

Genome partitioning of genetic variation for complex traits using common SNPs

Jian Yang^{1*}, Teri A Manolio², Louis R Pasquale³, Eric Boerwinkle⁴, Neil Caporaso⁵, Julie M Cunningham⁶, Mariza de Andrade⁷, Bjarke Feenstra⁸, Eleanor Feingold⁹, M Geoffrey Hayes¹⁰, William G Hill¹¹, Maria Teresa Landi¹², Alvaro Alonso¹³, Guillaume Lettre¹⁴, Peng Lin¹⁵, Hua Ling¹⁶, William Lowe¹⁷, Rasika A Mathias¹⁸, Mads Melbye⁸, Elizabeth Pugh¹⁶, Marilyn C Cornelis¹⁹, Bruce S Weir²⁰, Michael E Goddard^{21,22} & Peter M Visscher¹

We estimate and partition genetic variation for height, body mass index (BMI), von Willebrand factor and QT interval (QTi) using 586,898 SNPs genotyped on 11,586 unrelated individuals. We estimate that ~45%, ~17%, ~25% and ~21% of the variance in height, BMI, von Willebrand factor and QT_i, respectively, can be explained by all autosomal SNPs and a further ~0.5–1% can be explained by X chromosome SNPs. We show that the variance explained by each chromosome is proportional to its length, and that SNPs in or near genes explain more variation than SNPs between genes. We propose a new approach to estimate variation due to cryptic relatedness and population stratification. Our results provide further evidence that a substantial proportion of heritability is captured by common SNPs, that height, BMI and QT_i are highly polygenic traits, and that the additive variation explained by a part of the genome is approximately proportional to the total length of DNA contained within genes therein.

Genome-wide association studies (GWAS) have led to the discovery of hundreds of marker loci that are associated with complex traits, including disease and quantitative phenotypes¹, yet for most traits, the associated variants cumulatively explain only a small fraction of total heritability². GWAS have provided insight into biology through the discovery of pathways that were previously not known to be involved in the trait and the discovery of genes and pathways that are common to two or more complex traits³. As an experimental design, GWAS are hypothesis generating, and typically very stringent statistical thresholds are set to control false positive rates. This approach is at the expense of the false negative rate, that is, failure to detect loci that are associated with the trait but whose effect sizes are too small to reach genome-wide statistical significance. In addition, GWAS typically use common SNP markers. If ungenotyped causal variants have a lower allele frequency than the SNPs in the GWAS, they will be in low linkage disequilibrium (LD) with common SNPs, and the effect estimated

at the SNPs will be proportionally attenuated. That is, the proportion of heritability that can be captured with common SNPs depends on how well causal variants are tagged by these SNPs. For these reasons, the cumulative genetic variation accounted for by SNPs that reach genome-wide statistical significance is certain to be smaller than the total genetic variance.

An alternative to hypothesis testing is to focus on the estimation of the variance explained by all SNPs together. Recently, we showed how this may be done and estimated that ~45% of phenotypic variation for human height is accounted for by common SNPs from a sample of ~4,000 Australians with ancestry in the British Isles⁴. In a separate study, we partitioned additive variance for height onto chromosomes using within-family segregation, which captures the effects of all causal variants, and concluded that the variance was explained in proportion to chromosome length⁵. Here we take these studies further, using a much larger sample of 11,586 unrelated European Americans and considering a range of traits. We partitioned additive genetic variation for height, BMI, von Willebrand factor (vWF) and QT_i onto the autosomes, the X chromosome and genomic segments. vWF is a large adhesive glycoprotein that circulates in plasma and is essential in hemostasis, whereas QT_i is an important electrocardiographic measure related to ventricular arrhythmias and sudden death. We find that the genetic variation explained by a genomic segment is proportional to the length of DNA contained within genes in that segment. We estimate the proportion of variation due to population structure and report empirical results for the X chromosome that are consistent with full dosage compensation (X inactivation) in females in genes that affect these traits.

RESULTS

Variance explained by all autosomal SNPs

We selected 14,347 individuals from three population-based GWAS (the Health Professionals Follow-up Study (HPFS), the Nurses' Health Study (NHS) and the Atherosclerosis Risk in Communities (ARIC) study^{6–8}) and estimated the genetic relationship matrix (GRM) of all the individuals using 565,040 autosomal SNPs that passed quality control (Online Methods). We excluded one of each pair of individuals with an estimated genetic relationship >0.025 (that is, more related than third or fourth cousins) and retained a subset of 11,586 unrelated

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

Received 29 November 2010; accepted 7 April 2011; published online 8 May 2011; doi:10.1038/ng.823

Table 1 Estimates of the variance explained by all autosomal SNPs for height, BMI, vWF and QTl

Trait	<i>n</i>	No PC ^a		10 PCs ^b		Heritability ^d	GWAS ^e
		h_G^2 (s.e.) ^c	<i>P</i>	h_G^2 (s.e.)	<i>P</i>		
Height	11,576	0.448 (0.029)	4.5×10^{-69}	0.419 (0.030)	7.9×10^{-48}	80–90% ³²	~10% ²³
BMI	11,558	0.165 (0.029)	3.0×10^{-10}	0.159 (0.029)	5.3×10^{-9}	42–80% ^{25,26}	~1.5% ¹⁴
vWF	6,641	0.252 (0.051)	1.6×10^{-7}	0.254 (0.051)	2.0×10^{-7}	66–75% ^{33,34}	~13% ¹⁵
QTl	6,567	0.209 (0.050)	3.1×10^{-6}	0.168 (0.052)	5.0×10^{-4}	37–60% ^{35,36}	~7% ¹⁶

The traits vWF and QTl were available in the ARIC cohort only.

^aWithout principal component adjustment. ^bAdjustment with the first 10 principal components from principal component analysis.

