The American Journal of Human Genetics, *91*

Supplemental Data

A Multi-SNP Locus-Association Method Reveals a Substantial Fraction of the Missing Heritability

Georg B. Ehret, David Lamparter, Clive J. Hoggart, Genetic Investigation of Anthropometric Traits Consortium, John C. Whittaker, Jacques S. Beckmann, and Zoltán Kutalik

1 Derivation of the lower bound on the explained variance of the tagged causal marker

As mentioned in the main paper the causal variant may not be directly observed. However, several other variants can be observed (or imputed) at this locus, some of which may show some evidence for association, merely due to their linkage disequilibrium (LD) with the causal variant. Let ρ_i be the (unknown) correlation between the casual marker and the ith observed SNP. By definition of $\boldsymbol{g}$ being the causal variant, any SNP that is associated with the phenotype at this locus is associated via $\boldsymbol{g}$, hence not associated with ε. Let r_i^2 denote the explained variance and F_i the genotype vector of SNP i. Since the phenotype and all genotypes are standardized $\rho_i = \mathsf{F}_i^\top \boldsymbol{g}/n$, $r_i^2 = \beta_i^2$, and $r_g^2 = \beta_g^2$. We can then express the (square-root of the) explained variance of the ith SNP as

$$\beta_i = \frac{\mathsf{F}_i^\top \boldsymbol{y}}{\mathsf{F}_i^\top \mathsf{F}_i} = \frac{\mathsf{F}_i^\top (\beta_g \boldsymbol{g} + \varepsilon)}{n} = \frac{\beta_g \mathsf{F}_i^\top \boldsymbol{g}}{n} = r_g \frac{1}{n} \mathsf{F}_i^\top \boldsymbol{g} = r_g \rho_i \ . \quad (1)$$

Important to note that this equation is meant for infinitely large population ($n \to \infty$), where β_i and ρ_i are not estimates, but the true values.

Next we split the causal variant $\boldsymbol{g}$ into two parts. One is a projection of it onto the space spanned by F, the other is the component perpendicular to this space

$$\boldsymbol{g} = \mathsf{F}\boldsymbol{\alpha} + \boldsymbol{h} \ . \quad (2)$$

Since all F_is are orthogonal to $\boldsymbol{h}$, ρ_i simplifies to

$$\rho_i = \frac{1}{n} F_i^\top \boldsymbol{g} = \frac{1}{n} F_i^\top (F\boldsymbol{\alpha} + \boldsymbol{h}) = C_i^\top \boldsymbol{\alpha} \ , \tag{3}$$

where C is the correlation matrix of F and again C_i denotes the ith column of this matrix. By combining Eqs (1) and (3) we obtain

$$\boldsymbol{\beta} = r_g \cdot C\boldsymbol{\alpha} \ ,$$

hence we have the solution for α

$$\boldsymbol{\alpha} = \frac{1}{r_g} C^{-1}\boldsymbol{\beta} \ .$$

Substituting this solution to Eq (2) yields

$$\boldsymbol{g} = \frac{1}{r_g} FC^{-1}\boldsymbol{\beta} + \boldsymbol{h} \ .$$

Due to the fact that $\mathrm{Var}(\boldsymbol{g})$ was set to 1, we have

$$1 = \mathrm{Var}(\boldsymbol{g}) = \frac{1}{n}\boldsymbol{g}^\top \boldsymbol{g} = \frac{1}{n}\frac{1}{r_g^2}\boldsymbol{\beta}^\top C^{-1} F^\top F C^{-1}\boldsymbol{\beta} + \mathrm{Var}(\boldsymbol{h}) = \frac{1}{r_g^2}\boldsymbol{\beta}^\top C^{-1}\boldsymbol{\beta} + \mathrm{Var}(\boldsymbol{h}) \ .$$

Clearly, $\mathrm{Var}(\boldsymbol{h})$ is non-negative, thus

$$1 \geq \frac{1}{r_g^2}\boldsymbol{\beta}^\top C^{-1}\boldsymbol{\beta} \ .$$

hence

$$r_g^2 \geq \boldsymbol{\beta}^\top C^{-1}\boldsymbol{\beta} \ . \tag{4}$$

2 Unbiased estimator of the lower bound

To derive an unbiased estimate for $r_{locus}^2 = \boldsymbol{\beta}^\top C^{-1}\boldsymbol{\beta}$ we need to estimate the quantities in Eq. (4). The correlation matrix C can directly be estimated from the genotype data and $\boldsymbol{\beta}$ can be estimated by least squares regression

$$\widehat{\boldsymbol{\beta}}_i = \frac{1}{n} F_i^\top \boldsymbol{y} \ .$$

It is easy to show that under the normal linear model assumption

$$\widehat{\boldsymbol{\beta}} = \frac{1}{n} F^\top \boldsymbol{y} = \frac{1}{n} F^\top \left((F^\top F)^{-1} F^\top \boldsymbol{y} + \varepsilon_{locus} \right) = \boldsymbol{\beta} + \frac{1}{n} F^\top \varepsilon_{locus} \sim \mathcal{N}\left(\boldsymbol{\beta}, \frac{1 - r_{locus}^2}{n} C \right) \ . \tag{5}$$

If we define $\widehat{\boldsymbol{q}}$ as

$$\widehat{\boldsymbol{q}} = \mathsf{C}^{-1/2}\widehat{\boldsymbol{\beta}}\ ,$$

it can be readily seen that

$$\widehat{\boldsymbol{q}} \sim \mathcal{N}\left(\mathsf{C}^{-1/2}\boldsymbol{\beta}, \frac{1-r^2_{locus}}{n}\mathsf{I}\right)\ .$$

Therefore

$$\mathrm{E}\left(\widehat{\boldsymbol{\beta}}^\top \mathsf{C}^{-1}\widehat{\boldsymbol{\beta}}\right) = \mathrm{E}\left(\widehat{\boldsymbol{q}}^\top\widehat{\boldsymbol{q}}\right) = \boldsymbol{\beta}^\top \mathsf{C}^{-1}\boldsymbol{\beta} + (1-r^2_{locus})\frac{m}{n}\ .$$

This finally enables us to provide an unbiased estimate for $\boldsymbol{\beta}^\top \mathsf{C}^{-1}\boldsymbol{\beta}$ and therefore a lower bound on r_g^2 estimated from the data:

$$r_g^2 \geq \mathrm{E}\left(\widehat{\boldsymbol{\beta}}^\top \mathsf{C}^{-1}\widehat{\boldsymbol{\beta}}\right) - (1-r^2_{locus})\frac{m}{n}\ ,$$

subsequently

$$r_g^2 \geq \frac{n}{n-m}\left(\mathrm{E}\left(\widehat{\boldsymbol{\beta}}^\top \mathsf{C}^{-1}\widehat{\boldsymbol{\beta}}\right) - \frac{m}{n}\right)\ .$$

Note that the inequality is sharp only when $\boldsymbol{h}$ is not associated with the phenotype. This final equation guarantees that

$$\widehat{r^2}_{locus} = \frac{n}{n-m}\left(\widehat{\boldsymbol{\beta}}^\top \mathsf{C}^{-1}\widehat{\boldsymbol{\beta}} - \frac{m}{n}\right) = \frac{n}{n-m}\left(\widehat{\boldsymbol{q}}^\top\widehat{\boldsymbol{q}} - \frac{m}{n}\right) \quad (6)$$

is an unbiased estimator of the lower bound on r_g^2. As we saw above $\widehat{\boldsymbol{q}}$ consists of independent normally distributed variables, hence

$$\frac{n}{1-r^2_{locus}}\cdot\widehat{\boldsymbol{q}}^\top\widehat{\boldsymbol{q}} \sim \chi^2_{m,\lambda} \quad (7)$$

i.e. it follows a non-central chi-square distribution, with non-centrality parameter $\lambda = \frac{n}{1-r^2_{locus}}\cdot\boldsymbol{\beta}^\top \mathsf{C}^{-1}\boldsymbol{\beta}$. Note that we reduce the test statistic by substituting $r_{locus} = 0$, hence making the test more conservative. Thus, to test this statistic under the null (i.e. $r_{locus} = 0$) needs no knowledge of the true explained variance of the causal variant r_g^2. As a consequence

$$\mathrm{Var}(\widehat{r^2}_{locus}) = \left(\frac{n}{n-m}\right)^2\cdot\frac{1-r^2_{locus}}{n}\cdot\left(4\cdot\boldsymbol{\beta}^\top \mathsf{C}^{-1}\boldsymbol{\beta} + 2m\cdot\frac{1-r^2_{locus}}{n}\right)\ .$$

Therefore, an unbiased estimate for the variance is

$$\widehat{\mathrm{Var}(\widehat{r^2}_{locus})} = \left(\frac{n}{n-m}\right)^2\cdot\frac{1-r^2_{locus}}{n}\cdot\left(4\cdot\widehat{\boldsymbol{\beta}}^\top \mathsf{C}^{-1}\widehat{\boldsymbol{\beta}} - 2m\cdot\frac{1-r^2_{locus}}{n}\right)\ . \quad (8)$$

Although r^2_{locus} is not known, we can use $\widehat{r^2}_{locus}$ as an approximation for it.

