Title: A method to estimate the contribution of regional genetic associations to complex traits from summary association statistics

Guillaume Pare[1,2,3,4,*], Shihong Mao[3], Wei Q. Deng[5]

1 Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada, 2 Population Genomics Program, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada, 3 Population Health Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada, 4 Thrombosis and Atherosclerosis Research Institute, Hamilton, Canada, 5 Department of Statistical Sciences, University of Toronto, Toronto, Canada

*Corresponding author: pareg@mcmaster.ca

Abstract:

Despite considerable efforts, known genetic associations only explain a small fraction of predicted heritability. Regional associations combine information from multiple contiguous genetic variants and can improve variance explained at established association loci. However, regional associations are not easily amenable to estimation using summary association statistics because of sensitivity to linkage disequilibrium (LD). We now propose a novel method to estimate phenotypic variance explained by regional associations using summary statistics while accounting for LD. Our method is asymptotically equivalent to multiple regression models when no interaction or haplotype effects are present. It has multiple applications, such as ranking of genetic regions according to variance explained and derivation of regional gene scores (GS). We show that most genetic variance lies in a small proportion of the genome, and that GS derived from regional associations can improve trait prediction above optimal polygenic scores. Our results also suggest regional associations underlie known linkage peaks.

Introduction

Currently known genetic associations only explain a relatively small proportion of complex traits variance. In accordance with the widely accepted polygenic nature of complex traits, it has been proposed that weak, yet undetected, associations underlie complex trait heritability[1]. We have recently shown that the joint association of multiple weakly associated variants over large chromosomal regions contributes to complex traits variance[2]. Such regional associations are not easily amenable to estimation using summary-level data such as association statistics because of sensitivity to linkage disequilibrium (LD). Nonetheless, only large meta-analyses have the necessary power to identify weakly associated variants. In this report, we propose a novel method to assess the contribution of regional associations to complex traits variance using summary association statistics. Estimation of regional associations can identify key genomic regions involved in the determination of complex traits, and can improve predictiveness of gene scores.

Clustering of weak associations within defined chromosomal regions has been previously suggested[3] and can increase variance explained at established association loci as compared to genome-wide significant SNPs alone[4]. Such regional associations extended up to 433.0 Kb from genome-wide significant SNPs[2], a distance compatible with long-range *cis* regulation of gene expression[5,6]. Furthermore, regional associations appeared to be the results of multiple weak associations rather than one or a few very significant univariate associations. These results point towards the existence of key regulatory regions where functional genetic variants aggregate, the identification of which can lead to novel biological insights and a better understanding of complex traits genetics.

Several methods have been described to estimate the overall contribution of common genetic variants to complex traits variance, but no method exists to estimate

regional association using summary association statistics. For instance, a popular approach is based on variance component models using genetic relatedness as the variance-covariance matrix of the random effect. An implementation of this approach uses REML[1] to estimate the genetic effect, and variations have been reported to either take into account LD between SNPs[7] or to handle large datasets[8]. While very useful and informative, all of these approaches require individual-level data. This latter pitfall has been overcome by development of LDScore[9], which uses summary-level association statistics as input and has been shown to be equivalent to Elston regression[10]. However, LDScore is ill suited for estimation of regional heritability because it requires regressing genetic effect as function of linkage disequilibrium over large number of SNPs. There is thus a need for a method to estimate the regional contribution of common genetic variants using summary association statistics.

Results

Comparison of regional variance explained using summary statistics to multiple regression and variance component models

Our method estimates regional contribution to complex trait variance using summary data by adjusting the variance explained by each SNP for its LD with neighboring SNPs (Online Methods). As the method is equivalent to multiple regression models when no haplotype or interaction effects are present, we sought to compare overall variance explained by the novel method to multiple regression and variance component models. To do so, we applied our method to BMI and height in the Health Retirement Study (HRS; N = 7,776) for which individual-level genotypes were available. We first divided the genome in SNP blocks minimizing inter-block LD and tested all three methods, combining adjusted R^2 from all blocks to derive genome-wide multiple regression estimates. Estimates of phenotypic variance explained using summary association statistics and variance components models were highly consistent (Figure 1). In contrast, estimates derived from a multiple linear regression model were higher and

dependent on SNP block size, with smaller blocks having higher estimates because of "spillage" association.

Adopting a median SNP block size of 1 Mb by setting minimum block size at 195 SNPs and maximum size at 205 SNPs, the overall genetic contribution was estimated at 0.12 (95% CI 0.00-0.24) for BMI using our novel method and 0.43 (95% CI 0.19-0.66) using a multiple linear regression model. Corresponding estimates for height were 0.27 (95% CI 0.15-0.39) and 0.39 (95% CI 0.15-0.63). Estimates of genetic variance explained at the SNP block level were moderately correlated between both methods (r^2=0.26 for BMI and r^2=0.24 for height). Nonetheless, genetic variance was significantly higher with multiple regression for BMI, with a difference of 0.31 (95% CI 0.09-0.51; p = 0.0047). No significant difference genetic variance was observed for height. In comparison, the overall genetic variance was estimated at 0.13 (95% CI 0.03-0.24) and 0.28 (95% CI 0.17-0.38) using variance component models[11] for BMI and height. We also explored the impact of adjustment for genetic principal components. The first 20 components provided adequate protection against population stratification while inclusion of fewer components led to inflation in genetic variance, especially for height.