^cEstimate of variance explained by all autosomal SNPs. ^dNarrow sense heritability estimate from family or twin studies from the literature. ^eVariance explained by GWAS associated loci from the literature. PC, principal component; s.e., standard error.

individuals. The reason for excluding related pairs is to avoid the possibility that the phenotypic resemblance between close relatives could be because of non-genetic effects (for example, shared environment) and causal variants not tagged by SNPs but captured by pedigree^{9,10}. We then fitted the GRM in a mixed linear model (MLM) to estimate the proportion of variance explained by all the autosomal SNPs (h_G^2) for height, BMI, vWF and QTl in each cohort and the combined data where applicable (Online Methods, **Table 1** and **Supplementary Table 1**). Data on vWF and QTl were available from the ARIC sample only. We show that 44.8% (s.e. 2.9%) of the phenotypic variance for height can be explained by all the autosomal SNPs, which is in line with an estimate of 44.5% (s.e. 8.3%) from a similar analysis of an Australian cohort (3,925 unrelated individuals genotyped by 294,831 SNPs on Illumina arrays, in contrast to the Affymetrix arrays used in the present study)⁴. We show for the first time that 16.5% (s.e. 2.9%), 25.2% (s.e. 5.1%) and 20.9% (s.e. 5.0%) of variances for BMI, vWF and QTl, respectively, can be explained by all the autosomal SNPs, which is approximately tenfold, twofold and threefold larger than the variance explained by all known validated loci found by GWAS for BMI^{11–14}, vWF¹⁵ and QTl¹⁶, respectively. We note that the *ABO* blood group locus on chromosome 9 is known to explain approximately 10% of phenotypic variation for vWF¹⁵ through modification of the amount of H antigen expression on the circulating vWF glycoprotein^{17,18}. The estimate of h_G^2 for weight is 18.6% (s.e. 2.8%). Because of the high phenotypic correlation between BMI and weight ($r = 0.92$), results for these two traits are very similar. We therefore report results for BMI in the following sections and for completion give all results for weight in the supplementary online material (**Supplementary Figs. 1–7** and **Supplementary Tables 1–13**).

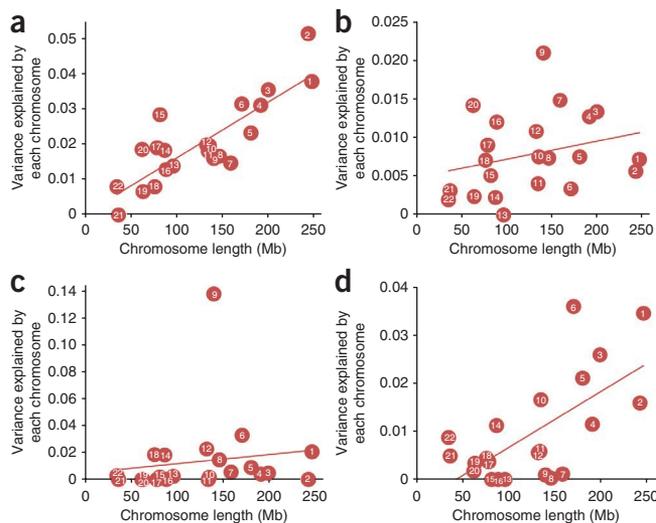
Genome partitioning of genetic variation

Next, we estimated the GRM from the SNPs on each autosome and partitioned the total genetic variance onto individual chromosomes by fitting the GRMs of all the chromosomes simultaneously in a joint analysis (Online Methods). We observed a strong linear relationship between the estimate of variance explained by each chromosome (h_C^2) and chromosome length (L_C , in Mb units) for height ($P = 1.4 \times 10^{-6}$ and $R^2 = 0.695$) and QTl ($P = 1.1 \times 10^{-3}$ and $R^2 = 0.422$) (**Fig. 1** and **Supplementary Tables 2** and **3**). We mapped SNPs to 17,787

Figure 1 Variance explained by chromosomes. Shown are the estimate of the variance explained by each chromosome for (a) height (combined), (b) BMI (combined), (c) vWF (ARIC) and (d) QTl (ARIC) by joint analysis using 11,586 unrelated individuals against chromosome length. The numbers in the circles and squares are the chromosome numbers. The regression slopes and R^2 were 1.6×10^{-4} ($P = 1.4 \times 10^{-6}$) and 0.695 for height, 2.3×10^{-5} ($P = 0.214$) and 0.076 for BMI, 6.9×10^{-5} ($P = 0.524$) and 0.021 for vWF, and 1.2×10^{-4} ($P = 1.1 \times 10^{-3}$) and 0.422 for QTl, respectively.

genes according to positions on the UCSC Genome Browser hg18 assembly¹⁹, 17,652 of which had at least one SNP within ± 50 kb of the 5' and 3' untranslated regions (UTRs). There was also a significant correlation between the estimate of h_C^2 and the number of genes on each chromosome ($N_{g(C)}$) for height ($P = 7.9 \times 10^{-3}$) and QTl ($P = 8.1 \times 10^{-4}$) (**Supplementary Table 3**). Because L_C and $N_{g(C)}$ are correlated ($r = 0.628$), we performed a multiple regression analysis of the estimate of h_C^2 on L_C and $N_{g(C)}$ and fitted models in which chromosome length was fitted after the number of genes and vice versa. When including both L_C and $N_{g(C)}$ in the regression model, $N_{g(C)}$ was not significant and L_C was still significant for height ($P = 8.8 \times 10^{-5}$) and QTl ($P = 2.8 \times 10^{-4}$) (**Supplementary Table 3**). The regression of the estimate of h_C^2 on either L_C or $N_{g(C)}$ was not significant for BMI and vWF. These results are consistent with the variance explained by each chromosome for height and QTl (but less so for BMI and vWF) being proportional to the proportion of the genome being considered. Although longer chromosomes harbor more genes that are implicated in abnormal growth or skeletal development, the relationship between variance explained for height and chromosome length remained significant ($P = 0.016$) after fitting the number of such genes (**Supplementary Fig. 1**). We provide evidence that the linear relationship between the estimate of h_C^2 and L_C cannot be attributed to the fact that longer chromosomes have more SNPs and thereby smaller sampling errors when estimating genetic relationships between individuals (**Supplementary Note** and **Supplementary Figs. 2** and **3**).

However, genes vary greatly in size, and when we considered the length of the genes, we observed that the estimate of h_C^2 for height and QTl was also proportional to the total length of genes on each chromosome ($L_{g(C)}$), where gene length is defined as the physical distance between the beginning and end of the UTRs (**Supplementary Fig. 4**). Because the correlation between L_C and $L_{g(C)}$ is extremely high ($r = 0.97$), we were unable to discriminate whether L_C or $L_{g(C)}$ is causative by multiple regression: the regression of h_C^2 on L_C was not significant after being fitted for $L_{g(C)}$ and vice versa (**Supplementary Table 3**). Therefore, a different analysis was required. We asked whether we could still observe a significant regression of h_C^2 on $L_{g(C)}$ when chromosome length was held constant. We investigated this by dividing the genome into segments with the same length of either 50 or 30 Mb