3 Hypothesis testing

Next we use the lower bound estimate $\widehat{r^2}_{locus}$ to test if (i) the *multi-SNP* association is significant and (ii) how the magnitude of TEV by *multi-SNP* compares to estimations by the lead SNPs only. Once nominal P-values are calculated for each locus we used false discovery rate control to adjust for multiple testing Benjamini and Hochberg (1995).

The first null hypothesis is simply formulated as $\boldsymbol{\beta}^\top \boldsymbol{C}^{-1}\boldsymbol{\beta} = 0$. Our test statistic is defined as $T_1 = (n/(1-\widehat{r^2}_{locus})) \cdot \widehat{\boldsymbol{\beta}}^\top \boldsymbol{C}^{-1}\widehat{\boldsymbol{\beta}}$, which follows a chi-square distribution with m degrees of freedom (see Eq. (7)). When testing the null hypothesis we set $\widehat{r^2}_{locus} = 0$ in T_1.

The second null hypothesis can formally be written as $\boldsymbol{\beta}^\top \boldsymbol{C}^{-1}\boldsymbol{\beta} = r_j^2$, where j is the index of the SNP with the best P-value in the discovery set. The test statistic for this hypothesis is $T_2 = (n/(1-\widehat{r^2}_{locus})) \cdot \left(\widehat{\boldsymbol{\beta}}^\top \boldsymbol{C}^{-1}\widehat{\boldsymbol{\beta}} - \widehat{r_j^2}\right)$, where the estimates are coming from the validation sample. We can obtain (see Section 6) that under the null T_2 follows a chi-square distribution with $(m-1)$ degrees of freedom. For simplicity, we again set $\widehat{r^2}_{locus} = 0$, resulting in a conservative test.

4 Estimating the fraction of null SNPs

Assume that we identified m variants, an α fraction of this are false positives. For an identified variant i we obtained an unbiased estimate s_i for its true explained variance r_i^2 as described in the main paper. We then group these explained variance estimates into disjoint bins of $I_1, I_2, \ldots, I_K$, with bin centers $c_1, c_2, \ldots, c_K$ and such that $\cup_i I_i = [0,1]$. Under the null, each s_i are independent and follow a $-1/n$-shifted chi-square distribution with 1 degree of freedom.

We have two ways to estimate their total explained variance: (i) by simply summing them up $\sum_i \widehat{s_i}$ is an unbiased estimate of the true total explained variance; or (ii) If for each bin I_j we can estimate the bin specific true discovery rate (TDR) as t_j, the total explained variance is simply approximated as

$$\sum_{j=1}^{K} c_j \cdot |\{i : s_i \in I_j\}| \cdot t_j$$

The TDR can be expressed as

$$\begin{aligned} t_j = Pr(i \in H_1 | r_i^2 \in I_j) &= \left(1 - Pr(i \in H_0 | r_i^2 \in I_j)\right) \\ &= \left(1 - Pr(r_i^2 \in I_j | i \in H_0) \cdot \frac{Pr(i \in H_0)}{Pr(r_i^2 \in I_j)}\right) \\ &= 1 - \frac{Pr(r_i^2 \in I_j | i \in H_0) \cdot Pr(i \in H_0)}{Pr(r_i^2 \in I_j)} \end{aligned}$$

and hence can be estimated by

$$\widehat{t_j} \;=\; 1 - Pr(S^2_{(0)} \in I_j + 1/n) \cdot \frac{\alpha \cdot m}{|\{i : s_i \in I_j\}|}$$

where $S_{(0)} \sim \chi^2_1$. Therefore,

$$|\{i : s_i \in I_j\}| \cdot \widehat{t_j} \;=\; |\{i : s_i \in I_j\}| - \alpha \cdot m \cdot \int_{I_j + 1/n} \frac{1}{2^{1/2}\Gamma(1/2)} x^{-1/2} \mathrm{e}^{-x/2} dx \ .$$

Let h_j denote $\int_{I_j+1/n} \frac{1}{2^{1/2}\Gamma(1/2)} x^{-1/2}\mathrm{e}^{-x/2}dx$. Since the two estimates have equal expectation

$$\sum_i s_i = \sum_{j=1}^{K} c_j \cdot (|\{i : s_i \in I_j\}| - \alpha \cdot (m \cdot h_j)) = \sum_{i:s_i>0} s_i - \alpha \cdot m \cdot \sum_j c_j \cdot h_j \ .$$

This yields an estimate for the fraction of false positive SNPs α

$$\widehat{\alpha} = \frac{-\sum_{i:s_i<0} s_i}{m \cdot \sum_j c_j \cdot h_j} \ .$$

5 Decomposition of the multi-SNP

Here we demonstrate how the *multi-SNP* association derived from F can be decomposed into *multi-SNP* associations derived from only a subset of the available SNPs at the locus. In other words, we split F into two parts $[\mathsf{F}_u|\mathsf{F}_v]$ and also partition $\widehat{\boldsymbol{\beta}}^\top$ to $[\widehat{\boldsymbol{u}}^\top|\widehat{\boldsymbol{v}}^\top]$ and $\boldsymbol{\beta}^\top$ to $[\boldsymbol{u}^\top|\boldsymbol{v}^\top]$. The covariance matrix C is then split up to

$$\mathsf{C} = \begin{bmatrix} \mathsf{A} & \mathsf{B} \\ \mathsf{B}^\top & \mathsf{D} \end{bmatrix} \ .$$

One can verify that

$$\mathsf{C}^{-1} = \begin{bmatrix} (\mathsf{A} - \mathsf{B}\mathsf{D}^{-1}\mathsf{B}^\top)^{-1} & -\mathsf{A}^{-1}\mathsf{B}(\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1} \\ -(\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}\mathsf{B}^\top\mathsf{A}^{-1} & (\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1} \end{bmatrix} \ . \tag{9}$$

Hence,

$$\begin{aligned} t &= \boldsymbol{\beta}^\top \mathsf{C}^{-1} \boldsymbol{\beta} \\ &= \boldsymbol{u}^\top (\mathsf{A} - \mathsf{B}\mathsf{D}^{-1}\mathsf{B}^\top)^{-1} \boldsymbol{u} - 2 \cdot \boldsymbol{u}^\top \mathsf{A}^{-1}\mathsf{B}(\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}\boldsymbol{v} \\ &+ \boldsymbol{v}^\top (\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}\boldsymbol{v} \\ &= \boldsymbol{u}^\top (\mathsf{A}^{-1} + \mathsf{A}^{-1}\mathsf{B}(\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}\mathsf{B}^\top\mathsf{A}^{-1})\boldsymbol{u} \\ &- 2 \cdot \boldsymbol{u}^\top \mathsf{A}^{-1}\mathsf{B}(\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}\boldsymbol{v} \\ &+ \boldsymbol{v}^\top (\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}\boldsymbol{v} \\ &= \boldsymbol{u}^\top \mathsf{A}^{-1}\boldsymbol{u} + (\mathsf{B}^\top\mathsf{A}^{-1}\boldsymbol{u} - \boldsymbol{v})^\top (\mathsf{D} - \mathsf{B}^\top\mathsf{A}^{-1}\mathsf{B})^{-1}(\mathsf{B}^\top\mathsf{A}^{-1}\boldsymbol{u} - \boldsymbol{v}) \qquad (10) \\ &\geq \boldsymbol{u}^\top \mathsf{A}^{-1}\boldsymbol{u} \ . \end{aligned}$$

6 Distribution of T_2 under the null

We defined our test statistic as $T_2 = (n/(1-r^2_{locus})) \cdot \left(\widehat{\boldsymbol{\beta}}^\top C^{-1}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{u}}^\top A^{-1}\widehat{\boldsymbol{u}}\right)$ and establish here its distribution under the null. To expand T_2 we will use Eq. (10), but replace $\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}$ with their estimates $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}$ as follows

$$T_2 = \frac{n}{1-r^2_{locus}} \cdot \left(\widehat{\boldsymbol{\beta}}^\top C^{-1}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{u}}^2\right) = \frac{n}{1-r^2_{locus}} \cdot (B^\top\widehat{\boldsymbol{u}} - \widehat{\boldsymbol{v}})^\top (D - B^\top B)^{-1}(B^\top\widehat{\boldsymbol{u}} - \widehat{\boldsymbol{v}})$$

To establish its distribution under the null we define $\boldsymbol{w}$ as

$$\boldsymbol{w} = (D - B^\top B)^{-1/2}(B^\top\widehat{\boldsymbol{u}} - \widehat{\boldsymbol{v}}) \ .$$

As any affine transformation of $\widehat{\boldsymbol{\beta}}$, $\boldsymbol{w}$ also follows a multivariate normal distribution. Under the null $\mathrm{E}(\boldsymbol{w}) = \mathbf{0}$ and its variance-covariance matrix is

$$\begin{aligned}\mathrm{Var}(\boldsymbol{w}) &= (D - B^\top B)^{-1/2}\left(B^\top \mathrm{Var}(\widehat{\boldsymbol{u}})B - 2\cdot B^\top \mathrm{Cov}(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) + \mathrm{Var}(\widehat{\boldsymbol{v}})\right)(D - B^\top B)^{-1/2} \\ &= \frac{1-r^2_{locus}}{n} \cdot (D - B^\top B)^{-1/2}\left(D - B^\top B\right)(D - B^\top B)^{-1/2} = \frac{1-r^2_{locus}}{n} \cdot I \ .\end{aligned}$$

Therefore, $\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \frac{1-r^2_{locus}}{n} \cdot I\right)$. Since $T_2 = \frac{n}{1-r^2_{locus}} \cdot \sum_i w_i^2$, we have shown that T_2 follows a chi-square distribution under the null with $(m-1)$ degrees of freedom.