Estimation of genetic variance using GIANT summary association statistics

We applied our method to summary association statistics from the GIANT consortium for BMI ($N_{effective}$=229,031) and height ($N_{effective}$=240,773). Genome-wide genetic variance was estimated at 0.098 (95% CI 0.095-0.102) for BMI and 0.713 (95% CI 0.709-0.717) for height. We next sought to determine if SNP blocks could be ranked for genetic variance using summary association statistics by calculating the variance explained by each block in GIANT and sorting them in a decreasing order. We then tested these same blocks in HRS, successively adding a higher proportion of top GIANT blocks but estimating genetic variance solely with HRS data. As illustrated in Figure 2, a relatively small proportion of blocks contributed disproportionally to genetic variance. When dividing the genome in 1 Mb blocks, corresponding to 195-205 contiguous SNPs, the top 25% GIANT blocks explained 0.91 of BMI genetic variance (as estimated by our

proposed approach i.e. =0.111/0.121) and 0.71 of height genetic variance. These results could potentially be explained by the presence of one or more very strong associations in each of these top SNP blocks. To explore this possibility, we recorded the minimum univariate association SNP *p*-value for each block in both GIANT and HRS (Figure 3). Median minimum univariate *p*-value was 2.7×10^{-4} for BMI in GIANT, with 0.09 of blocks having one or more genome-wide significant associations ($p < 5 \times 10^{-8}$). On the other hand, the median minimum *p*-value was 1.0×10^{-10} for height in GIANT, with 0.70 of blocks having one or more genome-wide significant associations. The median minimum univariate *p*-values were 4.3×10^{-3} and 3.6×10^{-3} for BMI and height in HRS. No SNP reached genome-wide significance in HRS.

Gene score analysis

Informed by previous results, we tested whether inclusion of regional associations into gene scores could improve prediction as characterized by the adjusted R^2 for goodness of fit. To do so, we first ranked SNP blocks (of median size 1 Mb) according to their genetic variance explained in GIANT, as done above. We then successively added top SNP blocks to a gene score, deriving gene score regression coefficients from GIANT summary association statistics (see Online Methods) but testing in the independent sample of HRS. Including all SNP blocks in the gene score yielded a prediction R^2 of 0.051 for BMI and 0.052 for height (Figure 4). When including the optimal proportion of top blocks, prediction R^2 was 0.053 for BMI (top 0.87 blocks) and 0.083 for height (top 0.25 blocks). Since our approach is complementary to polygenic scores, we next added regional association gene scores to the optimal polygenic gene score. Optimal polygenic scores were derived using LD-based clumping procedures to remove correlated SNPs and were determined to have a *p*-value threshold of 10^{-1} for BMI and 10^{-4} for height, with corresponding prediction R^2 of 0.052 and 0.101. Addition of optimal regional association gene scores improved prediction R^2 by 13% to 0.059 for BMI and by 4% to 0.105 for height as compared to polygenic score alone (*p* for improvement $< 10^{-8}$ in both cases).

We also sought to directly compare regional association gene scores derived using our novel approach to gene scores based on a multiple linear regression. As individual-level data is needed to calculate the multiple linear regression coefficients, this analysis was performed in HRS using 5-fold cross-validation. In other words, the HRS dataset was randomly divided in 5 parts, with 1 part successively used for validation of gene scores derived from the 4 remaining parts. In keeping with previous analyses, SNP blocks were included in gene scores according to GIANT ranking. Gene scores derived using our novel approach outperformed multiple linear regression gene scores for height but not for BMI (Figure 5). For instance, when including the top 0.05 GIANT blocks mean prediction R^2 was 0.017 with our approach and 0.006 with multiple linear regression gene scores (Wilcoxon p for difference 0.008). No difference was observed when including all blocks.

Analysis of known linkage peaks

A unique application of our method is the estimation of genetic variance over extended genomic regions using summary association statistics data. We therefore tested the hypothesis that previously identified linkage peaks are enriched for regional associations. Based on the largest single linkage study of height and BMI[12], 3 peaks with suggestive (LOD>2.0) evidence of linkage in Europeans were identified, all for height. The only peak with LOD>3.0 showed a significant (p=0.002) enrichment in regional association within a distance of +/- 7.5 Mb of the linkage marker, corresponding to an estimated excess regional genetic variance of 0.0044 as compared to genome-wide average (Table 1). Upon closer inspection, the region contained several sub-regions with genome-wide significant associations.

Discussion

We propose a novel method to estimate regional genetic variance from summary association statistics. Using this method, we confirmed a major role of regional associations in complex trait heritability, whereby the aggregation of genetic associations

contributes disproportionally to phenotypic variance. Selecting the top SNP blocks from the GIANT meta-analysis, we showed that 25% of the genome is responsible for up to 0.91 and 0.71 of BMI and height genetic variance. A large proportion of these blocks had unremarkable minimum univariate *p*-values, suggesting the presence of multiple weak associations underlies their impact on phenotypic variance, especially for BMI. Concentration of genetic associations within these regions also supports the existence of critical nodes in the genetic regulation of complex traits such as height and BMI, with implications not only for association testing but also for population genetics and natural selection. Our proposed method has several advantages. First, it provides results that are highly consistent with the widely used variance component models requiring individual-level data. Second, it is immune to "spillage" association caused by LD between SNP blocks, thus enabling blocks to be freely defined. Third, it is computationally straightforward and can be intuitively adapted into gene scores that are complementary to polygenic scores and superior to a multiple linear regression approach. Fourth, it is agnostic and therefore complementary to functional annotations of the genome.

Our results suggest a combination of strong genetic associations and regional associations contribute to complex traits variance, with the relative proportions varying across traits. For instance, a higher proportion of genetic variance was found in the top 25% blocks for BMI yet these blocks had less significant minimum univariate *p*-value than height. On the other hand, the addition of regional gene scores to optimal polygenic score did not improve prediction R^2 as much for height (4%) as for BMI (13%), despite polygenic score *p*-value threshold being set at 10^{-1} for BMI and 10^{-4} for height. These observations are consistent with the complementarity of polygenic and regional association scores. Indeed, polygenic scores capture significant associations that might not necessarily be in significant regions whereas regional genetic scores can capture significant regions that might not necessarily contain strong associations. In addition, regional gene scores can average association signals over many SNPs when they are in LD, which could provide added robustness even for strong associations.