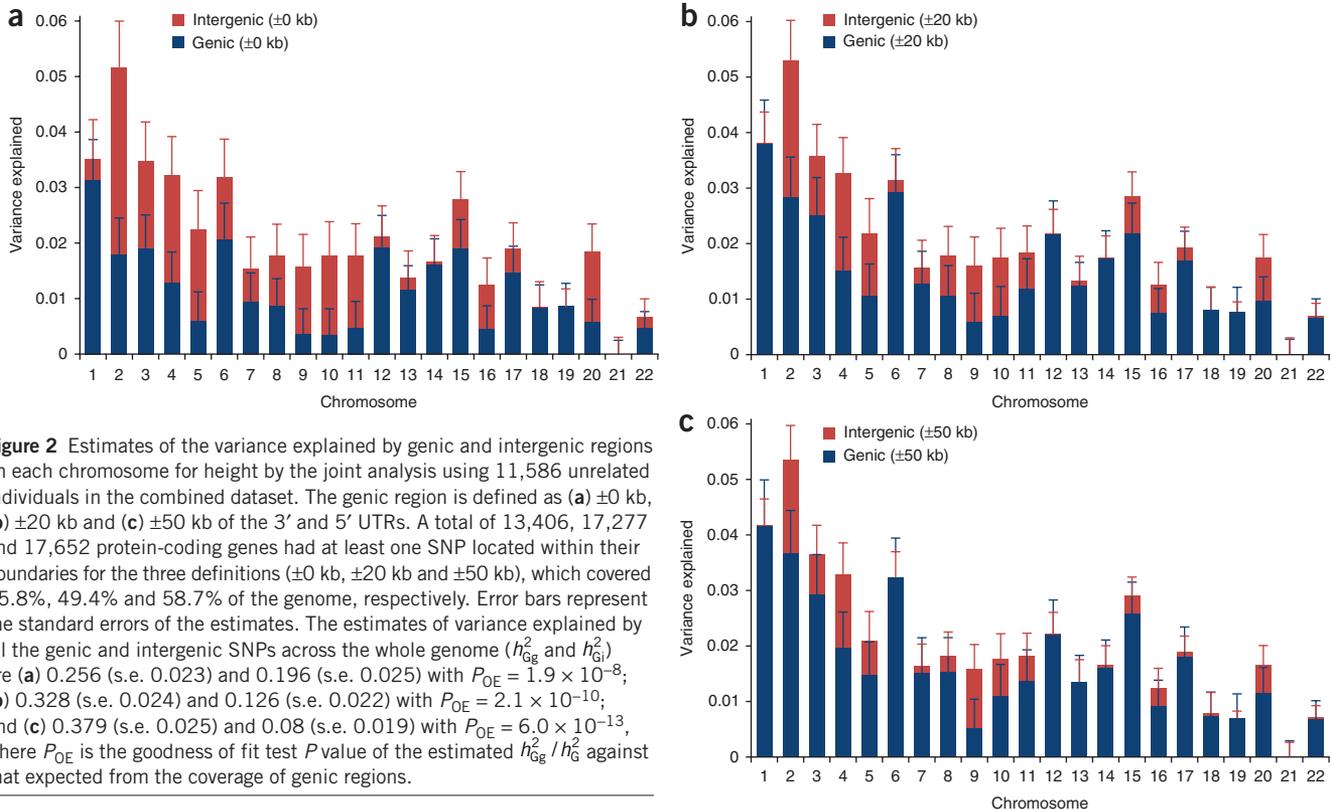


Figure 2 Estimates of the variance explained by genic and intergenic regions on each chromosome for height by the joint analysis using 11,586 unrelated individuals in the combined dataset. The genic region is defined as (a) ± 0 kb, (b) ± 20 kb and (c) ± 50 kb of the 3' and 5' UTRs. A total of 13,406, 17,277 and 17,652 protein-coding genes had at least one SNP located within their boundaries for the three definitions (± 0 kb, ± 20 kb and ± 50 kb), which covered 35.8%, 49.4% and 58.7% of the genome, respectively. Error bars represent the standard errors of the estimates. The estimates of variance explained by all the genic and intergenic SNPs across the whole genome (h_{Gg}^2 and h_{Gi}^2) are (a) 0.256 (s.e. 0.023) and 0.196 (s.e. 0.025) with $P_{OE} = 1.9 \times 10^{-8}$; (b) 0.328 (s.e. 0.024) and 0.126 (s.e. 0.022) with $P_{OE} = 2.1 \times 10^{-10}$; and (c) 0.379 (s.e. 0.025) and 0.08 (s.e. 0.019) with $P_{OE} = 6.0 \times 10^{-13}$, where P_{OE} is the goodness of fit test P value of the estimated h_{Gg}^2/h_{Gi}^2 against that expected from the coverage of genic regions.

and then estimated the variance explained by each segment (h_s^2) in a joint analysis (Online Methods). We found that the regression of h_s^2 on the total gene length per segment ($L_{g(S)}$) remained significant for height, with $P = 1.7 \times 10^{-3}$ for 50-Mb segments and $P = 1.2 \times 10^{-4}$ for 30-Mb segments (Supplementary Fig. 5). The regressions of h_s^2 on the number of genes, the total length of exons and the number of exons on each segment were also significant in some cases, but none of the regressions were significant when fitted after $L_{g(S)}$, whereas $L_{g(S)}$ was always significant fitted after any of them (Supplementary Table 4). These results suggest that, at least for height, genic regions explain variation in proportion to their genic content.

To quantify these effects genome wide, we partitioned the variance explained by all the SNPs onto genic (h_{Gg}^2) and intergenic (h_{Gi}^2) regions of the whole genome (Online Methods). We defined the gene boundaries as ± 0 kb, ± 20 kb and ± 50 kb of the 3' and 5' UTRs. A total of 213,509, 282,058 and 336,127 SNPs were located within the boundaries of 13,406, 17,277 and 17,652 protein-coding genes for the three definitions (± 0 kb, ± 20 kb and ± 50 kb), respectively, which covered 35.8%, 49.4% and 58.7% of the genome. Some genes did not have any SNPs within them, especially if we used the most stringent definition of gene boundary (± 0 kb). We tested the estimates of h_{Gg}^2 and h_{Gi}^2 against the expected values from the genic and intergenic coverages by a goodness of fit test. We found strong evidence for height and vWF, and less so for BMI and QT_i, that genic regions proportionally explain more variation than intergenic regions (see legends of Fig. 2 and Supplementary Fig. 6). As an example, we considered the case of genes ± 20 kb of the 3' and 5' UTRs, where genic and intergenic coverages are roughly equal (49.4% compared to 50.6%). The estimates of h_{Gg}^2 compared to h_{Gi}^2 were 32.8% versus 12.6% ($P_{OE} = 2.1 \times 10^{-10}$) for height, 22.7% versus 4.0% ($P_{OE} = 5.1 \times 10^{-4}$) for vWF, 11.7% versus 4.7% ($P_{OE} = 0.022$) for BMI and 13.5% versus 7.5% ($P_{OE} = 0.251$) for QT_i, where P_{OE} is the goodness of fit test P value of the estimated

h_{Gg}^2/h_{Gi}^2 against that expected from the coverage of genic regions. We further partitioned the genetic variance onto the genic and intergenic regions of each chromosome (Online Methods). In general, the results agree with those of the whole-genome partitioning analysis in that the genic regions proportionally explained more variation (Fig. 2 and Supplementary Fig. 6). The variance attributable to chromosome 9 for vWF is dominated by the genic regions, which is expected because *ABO* on this chromosome explains $\sim 10\%$ of its variance¹⁵. However, there appear to be exceptions, for example, the intergenic regions of chromosome 2 and chromosome 5 seemed to be more important for BMI and QT_i, respectively. These results are not conclusive because the standard errors of the estimates are large. Despite these special cases, overall, the results are consistent with causal variants being more likely to be located in the vicinity of functional genes.