7 Alternative validation step

The proposed *multi-SNP* association can be readily implemented in the current GWAS meta-analysis work-flow where each study applied a univariate analysis and external data is used to approximate the correlation matrix. The validation step, however, can be made more exact by each validation cohort using locus-by-locus a *multivariate* regression for the SNPs carried forward from the discovery. For a given locus, validation cohort j is then asked to report their sample size n_j, and the estimate for total explained variance of the multivariate model $\widehat{r_j}^2$. This quantity needs to be computed by the individual cohort analyst, using for example the following formula:

$$\widehat{r_j}^2 = \widehat{\boldsymbol{y}}^\top\widehat{\boldsymbol{y}}/\boldsymbol{y}^\top\boldsymbol{y} \quad \text{where} \quad \widehat{\boldsymbol{y}} = F\widehat{\boldsymbol{\beta}_m^{(j)}} \ ,$$

where $\widehat{\boldsymbol{\beta}_m^{(j)}}$ is the effect size estimate of the selected SNPs in a multivariate regression. Note that in our notation we assume normalized (i.e. zero-mean, unit-variance) phenotype and genotype coding.

As shown before (Eq. (6)) the $\widehat{r_j}^2$ estimate is biased, but its modification

$$s_j = \frac{n_j}{n_j - m} \cdot \left(\widehat{r_j}^2 - \frac{m}{n_j}\right)$$

is an unbiased estimate for the lower bound of the total explained variance of the locus (encompassed in a *multi-SNP*) in validation cohort j.

In the next step, the unbiased total explained variance estimates (s_j) are meta-analyzed using inverse-variance weighting, i.e.

$$s = \sum_j w_j \cdot s_j \tag{11}$$

where

$$\begin{aligned} w_j &\propto \mathrm{Var}(s_j)^{-1} \\ w_j &\geq 0 \\ \sum_j w_j &= 1 \end{aligned}$$

Utilizing the fact that $\frac{n_j}{1-r^2_{locus}} \cdot \widehat{r^2_j} \sim \chi^2_{m,\lambda_j}$ – where $\lambda_j = n_j \cdot \frac{r^2_j}{1-r^2_{locus}}$ – the weights can be obtained as

$$w_j^{-1} \propto \mathrm{Var}(s_j) = 2 \cdot \frac{(1-r^2_{locus})^2}{(n_j - m)^2} \cdot \left(m + 2n_j \cdot \frac{r^2_j}{1-r^2_{locus}} \right) \ . \tag{12}$$

As can be noted, the calculation of w_j requires the knowledge of r^2_{locus}, or at least its unbiased estimate s. Conversely, to calculate s one needs to know w_j. Therefore, we apply an iterative procedure where we first initialize $r^2_{locus} = 0$, substitute it into Eq. (12), then we update s according to Eq. (11) using the newly obtained w_j values. The two steps are repeated until convergence.

Moreover, it can be easily derived that

$$s + \sum_j w_j \cdot \frac{m}{n_j - m} \sim \chi^2_{m,\lambda} \quad \text{where} \quad \lambda = \sum_j \lambda_j \ .$$

Consequently, the variance of s can be readily obtained

$$\mathrm{Var}(s) = 2 \cdot (m + 2\lambda) \cdot \left(\frac{1-r^2_{locus}}{n-m} \right)^2 \ ,$$

where $n = \sum_j n_j$ is the total validation sample size and s is used for r^2_{locus}. Also, by setting $r^2_{locus} = 0$, the null hypothesis can be tested via a simple (central) chi-square test.

Additionally, if multivariate effect sizes and variance-covariance matrices are provided by the validation cohorts, one can shed light on the individual contribution of each selected SNP to the *multi-SNP* association by meta-analyzing the effects $(\boldsymbol{\beta}^{(j)}_m)$ using the previously obtained weights w_j.

8 In-silico testing of the method to detect imperfect tagging

We simulated *in-silico* phenotype data to test the advantage of our locus-association method. The details of these simulation are presented in the main paper. In addition, we extensively explored different combinations of LD-pruning thresholds between (r^2 of) 0.1 to 1 and discovery P-value cut-offs $\alpha = 2 \cdot 10^{-5}, \ldots, 0.1$, while fixing $r_g^2 = 3 \cdot 10^{-2}$, $\rho^2 = 0.2$. We observed that while the explained variance estimate is increasing its standard error does so too (Fig. S1a-b). For each set of parameters there is an optimal (but in practice unknown) α and LD pruning threshold that yields the best *multi-SNP* association P-value (Fig. S1c).

9 Genotype data used to estimate LD

The genotype data sets used in this study were described elsewhere: CoLaus (n=5'435) and Hypergenes (n=3'615) (Lango Allen et al. 2010), Swiss Hepatitis C Cohort Study (SCCS, n=1,068) (Rauch et al. 2010), an Australian Hepatitis C cohort (n=302) (Suppiah et al. 2009), a French Hepatitis C cohort (n=467) (Nalpas et al. 2010), and the 1000 Genomes project 2010 November release (n=381) (1000 Genomes Project Consortium 2010). Note that the LD data is used twice, once for SNP selection in the discovery cohort, and then in the formula to estimate TEV in the validation cohort.

10 Lipid association

For the lipid association data we selected the lead SNPs using the same (combined) P-value threshold as we did for the GIANT data. However, as opposed to the GIANT data, for lipid associations we do not have access to individual cohort summary statistics, only the overall meta-analysis results (Teslovich et al. 2010). Thus, we cannot split association results into discovery and validation sets. For this reason we only aimed at discovering individual loci with significant evidence for allelic heterogeneity. To achieve this goal we modified the SNP selection procedure: once the lead SNP is found the remaining markers were not filtered based on their P-value (to avoid bias), but were simply pruned based on LD. This way the difference statistic (T_2), defined as the additional explained variance on top of the lead SNP, preserves its properties (such as its distribution under the null) described in the main paper's Methods section. Although the method is less efficient here, we still observed modest increased TEV values genome-wide. The (severely underestimated) gain in TEV varied within the 4-8% range for the different lipid traits (see Table S1).

We also looked at evidence of allelic heterogeneity at the individual locus level at

5% FDR. We found 100 such loci for LDL, 51 for HDL, 62 for TG, and 60 for TC. A detailed list of such loci can be found in Tables S5-S8. Here we only present one example for triglyceride association at the 4q31.3 locus (Fig. S2). While the lead SNP points to *LRAT* (MIM 604863), the *multi-SNP* association reveals an even stronger link with *DCHS2* (MIM 612486) implicating both synonymous and non-synonymous changes. DCHS2 is a cellular adhesion protein that belongs to the cadherin superfamily along with beta-catenin, whose knock-out mouse showed defective cholesterol and bile acid metabolism in the liver (Behari et al. 2010). Thus we can hypothesize that *DCHS2* might influence cholesterol metabolism via altered cell adhesion.

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Behari, J., Yeh, T. H., Krauland, L., Otruba, W., Cieply, B., Hauth, B., Apte, U., Wu, T., Evans, R., and Monga, S. P. (2010). Liver-specific beta-catenin knockout mice exhibit defective bile acid and cholesterol homeostasis and increased susceptibility to diet-induced steatohepatitis. *Am J Pathol*, 176(2):744–753.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57:289–300.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., and Raychaudhuri, S. *et al.*. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.

Nalpas, B., Lavialle-Meziani, R., Plancoulaine, S., Jouanguy, E., Nalpas, A., Munteanu, M., Charlotte, F., Ranque, B., Patin, E., Heath, S., Fontaine, H., Vallet-Pichard, A., Pontoire, D., Bourlière, M., Casanova, J. L., Lathrop, M., Bréchot, C., Poynard, T., Matsuda, F., Pol, S., and Abel, L. (2010). Interferon gamma receptor 2 gene variants are associated with liver fibrosis in patients with chronic hepatitis c infection. *Gut*, 59(8):1120–1126.

Rauch, A., Kutalik, Z., Descombes, P., Cai, T., Di Iulio, J., Mueller, T., Bochud, M., Battegay, M., Bernasconi, E., Borovicka, J., Colombo, S., Cerny, A., Dufour, J. F., Furrer, H., Günthard, H. F., Heim, M., Hirschel, B., Malinverni, R., Moradpour, D., Müllhaupt, B., Witteck, A., Beckmann, J. S., Berg, T., Bergmann, S., Negro, F., Telenti, A., Bochud, P. Y., Swiss Hepatitis C Cohort Study, and Swiss HIV Cohort Study (2010). Genetic variation in il28b is associated with chronic hepatitis c and treatment failure: a genome-wide association study. *Gastroenterology*, 138(4):1338–1345.