Our method can also be used to estimate the genetic variance explained by extended regions. We therefore tested the hypothesis some of the previously identified linkage peaks are the result of regional associations. The only peak with LOD>3.0 from the largest linkage study of height showed a marked and significant enrichment in regional association[12]. This region had been previously identified in other linkage studies[13]. Genetic variance explained by the region was estimated at 0.0044, which is unlikely to explain the linkage peak by itself. Nonetheless, the juxtaposition of linkage and regional associations points towards concentration of functional variants as a potential explanation for the observed linkage. The lack of regional association at other peaks can be explained by false-positive linkage results, the possibility rare variants not captured in GWAS studies underlie linkage peaks, or by differences in genetic architecture between studied populations.

A few limitations are worth mentioning. First, the assumption that SNPs contribute to genetic variance without any interaction or haplotype effects can lead to an underestimation of genetic variance. Indeed, the "loss" in variance explained of 0.31 for BMI as compared to multiple linear regression could be due to either better capturing of gene x gene interaction effects[2] or "tagging" of untyped rare variants[14] by multiple linear regression models. Nonetheless, our estimates were consistent with variance component models and gene scores derived using our proposed approach were equal or superior to gene scores derived from multiple linear regression models. These results emphasize the need for statistical models that more fully capture genetic variance, especially when strong haplotype effects are expected[15]. Second, estimation of overall genome-wide genetic variance using summary association statistics is dependent on both the accuracy of SNP effect size estimates and correct specification of LD structure. For example, differences in adjustment for population stratification (e.g. principal components) in individual studies participating to a meta-analysis could potentially affect results. Indeed, estimates of BMI genetic variance explained were similar between HRS (0.12) and GIANT (0.10), but corresponding estimates for height were markedly different (i.e. 0.27 and 0.71). While the latter results might represent true population differences in genetic architecture, they should be interpreted with caution and emphasis placed on the relative

contribution of regions rather than the overall variance explained. Third, we have only tested continuous traits. However, our method can be easily adapted to other outcome types through the use of generalized linear models.

In this report, we establish a novel method to estimate the regional contribution of common variants to complex traits variance using summary association statistics. Beyond estimation of overall genetic variance using summary association statistics, our method has many applications. For instance, by ranking genetic regions we showed regional enrichment in genetic associations for BMI and height, and leveraged these observations into improved gene scores. Identification of key genetic regions is also important for future fine-mapping studies. Indeed, our method can be used to perform network analysis using summary association statistics, or to combine summary association statistics with other types of genetic annotations such as linkage studies results, as we have done. Finally, our method can provide insights into patterns of genetic associations, such as the observed dissimilarities between BMI and height with respect to distribution of regional associations and response to regional gene scores.

Online Methods

Methods overview

We have previously shown that large regions joint associations, where multiple genetic variants are included as independent variables in a linear model, combine robustness to linkage disequilibrium and non-additive effects while retaining adequate power as compared to other association models[2]. However, such regional joint associations are not easily amenable to estimation using summary data because of sensitivity to linkage disequilibrium. To evaluate the contribution of large regions joint associations to variance of complex traits, we first devised an algorithm to divide the genome in blocks of SNPs in such a way as to minimize inter-block linkage disequilibrium and thus "spillage" associations. We then derived a method to estimate regional contribution to complex trait variance using summary data and showed this method to be equivalent to multiple regression models when genetic effects are strictly additive (i.e. no haplotype or interaction effect). Using regional variance estimates from summary-level association statistics for height and BMI from the GIANT consortium, we estimated the distribution of genetic effects across the genome. Finally, we used the latter results to derive genetic scores for prediction of height and BMI in the Health Retirement Study (HRS).

Dividing the genome into SNP blocks

We first divided the genome into regions of contiguous SNPs varying in size (e.g. from 195 SNPs to 205 SNPs), herein referred as SNP blocks and used as units for regional associations. To minimize inter-block LD and thus "spillage" associations, we devised a greedy algorithm optimizing choice of block boundary sequentially from one end of a chromosome to the other. Briefly, using user-defined minimum and maximum block size (in number of SNPs) and starting at one end of a chromosome arm, each possible "cut-point" between the first and second block are tested and maximal LD (r^2) between pairs of SNPs crossing block boundary is calculated. The cut-point that

minimizes maximal LD is chosen, thus defining the first block, and the procedure is repeated for each subsequent block until all SNPs on a chromosome arm have been assigned to a block. We empirically determined that SNP blocks of size 195 SNPs to 205 SNPs (median 200 SNPs) had a median physical size of 1 Mb.

Estimating the contribution of regional genetic associations with individual-level genotypes

Use of adjusted R^2 lends itself nicely to estimation of regional variance explained when individual-level genotypes are available[2]. In this context, SNPs comprised in a given SNP block are included as independent variables in a multiple linear regression model and the goodness of fit statistic, adjusted R^2 calculated. Because adjusted R^2 accounts for the number of SNPs included in each block, expected adjusted R^2 is zero under the null of no association and the expected sum of adjusted R^2 over all SNP blocks is also zero. The overall contribution of regional associations to complex traits variance can be estimated by simply summing adjusted R^2 over all (or selected) SNP blocks. Furthermore, the distribution of adjusted R^2 under null has been previously described[16] and can be used to derive the distribution of the sum of adjusted R^2.

