nature genetics

Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index

Jian Yang[1,2,24], Andrew Bakshi[1], Zhihong Zhu[1], Gibran Hemani[1,3], Anna A E Vinkhuyzen[1], Sang Hong Lee[1,4], Matthew R Robinson[1], John R B Perry[5], Ilja M Nolte[6], Jana V van Vliet-Ostaptchouk[6,7], Harold Snieder[6], The LifeLines Cohort Study[8], Tonu Esko[9–12], Lili Milani[9], Reedik Mägi[9], Andres Metspalu[9,13], Anders Hamsten[14], Patrik K E Magnusson[15], Nancy L Pedersen[15], Erik Ingelsson[16,17], Nicole Soranzo[18,19], Matthew C Keller[20,21], Naomi R Wray[1], Michael E Goddard[22,23] & Peter M Visscher[1,2,24]

We propose a method (GREML-LDMS) to estimate heritability for human complex traits in unrelated individuals using whole-genome sequencing data. We demonstrate using simulations based on whole-genome sequencing data that ~97% and ~68% of variation at common and rare variants, respectively, can be captured by imputation. Using the GREML-LDMS method, we estimate from 44,126 unrelated individuals that all ~17 million imputed variants explain 56% (standard error (s.e.) = 2.3%) of variance for height and 27% (s.e. = 2.5%) of variance for body mass index (BMI), and we find evidence that height- and BMI-associated variants have been under natural selection. Considering the imperfect tagging of imputation and potential overestimation of heritability from previous family-based studies, heritability is likely to be 60–70% for height and 30–40% for BMI. Therefore, the missing heritability is small for both traits. For further discovery of genes associated with complex traits, a study design with SNP arrays followed by imputation is more cost-effective than whole-genome sequencing at current prices.

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of human complex traits and diseases[1]. However, genome-wide significant SNPs often explain only a small proportion of the heritability estimated from family-based studies, in the so-called 'missing heritability' problem[2]. Recent studies have shown that the total variance explained by all common SNPs is a large proportion of the heritability for complex traits and diseases[3,4]. This implies that much of the missing heritability is due to variants whose effects are too small to reach the level of genome-wide significance. This conclusion is supported by recent findings that complex traits and diseases such as height, BMI, age at menarche, inflammatory bowel diseases and schizophrenia are influenced by hundreds or even thousands of genetic variants of small effect[5–9]. Nevertheless, the genetic variance accounted for by all common SNPs is still less than that expected from family-based studies, and there has not been a consensus explanation for the missing heritability problem[2]. There are three major hypotheses. The first hypothesis is that missing heritability is largely due to rare variants of large effect, which are neither on the currently available commercial SNP arrays nor well tagged by the SNPs on the arrays. Here we define rare variants as variants with a minor allele frequency (MAF) of ≤0.01. To genotype rare variants with reasonably high accuracy, whole-genome sequencing with sufficiently high coverage in a large sample is required. The second hypothesis is that the majority of heritability is attributable to common variants (MAF >0.01) of small effect, such that many variants are not detected at the level of genome-wide significance; most of these common variants are either well tagged by genotyped SNPs through linkage disequilibrium (LD) or can be imputed with reasonably high accuracy from whole-genome sequencing reference panels. If the second hypothesis is true, increasing sample size will be more important than extending variant coverage for continued progress in genetic association studies. The third hypothesis is that heritability estimates from family-based studies are biased upward, as a result, for instance, of shared environmental effects. Therefore, quantifying the relative contributions of rare and common variants to trait variation is critical to inform the design of future experiments and to disentangle the genetic architecture of complex traits and diseases. In this study, we seek to quantify the proportion of variation at common and rare sequence variants that can be captured by SNP array genotyping followed by imputation, and we subsequently estimate the proportion of phenotypic variance for the model complex traits height and BMI that can be explained by all imputed variants.

RESULTS

Unbiased estimate of heritability using whole-genome sequence data

Let h^2_{WGS} denote the narrow-sense heritability (h^2) for a complex trait captured by the sequence variants from whole-genome sequencing and h^2_{1KGP} denote the heritability captured by all variants from

A full list of affiliations appears at the end of the paper.

Received 10 February; accepted 31 July; published online 31 August 2015; doi:10.1038/ng.3390

imputing the SNP array data to the 1000 Genomes Project reference panels[10], where $h^2_{\mathrm{WGS}} > h^2_{\mathrm{1KGP}}$ because of the loss of tagging due to imperfect imputation. We previously developed the single-component GREML analysis (GREML-SC) method (based on a single genetic relationship matrix, or GRM) as implemented in GCTA[11] to estimate the proportion of variance explained by all common SNPs in a GWAS sample of unrelated individuals[12]. To quantify the amount of variation at sequence variants that could be captured by 1000 Genomes Project imputation, we first needed to investigate whether this approach could provide an unbiased estimate of heritability using whole-genome sequencing data. We performed extensive simulations based on a whole-genome sequencing data set from the UK10K project[13] (UK10K-WGS), which comprised 17.6 million genetic variants (excluding singletons and doubletons) in 3,642 unrelated individuals after quality controls (Online Methods). The simulation results showed that, if causal variants were a random subset of all the sequence variants (52.7% rare), the GREML-SC estimate of h^2_{WGS} using all variants (including the causal variants) was unbiased (**Fig. 1**), consistent with our theoretical derivation (**Supplementary Note**). By 'unbiased', we mean that the mean estimate of h^2_{WGS} from 200 simulation replicates was not significantly different from the h^2 parameter used for simulation. We could also expect from the theoretical derivation that, if causal variants had a different MAF spectrum than the variants used in the analysis, the GREML-SC estimate of h^2_{WGS} would be biased. This was demonstrated using simulations (Online Methods): if we randomly sampled disproportionally rare (or common) variants as causal variants, the estimate of h^2_{WGS} was biased downward (or upward) (**Fig. 1**). This problem has been discussed previously[12] and can be solved by MAF-stratified GREML (GREML-MS) analysis[14] (Online Methods). We show by simulations that the estimate of h^2_{WGS} from GREML-MS was unbiased, irrespective of the MAF spectrum of the causal variants (**Fig. 1**).