Suppiah, V., Moldovan, M., Ahlenstiel, G., Berg, T., Weltman, M., Abate, M. L., Bassendine, M., Spengler, U., Dore, G. J., Powell, E., Riordan, S., Sheridan, D., Smedile, A., Fragomeli, V., Müller, T., Bahlo, M., Stewart, G. J., Booth, D. R., and George, J. (2009). Il28b is associated with response to chronic hepatitis c interferon-alpha and ribavirin therapy. *Nat Genet*, 41(10):1100–1104.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., and Willer, C. J. *et al.*. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713.

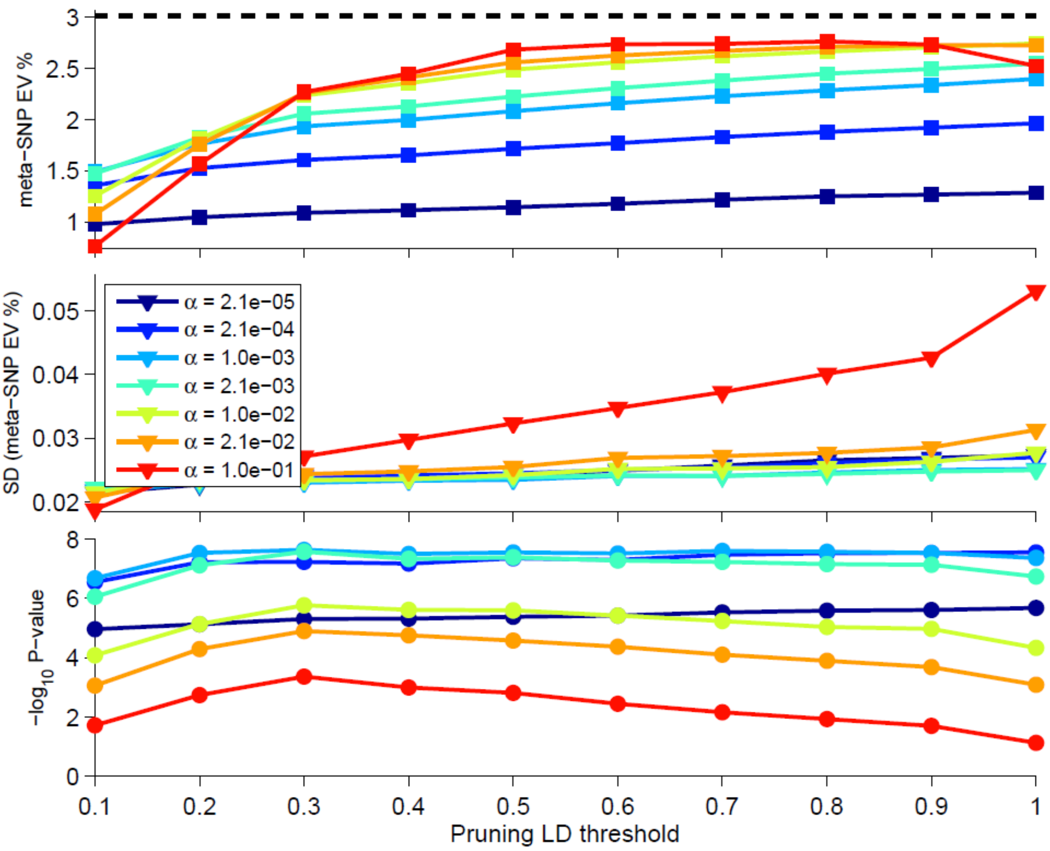

Figure S1. ***Multi-SNP*** **Properties as the Function of LD Pruning**

(A) Explained variance estimate.

(B) Standard error of the explained variance estimate.

(C) P-value of the *multi-SNP* association.

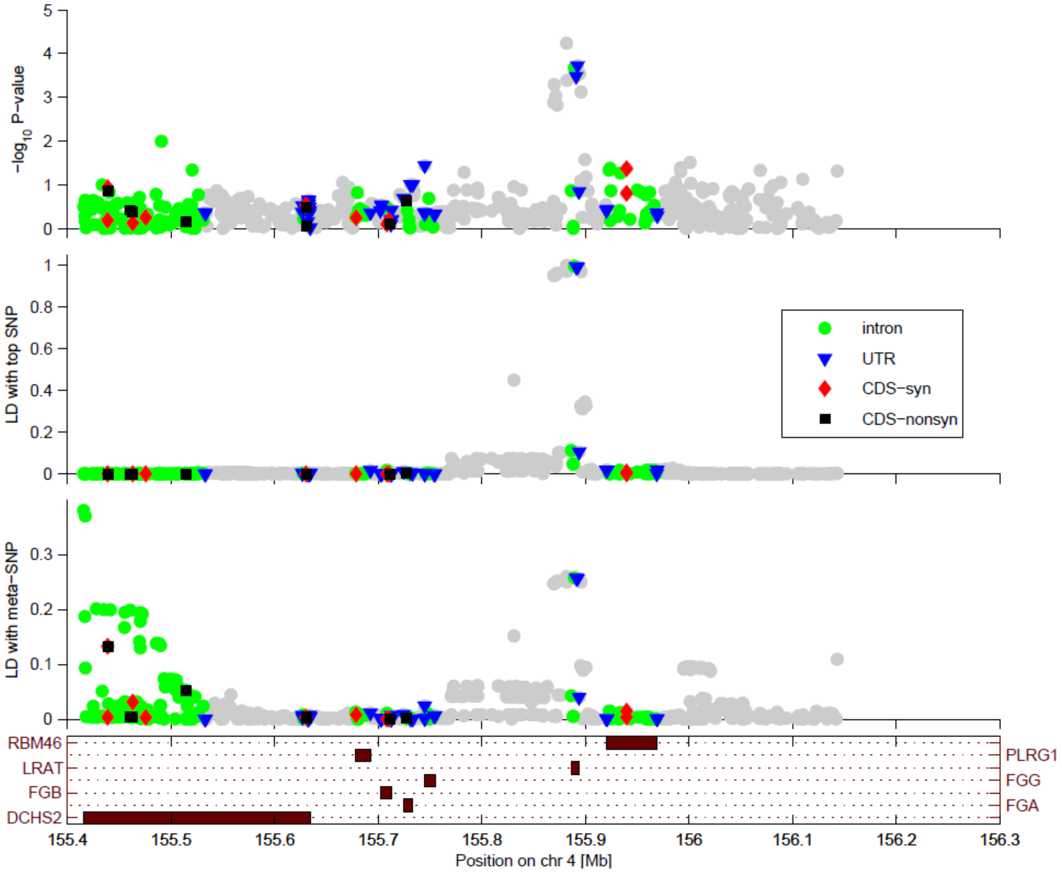

Figure S2. Example of a Triglyceride Association with Allelic Heterogeneity

Here again, important non-synonymous changes are ignored by looking at the lead SNP only, which are picked up by the *multi-SNP* association.

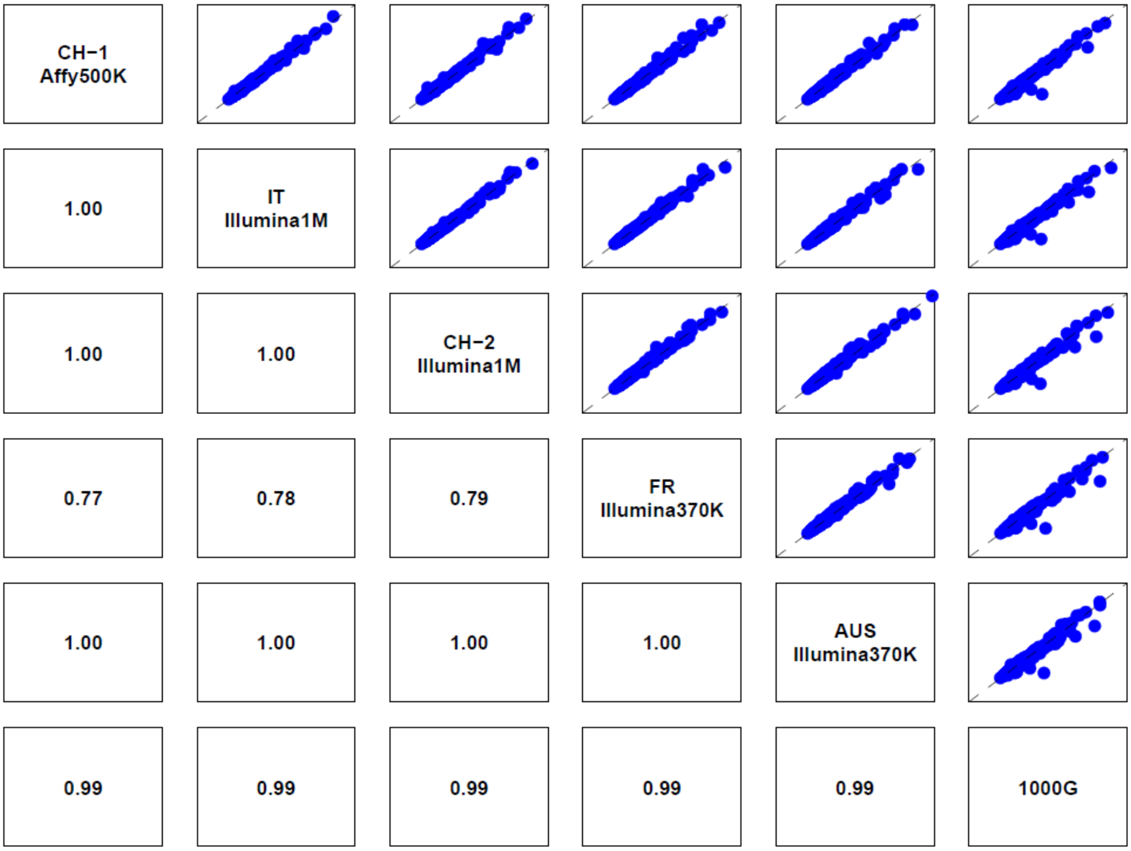

Figure S3. We Obtained Explained Variance Estimates of the Examined 1,628 Height Loci

The estimates depend on the data set used to derive the LD structure. The figure demonstrates that the estimation procedure is robust to slight deviations in the LD structure, such as different genotyping platform, or different (Caucasian) ethnicity. Numbers represent the pair-wise Pearson correlation between the explained variances. CH-1: CoLaus, IT: Hypergenes, CH-2: Swiss Hepatitis C Study, FR: French Hepatitis C cohort, AUS: Australian Hepatitis C cohort, 1000G: 1000 Genomes Project.