Quantifying the effect of population structure

To quantify the effect of population structure, we estimated the variance for each chromosome when analyzed individually (h_C^2 (sep)) and when analyzed jointly (h_C^2) in the entire sample of 14,347 individuals (without removing cryptic relatives) and regressed the difference between these estimates on chromosome length (Online Methods). The intercept of this regression (b_0) appears to be due to cryptic relatedness because when we eliminated relatives with a relationship > 0.025 , b_0 declined to zero (Fig. 3). We therefore predicted that cryptic relatedness accounted for 1.5%, 0.084%, 0.22% and 0.065% (not significant) of the phenotypic variance for height, BMI, vWF and QT_i, respectively, in the entire sample. The variance attributed to cryptic relatedness is irrespective of chromosome length because it does not require very many SNPs per chromosome to detect close relatives. Conversely, the regression slope b_1 appears to be due to population stratification because longer chromosomes are likely to have more ancestry informative markers (AIMs), assuming that the

Figure 3 Variance due to cryptic relatedness and population stratification. Shown is the difference between the estimates of variance explained by each chromosome by the separate ($h_c^2(\text{sep})$) and joint (h_c^2) analyses for (a) height (combined), (b) BMI (combined), (c) vWF (ARIC) and (d) QTl (ARIC) against chromosome length. All, using all the individuals in the entire sample. Unrelated, using unrelated individuals after excluding one of each pair of individuals with an estimate of genetic relationship >0.025 . The intercept and slope are 0.015 ($P = 5.5 \times 10^{-10}$) and 6.9×10^{-5} ($P = 3.4 \times 10^{-7}$) for height; 8.4×10^{-4} ($P = 0.046$) and 7.2×10^{-6} ($P = 0.020$) for BMI; 2.2×10^{-3} ($P = 0.025$) and -1.9×10^{-6} ($P = 0.779$) for vWF; and 6.5×10^{-4} ($P = 0.401$) and 2.3×10^{-5} ($P = 4.1 \times 10^{-4}$) for QTl in the entire sample and are 0.002 ($P = 0.070$) and 5.6×10^{-5} ($P = 5.5 \times 10^{-7}$) for height; 2.9×10^{-4} ($P = 0.556$) and 7.1×10^{-6} ($P = 0.054$) for BMI; 1.7×10^{-3} ($P = 0.179$) and 1.1×10^{-6} ($P = 0.901$) for vWF; and 5.9×10^{-4} ($P = 0.523$) and 2.4×10^{-5} ($P = 0.001$) for QTl in unrelated individuals.

AIMs are randomly distributed across the genome. We then predicted that population stratification accounted for $6.9 \times 10^{-5} L_C$, $7.2 \times 10^{-6} L_C$, $-1.92 \times 10^{-6} L_C$ (not significantly different from zero) and $2.3 \times 10^{-5} L_C$ of variance for height, BMI, vWF and QTl, respectively, in the entire sample and a similar amount in the dataset of unrelated individuals (Fig. 3). The difference between $h_c^2(\text{sep})$ and h_c^2 represents the overall effect of all the other 21 chromosomes on one chromosome. Therefore, the proportion of variance attributed to population structure (cryptic relatedness and population stratification) across

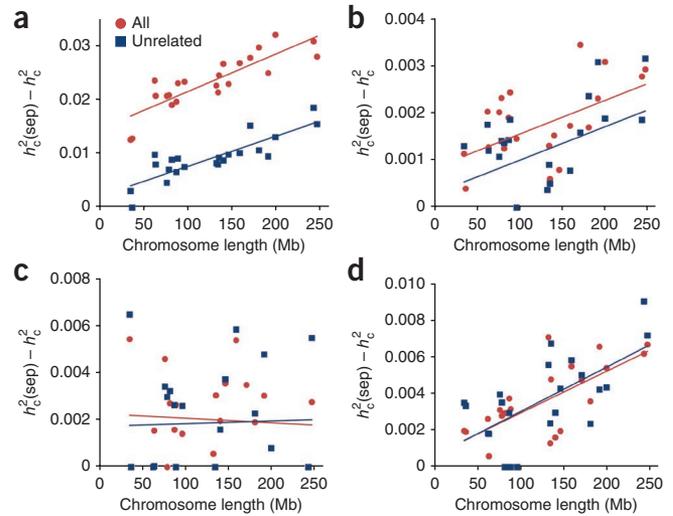
the whole genome is approximately equal to $b_0 22 / 21 + b_1 \sum_{C=1}^{22} L_C / 21$,

which is (1.6% + 0.91%), (0.088% + 0.095%), (0.23% + 0.0%) and (0.068% + 0.30%) for height, BMI, vWF and QTl, respectively, in the entire sample. Hence, we provide a simple approach to estimate and partition the variance attributed to population structure for complex traits. The variances due to cryptic relatedness and population stratification depend on the data structure in the sample. Therefore, the estimates we present above are specific for the data in this study.

It is common to fit eigenvectors (principal components) from principal component analysis in single SNP association studies to correct for possible population structure^{20,21}. We show that fitting the first ten principal components and one chromosome at a time or fitting all chromosomes simultaneously without fitting principal components led to similar estimates of the variance explained by each chromosome (Supplementary Fig. 7), which suggests that the majority of variance attributed to population structure is well captured by the first ten principal components in these data.