We know from the theoretical derivation (**Supplementary Note**) that GREML-SC is biased if causal variants have a different LD property than the variants used in the analysis. A difference in LD can be caused by a difference in the MAF spectrum, which can be corrected for using the GREML-MS approach, as shown above. However, GREML-MS is unable to correct for the region-specific LD heterogeneity across the genome (**Supplementary Fig. 1**). That is, if causal variants tend to be enriched in genomic regions with higher or lower LD than average, the estimate of h^2 from either GREML-SC or GREML-MS will be biased. We confirmed this using simulations where, if all causal variants were sampled from the variants at DNase I–hypersensitive sites (DHSs) (Online Methods), which have systematically lower LD than average[15], the GREML-MS estimate of h^2_{WGS} was biased downward (**Fig. 1**). Methods have been developed to adjust for LD heterogeneity, for example, the LDAK approach[16], which gives each variant a weight according to the LD r^2 value between the variant and all other variants in the region, and the LD residual (LDres) approach[15,17], which uses the residuals from a linear regression of each variant on a set of LD-pruned variants in the region. However, the LDAK adjustment resulted in a substantial overestimation of h^2_{WGS}, regardless of whether the variants were stratified by MAF (**Fig. 1**). This is because the LDAK adjustment created a strong negative correlation between the weights and MAFs of the variants (**Supplementary Fig. 2**), such that rare variants, which tend to have lower LD with surrounding variants, received too much weight. We also observed small biases using LDres and MAF-stratified LDres (LDres-MS) (**Fig. 1**). We propose a method (Online Methods), termed LD- and MAF-stratified GREML (GREML-LDMS), which uses a sliding window approach to fit the region-specific LD heterogeneity across the genome (**Fig. 2**). We demonstrate by analyzing simulated data, under four different scenarios, that the GREML-LDMS estimates of h^2_{WGS} are unbiased, regardless of the MAF and LD properties of causal variants (**Fig. 1**) and the number of LD and MAF groups (**Supplementary Fig. 3**). The heritability parameter used in all the simulations above was 0.8. We show that all the conclusions hold irrespective of the size of the heritability parameter used for simulation (**Supplementary Table 1**).

Variation at whole-genome sequence variants captured by 1000 Genomes Project imputation

We have quantified above the (un)biased nature of GREML methods in estimating h^2_{WGS} under different simulation scenarios. In practice, however, there have not been whole-genome sequencing data available with a sample size that is sufficiently large to estimate h^2_{WGS} with useful precision. However, there are a large number of GWAS samples available that have been imputed to the 1000 Genomes Project reference panels. We therefore addressed the question of how much variation at sequence variants could be captured by imputing GWAS genotype data to the 1000 Genomes Project reference panels. From UK10K-WGS data, we extracted the genotypes of variants represented on the Illumina CoreExome array and then imputed the genotype data to the 1000 Genomes Project reference panels (Online Methods). We used the GREML-MS approach (seven genetic components) to estimate the variance explained by the 1000 Genomes Project–imputed variants (h^2_{1KGP}) for the simulated phenotype (1,000 causal variants randomly sampled from all sequence variants) (Online Methods). We know from the simulation results presented above that, under this scenario (that is, where causal variants are sampled completely at random), all three GREML methods—GREML-SC, GREML-MS and GREML-LDMS—are unbiased. We chose to use GREML-MS because it is able to provide estimates of variance explained for variants in different MAF groups with standard errors smaller than those from GREML-LDMS (**Supplementary Table 1**). The results showed that the proportion of variation at variants from whole-genome sequencing captured by 1000 Genomes Project imputation decreased with more stringent thresholds for imputation accuracy (the INFO metric

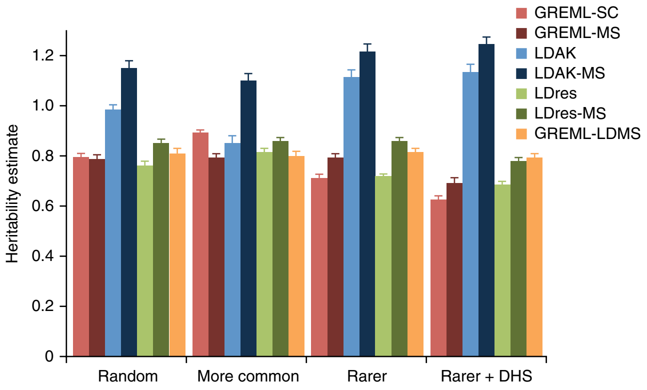

Figure 1 Estimates of heritability using sequence variants under different simulation scenarios based on the UK10K-WGS data set. Each column represents the mean estimate from 200 simulations. Error bars, s.e.m. The true heritability parameter was 0.8 for the simulated trait. Four simulation scenarios are shown: random causal variants are sampled at random; more common causal variants are more frequent than random variants; rarer causal variants are less frequent than random variants; rarer + DHS causal variants are all in DHSs and are less frequent than random variants (see the Online Methods for more details on the simulation scenarios).

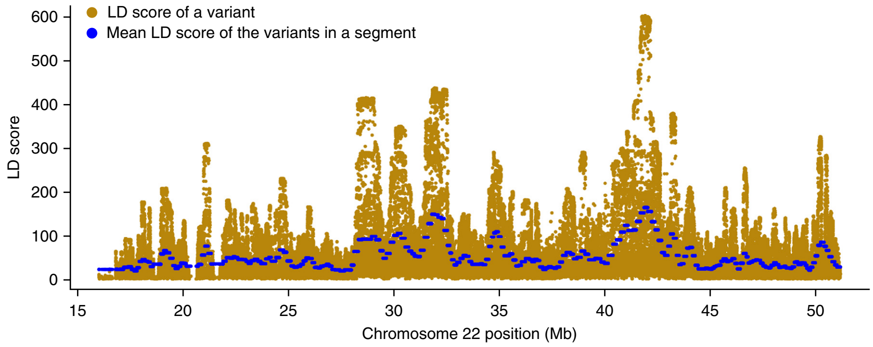

Figure 2 Fitting region-specific LD heterogeneity in the genome using a sliding window approach. Shown are the results for chromosome 22 using the UK10K-WGS data set as an example. The LD score of each variant is defined as the sum of the LD r^2 values for the target variant and all variants (including the target variant) within the 20-Mb region centered on the target variant. In GREML-LDMS analysis, the region-specific LD heterogeneity is fitted by segments with an average length of 100 kb (blue dots) using a sliding window approach (Online Methods).

from IMPUTE2) used for variant filtering and that all the 1000 Genomes Project–imputed variants (without filtering of variants for IMPUTE-INFO score) captured ~96% and ~73% of variation at common and rare whole-genome sequencing variants, respectively (**Fig. 3a**). We then repeated the imputation on the basis of four other types of SNP arrays, namely the Affymetrix 6.0, Affymetrix Axiom, Illumina OmniExpress and Illumina Omni2.5 platforms. The results were remarkably consistent irrespective of the types of SNP arrays used for baseline genotyping (**Fig. 3b**). On average, across the five different SNP arrays used for baseline genotyping, ~97% and ~72% of variation at common and rare sequence variants could be captured by 1000 Genomes Project–imputed variants, respectively. Surprisingly, despite the Illumina Omni2.5 array having ~933,000 and ~1.2 million more variants, respectively, than the OmniExpress and CoreExome arrays used for imputation (**Supplementary Table 2**), the proportions of variation at sequence variants captured by 1000 Genomes Project–imputed variants based on these three types of SNP arrays were almost identical (**Fig. 3b**). We further performed simulations under 4 scenarios (Online Methods) and analyzed the simulated data using the GREML-LDMS approach (28 genetic components). The results (**Supplementary Fig. 4**) show that the proportion of variation at sequence variance captured by 1000 Genomes Project imputation is almost independent of genetic architecture. On average, across SNP arrays and simulation scenarios, ~97% of variation at common variants and ~68% of variation at rare variants could be captured by 1000 Genomes Project imputation.