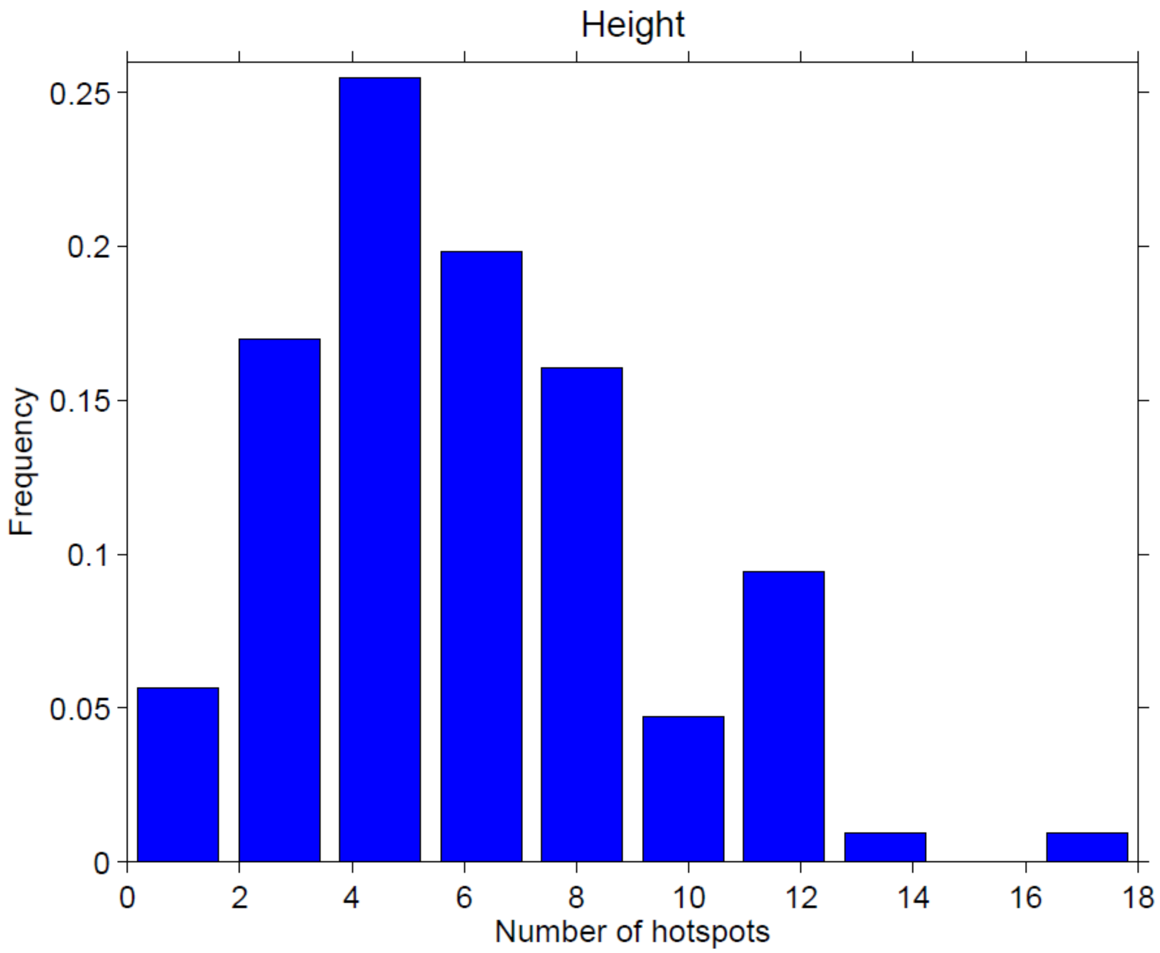

Figure S4. Distribution of the Number of Total Recombination Hotspots that Lie within the Constituent Markers for the (Statistically Significant) *Multi-SNP* Associations for Height

In cases when the causal variant is just a single unobserved SNP and the *multi-SNP* is crossing several recombination hotspots, it can be suspected that the causal SNP is a rather newly emerged variant. We observed that our *multi-SNP* associations are composed of SNPs that are often separated by recombination hotspots. This observation suggests that most *multi-SNP*s we find are either not tagging a single unobserved SNP or if they do the unobserved causal SNP is relatively recent. This phenomenon however may be an artefact of the pruning step.

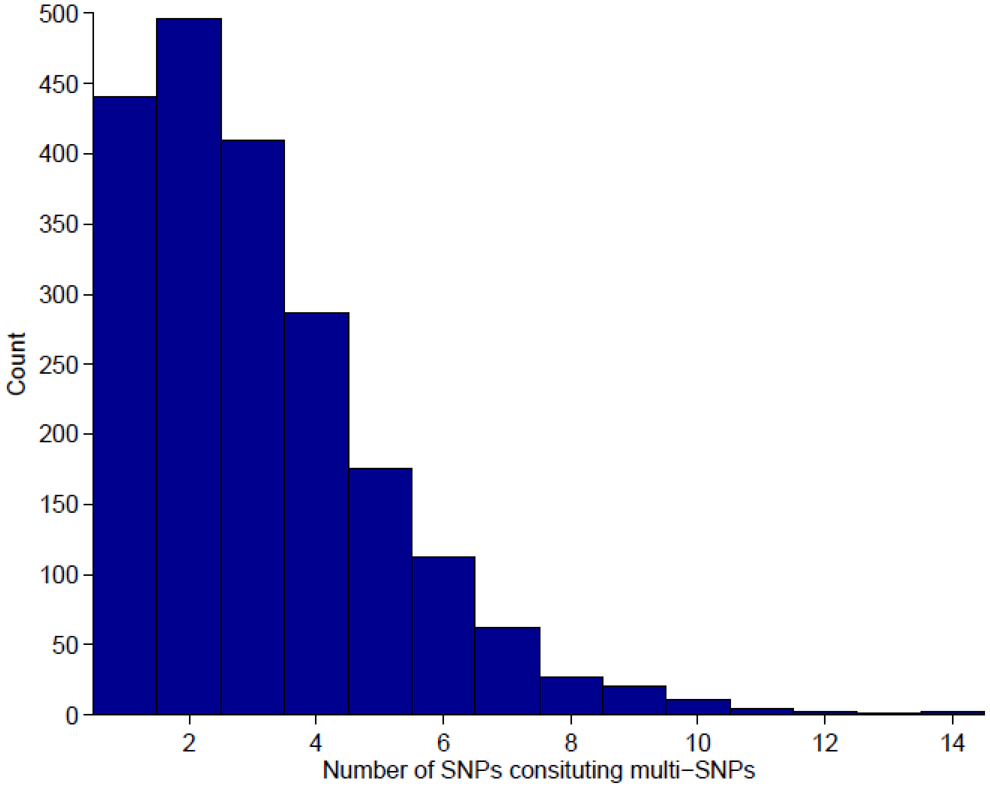

Figure S5. Distribution of the Number of SNPs Constituting the 2,073 *Multi-SNP*s Obtained for Height

The distribution looks very similar for the other examined traits.