Estimation of variance explained by the X chromosome

We estimated the GRM for the X chromosome and parameterized it under three assumptions of dosage compensation⁹: (i) equal X-linked variance for males and females; (ii) no dosage compensation (both X chromosomes are active for females); and (iii) full dosage compensation (one of the X chromosomes is completely inactive for females). We fitted the parameterized GRMs for the X chromosome in an MLM while simultaneously estimating h_c^2 in the model to capture the genetic variation on the autosomes and variation due to possible population structure. For all the traits, the full-dosage compensation model fits the data best and the no-dosage compensation model is the worst, with the equal-variance model being in between (Supplementary Table 5). However, the differences in estimates were relatively small and none of them were statistically significant. Larger datasets will be required to distinguish such small differences. Under the assumption of full dosage compensation, the variance attributable to the X chromosome for females was 0.61% (s.e. 0.32%), 0.82%



(s.e. 0.35%), 0.57% (s.e. 0.52%) and 0.0% (s.e. 0.48%) for height, BMI, vWF and QTl, respectively. To verify those results, we detected heterogeneous variances on the X chromosome rather than autosomal variance differences between males and females, and we fitted the same dosage compensation models for the autosomes. The equal variance model fitted the data best and the full dosage compensation model was the worst fit for all the traits (Supplementary Table 6). Therefore, the data are consistent with twice as much additive genetic variation for height, BMI and vWF on the X chromosome in males as in females, which is predicted from theory under the assumption of random X inactivation²². Although there are syndromic examples illustrating the phenotypic effect of the Lyon hypothesis (for example, Turner's syndrome and Klinefelter syndrome), to our knowledge, this is the first empirical evidence from genotype-phenotype associations on complex traits that the amount of genetic variation on the X chromosome appears consistent with X-chromosome inactivation. However, the evidence is indirect and not overwhelming. Larger samples sizes and the detection of multiple associated loci on the X chromosome will be necessary to investigate the expression of genes on the X chromosome that affect the traits studied.

Comparison with known associated variants

To quantify the effect of known associated variants on the results, we included the *FTO* SNP rs939609 on chromosome 16 for BMI and the *ABO* SNP rs612169 on chromosome 9 for vWF as a covariate when estimating h_c^2 by the joint analysis of all autosomes. *FTO* was the first locus to be detected through GWAS that is associated with BMI¹³, and *ABO* is a major determinant of vWF¹⁸. When compared to the result without adjustment, the estimate of variance due to chromosome 16 (h_{16}^2) for BMI decreased from 1.19% to 0.61%, which is in line with an estimate of ~0.34% to ~1% of variance explained by the *FTO* locus for BMI in previous GWAS^{11,13,14} and an estimate of ~0.46% from the association analysis in the present study; the estimate of h_9^2 for vWF decreased by 11.8%, which is consistent with an estimate of ~10% of variance for vWF explained by the *ABO* locus in GWAS¹⁵; and the estimates for the other chromosomes remained the same (Supplementary Fig. 8).

The meta-analysis of ~133,000 individuals by the GIANT consortium has identified 180 independent loci associated with genetic variation of height²³. The estimate of h_c^2 by a joint analysis in our study shows a high correlation ($r = 0.715$ and $P = 1.8 \times 10^{-4}$) with the sum of the variance explained at the associated loci on each chromosome from the GIANT meta-analysis (Fig. 4).

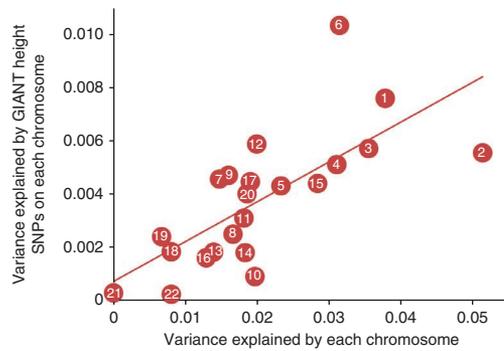


Figure 4 The sum of variance explained by the GWAS associated SNPs on each chromosome in the GIANT meta-analysis of height²³ against the estimate of variance explained by each chromosome for height by the joint analysis using the combined data of 11,586 unrelated individuals in the present study. We calculated the variance explained by GWAS loci in the GIANT meta-analysis based on the result of its replication study. The regression R^2 is 0.511 ($P = 1.8 \times 10^{-4}$).

Additional models

We fitted a number of other models to quantify the effect of having multiple phenotypic observations per individual and to test for genotype-sex interaction effects and for the effect of sample ascertainment. We also estimated the genetic correlation between height and weight. These additional models exemplify the versatility of the linear mixed model methodology used in this study. Results are shown in the **Supplementary Note**.

DISCUSSION

In this study, we estimate that ~45%, ~17%, ~25% and ~21% of phenotypic variation for height, BMI, vWF and QT_i, respectively, is tagged by common SNPs, and we partition this variation onto autosomes, chromosome segments and the X chromosome. We find that chromosome segments explain variation in approximate proportion to the total length of genes contained therein. Although this suggests that there are very many polymorphisms affecting these traits, the linear relationship between the estimate of variance explained and genomic length is not perfect, especially for BMI and vWF. Chromosomes with similar (genic) lengths can explain different amounts of variation (Fig. 1 and **Supplementary Fig. 4**), and the estimates of variance explained by genomic segments with equal length also show large variability (**Supplementary Fig. 5**), suggesting some granularity in the distribution of causal variants. The genetic architecture of vWF is distinct from the other traits we analyzed, as a large proportion of variance is explained by a common SNP in a single gene (*ABO*). We show that the variance attributed to a single major gene can be captured by all the SNPs on that chromosome or the whole genome, showing that our whole-genome and chromosome estimation approach is independent of the distribution of effect sizes. Our results provide further evidence for the highly polygenic nature of complex trait variation and that a substantial proportion of genetic variation is tagged by common SNPs^{4,24}. These results have implications for the experimental design to detect additional variation and are informative with respect to the nature of complex trait variation.

Of the four traits studied, the largest proportion of phenotypic variance explained by the SNPs was for height and the smallest was for BMI. Why are the results for height and BMI so different? Heritability of height is approximately 80%, and we estimate that more than half of this variation ($45/80 = 0.56$) is tagged by common SNPs. Estimates of the narrow sense heritability of BMI appear to

be more variable, ranging from 42–58% when estimated from the correlation of full brothers and fathers and sons²⁵ to 60–80% from twin studies²⁶. Nevertheless, even if we assume that the narrow sense heritability for BMI is 50%, then only $17/50 = 0.34$ of additive genetic variation is explained by common SNPs. Given these assumptions and the standard errors listed in **Table 1**, the standard error of the difference in the proportion of genetic variance explained for height and BMI is approximately 0.07, so the observed difference of 0.22 appears statistically significant. These results are consistent with the proportion of phenotypic variation for height and BMI explained by genome-wide significant SNPs in that for height, about 10% of the phenotypic variance is explained, yet for BMI the phenotypic variance explained is less than 2%^{14,23}, despite similar and large experimental sample sizes. These results imply that causal variants for BMI are in less LD with common SNPs than causal variants for height, possibly because, on average, causal variants for BMI have a lower minor allele frequency than causal variants for height. Both observations from GWAS and our analyses are consistent with the allelic architecture for BMI being different from that for height. Different evolutionary pressures on obesity (or leanness) and height could account for such differences because natural selection will result in low frequencies of alleles that are correlated with fitness²⁷. However, we do not provide direct evidence to support this hypothesis.