Estimating the contribution of regional genetic associations with summary association statistics data

Estimating the contribution of regional genetic associations from summary association statistics is challenging when the exact SNP linkage disequilibrium structure of source populations is unknown. While approaches have been described to perform joint or conditional associations[15,17] using estimated SNP covariance matrices, they do not perform well when estimating regional variance explained because of sensitivity to misspecification of linkage disequilibrium (data not shown) and ensuing overestimation of regional associations. We therefore created a simple procedure to estimate regional variance explained from summary association statistics data, adjusting for linkage disequilibrium. Without loss of generalizability, we assume a quantitative trait (Y)

standard normally distributed and genotypes normalized to have mean=0 and standard deviation=1 throughout. Given an n x m genotype matrix $\boldsymbol{X}$ representing genotypes at m SNPs in n individuals and the pairwise linkage disequilibrium (r^2) between two SNPs k and l as $r^2_{k,l}$, for a SNP d, the following LD adjustment (η_d) can be defined as the summation of LD between the d^{th} SNP and 100 SNPs upstream and downstream:

$$\eta_d = \sum_{e=d-100}^{e=d+100} r^2_{d,e}$$

with a distance of 100 SNPs assumed sufficient to ensure linkage equilibrium (other values might be used). Only including SNPs with summary GWAS statistics in the sum, variance explained by each SNP d is given by:

$$R^2_d = \frac{b^2_d}{\eta_d}$$

where b_d denotes the univariate regression coefficient commonly reported in GWAS results. Regional variance explained is then given by the sum of R^2_d over SNPs comprised in a given region. Assuming a strictly additive genetic model where each SNP contributes additively to a trait without any interaction or haplotype effects, we demonstrate the total variance explained over a region $\sum_d R^2_d$, is approximately equal to the multiple linear regression variance explained R^2 when the sample size is sufficiently large.

Suppose the true genetic effect is a fixed vector β, whose individual components are random over all SNPs, $i = 1, 2, \cdots, m$, with mean 0 and variance σ^2. The genetic model can be expressed as:

$$Y = X\beta + \varepsilon$$

where ε is a vector of standard normal error with identity variance covariance matrix. Then, the vector multiple linear regression coefficient B is given by:

$$B = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

On the other hand, the univariate regression coefficients, denoted by lower case b and which can be obtained from GWAS summary statistics, are given by

$$b = \frac{X'Y}{N} = \frac{1}{N}X'X\beta + \frac{1}{N}X'\varepsilon$$

The multivariate variance explained R^2 can be written in terms of the true effect and the error term:

$$\begin{aligned} R^2 &= \frac{1}{N}(XB)'(XB) \\ &= \frac{1}{N}(X\beta)'(X\beta) + \frac{1}{N}(X(X'X)^{-1}X'\varepsilon)'(X(X'X)^{-1}X'\varepsilon) \\ &= \frac{1}{N}(X\beta)'(X\beta) + \frac{1}{N}\varepsilon'(X(X'X)^{-1}X')\varepsilon \end{aligned}$$

and since the error term has identity variance covariance and the true effect β has variance covariance matrix $\sigma^2 I$, the expected variance explained is simplified to

$$\begin{aligned} E[R^2] &= \frac{1}{N}E(\beta'X'X\beta + \varepsilon'(X(X'X)^{-1}X')\varepsilon) \\ &= \sigma^2 + \frac{tr(X(X'X)^{-1}X')}{N} = \sigma^2 + \frac{m}{N} \end{aligned}$$

and the variance can be calculated according to the fact that that ε_j^2 has a chi-squared distribution with 1 degree of freedom:

$$Var(R^2) = \frac{2m}{N^2}.$$

The total variance explained over a region $\sum_d R_d^2$ can be calculated using only the univariate regression coefficients from GWAS:

$$\sum_d R_d^2 = \sum_d \frac{b_d^2}{\eta_d}$$

$$= b' \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} b$$

$$= \frac{1}{N^2}(X'X\beta)' \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} (X'X\beta) + \frac{1}{N^2}(X'\varepsilon)' \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} (X'\varepsilon)$$

Replacing $X'X$ by D such that $X'X = D = \begin{bmatrix} D_1 & D_2 & \cdots & D_m \end{bmatrix}$, where D_k is a 1 x m vector whose entries represent the k^{th} row of D:

$$\sum_d R_d^2 = \frac{1}{N^2}\beta' \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{bmatrix} \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} \begin{bmatrix} D_1 & D_2 & \cdots & D_m \end{bmatrix} \beta + \frac{1}{N^2}\sum_{d=1}^{m} (X_d'\varepsilon)'(X_d'\varepsilon) \Big/ \eta_d$$

$$= \frac{1}{N^2}\beta' \begin{bmatrix} D_1D_1' \Big/ \eta_1 \\ D_2D_2' \Big/ \eta_2 \\ \vdots \\ D_mD_m' \Big/ \eta_m \end{bmatrix} \beta + \frac{1}{N^2}\sum_{d=1}^{m} (\varepsilon' X_d X_d' \varepsilon) \Big/ \eta_d$$

$$\sim \beta'\beta + \frac{1}{N^2}\sum_{d=1}^{m} (\sum_{j=1}^{N} x_{j,d}\varepsilon_j)^2 \Big/ \eta_d$$

Since we can assume that $D_kD_k' \Big/ \eta_k = \frac{\sum_{d=1}^{m} N^2 r_{k,d}^2}{\eta_k} = \frac{\sum_{d=1}^{m} N^2 r_{k,d}^2}{\sum_{d=k-100}^{k+100} r_{k,d}^2} \sim N^2$ as LD tends to be weak 100 SNPs upstream and downstream away from the index SNP. In addition,

$$E\left[(X_k'\varepsilon)'(X_k'\varepsilon)\right] = N$$

such that we can simplify the expected total variance explained using the summary statistics to:

$$E(\sum_d R_d^2) = \sigma^2 + \frac{1}{N}\sum_d 1 \Big/ \eta_d .$$

The variance of $\sum_d R_d^2$ can be similarly derived since

$$\varepsilon' X \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} X' \varepsilon$$

is a quadratic form. By using the cyclic property of trace and the fact that

$$D_k D_k' \Big/ \eta_k = \frac{\sum_{d=1}^{m} N^2 r_{k,d}^2}{\eta_k} = \frac{\sum_{d=1}^{m} N^2 r_{k,d}^2}{\sum_{d=k-100}^{k+100} r_{k,d}^2} \sim N^2$$

the variance can be expressed as

$$Var(\sum\nolimits_d R_d^2) = \frac{1}{N^4} Var(\varepsilon' X \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} X' \varepsilon)$$

$$= \frac{2}{N^4} tr(X \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} X' X \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} X').$$

$$= \frac{2}{N^4} tr(X' X \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix} X' X \begin{bmatrix} 1/\eta_1 & 0 & \cdots & 0 \\ 0 & 1/\eta_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 1/\eta_m \end{bmatrix})$$

$$= \frac{2}{N^4} \left\{ \sum\nolimits_{d=1}^{m} \frac{D_d D_d'}{\eta_d^2} \right\} \sim \frac{2}{N^2} \left\{ \sum\nolimits_{d=1}^{m} \frac{1}{\eta_d} \right\}$$

Variance is thus similar to the corresponding variance for linear multiple regression models, with the notable exception the number of SNPs m in $Var(R^2) = \frac{2m}{N^2}$ is replaced

by the "effective" number of genetic markers $\sum_{d=1}^{m} \frac{1}{\eta_d}$. In other words, $Var(\sum_d R_d^2)$ is expected to be equal or lower than the corresponding $Var(R^2)$.