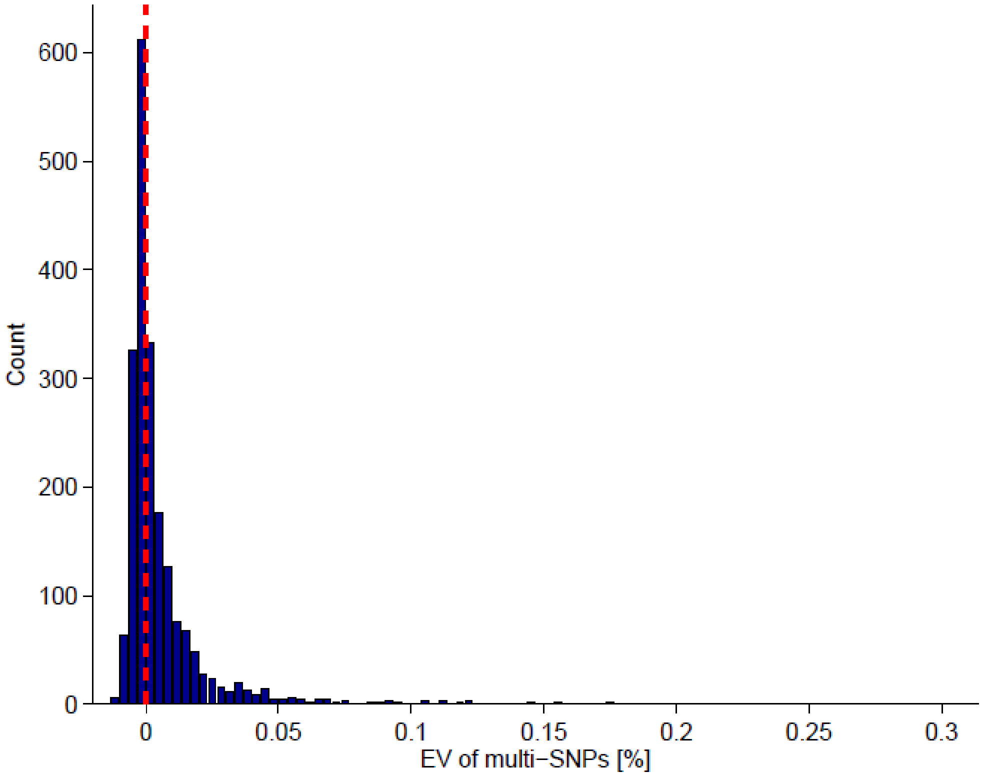

Figure S6. Distribution of Estimated Explained Variances of *Multi-SNP*s for Height

Note that in total 6,458 SNPs were used to build *multi-SNP*s for height, 5,199 for BMI and 4,693 for WHR.

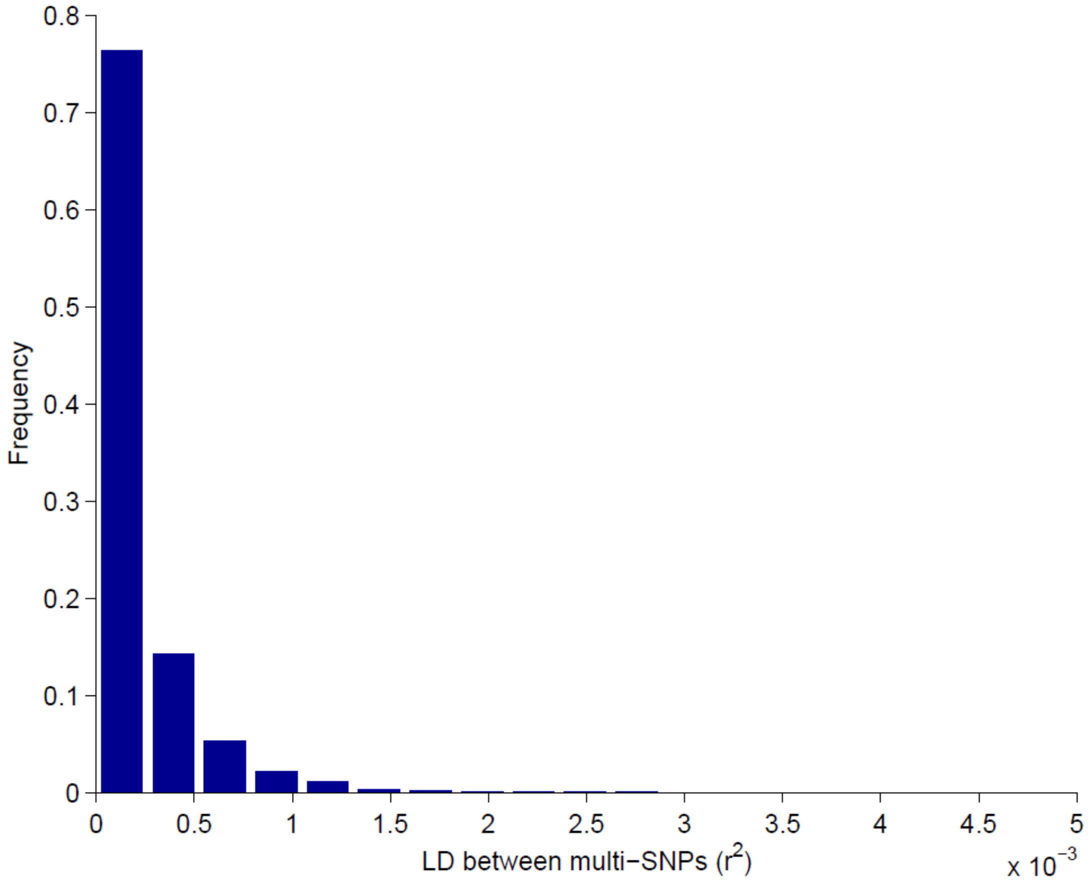

Figure S7. Histogram of LDs between Neighboring *Multi-SNP*s

We verified that the loci showing significant *multi-SNP* association are independent of each other, hence explained variance estimates can be indeed summed up locus-by-locus. To this end, we calculated the pairwise LD between all *multi-SNP*s obtained for height. The distribution shows that all loci are completely independent $r^2 < 0.01$.

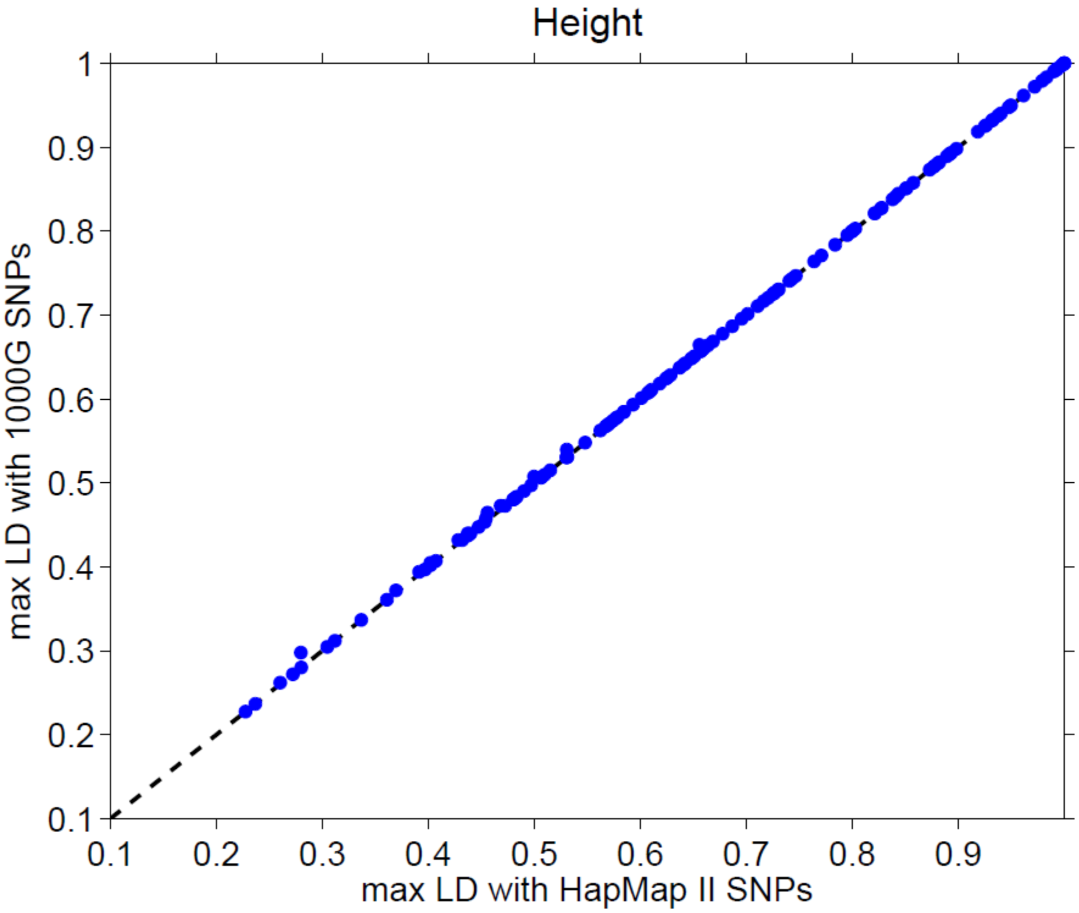

Figure S8. Correlation between *Multi-SNP*s and SNPs in HapMap and 1000 Genomes Project (2010 November release)

Our methodology cannot clearly distinguish between true allelic heterogeneity and multiple independent signals tagging an unobserved variant. We asked, nevertheless, if any discovered *multi-SNP* (composed of HapMap SNPs) could be tagging a single SNP present only in the 1000 Genomes catalogue. This comparison did not identify such *multi-SNP*, indicating that imperfect tagging may be less of an issue for common variant associations. Note however that the LD-pruning step in our procedure reduces the chance of detecting imperfect tagging scenario in set-ups where only association summary statistics are available.