Quantifying the missing heritability for height and BMI

We next applied the GREML-LDMS approach to estimate the proportion of phenotypic variance explained by the 1000 Genomes Project–imputed variants (h^2_{1KGP}) for height and BMI. We combined data from 7 GWAS cohorts (Online Methods), comprising ~17 million 1000 Genomes Project–imputed variants on 44,126 unrelated individuals after quality control. Results from GREML-LDMS analyses (Online Methods) showed that all 1000 Genomes Project–imputed variants explained 55.5% (s.e. = 2.3%) of phenotypic variance for height and 27.4% (s.e. = 2.5%) of phenotypic variance for BMI, with common variants explaining 47.6% (s.e. = 1.2%) and 25.0% (s.e. = 1.2%) of the variance, respectively (**Supplementary Fig. 5** and **Supplementary Table 3**). The results also showed that 1000 Genomes Project–based imputation captured a significant amount of genetic variation at rare variants for height, with $\hat{h}^2_{\text{1KGP}}$ = 8.4% (s.e. = 1.9%, $P = 6.1 \times 10^{-6}$) for variants with $0.001 < \text{MAF} \leq 0.01$ and, for BMI, with $\hat{h}^2_{\text{1KGP}}$ = 3.8% (s.e. = 1.8%, P = 0.032) for variants with $2.5 \times 10^{-5} < \text{MAF} \leq 0.001$ (**Supplementary Table 3**). The (co)variance matrices of the estimates of variance components from GREML-LDMS are illustrated in **Supplementary Figure 6**. We also performed the analyses using GREML-MS. The results were similar, with $\hat{h}^2_{\text{1KGP}}$ = 7.9% (s.e. = 1.7%, $P = 2.6 \times 10^{-6}$) for variants with $0.001 < \text{MAF} \leq 0.01$ for height and $\hat{h}^2_{\text{1KGP}}$ = 4.1% (s.e. = 1.6%, P = 0.011) for variants with $2.5 \times 10^{-5} < \text{MAF} \leq 0.001$ for BMI (**Supplementary Fig. 7**).

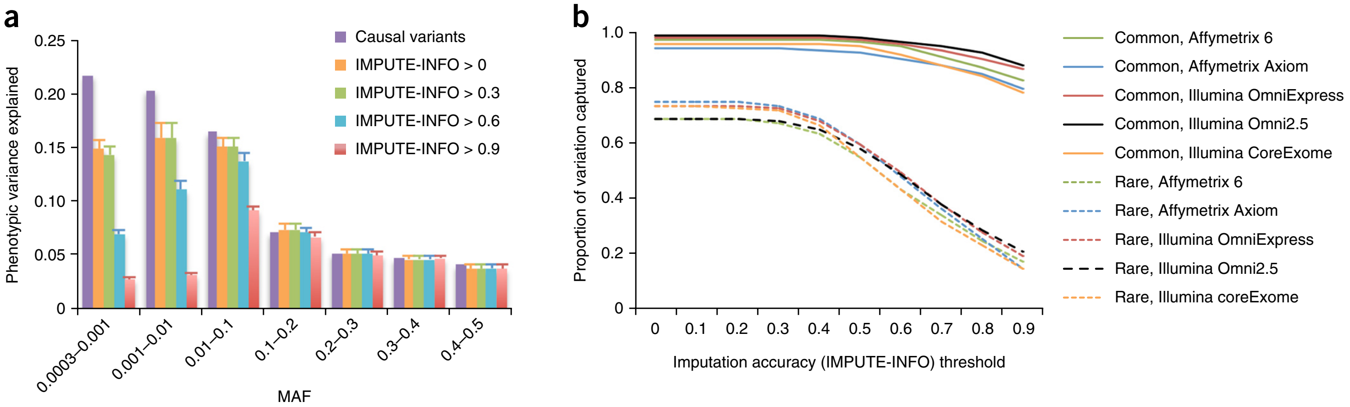

Figure 3 Proportion of variation at sequence variants captured by 1000 Genomes Project imputation in the UK10K-WGS data set. The results are the averages from 200 simulations (Online Methods). (**a**) Estimates of the proportion of phenotypic variance explained by 1000 Genomes Project–imputed variants in different MAF groups from GREML-MS. The 1000 Genomes Project imputation was based on variants on the Illumina CoreExome array extracted from UK10K-WGS data. Columns in purple represent the variance explained by the causal variants. The four other columns represent the estimates using 1000 Genomes Project–imputed variants filtered at three thresholds for imputation accuracy (IMPUTE-INFO score) or not filtered. Error bars, s.e.m. Without filtering variants for IMPUTE-INFO score (columns in orange), the sum of the estimates was 96.2% for common variants and 73.4% for rare variants. (**b**) Estimates of the proportion of variation at sequence variants captured by 1000 Genomes Project imputation (the estimate of the phenotypic variance explained by the 1000 Genomes Project–imputed variants summed over the MAF groups divided by the variance explained by the causal variants) based on the different types of SNP genotyping arrays. Common variants, MAF > 0.01; rare variants, 0.01 ≥ MAF > 0.0003.