If genetic variation is a function of the length of a chromosome segment occupied by genes, then this implies that causal variants are more likely to occur in the vicinity of the genes than in intergenic regions (Fig. 2 and **Supplementary Fig. 6**). These causal variants could either change the protein structure or regulate the expression of the gene in *cis*. However, regulatory elements sometimes occur a long distance away from the gene they regulate, and our results show that SNPs situated >50 kb from any gene still explain some of the variance, although they explain less than SNPs nearer to a gene. These results are consistent with analyses of published genome-wide significant SNPs for complex traits in that a substantial proportion is found in intergenic regions¹.

GWAS for height, BMI, vWF and QT_i to date have identified individual genetic variants that cumulatively explain about 10%, 1.5%, 13% and 7% of phenotypic variation, respectively^{14–16,23}. In contrast, we show that 45%, 17%, 25% and 21%, respectively, of the variance is explained by common SNPs (**Table 1**). The difference between these two sets of figures is caused by SNPs that are associated with the traits but do not reach genome-wide significance. The proportion of variance explained by all the SNPs is less than the heritability because of incomplete LD between the causal polymorphisms and the SNPs. Therefore, experiments to find SNPs that pass the genome-wide significance threshold can focus on the proportion of variation that is tagged by common SNPs by increasing sample size or focus on the proportion of variation that is not tagged, for example, by considering less common variants. The former approach has been successfully done by the GIANT consortium, which reported that 10% and 1.5% of variation for height and BMI, respectively, can be accounted for by common SNPs using sample sizes of more than 100,000 (refs. 14,23). The latter will be facilitated by the 1000 Genomes Project²⁸ and independently by efforts to sequence exomes and whole genomes. Experimental designs to discover causal variants that are in LD with common SNPs and those that interrogate less common or rare variants are complementary, and recent publications that suggest that all or most variation for disease is to be found in less common or rare (coding) variants^{29,30} are not consistent with empirical data, at least for a range of complex traits, including height, BMI, lipids and schizophrenia^{14,23,24,31}. For those causal variants that are rare in

the population (for example, with a frequency of less than 1%), an important but unanswered question is whether their effect sizes are large enough to be detected through conventional association analysis. The power of detection for a rare variant is proportional to the product of its frequency (which is small) and the square of its effect size. Hence, rare variants will be detected only if their effect sizes are large enough given their low frequency. Our results imply that there are many chromosomal regions that contain causal variants and so most must explain a small proportion of total variance. Such small contributions can be due to loci with very low minor allele frequency and large effect sizes, but our ability to detect them by association is limited by the amount of variance explained.

Genome partitioning methods such as applied here help us further understand the genetic architecture of complex traits. All the methods and analyses presented in this paper have been implemented in the GCTA software⁹. With ever larger samples sizes, the methods we have used and those that are based upon traditional GWAS analyses will converge in inference in that we will be able to partition variation to individual loci.

URLs. UCSC Genome Browser, <http://genome.ucsc.edu/>; GCTA, <http://gump.qimr.edu.au/gcta/>.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturegenetics/>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Funding support for the Gene, Environment Association Studies (GENEVA) project has been provided through the US National Institutes of Health Genes, Environment and Health Initiative. For the ARIC project, support was from U01 HG 004402 (PI: E.A. Boerwinkle). For the NHS and HPFS support is from U01 HG 004399 and U01 HG 004728 (PIs: F.B. Hu and L.R. Pasquale). The genotyping for the ARIC, NHS and HPFS studies was performed at the Broad Institute of MIT and Harvard with funding support from U01 HG04424 (PI: S. Gabriel). The GENEVA Coordinating Center receives support from U01 HG 004446 (PI: B.S. Weir). Assistance with GENEVA data cleaning was provided by the National Center for Biotechnology Information. D. Crosslin and C. Laurie of the GENEVA project assisted in making the data available for analysis. A Physician Scientist Award from Research to Prevent Blindness in New York City also supports L.R.P. M.C.C. is a recipient of a Canadian Institutes of Health Research Fellowship. We acknowledge funding from the Australian National Health and Medical Research Council (NHMRC grants 389892 and 613672) and the Australian Research Council (ARC grants DP0770096 and DP1093900). We thank D. Posthuma for discussions and the referees for constructive comments.

AUTHOR CONTRIBUTIONS

P.M.V., M.E.G., B.S.W. and T.A.M. designed the study. J.Y. performed all statistical analyses. J.Y. and P.M.V. wrote the first draft of the paper. L.R.P., E.B., N.C., J.M.C., M.d.A., B.F., E.F., M.G.H., W.G.H., M.T.L., A.A., G.L., P.L., H.L., W.L., R.A.M., M.M., E.P. and M.C.C. contributed by providing genotype and phenotype data, by giving advice on analyses and interpretation of results and/or by giving advice on the contents of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

- Hindorf, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
- Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

- Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
- Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
- Visscher, P.M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am. J. Hum. Genet.* **81**, 1104–1110 (2007).
- Rimm, E.B. *et al.* Prospective study of alcohol consumption and risk of coronary disease in men. *Lancet* **338**, 464–468 (1991).
- Colditz, G.A. & Hankinson, S.E. The Nurses' Health Study: lifestyle and health among women. *Nat. Rev. Cancer* **5**, 388–396 (2005).
- Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ. Cardiovasc. Genet.* **2**, 73–80 (2009).
- 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).
- Visscher, P.M., Yang, J. & Goddard, M.E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang *et al.* (2010). *Twin Res. Hum. Genet.* **13**, 517–524 (2010).
- Willer, C.J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
- Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41**, 18–24 (2009).
- Frayling, T.M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
- 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).
- Smith, N.L. *et al.* Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: the CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) consortium. *Circulation* **121**, 1382–1392 (2010).
- Shah, S.H. & Pitt, G.S. Genetics of cardiac repolarization. *Nat. Genet.* **41**, 388–389 (2009).
- Preston, A.E. & Barr, A. The plasma concentration of factor viii in the normal population. II. The effects of age, sex and blood group. *Br. J. Haematol.* **10**, 238–245 (1964).
- O'Donnell, J., Boulton, F.E., Manning, R.A. & Laffan, M.A. Amount of H antigen expressed on circulating von Willebrand factor is modified by ABO blood group genotype and is a major determinant of plasma von Willebrand factor antigen levels. *Arterioscler. Thromb. Vasc. Biol.* **22**, 335–341 (2002).
- Liu, J.Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
- Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
- Bulmer, M.G. *The Mathematical Theory of Quantitative Genetics* (Oxford University Press, New York, New York, USA, 1985).
- Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
- Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
- Magnusson, P.K. & Rasmussen, F. Familial resemblance of body mass index and familial risk of high and low body mass index. A study of young men in Sweden. *Int. J. Obes. Relat. Metab. Disord.* **26**, 1225–1231 (2002).
- Schousboe, K. *et al.* Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Res.* **6**, 409–421 (2003).
- Eyre-Walker, A. 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).
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
- McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
- Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
- Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
- Orstavik, K.H. *et al.* Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. *Am. J. Hum. Genet.* **37**, 89–101 (1985).
- de Lange, M., Snieder, H., Ariens, R.A., Spector, T.D. & Grant, P.J. The genetics of haemostasis: a twin study. *Lancet* **357**, 101–105 (2001).
- Dalageorgou, C. *et al.* Heritability of QT interval: how much is explained by genes for resting heart rate? *J. Cardiovasc. Electrophysiol.* **19**, 386–391 (2008).
- Russell, M.W., Law, I., Sholinsky, P. & Fabsitz, R.R. Heritability of ECG measurements in adult male twins. *J. Electrocardiol.* **30** Suppl, 64–68 (1998).