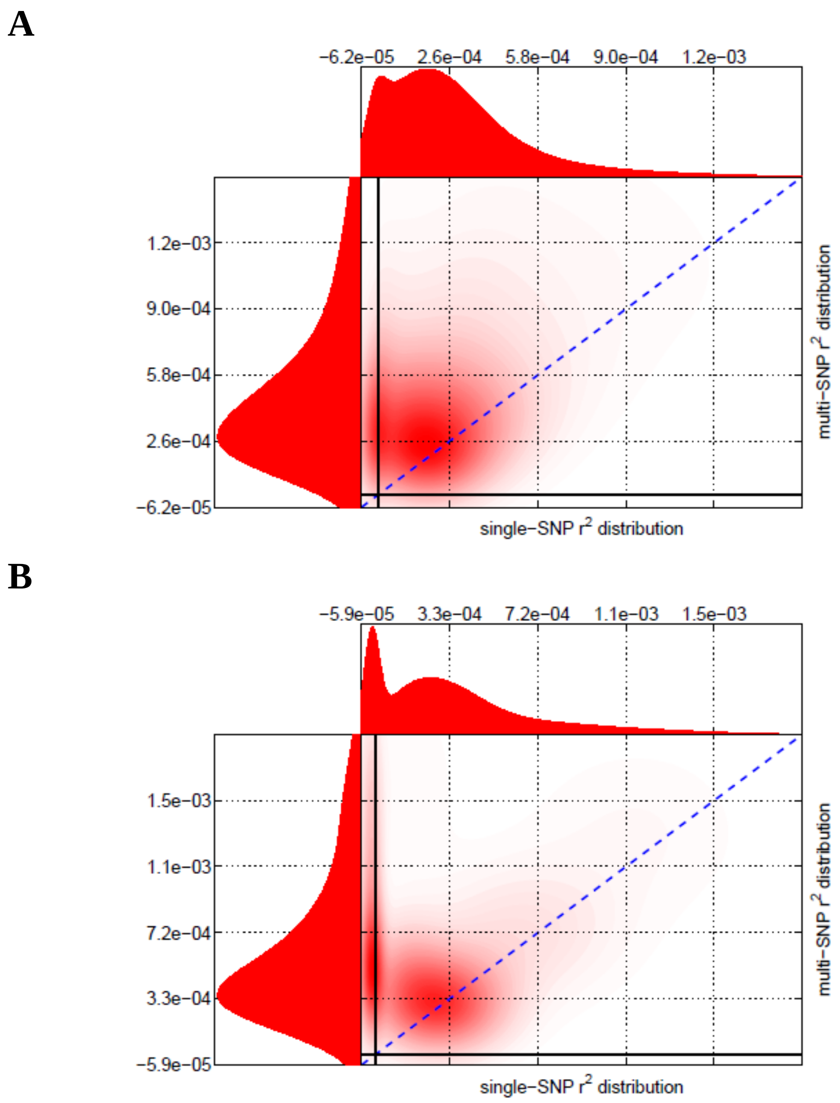

Figure S9. Contour Plot of the Joint Distribution of Single SNP and *Multi-SNP* Explained Variances for (A) Height (B) BMI

The bulk of the points lie on the diagonal, i.e. these *multi-SNP*s do not differ from single SNPs. However, an additional cluster can be observed above the diagonal, which represents loci with substantial allelic heterogeneity. As visible on the marginal distribution plots, a substantial fraction of the estimates are centered closely around zero, indicating no association.

Table S1. Total Explained Variance (TEV) of Single SNPs and Loci for Each Lipid Phenotype

Trait	TEV-single SNPs	TEV-loci
LDL	30.60%	38.13%
HDL	29.12%	33.04%
TG	30.10%	35.86%
TC	29.76%	33.45%

Only those loci were selected whose lead SNP had a P-value < 10^{-2}.

Table S2. Marker Density vs. Allelic Heterogeneity

Trait	OR	P-value	#SNPs
ht	1.825792	0.001	N=806
bmi	1.410130	0.608	N=625
whr	1.784330	0.435	N=428
ldl	1.3652	0.144	N=545
hdl	0.9500	0.825	N=538
tg	1.1821	0.425	N=523
tc	1.3343	0.216	N=555
Combined	1.3451	0.002	N=4020

We found that loci with higher marker density are slightly more prone to harbor allelic heterogeneity (P = 0.002).

Table S3. Marker Conservation vs. Allelic Heterogeneity

Trait	OR	P-value	#SNPs
ht	2.51e+03	8.50e-02	N=806
bmi	1.67e+08	1.28e-01	N=625
whr	4.17e-04	6.46e-01	N=428
ldl	1.29e+03	1.46e-01	N=545
hdl	1.39e+02	3.36e-01	N=538
tg	2.63e+04	4.38e-02	N=523
tc	1.09e+04	7.73e-02	N=555
Combined	2.82e+03	2.47e-04	N=4020

We found evidence that more conserved loci exhibit more allelic heterogeneity ($P = 2.5 \times 10^{-4}$).

Table S4. Explained Variance (EV) of Single SNPs and Loci for Each Anthropometric-Trait Phenotype

Trait	rs	chr	pos	gene	SNP		Locus			P_{diff}	pub	dist (kb)
					EV	*P*	EV	*P*	*m*			
ht	rs7689420	4	145787801	*HHIP*	0.09%	2.1e-11	0.24%	2.8e-24	5	1.7e-15	*	46
ht	rs143384	20	33489169	*GDF5*	0.15%	1.2e-16	0.31%	6.3e-29	7	2.8e-15		
ht	rs17499838	4	82517775	*PRKG2*	0.06%	1.3e-07	0.20%	1.3e-18	8	1.8e-13		
ht	rs310405	6	81857080		0.03%	4.1e-05	0.17%	4.3e-16	8	3.1e-13		
ht	rs10283100	8	120665203	*ENPP2*	0.01%	2.5e-02	0.15%	2.0e-12	11	6.5e-12		
ht	rs4540689	1	170325968	*hsa-mir-214*	0.03%	4.4e-05	0.15%	1.9e-13	10	1.1e-10	*	182
ht	rs12531256	7	17248344	*AHR*	0.00%	2.5e-01	0.09%	1.1e-08	4	4.8e-09		
ht	rs17704359	15	49429780	*GLDN*	0.00%	1.1e-01	0.10%	9.0e-09	10	1.0e-08		

ht	rs1173727	5	32866277	*C5orf23*	0.03%	3.3e-05	0.12%	2.9e-11	6	2.6e-08	*	62
ht	rs817300	9	97420042	*PTCH1*	0.04%	7.6e-05	0.12%	4.4e-10	11	1.5e-07		
bmi	rs3843918	8	53172106	*ST18*	-0.00%	5.4e-01	0.15%	3.1e-13	5	8.3e-14		
bmi	rs2902438	18	25463018	*AC091321.1*	0.01%	3.1e-02	0.16%	3.0e-11	6	7.3e-11		
bmi	rs7688282	4	3036298	*HTT*	-0.00%	5.5e-01	0.09%	6.6e-08	6	2.4e-08		
bmi	rs12682967	9	28197636	*LINGO2*	-0.00%	4.4e-01	0.06%	1.0e-05	4	3.9e-06		
bmi	rs4361395	4	101556552	*EMCN*	-0.00%	5.8e-01	0.05%	1.1e-04	5	4.7e-05		
bmi	rs6671066	1	74875615	*C1orf173*	0.01%	9.8e-03	0.05%	2.7e-05	3	1.9e-04		
bmi	rs17039772	2	50165509	*NRXN1*	0.01%	1.5e-01	0.05%	2.1e-04	6	2.1e-04		
whr	rs9828546	3	173571977	*FNDC3B*	-0.00%	7.8e-01	0.35%	4.5e-11	4	1.0e-11		
whr	rs729761	6	43912548	*VEGFA*	-0.00%	6.4e-01	0.09%	5.6e-05	4	1.9e-05		

Only loci with strong evidence of allelic heterogeneity are listed and –due to space constraint – truncated at the top ten when needed. Abbreviations used: rs = rs number, chr = chromosome, pos = position, $m =$ number of SNPs constituting the given *multi-SNP*, $P_{diff} =$ the likelihood ratio test P-value for locus vs (single) SNP association, pub = already published locus for allelic heterogeneity for the given trait, dist (kb) = distance from the published locus in kb.