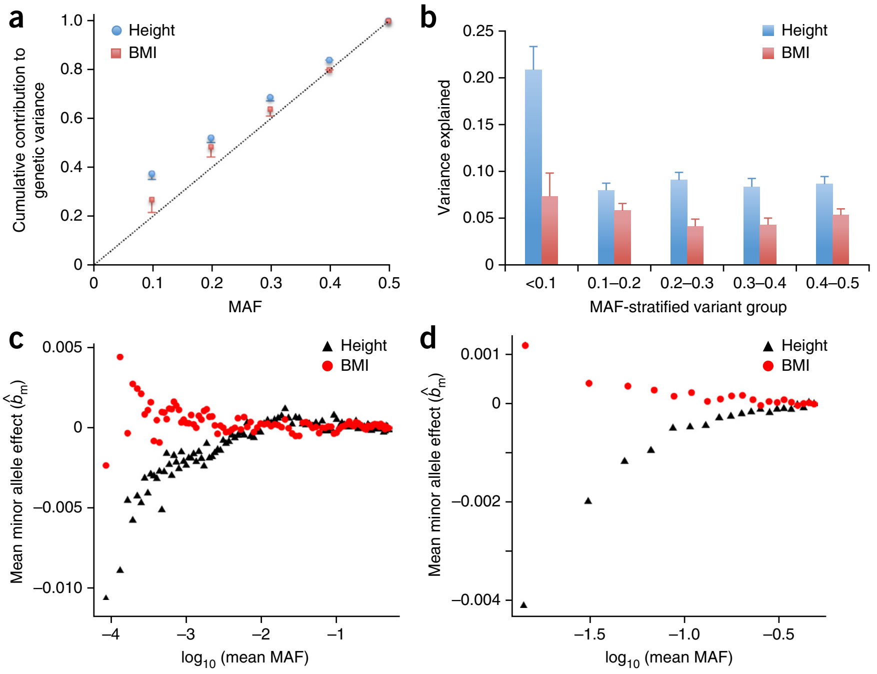

Figure 4 Evidence for height- and BMI-associated genetic variants being under natural selection. (**a**,**b**) Results from GREML-LDMS analyses (Online Methods). (**a**) Estimate of the cumulative contribution of variants with MAF ≤θ to the genetic variance, that is, $\hat{\sigma}_v^2(\text{MAF} \le \theta)/\hat{\sigma}_v^2(\text{MAF} \le 0.5)$. The dotted line represents the expectation under a neutral evolutionary model. (**b**) Estimates of $\hat{h}^2_{\text{1KGP}}$ for variants in each MAF group. Error bars, standard error of the estimates. (**c**) Results from genome-wide association analyses in the combined data set (Online Methods). b_m is defined as the effect size of the minor allele of a variant. Variants are stratified into 100 MAF bins (100 quartiles of the MAF distribution). Plotted is the mean of $\hat{b}_m$ against $\log_{10}$ (mean MAF) in each bin. The correlation between mean $\hat{b}_m$ and $\log_{10}$ (mean MAF) is 0.77 ($P < 1.0 \times 10^{-6}$) for height and –0.39 ($P = 8.0 \times 10^{-6}$) for BMI. (**d**) Results from the latest GIANT Consortium meta-analyses for height[5] and BMI[22] (see URLs) for common SNPs (MAF > 0.01). There are ~2.5 million SNPs stratified into 20 MAF bins. The correlation between mean $\hat{b}_m$ and $\log_{10}$ (mean MAF) is 0.89 ($P_{\text{permutation}} < 1.0 \times 10^{-6}$) for height and –0.87 ($P_{\text{permutation}} < 1.0 \times 10^{-6}$) for BMI. The mean $\hat{b}_m$ seems smaller in **c** than in **d** because of the smaller MAF range of each bin and larger number of variants in each bin in **c** in comparison to **d**.

Under a model of neutral evolution, most variants segregating in the population are rare, whereas most genetic variation underlying traits is due to common variants[18]. The neutral evolutionary model predicts that the cumulative contribution of variants with MAF ≤θ to the total genetic variance is linearly proportional to θ, where θ is a MAF threshold. However, our observed results for height strongly deviated from this model (**Fig. 4a**), suggesting that height-associated variants have been under natural selection. Such deviation would be even stronger with whole-genome sequencing data because variation at rare sequence variants is less well captured by 1000 Genomes Project imputation than that at common variants (**Fig. 3** and **Supplementary Fig. 4**). The deviation from the neutral evolutionary model was smaller for BMI, and the standard error of the estimate of cumulative contribution (see the **Supplementary Note** for the method used to calculate standard error) was much larger for BMI than for height. Equivalently, the neutral evolutionary model also predicts that variance explained is uniformly distributed as a function of MAF[18], such that the variance explained by variants with MAF ≤0.1 equals that of variants with MAF >0.4. However, we observed that, although the variance explained per variant (defined as $\hat{h}^2_{\text{1KGP}}/m$, with m being the number of variants) for rare variants was much smaller than that for common variants for both height and BMI (**Supplementary Fig. 8**), the variants with MAF ≤0.1 in total explained a significantly larger proportion of variance than those with MAF >0.4 (21.0% versus 8.8%, $P_{\text{difference}} = 9.2 \times 10^{-7}$) for height (**Fig. 4b** and **Supplementary Table 3**), consistent with height-associated variants being under selection. The difference was much smaller (7.4% versus 5.4%) and not significant ($P_{\text{difference}} = 0.45$) for BMI. These results were not driven by population stratification (**Supplementary Fig. 9** and **Supplementary Note**).

Theoretical studies on variation in complex traits based on models of natural selection suggest that rare variants only explain a substantial amount of variance under strong assumptions about the relationship between effect size and selection strength[19–21]. We performed genome-wide association analyses for height and BMI in the combined data set (Online Methods) and found that the minor alleles of variants with lower MAF tended to have stronger and negative effects on height and stronger but positive effects on BMI (**Fig. 4c**). The correlation between minor allele effect and MAF was highly significant for both height ($P < 1.0 \times 10^{-6}$) and BMI ($P = 8.0 \times 10^{-5}$) and was even stronger for both traits in the data from the latest GIANT Consortium meta-analyses[5,22] (**Fig. 4d**); these correlations were not driven by population stratification (**Supplementary Fig. 10**). All these results suggest that height- and BMI-associated variants have been under selection. These results are consistent with the hypothesis that new mutations that decrease height or increase obesity tend to be

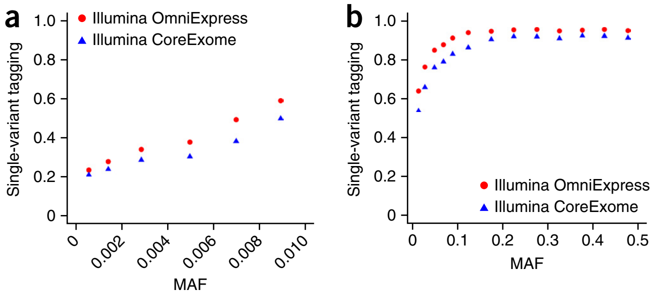

Figure 5 Single-variant tagging of sequence variants by 1000 Genomes Project–imputed variants. Single-variant tagging is defined as the squared correlation (r^2_{max}) between a sequence variant and the best tagging variant from 1000 Genomes Project imputation within the 2-Mb region centered on the variant. Shown are the average r^2_{max} values of variants in MAF bins for 10,000 sequence variants randomly sampled from UK10K-WGS data. The 1000 Genomes Project imputation analyses are based on variants on the Illumina OmniExpress (red) and Illumina CoreExome (blue) arrays extracted from UK10K-WGS data (see the Online Methods for details about the imputation analyses based on UK10K-WGS data). (**a**) Rare variants. (**b**) Common variants.

deleterious to fitness and are hence kept at low frequencies in the population by purifying selection.

DISCUSSION

We have shown using simulations based on whole-genome sequencing data that the GREML-SC estimate of heritability is unbiased if causal variants are a random subset of the sequence variants used in the analysis, that GREML-MS is unbiased if there is no region-specific LD heterogeneity and that GREML-LDMS is unbiased regardless of the MAF and LD properties of causal variants (**Fig. 1**). Although we describe the methods for quantitative traits, they can also be applied to case-control studies by analyzing data on the observed disease status scale (0 or 1) and interpreting estimates on an underlying liability scale by a linear transformation[4,23]. As suggested in previous studies[4,24], more stringent quality controls than those used in GWAS are required to apply the GREML approaches in case-control data. Because GREML-LDMS fits a large number of components (for example, we fitted 28 genetic components in the analyses of height and BMI data), the sampling variance of $\hat{h}^2_{\mathrm{1KGP}}$ is much larger than that from GREML-SC. We quantified empirically that the standard error of $\hat{h}^2_{\mathrm{1KGP}}$ from GREML-LDMS was approximately $927/n$ (**Supplementary Fig. 11**), inversely proportional to the sample size (n) and independent of $\hat{h}^2_{\mathrm{1KGP}}$. For a GREML-LDMS analysis using either whole-genome sequencing data or combined imputed and whole-genome sequencing data, a sample of at least 18,540 unrelated individuals is therefore required to obtain s.e. <0.05. We have further demonstrated that the GREML-LDMS approach is robust to the model assumption about the relationship between effect size and MAF (**Supplementary Figs. 12** and **13**, and **Supplementary Note**).