Gene scores

Our approach lends itself naturally to derivation of gene scores. Adjusting GWAS univariate regression coefficients using:

$$b_d' = \frac{b_d}{\sqrt{\eta_d}}$$

It follows that:

$$R_d^2 = \frac{b_d^2}{\eta_d} = b'^2_d$$

And

$$\sum_d R_d^2 = \sum_d b'^2_d$$

Such that the new adjusted regression coefficients b_d' directly capture the contribution of each SNP to regional variance explained adjusting for LD.

Health Retirement Study

We conducted large region joint association analysis for height using genome-wide data from the publicly available Health Retirement Study (HRS; dbGaP Study Accession: phs000428.v1.p1). HRS quality control criteria were used for filtering of

both genotype and phenotype data, namely: (1) SNPs and individuals with missingness higher than 2% were excluded, (2) related individuals were excluded, (3) only participants with self-reported European ancestry genetically confirmed by principal component analysis were included, (4) SNPs with Hardy-Weinberg equilibrium $p < 1x10^{-6}$ were excluded, (5) individuals for whom the reported sex does not match their genetic sex were excluded, (6) SNPs with minor allele frequency lower than 0.02 were removed. After further pruning SNPs for LD using PLINK v.1.07[18] with window size = 100 SNPs, step size = 50 SNPs and LD r^2 = 0.80, the final dataset included 7,776 European participants genotyped for 441,351 SNPs. This latter step was performed to enable calculation of regional variance explained using multiple linear regression. Height and BMI was adjusted for age and sex in all analyses. To mitigate the effect of outliers, we have removed values outside the 1st and 99th percentile range for each of height and BMI. All analyses are adjusted for the first 20 genetic principal components unless stated otherwise. Polygenic scores were derived using the "clump" function of PLINK with a LD r^2 threshold of 0.2 and testing p-value thresholds of 10^{-8}, 10^{-7}, 10^{-6}, … , 10^{-1} and 1. All LD estimates used throughout the manuscript were derived from HRS genotypes. HRS was not part of the Genetic Investigation of Anthropometric Traits (GIANT) meta-analysis of height [19,20].

References

1. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565--569 (2010).
2. Pare, G., Asma, S. & Deng, W.Q. Contribution of large region joint associations to complex traits genetics. *PLoS Genet* **11**, e1005103 (2015).
3. Beyene, J., Tritchler, D., Asimit, J.L. & Hamid, J.S. Gene- or region-based analysis of genome-wide association studies. *Genet Epidemiol* **33 Suppl 1**, S105-10 (2009).
4. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet* **9**, e1003993 (2013).
5. Cheung, V.G. & Spielman, R.S. Genetics of human gene expression: mapping DNA variants that influence gene expression. **10**, 595--604 (2009).
6. Consortium, T.E.P. An integrated encyclopedia of DNA elements in the human genome. **489**, 57--74 (2012).
7. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-21 (2012).
8. Loh, P.-R. *et al.* Contrasting regional architectures of schizophrenia and other complex diseases using fast variance components analysis. *bioRxiv* (2015).
9. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
10. Bulik-Sullivan, B. Relationship between LD Score and Haseman-Elston Regression. *bioRxiv* (2015).
11. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
12. Sammalisto, S. *et al.* Genome-wide linkage screen for stature and body mass index in 3.032 families: evidence for sex- and population-specific genetic effects. *Eur J Hum Genet* **17**, 258-66 (2009).
13. Perola, M. *et al.* Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet* **3**, e97 (2007).
14. Bhatia, G. *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases. *bioRxiv* (2015).
15. Vilhjalmsson, B. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *bioRxiv* (2015).
16. Ohtani, K. & Tanizaki, H. Exact Distributions of R2 and Adjusted R2 in a Linear Regression Model with Multivariate Error Terms. *Journal Of The Japan Statistical Society* **34**, 101-109 (2004).
17. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
18. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
19. Berndt, S.I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45**, 501-12 (2013).

20. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).

Figure 1: Total genetic variance estimated by summary association statistics, multiple linear regression and variance component models.

The overall genetic variance estimated by the three approaches is illustrated as a function of SNP block size (Panels a and c) and number of principal components (Panels b and d) for BMI (Panels a and b) and height (Panels c and d). Blue lines represent estimates of genetic variance using summary association statistics; with 95% confidence intervals illustrated as blue shaded area. Corresponding estimates for variance component models are in green and estimates using multiple linear regression models in red.