¹Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, Brisbane, Australia. ²Office of Population Genomics, National Human Genome Research Institute (NHGRI), Bethesda, Maryland, USA. ³Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts, USA. ⁴Human Genetics Center and Division of Epidemiology, University of Texas Health Science Center, Houston, Texas, USA. ⁵Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), Bethesda, Maryland, USA. ⁶Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. ⁷Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA. ⁸Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark. ⁹Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. ¹⁰Division of Endocrinology, Metabolism and Molecular Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA. ¹¹Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. ¹²Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland, USA. ¹³Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA. ¹⁴Montréal Heart Institute, Université de Montréal, Montréal, Quebec, Canada. ¹⁵Human and Statistical Genetics Program, School of Medicine, Washington University, St. Louis, Missouri, USA. ¹⁶Center for Inherited Disease Research, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. ¹⁷Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA. ¹⁸Department of General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. ¹⁹Department of Nutrition, School of Public Health, Harvard University, Boston, Massachusetts, USA. ²⁰Department of Biostatistics, University of Washington, Seattle, Washington, USA. ²¹Department of Food and Agricultural Systems, University of Melbourne, Victoria, Australia. ²²Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria, Australia. Correspondence should be addressed to P.M.V. (peter.visscher@qimr.edu.au).

ONLINE METHODS

GWAS samples and quality control. Details of the HPFS, NHS and ARIC cohorts have been described previously^{6–8}. The GWAS data in terms of study design, sample selection and genotyping have been detailed for the HPFS and NHS³⁷ cohorts and for the ARIC cohort⁸. All three cohorts have been studied as part of the GENEVA (the Gene, Environment Association Studies) project³⁸, and this study has benefitted from using data from the consortium that have been generated and cleaned using a common protocol. We selected 6,293 individuals (2,745 cases with type 2 diabetes and 3,148 controls) from the NHS and HPFS cohorts and 15,792 individuals from the ARIC cohort. All of these selected individuals were genotyped using the Affymetrix Genome-Wide Human 6.0 array.

Of the 909,622 SNP probes, 874,517 (HPFS), 879,071 (NHS) and 841,820 (ARIC) passed quality control analysis performed by the Broad Institute and the GENEVA Coordinating Center (excluding SNPs with missing call rate ≥5% or plate association $P < 1 \times 10^{-10}$)³⁹. We further excluded SNPs with missing rate ≥2%, >1 discordance in the duplicated samples, Hardy-Weinberg equilibrium $P < 1 \times 10^{-3}$ or minor allele frequency <0.01. A total of 687,398 (27,578), 665,163 (24,108) and 593,521 (23,664) autosomal (X chromosome) SNPs were retained for the HPFS, NHS and ARIC cohorts, respectively, 565,040 (21,858) of which were in common across the three cohorts.

We included only one of each set of duplicated samples and one of each pair of samples that were identified as full siblings by an initial scan of relatedness in PLINK⁴⁰. We investigated population structure by PCA of all the autosomal SNPs that passed quality control and included only samples of European ancestry (Supplementary Fig. 9). We excluded samples with gender misidentification by examining the mean of the intensities of SNP probes on the X and Y chromosomes. We also excluded samples with missing call rate ≥2% and samples on two plates that showed an extremely high level of mean inbreeding coefficients. A total of 2,400 (HPFS), 3,265 (NHS) and 8,682 (ARIC) samples were retained for analysis with a combined set of 14,347 samples.

Phenotypes. Summary statistics of the phenotypes of height, weight, BMI, vWF and QT_i are shown in Supplementary Table 7. There are three measures of weight and a single measure of height in both the HPFS and NHS cohorts, four measures of weight and three measures of height in the ARIC cohort, and single measures of vWF and QT_i in the ARIC cohort. For height, weight and BMI, we used the mean of repeated measures in all the analyses except for the analysis of the repeatability model. We adjusted the phenotypes (or the mean phenotype) for age and standardized it to a z score in each gender group in each of the three cohorts separately.

Statistical analysis. We estimated the GRM of all individuals in the combined data from all the autosomal SNPs using the method we recently developed^{4,9} and excluded one of each pair of individuals with an estimated genetic relationship >0.025. We then estimated the variance explained by all autosomal SNPs by restricted maximum likelihood analysis of an MLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}_G + \boldsymbol{\epsilon}$, where \mathbf{y} is a vector of phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects (for example, the first ten principal components) with its incidence matrix \mathbf{X} , \mathbf{g}_G is a vector of aggregate effects of all autosomal SNPs with $\text{var}(\mathbf{g}_G) = \mathbf{A}_G \boldsymbol{\sigma}_G^2$, and \mathbf{A}_G is the GRM estimated from all autosomal SNPs. The proportion of variance explained by all autosomal SNPs is defined as $h_G^2 = \boldsymbol{\sigma}_G^2 / \boldsymbol{\sigma}_P^2$, with $\boldsymbol{\sigma}_P^2$ being the phenotypic variance.

Furthermore, we estimated the GRM from the SNPs on each chromosome (\mathbf{A}_C) and estimated the variance attributable to each chromosome by fitting the GRMs of all the chromosomes simultaneously in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{C=1}^{22} \mathbf{g}_C + \boldsymbol{\epsilon}; \text{ where } \mathbf{g}_C \text{ is a vector of genetic effects attributable to each}$$

chromosome and $\text{var}(\mathbf{g}_C) = \mathbf{A}_C \boldsymbol{\sigma}_C^2$ (joint analysis). The proportion of variance explained by each chromosome is defined as $h_C^2 = \boldsymbol{\sigma}_C^2 / \boldsymbol{\sigma}_P^2$. We also fitted one chromosome at a time in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}_C + \boldsymbol{\epsilon}$ (separate analysis). If there is an effect of population structure, SNPs on one chromosome will be correlated with the SNPs on the other chromosomes such that h_C^2 (sep) will be overestimated in the separate analysis.