Using the GREML-LDMS approach, we estimated that all the 1000 Genomes Project–imputed variants explained 56% (s.e. = 2.3%) and 27% (s.e. = 2.5%) of phenotypic variance for height and BMI, respectively. These estimates are still lower than the frequently quoted estimates of narrow-sense heritability (h^2) for height (80%) and BMI (40–60%) from family and twin studies. Therefore, it seems that heritability is still 'missing'. There are two possible explanations for the remaining missing heritability. The first is that there are a large number of extremely rare causal variants, not polymorphic in the 1000 Genomes Project–imputed data or removed by imputation quality control. For example, there are >40 million and >45 million variants in the 1000 Genomes Project[10] and UK10K[13] databases, respectively, whereas 17 million imputed variants were used in the GREML analyses for height and BMI. Complex DNA variations such as copy number variations are also not well represented by current sequencing methods[25,26]. The second explanation is that heritability is overestimated in family studies, owing to effects from factors such as common environment and assortative mating that are not properly modeled[27]. Results from a previous study show that the phenotypic correlation for height between distant relatives (for example, cousins) is larger than what would be expected given $h^2 = 0.8$ under an additive model[28], suggesting substantial confounding in the family-based estimate of h^2 but not supporting an important role for non-additive genetic variance. A recent study[29] that used extended genealogy in a large sample ($n = 38,167$) provided very precise estimates of the heritability for height ($\hat{h}^2 = 0.69$, s.e. = 0.016) and BMI ($\hat{h}^2 = 0.42$, s.e. = 0.018). These estimates can be regarded as the upper limits of heritability for height and BMI because shared environmental effects were not explicitly fitted in the model and these estimates could therefore still be inflated to some extent. The estimates from a within-family analysis that was free of confounding from shared environmental effects are highly consistent with heritability being 0.69 (s.e. = 0.14) for height and 0.42 (s.e. = 0.17) for BMI, but the standard errors are too large to draw strong inferences[30]. There has also been evidence suggesting that a population-based heritability estimate is likely to be lower than that from pedigrees[31]. If we extrapolate from the GREML-LDMS estimates (**Supplementary Table 3**) by taking into account the imperfect tagging of 1000 Genomes Project imputation (on average, across five different types of SNP arrays and four simulation scenarios, ~97% and ~68% of variation at common and rare variants is captured by 1000 Genomes Project imputation, respectively; **Supplementary Fig. 4**), the adjusted estimate of heritability would be 0.61 (s.e. = 0.045) for height and 0.29 (s.e. = 0.47) for BMI (see the **Supplementary Note** for the adjustment method). These estimates can be regarded as the lower limits of narrow-sense heritability for height and BMI. Our results suggest that heritability is likely between 0.6 and 0.7 for height and between 0.3 and 0.4 for BMI. Therefore, there is little missing heritability for these traits. These results also suggest that there is little room for the other possible sources of missing heritability (**Supplementary Note**).

We know from the simulations (**Fig. 1** and **Supplementary Table 1**) that the GREML-SC and GREML-MS methods can be biased depending on the MAF and LD properties of causal variants. For completion of the analysis, we also performed GREML-SC and GREML-MS in the combined GWAS data set using 1000 Genomes Project–imputed variants. The GREML-SC estimate of h^2_{1KGP} was 0.78 (s.e. = 0.017) for height and 0.40 (s.e. = 0.018) for BMI, with the estimates larger than those from GREML-LDMS for both traits. This is because the proportion of loss of tagging in 1000 Genomes Project imputation for rare variants is larger than that for common variants (**Fig. 3**), analogous to the situation where common variants explain a disproportionally higher proportion of the variance (simulation scenario II in the Online Methods), resulting in an overestimation of h^2_{WGS} (**Fig. 1** and **Supplementary Table 1**). This is an important caveat. If we were to draw inferences on the basis of these results, we would conclude that all the heritability for height and BMI has been captured by 1000 Genomes Project imputation, which is obviously not true. The GREML-MS estimate of h^2_{1KGP} was 0.523 (s.e. = 0.021) for height and 0.261 (s.e. = 0.022) for BMI, and these estimates were not very dissimilar from those from GREML-LDMS, especially for BMI. Therefore, it is more important to correct for difference in the MAF spectrum than for LD heterogeneity (**Supplementary Fig. 14** and **Supplementary Note**). There are also other scenarios under which the GREML methods can be misused and thus may lead to wrong conclusions, for example, in the whole-genome sequencing study by Morrison *et al.*[32] (**Supplementary Note**).

More generally, we have shown by simulations that, for variants with MAF >0.0003 in the UK10K-WGS data set, 96% and 73% of variation at common and rare whole-genome sequencing variants, respectively, can be captured by 1000 Genomes Project imputation, using Illumina CoreExome arrays for baseline genotyping (**Fig. 3**). These percentages of variation need to be interpreted as multivariant tagging, analogous to the multiple correlation squared (r^2_{multi}) between a sequence variant and all the 1000 Genomes Project–imputed variants in the region, which are not comparable to the single variant–based imputation accuracy (variation at a sequence variant tagged by a single imputed variant) quantified in previous studies[33,34]. In GWAS, analysis is usually performed on the basis of a single-variant model. The statistical power of a single-variant analysis using imputed data depends on the squared correlation between a causal variant and its best-tagging imputed variant (r^2_{max}). We show that, on average, 81% (s.e.m. = 0.4%) and 25% (s.e.m. = 0.4%) of variation at common and rare whole-genome sequencing variants, respectively, can be captured by the best-tagging variants from

1000 Genomes Project imputation based on Illumina CoreExome arrays (**Fig. 5**). These results are comparable with the single-variant imputation accuracy quantified previously. The single-variant tagging is slightly stronger for 1000 Genomes Project imputation based on Illumina OmniExpress arrays, which is 88% (s.e.m. = 0.3%) for common variants and 29% (s.e.m. = 0.4%) for rare variants. These results suggest that, for GWAS analyses using 1000 Genomes Project–imputed data, there is great potential to gain power using a multivariant or haplotype-based association analysis approach, for rare variants in particular. Therefore, with the increasingly large amount of summary-level data from large-scale meta-analyses of 1000 Genomes Project–imputed GWAS data that are becoming available, there is a great demand to develop powerful and efficient methods for multivariant or haplotype-based association analysis using summary data, which take into account both common and rare variants.