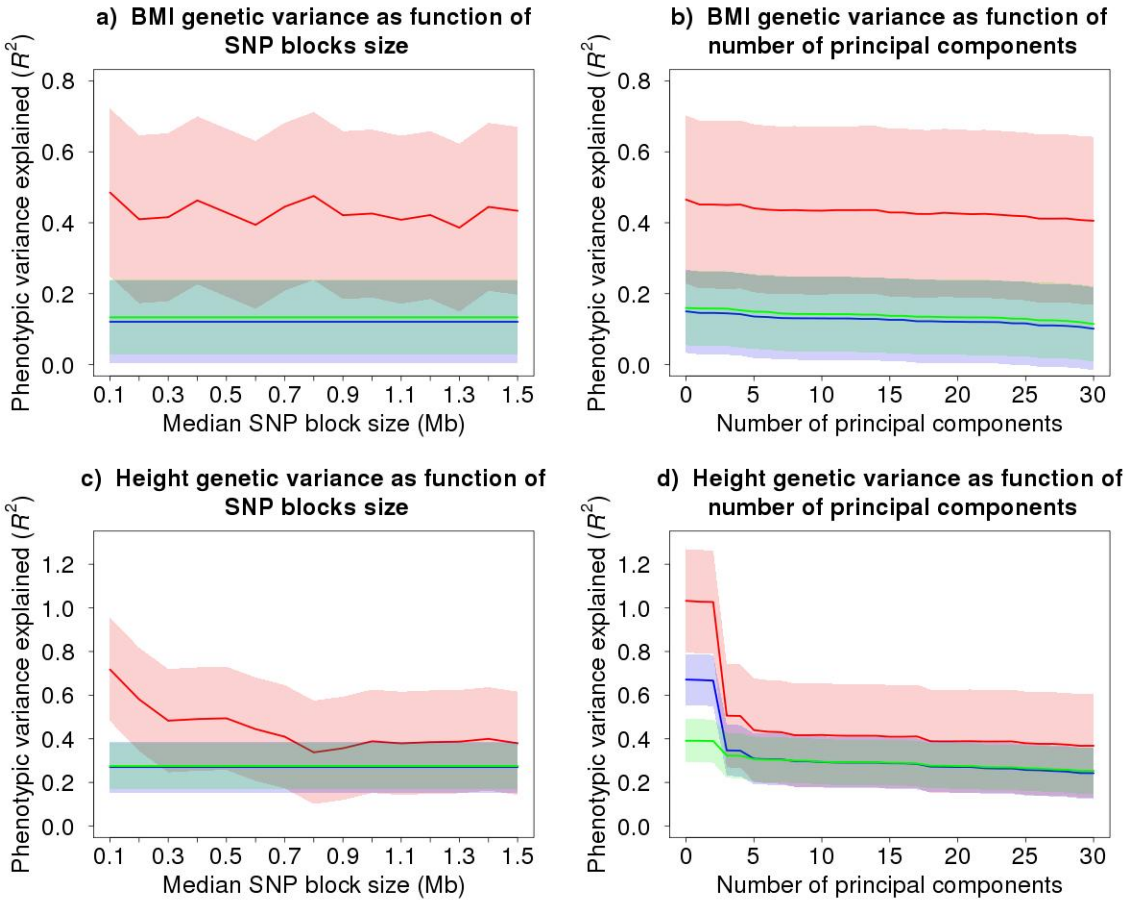

Figure 2: Phenotypic variance explained in HRS as a function of proportion of top GIANT SNP blocks included

Phenotypic variance explained estimated using summary association statistics (blue), variance component (green) and multiple linear regression models (red), with shaded areas representing 95% confidence intervals. The median SNP block size was set at 0.5 Mb (i.e. 95-105 SNPs; Panels a and b), 1.0 Mb (i.e. 195-205 SNPs; Panels c and d) and 1.5 Mb (i.e. 295-305 SNPs; Panels e and f) for both BMI (Panels a, c and e) and height (Panels b, d and f).