We extended the joint analysis of chromosomes to that of genomic segments. We divided the genome evenly into N_S segments with each of d_S Mb

length and then estimated the GRM using the SNPs on each segment. We estimated the variance explained by each segment (h_C^2) by fitting the GRMs

of all the segments in an MLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{S=1}^{N_S} \mathbf{g}_S + \boldsymbol{\epsilon}$ where \mathbf{g}_S is a vector of genetic effects attributable to each segment.

We further partitioned the variance explained by all the SNPs onto genic and intergenic regions of the whole genome (h_{Gg}^2 and h_{Gi}^2) as well as that of each chromosome (h_{Cg}^2 and h_{Ci}^2). The gene boundaries were defined as $\pm d_g$ kb away from the 3' and 5' UTRs. We estimated h_{Gg}^2 and h_{Gi}^2 by fitting all the genic and intergenic SNPs in an MLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}_{Cg} + \mathbf{g}_{Gi} + \boldsymbol{\epsilon}$, and estimated h_{Cg}^2 and h_{Ci}^2 by fitting the genic and nongenic SNPs on individual

chromosomes in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{C=1}^{22} \mathbf{g}_{Cg} + \sum_{C=1}^{22} \mathbf{g}_{Ci} + \boldsymbol{\epsilon}$.

We estimated the variance attributable to the X chromosome using the method we recently developed⁹. In brief, we estimated the GRM for the X chromosome (\mathbf{A}_X) using the following equations

$$\hat{A}_{jk}^M = \sum_{i=1}^N \frac{(x_{ij}^M - p_i)(x_{ik}^M - p_i)}{p_i(1 - p_i)}$$

for a male-male pair,

$$\hat{A}_{jk}^F = \sum_{i=1}^N \frac{(x_{ij}^F - 2p_i)(x_{ik}^F - 2p_i)}{2p_i(1 - p_i)}$$

for a female-female pair and

$$\hat{A}_{jk}^{MF} = \sum_{i=1}^N \frac{(x_{ij}^M - p_i)(x_{ik}^F - 2p_i)}{\sqrt{2}p_i(1 - p_i)}$$

for a male-female pair, where x_{ij}^M and x_{ij}^F are the number of copies of the reference allele for an X chromosome SNP for a male and a female, respectively, p_i is the frequency of the reference allele and N is the number of SNPs. Assuming the male-female genetic correlation to be 1, the X-linked phenotypic covariance is $\text{cov}_X(y_j^M, y_k^M) = A_{jk}^M \boldsymbol{\sigma}_{X(M)}^2$ for a male-male pair, $\text{cov}_X(y_j^F, y_k^F) = A_{jk}^F \boldsymbol{\sigma}_{X(F)}^2$ for a female-female pair or $\text{cov}_X(y_j^M, y_k^F) = A_{jk}^{MF} \boldsymbol{\sigma}_{X(M)} \boldsymbol{\sigma}_{X(F)}$ for a male-female pair^{22,41}, where $\boldsymbol{\sigma}_{X(M)}^2$ and $\boldsymbol{\sigma}_{X(F)}^2$ are X-linked genetic variances for males and females, respectively. Assumptions about inactivity of the X chromosome (dosage compensation) imposed a relationship between $\boldsymbol{\sigma}_{X(M)}^2$ and $\boldsymbol{\sigma}_{X(F)}^2$, which allow a single variance component $\boldsymbol{\sigma}_{X(F)}^2$ to account for the X-linked genetic variance for both sexes. Therefore, we can express the X-linked phenotypic covariances as $\text{cov}_X(y_j^M, y_k^M) = d^2 A_{jk}^M \boldsymbol{\sigma}_{X(F)}^2$, $\text{cov}_X(y_j^F, y_k^F) = A_{jk}^F \boldsymbol{\sigma}_{X(F)}^2$ and $\text{cov}_X(y_j^M, y_k^F) = d A_{jk}^{MF} \boldsymbol{\sigma}_{X(F)}^2$, where d is the lyonization coefficient, $\boldsymbol{\sigma}_{X(M)} = d \boldsymbol{\sigma}_{X(F)}$, which takes 1 under the hypothesis of equal X-linked genetic variance for both sexes, takes $1/\sqrt{2}$ under the hypothesis of no dosage compensation (both X chromosomes are active for females) and takes $\sqrt{2}$ under the hypothesis of full dosage compensation (complete inactivity of one X chromosome for females) (Supplementary Note). In the analysis of MLM, we took the lyonization coefficient into account by parameterizing the raw \mathbf{A}_X matrix, meaning $\mathbf{A}_X^p = d^2 \mathbf{A}_X$ for male pairs, $\mathbf{A}_X^p = \mathbf{A}_X$ for female pairs and $\mathbf{A}_X^p = d \mathbf{A}_X$ for male-female pairs. We estimated $\boldsymbol{\sigma}_{X(F)}^2$ under the three hypotheses by fitting the parameterized GRM for the X chromosome (\mathbf{A}_X^p) conditional on the GRM estimated from all autosomal SNPs in an MLM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}_X + \mathbf{g}_G + \boldsymbol{\epsilon}$, where \mathbf{g}_X is a vector of X-linked genetic effects with $\text{var}(\mathbf{g}_X) = \mathbf{A}_X^p \boldsymbol{\sigma}_{X(F)}^2$.

Variance attributed to population structure. Mixed linear model methods are useful to control for population structure in GWAS^{42,43}. Population structure in the data causes correlations of SNPs on different chromosomes. Consequently, fitting only one chromosome in the model (separate analysis) also captures some of the variance caused by other chromosomes so that the estimate of variance explained by each chromosome from the separate analysis (h_C^2 (sep)) is biased upwards. The joint analysis has the advantage of protecting against such inter-chromosomal correlations because the estimate of each h_C^2 is conditional on the other chromosomes in the model so that the estimates of variance explained by different chromosomes are independent of each other. We therefore can calculate

the variance attributable to population structure by comparing the estimates between $h_C^2(\text{sep})$ and h_C^2 . The inter-chromosomal SNP correlations occur for two reasons: (i) cryptic relatedness (for example, unexpected cousins), because closely related individuals will share SNPs identical by descent on more than one chromosome; or (ii) systematic difference in allele frequencies between sub-populations (population stratification). We modeled the variance attributed to these two forms of population structure as $h_C^2(\text{sep})-h_C^2 = b_0 + b_1L_C + \varepsilon$, where the slope b_1 allows for the possibility that longer chromosomes track population structure better than smaller chromosomes.

37. Qi, L. *et al.* Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum. Mol. Genet.* **19**, 2706–2715 (2010).
38. Cornelis, M.C. *et al.* The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet. Epidemiol.* **34**, 364–372 (2010).
39. Laurie, C.C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
40. 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).
41. Kent, J.W. Jr, Dyer, T.D. & Blangero, J. Estimating the additive genetic effect of the X chromosome. *Genet. Epidemiol.* **29**, 377–388 (2005).
42. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
43. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