With advances in genome sequencing technologies, it is now possible to sequence a human genome at high depth for $1,000, which, however, is still much more expensive than using a SNP array (for example, the Illumina CoreExome array). Given a fixed budget for genotyping and assuming that the genotyping cost using SNP arrays (for example, $50 per sample) is 20 times less than that for whole-genome sequencing (for example, $1,000 per sample), on average, 1000 Genomes Project imputation is currently at least 13 times more powerful than whole-genome sequencing using a multivariant association analysis approach (**Supplementary Fig. 15**). For a single variant–based association analysis, 1000 Genomes Project imputation is still at least 13 and 4 times more powerful than whole-genome sequencing in detecting common and rare variant associations, respectively. These results suggest that SNP array–based genotyping followed by imputation is now and in the near future will continue to be a more cost-effective strategy than whole-genome sequencing for GWAS of complex traits and diseases, even for rare variant associations. Nevertheless, the analyses above compared the average power for variants in a certain MAF range. There are a number of sequence variants (~10% rare and ~1% common) that are almost not tagged by any imputed variant individually (single-variant tagging $r^2_{\max} < 0.05$) in 1000 Genomes Project imputation based on the Illumina CoreExome array (**Supplementary Fig. 16**). For association analysis of such variants and those with extremely low frequency or unique to specific populations, high-coverage whole-genome sequencing or a haplotype-based method will be a more efficient strategy. In contrast, it has been suggested that extremely low-coverage whole-genome sequencing followed by imputation can be even more cost-effective than SNP array–based imputation for common variants[35], an interesting strategy that is worth being further investigated for its performance on rare variants.

With the latest imputation reference panel of large sample size (*n* = 31,000; Haplotype Reference Consortium, personal communication) and very large GWAS cohorts genotyped on the same type of SNP arrays (for example, the UK Biobank has genotyped >400,000 individuals using Affymetrix Axiom arrays) that are soon becoming available, we can expect a great improvement in imputation accuracy for rare variants. For complex traits and diseases that have a genetic architecture similar to that of height (enrichment of height-associated variants with MAF <0.1), we can expect to see a wave of discovery of trait- or disease-associated low-MAF variants in the near future, without the need for large-scale whole-genome sequencing.

URLs. GCTA-GREML-LDMS, http://cnsgenomics.com/software/gcta/greml_ldms.html; DNase I–hypersensitive site (DHS) annotation, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/; LD scores and GWAS summary data from the combined data set, http://cnsgenomics.com/software/data/yang_et_al_2015_ng.html; GIANT height and BMI summary data, http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research was supported by the Australian National Health and Medical Research Council (grants 1052684, 1078037 and 1050218), the Australian Research Council (grant 130102666), the US National Institutes of Health (R01MH100141), the Sylvia and Charles Viertel Charitable Foundation and the University of Queensland Foundation. This study makes use of data from the database of Genotypes and Phenotypes (dbGaP) available under accessions phs000090, phs000091 and phs000428 and the EGCUT, LifeLines, TwinGene and UK10K studies (see the **Supplementary Note** for the full set of acknowledgments for these data).

AUTHOR CONTRIBUTIONS

J.Y. and P.M.V. conceived and designed the study. J.Y. performed statistical analyses and simulations. M.E.G., J.Y. and P.M.V. derived the theory. A.B., Z.Z. and G.H. performed the imputation analysis. S.H.L., M.R.R., M.C.K. and N.R.W. provided statistical support. A.A.E.V., J.R.B.P., I.M.N., J.V.v.V.-O., H.S., the LifeLines Cohort Study, T.E., L.M., R.M., A.M., A.H., P.K.E.M., N.L.P., E.I. and N.S. contributed to data collection. J.Y. wrote the manuscript with the participation of all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
3. Yang, J. *et al.* Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLoS Genet.* **9**, e1003355 (2013).
4. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
5. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
6. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
7. Perry, J.R. *et al.* Parent-of-origin–specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
8. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
9. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
10. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
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. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
13. UK10K Consortium. The UK10K project: rare variants in health and disease. *Nature* (in the press).
14. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).
15. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
16. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
17. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).
18. Visscher, P.M., Goddard, M.E., Derks, E.M. & Wray, N.R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol. Psychiatry* **17**, 474–485 (2012).

19. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **107**, 1752–1756 (2010).
20. Simons, Y.B., Turchin, M.C., Pritchard, J.K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
21. Uricchio, L.H., Witte, J.S. & Hernandez, R.D. Selection and explosive growth may hamper the performance of rare variant association tests. *bioRxiv* doi:10.1101/015917 (2015).
22. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
23. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
24. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
25. Treangen, T.J. & Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
26. Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
27. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, 1998).
28. Visscher, P.M., McEvoy, B. & Yang, J. From Galton to GWAS: quantitative genetics of human height. *Genet. Res. (Camb.)* **92**, 371–379 (2010).
29. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
30. Hemani, G. *et al.* Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *Am. J. Hum. Genet.* **93**, 865–875 (2013).
31. Zaitlen, N. *et al.* Leveraging population admixture to characterize the heritability of complex traits. *Nat. Genet.* **46**, 1356–1362 (2014).
32. Morrison, A.C. *et al.* Whole-genome sequence–based analysis of high-density lipoprotein cholesterol. *Nat. Genet.* **45**, 899–901 (2013).
33. Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84**, 235–250 (2009).
34. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
35. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [2]University of Queensland Diamantina Institute, Translation Research Institute, Brisbane, Queensland, Australia. [3]Medical Research Council (MRC) Integrative Epidemiology Unit (IEU) at the University of Bristol, School of Social and Community Medicine, Bristol, UK. [4]School of Environmental and Rural Science, University of New England, Armidale, New South Wales, Australia. [5]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [6]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [7]Department of Endocrinology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [8]A full list of members and affiliations appears in the **Supplementary Note**. [9]Estonian Genome Center, University of Tartu, Tartu, Estonia. [10]Division of Endocrinology, Boston Children's Hospital, Cambridge, Massachusetts, USA. [11]Program in Medical and Populational Genetics, Broad Institute, Cambridge, Massachusetts, USA. [12]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [13]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. [14]Cardiovascular Genetics and Genomics Group, Atherosclerosis Research Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. [15]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [16]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [17]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. [18]Department of Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, UK. [19]Department of Haematology, University of Cambridge, Cambridge, UK. [20]Department of Psychology and Neuroscience, University of Colorado, Boulder, Colorado, USA. [21]Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado, USA. [22]Faculty of Veterinary and Agricultural Science, University of Melbourne, Parkville, Victoria, Australia. [23]Biosciences Research Division, Department of Economic Development, Jobs, Transport and Resources, Bundoora, Victoria, Australia. [24]These authors jointly supervised this work. Correspondence should be addressed to J.Y. (jian.yang@uq.edu.au).