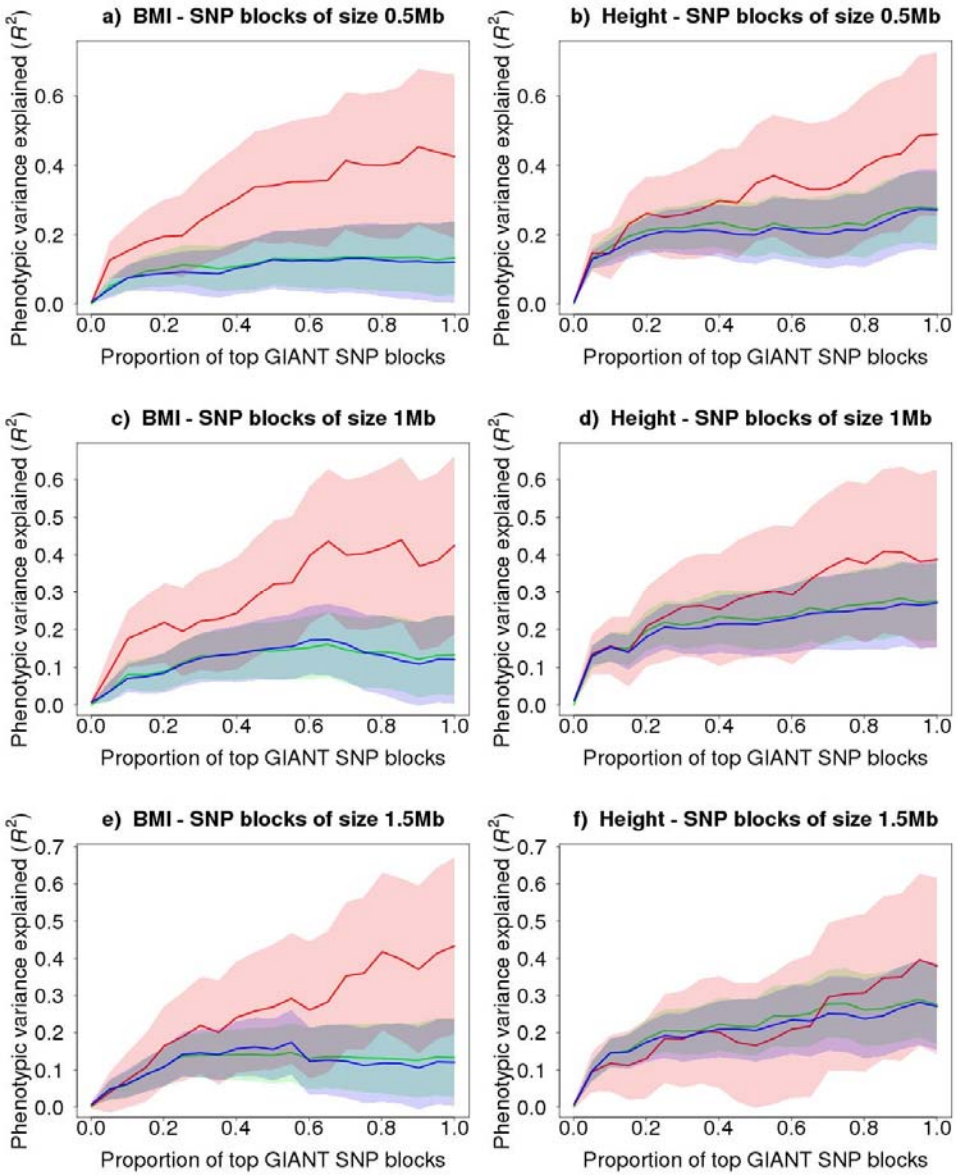

Figure 3: Minimum univariate SNP association *p*-value for each SNP block of median size 1 Mb

SNP block ranks are based on GIANT data throughout while the minimum univariate SNP association *p*-values were taken from GIANT (Panels a and c) or calculated in HRS (Panels b and d) for both BMI (Panels a and b) and height (Panels c and d).

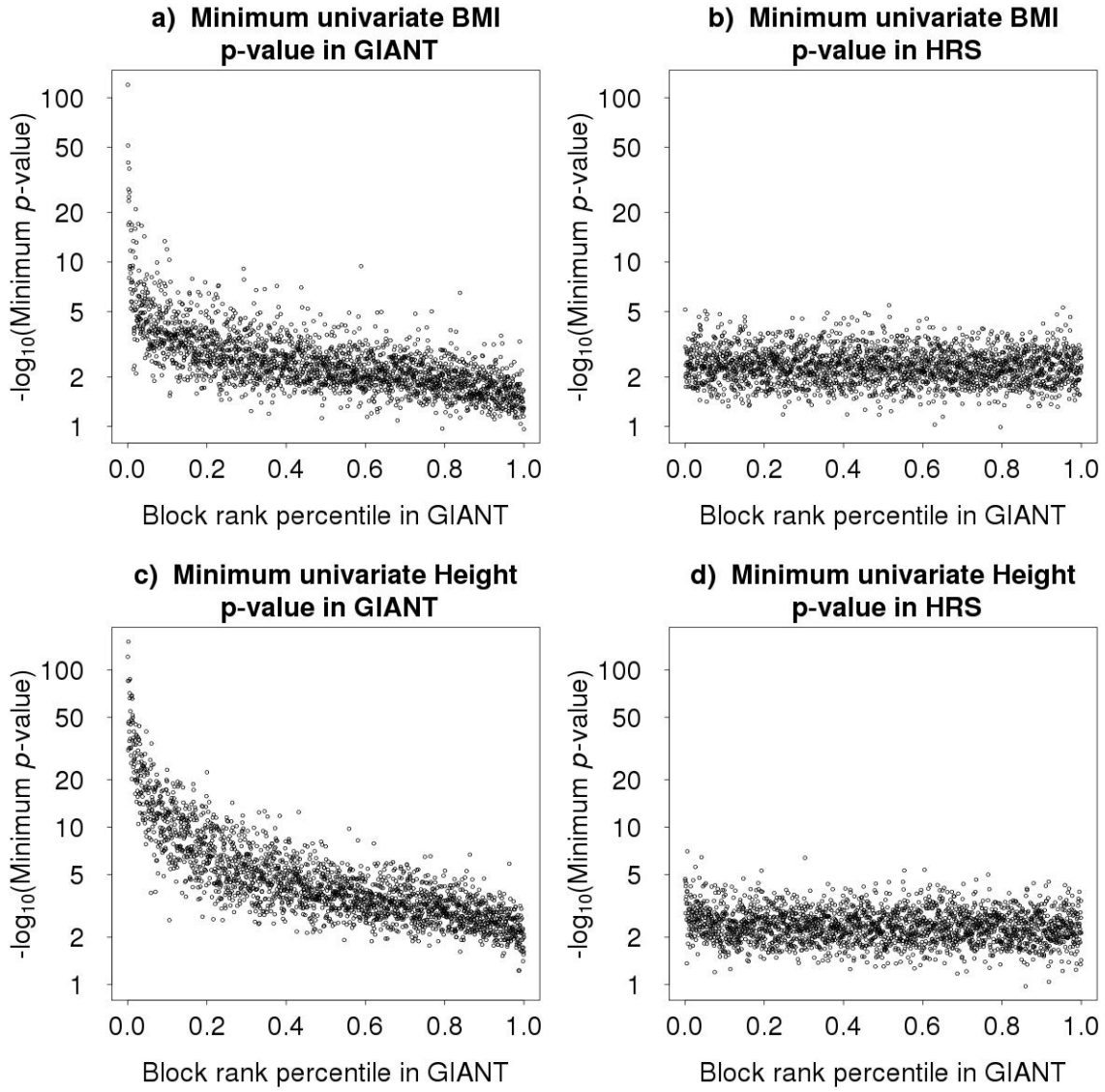

Figure 4: Gene score prediction R^2 as a function of proportion of top GIANT blocks included

Prediction R^2 for regional gene scores (blue line) as a function of proportion of top GIANT blocks is illustrated for BMI (Panel a) and height (Panel c). Prediction R^2 for genome-wide significant SNPs alone using LD clumping is illustrated as the turquoise horizontal line, while optimal polygenic gene score is illustrated as the orange horizontal line. In panels b and d, regional gene scores are added to optimal polygenic scores for BMI and height, respectively.