Table S5. HDL

rs	chr	pos	gene	SNP		Locus			P_{diff}
				EV	*P*	EV	*P*	*m*	
rs2821231	1	201785004		0.09%	7.6e-04	0.11%	7.2e-19	8	3.9e-17
rs6688343	1	4131616		0.01%	1.1e-03	0.11%	1.2e-16	19	4.5e-15
rs10275447	7	47202379		0.11%	2.6e-04	0.11%	5.7e-17	13	7.6e-15
rs6750325	2	211209577	*CPS1*	0.02%	4.5e-05	0.09%	2.4e-14	14	1.3e-11
rs8045908	16	73631311	*ZNRF1*	0.01%	3.8e-04	0.10%	2.0e-13	17	1.7e-11
rs7216000	17	5950679	*WSCD1*	0.04%	1.9e-03	0.08%	7.4e-13	11	1.9e-11
rs6543264	2	104636864		0.08%	2.5e-03	0.09%	1.2e-12	15	2.4e-11
rs881976	10	69641572	*MYPN*	0.01%	8.8e-04	0.08%	6.4e-13	12	3.1e-11
rs1883025	9	106704121	*ABCA1*	0.15%	2.0e-33	0.24%	4.1e-39	15	6.8e-11
rs3105630	10	30870597		0.01%	3.9e-03	0.10%	5.3e-12	15	7.3e-11
rs7270855	20	51627104	*ZNF217*	0.08%	1.8e-03	0.09%	2.8e-12	11	7.6e-11
rs717384	9	35495346	*RUSC2*	0.08%	4.1e-05	0.10%	5.0e-13	15	2.6e-10
rs17620787	13	100511387	*NALCN*	0.08%	9.7e-04	0.08%	9.0e-12	12	3.9e-10
rs13396033	2	74514619	*RTKN*	0.07%	2.6e-03	0.07%	3.7e-11	9	7.3e-10
rs12699614	7	14403857	*DGKB*	0.01%	6.5e-04	0.07%	1.4e-11	11	8.3e-10
rs9876578	3	30967586		0.07%	2.5e-03	0.08%	1.0e-10	12	2.1e-09
rs1515100	2	226837160		0.04%	2.0e-09	0.11%	3.5e-15	13	8.3e-09
rs17248301	22	49014223	*TUBGCP6*	0.04%	6.6e-03	0.06%	2.1e-09	10	2.0e-08
rs967768	2	436697		0.01%	1.8e-03	0.07%	1.2e-09	13	3.1e-08
rs7979878	12	61334332	*PPM1H*	0.01%	2.0e-04	0.09%	3.1e-10	17	3.8e-08

Table S6. LDL

rs	chr	pos	gene	SNP		Locus			P_{diff}	pub	dist (kb)
				EV	*P*	EV	*P*	*m*			
rs2479409	1	55277237	*PCSK9*	0.13%	2.1e-28	0.35%	9.2e-59	19	1.9e-34	*	453
rs4299376	2	43926079	*ABCG8*	0.23%	2.2e-47	0.37%	1.8e-63	16	2.3e-21	*	1
rs616334	13	45811213	*C13orf18*	0.01%	1.5e-03	0.11%	6.3e-15	10	2.0e-13		
rs7667003	4	91906344		0.01%	2.7e-03	0.09%	4.8e-13	17	8.8e-12		
rs12622910	2	42252421	*EML4*	0.01%	1.0e-03	0.09%	1.4e-12	16	5.6e-11		
rs2061944	11	38099712		0.01%	4.4e-03	0.12%	9.8e-12	24	1.1e-10		
rs13265741	8	2667320		0.08%	1.9e-03	0.08%	4.8e-12	16	1.2e-10		
rs8043572	16	54880732	*GNAO1*	0.01%	3.1e-03	0.10%	2.1e-11	19	3.3e-10		
rs9543624	13	73879043	*AL355390.1*	0.01%	8.6e-04	0.08%	7.1e-12	13	3.3e-10		
rs1481071	3	146700548		0.01%	1.9e-03	0.07%	9.1e-11	12	2.3e-09		
rs7167995	15	45571432		0.04%	2.9e-03	0.08%	1.7e-10	12	3.1e-09		
rs12632087	3	82344237		0.06%	2.1e-03	0.12%	3.8e-10	22	7.3e-09		
rs7660241	4	29925677		0.04%	6.0e-03	0.08%	7.4e-10	12	7.5e-09		
rs13122119	4	6968974	*TBC1D14*	0.01%	4.8e-03	0.05%	7.0e-10	5	8.3e-09		
rs10094246	8	4549684		0.05%	4.0e-03	0.06%	1.1e-09	7	1.6e-08		
rs9458378	6	162140536	*PARK2*	0.06%	4.8e-03	0.06%	2.5e-09	12	3.0e-08		
rs8127846	21	20072824		0.04%	4.2e-03	0.06%	2.3e-09	14	3.0e-08		
rs12733500	1	43980892	*ST3GAL3*	0.06%	7.4e-03	0.07%	4.3e-09	13	3.7e-08		
rs2078668	22	24169191		0.06%	4.8e-03	0.08%	3.8e-09	8	4.7e-08		
rs4747853	10	10412912		0.01%	4.8e-04	0.07%	7.7e-10	16	4.9e-08		

Table S7. TG

rs	chr	pos	gene	SNP		Locus			P_{diff}
				EV	*P*	EV	*P*	*m*	
rs4074448	15	61675547	*FBXL22*	0.02%	8.9e-07	0.13%	3.5e-20	14	6.0e-16
rs9600211	13	73460669	*KLF12*	0.09%	1.5e-05	0.15%	3.2e-18	14	4.7e-15
rs11198380	10	120134085		0.01%	9.8e-03	0.10%	1.2e-13	15	8.1e-13
rs2055014	11	29152307		0.01%	1.2e-04	0.11%	4.6e-15	22	9.1e-13
rs2392446	7	36514909	*AOAH*	0.08%	2.0e-03	0.10%	1.8e-13	16	4.1e-12
rs12634505	3	120566289	*ARHGAP31*	0.06%	3.1e-03	0.08%	2.4e-12	16	4.0e-11
rs16887883	7	77434104	*PHTF2*	0.01%	1.4e-03	0.07%	6.0e-12	13	2.0e-10
rs1360144	1	194849397	*KCNT2*	0.07%	3.0e-03	0.10%	3.0e-11	20	4.7e-10
rs7862588	9	87467411	*AGTPBP1*	0.03%	5.8e-03	0.07%	5.3e-11	13	5.5e-10
rs16856552	1	230373853		0.08%	8.2e-04	0.09%	1.3e-11	19	5.6e-10
rs12501328	4	155881780	*LRAT*	0.02%	5.8e-05	0.10%	2.6e-12	18	8.7e-10
rs11101342	10	49372158	*ARHGAP22*	0.01%	4.1e-04	0.09%	2.9e-11	13	2.4e-09
rs3755833	3	37839414	*ITGA9*	0.01%	1.6e-03	0.09%	1.2e-10	25	2.7e-09
rs11928774	3	72522838		0.05%	8.8e-03	0.06%	5.7e-10	7	4.2e-09
rs2247056	6	31373468		0.06%	1.6e-15	0.16%	1.2e-20	15	4.7e-09
rs1899227	5	16353092		0.07%	3.5e-04	0.08%	6.7e-11	14	6.1e-09
rs12598987	16	77222047	*WWOX*	0.01%	1.1e-03	0.10%	2.3e-10	14	8.2e-09
rs9939477	16	22103983	*SDR42E2*	0.01%	3.1e-04	0.06%	7.7e-11	5	1.0e-08
rs852058	20	17051888		0.08%	3.0e-03	0.07%	6.1e-10	13	1.0e-08
rs3858418	11	101276457	*ANGPTL5*	0.04%	1.9e-03	0.09%	6.1e-10	20	1.3e-08

Table S8. TC

rs	chr	pos	gene	SNP		Locus			P_{diff}	pub	dist (kb)
				EV	*P*	EV	*P*	*m*			
rs2479409	1	55277237	*PCSK9*	0.10%	4.0e-24	0.30%	1.0e-53	19	1.9e-33	*	393
rs4299376	2	43926079	*ABCG8*	0.21%	4.9e-45	0.34%	6.6e-62	16	5.1e-22	*	60
rs10920620	1	201601461		0.09%	1.7e-03	0.10%	1.2e-15	16	3.2e-14		
rs17050211	3	9293569		0.07%	4.6e-03	0.09%	6.9e-14	16	8.4e-13		
rs11696774	20	38351200		0.02%	2.5e-05	0.09%	1.6e-14	14	1.4e-11	*	263
rs1900310	2	209919219		0.01%	1.9e-03	0.09%	9.4e-13	19	2.1e-11		
rs1408194	13	45825430	*C13orf18*	0.02%	5.5e-03	0.09%	7.3e-12	11	7.9e-11		
rs17681539	2	42845324	*OXER1*	0.01%	1.0e-03	0.08%	5.0e-12	14	2.0e-10		
rs12635648	3	146827286		0.01%	5.5e-03	0.07%	1.5e-10	15	1.6e-09		
rs2063343	14	91270104	*CATSPERB*	0.02%	5.1e-03	0.07%	5.7e-10	18	6.2e-09		
rs1936471	6	96432003		0.01%	2.2e-03	0.07%	3.4e-10	21	6.4e-09		
rs11644777	16	22860124		0.01%	5.3e-04	0.07%	1.9e-10	14	1.2e-08		
rs5770794	22	49227646	*SAPS2*	0.01%	3.5e-04	0.07%	1.6e-10	12	1.5e-08		
rs10094246	8	4549684		0.05%	4.6e-03	0.05%	3.2e-09	7	4.0e-08		
rs349588	5	103699765		0.02%	4.7e-05	0.08%	1.1e-10	16	4.5e-08		
rs17248550	4	91985633	*TMSL4*	0.01%	1.0e-04	0.07%	2.1e-10	16	4.6e-08		
rs11645345	16	54814499	*GNAO1*	0.01%	7.3e-04	0.07%	1.2e-09	19	5.2e-08		
rs16971446	18	34852986		0.04%	7.1e-04	0.07%	3.7e-09	20	1.5e-07		
rs973563	5	120039598	*PRR16*	0.03%	1.0e-03	0.07%	4.5e-09	13	1.7e-07		
rs9543624	13	73879043	*AL355390.1*	0.01%	2.4e-03	0.06%	9.3e-09	13	1.9e-07		