ONLINE METHODS

Simulating phenotypes based on whole-genome sequencing data. We used whole-genome sequencing data from the UK10K project[13] for simulations. The data set comprises 3,781 individuals from the TwinsUK and ALSPAC cohorts and ~45.5 million genetic variants called from whole-genome sequencing after quality control. Informed consent was obtained from all the subjects. Details on the cohorts, sequencing, variant calling and quality control have been provided elsewhere[13]. We performed additional quality control steps on the data, excluding SNPs with missingness >0.05, Hardy-Weinberg equilibrium test P value $<1 \times 10^{-6}$ or minor allele count (MAC) <3 (equivalent to MAF <0.0003) using PLINK[36]. We also excluded individuals with genotype missingness rate >0.05 and one of each pair of individuals with estimated genetic relatedness >0.05 using variants on HapMap phase 3 reference panels (HapMap 3) after quality control. We retained 3,642 unrelated individuals and 17.6 million variants (see **Supplementary Fig. 17** for the MAF distribution of the variants). We call this data set UK10K-WGS throughout the manuscript. We randomly sampled m_q variants from UK10K-WGS as causal variants. We generated the phenotype using the model $y = g + e$, where $g = \Sigma_i^{m_\mathrm{q}} w_i u_i$ and $w_i = (x_i - 2p_i)/\sqrt{2p_i(1-p_i)}$, with x_i being the genotype variable of the ith causal variant (coded as 0, 1 or 2) and p_i being the frequency of the coded allele; u_i is the effect size per standardized genotype and e is the residual. This model assumes larger per-allele effect sizes for variants with lower MAFs. We generated u from $N(0, 1)$ and then generated the residual e from $N(0, \mathrm{var}(g)(1/h^2 - 1))$. We sampled the causal variants under four scenarios: scenario I (random), 1,000 causal variants randomly sampled from all the sequence variants (52.7% rare); scenario II (more common), 1,000 random and 500 additional common (MAF >0.01) causal variants; scenario III (rarer), 1,000 random and 500 additional rare (MAF <0.01) causal variants; and scenario IV (rarer and DHS), 1,000 random and 500 additional rare causal variants all sampled from the variants in DHSs (see URLs). In the UK10K-WGS data, the mean LD score for variants in DHSs (59.3) was lower than that for variants not in DHSs (80.3), consistent with results from previous studies for common SNPs[15], where the LD score for a variant was defined as the sum of the LD r^2 values between the target variant and all the variants (including the target variant itself) within the 20-Mb centered on the variant (LD r^2 threshold = 0.01). Given a simulated heritability of 0.8, the proportion of variance by the causal variants (stratified by MAF) in each of the four scenarios is presented in **Supplementary Figure 18**. We repeated the simulations 200 times with the causal variants resampled in each replicate. We analyzed the simulated data using the GREML methods. We also performed the analyses using LDAK[16] v3.0 (options: minmaf = 1×10^{-6}, minvar = 1×10^{-6} and maxiter = 4×10^5) and LDres as implemented in EIGENSTRAT[37] v6.0.1 (options: ldposlimit = 5×10^5 and numoutlieriter = 0).

Quantifying the proportion of genetic variation captured by imputation. We extracted genotype data for the variants on the Illumina CoreExome array from the UK10K-WGS data set and imputed the genotype data to the 1000 Genomes Project reference panels using IMPUTE2 (ref. 38). This was to mimic a GWAS where the subjects were genotyped using a SNP array followed by 1000 Genomes Project imputation. We repeated the analysis for four other types of commonly used SNP arrays—Affymetrix 6.0, Affymetrix Axiom, Illumina OmniExpress and Illumina Omni2.5 arrays. The numbers of variants used in the 1000 Genomes Project imputation analyses for the five types of arrays are listed in **Supplementary Table 2**. We converted the dosage scores (x_dose) from imputation to hard genotype calls (x), that is, $x = 0$ if $x_\mathrm{dose} < 0.5$, $x = 2$ if $x_\mathrm{dose} > 1.5$ and $x = 1$ otherwise), and removed imputed variants with a Hardy-Weinberg equilibrium P value $<1 \times 10^{-6}$ or MAC <3 (equivalent to MAF $<2.5 \times 10^{-5}$). We further removed variants with imputation accuracy (the metric INFO from IMPUTE2 output) below a specified threshold. We chose a range of IMPUTE-INFO threshold values (from 0 to 0.9 at intervals of 0.1) to investigate the loss of tagging by removing variants with lower imputation accuracy. We then used the MAF-stratified GREML analysis to estimate the proportion of variance in the simulated phenotype (on the basis of the UK10K-WGS data under scenario I) that could be explained by the 1000 Genomes Project–imputed variants for the five types of SNP arrays with the imputed variants filtered using ten IMPUTE-INFO thresholds. This analysis was to quantify the proportion of genetic variance that could be captured by SNP array–based genotyping followed by 1000 Genomes Project imputation when causal variants are a random sample of the whole-genome sequencing variants.

GREML-MS and GREML-LDMS approaches. We have shown by theoretical derivation (**Supplementary Note**) that GREML-SC estimates of h^2 are biased if causal variants have a different MAF spectrum than the variants used in analysis. This problem can be solved by MAF-stratified GREML (GREML-MS)[14], in which the variants are stratified into groups by MAF and the GRMs computed from the variants in each of these MAF groups are fitted jointly in a multicomponent GREML analysis[11]. For a quantitative trait, the model of a multicomponent GREML analysis can be written as:

$$\mathbf{y} = \mathbf{Kc} + \sum_t^T \mathbf{g}_t + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{c}$ is a vector of the effects of the fixed covariates (for example, the first ten eigenvectors) with the corresponding coefficient matrix $\mathbf{K}$, $\mathbf{g}_t$ is a vector of the genetic values of the individuals attributed to the variants in the tth group where $\mathbf{g}_t \sim N(\mathbf{0}, \mathbf{A}_t \sigma^2_{v(t)})$, $\mathbf{A}_t = \{A_{ij(t)}\}$ is the GRM between individuals at the variants in the tth group and T is the number of groups (variance components), and $\mathbf{e}$ is a vector of residuals with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_e)$. The variance-covariance matrix of the phenotypes is $\mathrm{var}(\mathbf{y}) = \Sigma_t^T \mathbf{A}_t \sigma^2_{v(t)} + \mathbf{I}\sigma^2_e$. For variants in the tth group, the genetic relationship between individuals i and j is calculated as in ref. 12:

$$A_{ij} = \frac{1}{m_t} \sum_k^{m_t} \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1 - p_k)} \tag{2}$$