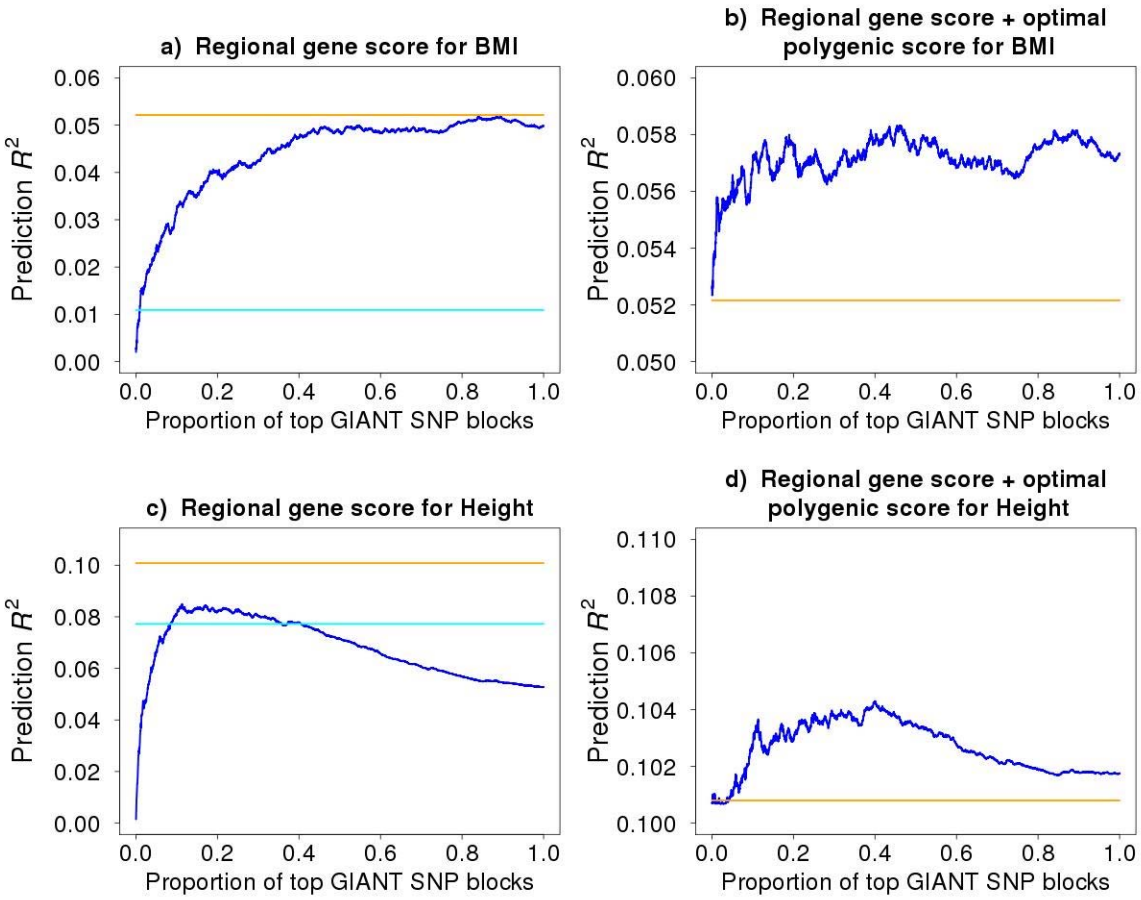

Figure 5: Prediction R^2 of gene scores derived in HRS using a 5-fold cross-validation

For each of BMI (Panel a) and height (Panel b), gene scores were derived using our novel method (blue) and a multiple linear regression (red) using a 5-fold cross-validation, and prediction R^2 plotted as a function of proportion of top GIANT blocks. Results of each individual fold validation are illustrated with thin lines whereas the thick lines represent the average of the 5 experiments.

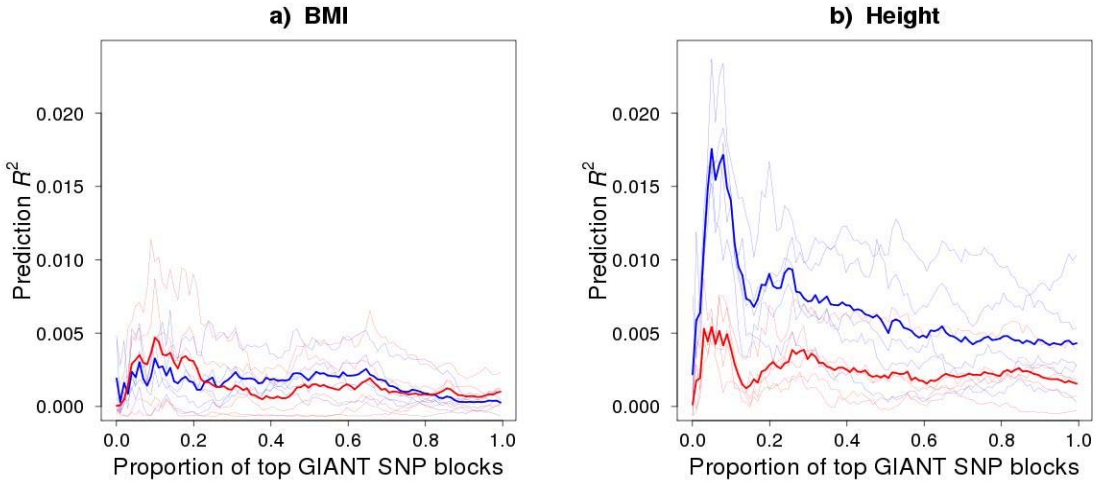

Table 1: Excess regional genetic variance at 3 suggestive (LOD>2.0) linkage peaks for height from GIANT summary association statistics

Chr.	Peak Marker	LOD Score	Excess genetic variance	95% CI Upper limit	95% CI lower limit	*p*-value
11	D11S2000	2.74	-0.0021	0.0002	-0.0044	0.07
12	D12S1301	2.07	-0.0005	0.0014	-0.0024	0.61
15	D15S655	3.00	0.0044	0.0072	0.0017	**0.002**

Regional genetic variance was calculated within +/- 7.5 Mb of each peak marker and compared to genome-wide average for regions of equivalent size. *P*-values are two-sided.