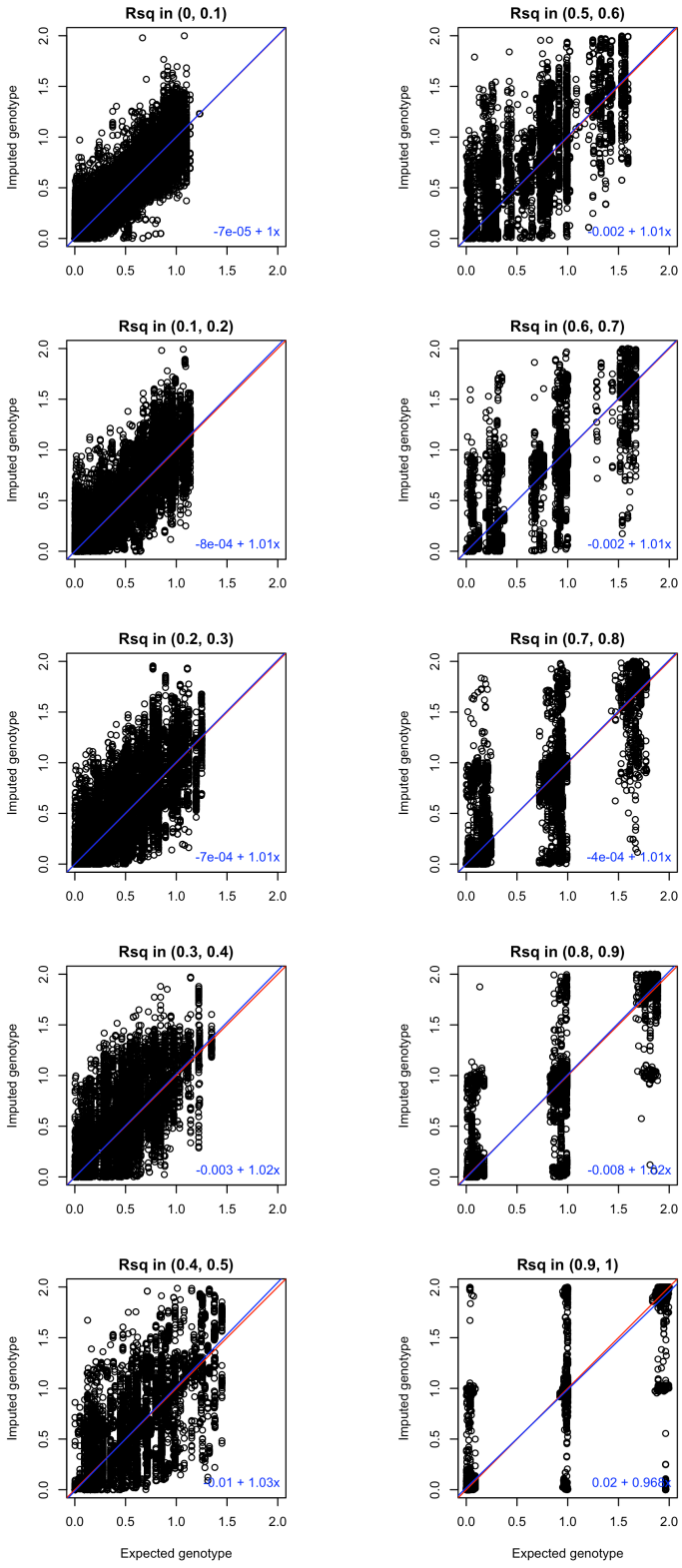

Figure S12. Scatter plots of imputed genotypes against the expected values given by equation (S5) for all variants binned into one of ten Rsq intervals in the simulation studies based on the WHI data. The diagonal line is shown in red. The blue line is the linear regression fit to all variants in each bin, with the formula shown on the bottom right.