where m_t is the number of variants, x is the genotype variable coded as 0, 1 or 2, and p is frequency of the coded allele. The variance components ($\sigma^2_{v(t)}$ and σ^2_e) can be estimated using the restricted maximum likelihood (REML) approach[11,39]. The variance explained by the variants in each group is defined as $h^2_{v(t)} = \sigma^2_{v(t)}/(\Sigma_t^T \sigma^2_{v(t)} + \sigma^2_e)$ and the variance explained by all the variants is defined as $h^2_v = \Sigma_t^T \sigma^2_{v(t)}/(\Sigma_t^T \sigma^2_{v(t)} + \sigma^2_e)$. Without loss of generality, we use the subscript v to represent a set of variants used in the analysis, with $h^2_v = h^2_\mathrm{SNP}$ for analysis using SNP array data, $h^2_v = h^2_\mathrm{WGS}$ for analysis using whole-genome sequencing data and $h^2_v = h^2_\mathrm{1KGP}$ for analysis using 1000 Genomes Project–imputed data. The standard errors of the estimates of $h^2_{v(t)}$ and h^2_v can be calculated by the delta method using sampling of the (co)variances of the estimates[27]. The multicomponent GREML approach has been applied previously to partition genetic variance into the contributions from variants on 22 chromosomes, variants in 5 MAF groups[24] or variants classified by functional annotation[15] using common SNPs. For whole-genome sequencing or imputed data from whole-genome sequencing panels, there are a large number of rare variants; we therefore stratified the variants into seven MAF groups, namely MAF ≤ 0.001, 0.001 < MAF ≤ 0.01, 0.01 < MAF ≤ 0.1, 0.1 < MAF ≤ 0.2, 0.2 < MAF ≤ 0.3, 0.3 < MAF ≤ 0.4 and 0.4 < MAF ≤ 0.5.

The GREML-MS approach corrects for the difference in MAF spectrum between causal variants and the variants used in the analysis. However, it does not take into account the region-specific heterogeneity in LD (**Supplementary Fig. 1**). If causal variants tend to be enriched in genomic regions with lower or higher LD than average, the GREML-MS estimate will still be biased (**Fig. 1** and **Supplementary Table 1**). We propose a method called LD- and MAF-stratified multicomponent GREML (GREML-LDMS). We first calculated the LD score for each SNP in the data set used for analysis, where the LD score of a variant is defined as the sum of the LD r^2 scores between the variant and all the variants (including the variant itself) within the 20-Mb region centered on the variant. We ignored LD r^2 values <0.01 to avoid chance correlations between SNPs that were not in LD. We used a sliding window approach to fit the region-specific LD heterogeneity by a large number of segments (**Fig. 2**). There were $2m_s$ variants in each segment and an overlap of m_s variants between two adjacent segments, where m_s is the average number of variants per 100 kb of a chromosome; that is, $m_s = 100 \times m/L$, with m being the total number of variants on a chromosome and L being the length of the chromosome in kilobases. We calculated the mean LD score of the variants in each segment and took the average of the mean LD scores of two adjacent segments for overlapped regions. Consequently, this process partitioned the genome into a large number of segments (m_s variants in each segment with an average length of 100 kb) with different mean LD scores to fit the region-specific LD heterogeneity across the genome (**Fig. 2**). We then stratified the segments into four groups by their mean LD scores corresponding to the first, second, third and fourth quartiles of the mean LD score distribution. We further stratified the variants in each of the 4 LD-stratified groups into 7 MAF groups as described for the GREML-MS analysis, resulting in 28 groups in total. The number of variants in each group is shown in **Supplementary Figure 4c**. All the methods described above have been implemented in GCTA[11]. For the analysis of real data, we performed the GREML-LDMS analysis with 28 variance components to investigate the variance explained as a function of MAF and LD. We have performed simulations to demonstrate the unbiased nature of the GREML-LDMS method in the analysis of different genetic architectures (**Fig. 1**, **Supplementary Fig. 12** and **Supplementary Table 1**) and different numbers of MAF and LD groups (**Supplementary Fig. 3**).

Analysis of GWAS data for height and BMI. We accessed GWAS data from seven cohorts—ARIC (dbGaP, phs000090), NHS (dbGaP, phs000091), HPFS (dbGaP, phs000091), TwinGene, HRS (dbGaP, phs000428), EGCUT and LifeLines. A summary description of the sample sizes, genotyping platforms and quality control criteria for the genotype data is listed in **Supplementary Table 4**. Informed consent was obtained from all subjects. The genotype data for each cohort after quality control were imputed to the 1000 Genomes Project using IMPUTE2, and the imputed dosage data were converted to hard genotype calls using the method described above. We combined the imputed data for all the cohorts and excluded SNPs with a Hardy-Weinberg equilibrium P value $<1 \times 10^{-6}$ or MAC <3. We did not filter variants for IMPUTE-INFO because our simulation results showed that removing variants using an IMPUTE-INFO threshold reduced tagging (**Fig. 3**). To avoid including close relatives in the sample, we estimated the genetic relatedness for pairs of individuals in the combined data set using ~1.2 million common variants on the HapMap 3 panel and removed one of each pair of individuals with estimated genetic relatedness >0.05. We retained 44,126 unrelated individuals and 17,007,473 variants in the combined data set for further analysis. We performed principal-component analyses[37] in the combined data set (i) using ~1.2 million common variants on the HapMap 3 panel and (ii) using all ~17 million variants. All the samples are of European descent, as demonstrated by the principal-component plot (**Supplementary Fig. 19**). In each sex group for each cohort, height phenotypes 5 s.d. and BMI phenotypes 7 s.d. away from the mean were not included in the analyses. The height and BMI phenotypes were adjusted for age and standardized to z scores in each sex group in each cohort, removing mean and variance difference between the sexes and between the cohorts. We then estimated the proportion of phenotypic variance that could be explained by all the imputed variants in the combined data set using the GREML-LDMS and GREML-MS approaches with GCTA[11]. We present the main results from GREML analyses with the first 10 principal components computed from ~1.2 million common variants on the HapMap 3 panel as fixed covariates and also show the results with principal components estimated from all ~17 million variants.

To investigate the relationship between minor allele effect (b_m) and MAF, we performed genome-wide association analyses for height and BMI in the combined data set. The first ten principal components computed from HapMap 3 variants were fitted as covariates in the association analyses. Variants were stratified into 100 MAF bins (100 quartiles) with ~17,000 variants in each bin. We define b_m as the effect size of the minor allele of a variant. We further defined $u_\mathrm{m} = b_\mathrm{m}\sqrt{2p(1-p)}$, with p being the MAF, which is interpreted as the effect size per standardized genotype of a variant. We calculated the means for $\hat{b}_\mathrm{m}$, $\hat{u}_\mathrm{m}$ and MAF in each bin and then calculated the correlation between mean $\hat{b}_m$ (or mean $\hat{u}_m$) and $\log_{10}$ (mean MAF) across bins. Because the MAF bins were not independent, we assessed the statistical significance of the correlation by comparing the observed value to the empirical distribution from 1 million resampling sets under the null hypothesis by randomly sampling the same number of variants for each MAF bin. In addition, we also performed the analyses above using ~2.5 million common SNPs from the GIANT meta-analyses for height and BMI. For the GIANT data, we stratified the variants into 20 MAF bins because of the smaller MAF range and smaller number of variants.

[36] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

[37] Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

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

[39] Patterson, H.D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).