nature genetics

Leveraging population admixture to characterize the heritability of complex traits

Noah Zaitlen[1], Bogdan Pasaniuc[2], Sriram Sankararaman[3,4], Gaurav Bhatia[3,5,6], Jianqi Zhang[7], Alexander Gusev[3,5,6], Taylor Young[3], Arti Tandon[3,4], Samuela Pollack[3,5,6], Bjarni J Vilhjálmsson[3,5,6], Themistocles L Assimes[8], Sonja I Berndt[9], William J Blot[10–12], Stephen Chanock[9], Nora Franceschini[13], Phyllis G Goodman[14], Jing He[7], Anselm J M Hennis[15–18], Ann Hsing[19,20], Sue A Ingles[7], William Isaacs[21], Rick A Kittles[22], Eric A Klein[23], Leslie A Lange[24], Barbara Nemesure[15], Nick Patterson[3], David Reich[3,4,25], Benjamin A Rybicki[26], Janet L Stanford[27], Victoria L Stevens[28], Sara S Strom[29], Eric A Whitsel[13], John S Witte[30], Jianfeng Xu[31], Christopher Haiman[7,32], James G Wilson[33], Charles Kooperberg[27], Daniel Stram[7], Alex P Reiner[34], Hua Tang[35,36] & Alkes L Price[3,5,6,36]

Despite recent progress on estimating the heritability explained by genotyped SNPs (h_g^2), a large gap between h_g^2 and estimates of total narrow-sense heritability (h^2) remains. Explanations for this gap include rare variants or upward bias in family-based estimates of h^2 due to shared environment or epistasis. We estimate h^2 from unrelated individuals in admixed populations by first estimating the heritability explained by local ancestry (h_γ^2). We show that $h_\gamma^2 = 2F_{STC}\theta(1-\theta)h^2$, where F_{STC} measures frequency differences between populations at causal loci and θ is the genome-wide ancestry proportion. Our approach is not susceptible to biases caused by epistasis or shared environment. We applied this approach to the analysis of 13 phenotypes in 21,497 African-American individuals from 3 cohorts. For height and body mass index (BMI), we obtained h^2 estimates of 0.55 ± 0.09 and 0.23 ± 0.06, respectively, which are larger than estimates of h_g^2 in these and other data but smaller than family-based estimates of h^2.

Understanding the genetic architecture of complex human phenotypes is a fundamental question in the field of genetics, with broad implications for identifying genes related to disease and predicting individual risk profiles[1–6]. A central element of this problem is estimating narrow-sense heritability (h^2), the fraction of phenotypic variation in a population determined by genetic variation under an additive model[7]. While the last decade of genome-wide association studies (GWAS) identified thousands of loci associated with hundreds of phenotypes[8], the sum of their effects (h_{GWAS}^2) explains only a small fraction of the estimated heritability for most phenotypes[5]. The gap between h_{GWAS}^2 and h^2 has been called the 'missing heritability', and several explanations for this difference have been posited, including upward bias in estimates of h^2 (refs. 4,9). The objective of this work is to develop a method for estimating h^2 (defined in the Online Methods) that (i) does not require closely related individuals, (ii) can be applied to both quantitative and case-control phenotypes, and (iii) is able to localize narrow-sense heritability to individual chromosomes or other genomic segments.

Current approaches to heritability estimation proceed by phenotyping many closely related individuals with a known genetic relationship, such as monozygotic and dizygotic twins[7]. Yang *et al.*[10] avoided the use of related individuals by applying linear mixed models to estimate the heritability explained by genotyped SNPs (h_g^2). h_g^2 corresponds to the fraction of phenotypic variation that could be captured by h_{GWAS}^2 under an additive model if GWAS sample sizes were infinitely large. Although current estimates of h_g^2 are often much larger than h_{GWAS}^2, they are typically only slightly more than half of h^2 (ref. 11). One reason h_g^2 is less than h^2 is because h_g^2 does not include the contribution of variants poorly tagged by the genotyping platform, such as rare variants. Another reason for the difference in the heritability estimates is that existing methods for estimating h^2 can be biased[12,13], as they rely on the use of related individuals. As a result, epistatic interactions between SNPs, gene environment interactions and the shared environmental factors of related individuals can all lead to inflated estimates of h^2 (refs. 12,13). We recently showed that, by jointly using related and unrelated individuals, it is possible to obtain less biased estimates of h^2 (ref. 11). However, the joint fit will still lead to inflated estimates of h^2 in the presence of shared environment[11] and cannot be applied to case-control phenotypes.

In this work, we propose a new approach for estimating h^2, which takes as input the phenotypes and genotypes of admixed individuals such as African Americans. We show via analytical derivation as well as extensive simulation over both simulated and real genotype data that heritability explained by local ancestry (h_γ^2) is related to the total narrow-sense heritability h^2 via the equation $h_\gamma^2 = 2F_{STC}\theta(1-\theta)h^2$, where F_{STC} is a specific measure of weighted allele frequency

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

Received 9 June; accepted 10 October; published online 10 November 2014; doi:10.1038/ng.3139

differences between ancestral populations at causal loci (Online Methods) and θ is the fraction of European ancestry[14,15]. Because our approach does not use closely related individuals, it is free from bias due to epistasis, gene environment interactions and shared environment effects. Unlike previous work in which h^2 estimates could not be obtained for case-control phenotypes[11], our current approach can obtain estimates of h^2 for both quantitative and case-control phenotypes, achieving goals (i) and (ii). Furthermore, unlike previous methods that provide genome-wide estimates, we are able to estimate h^2 for a particular genomic region, such as a chromosome, thereby achieving goal (iii). Our approach can be applied to all existing and future GWAS of admixed populations, without requiring additional expensive and time-consuming collections of large numbers of monozygotic and dizygotic twins.

We applied this approach to 21,497 African Americans from the NHLBI CARe (Candidate Gene Association Resource), WHI SHARe (SNP Health Association Resource) and AAPC (African-American Prostate Cancer) projects, analyzing 12 quantitative phenotypes and 1 case-control phenotype. For height and BMI, we obtained h^2 estimates of 0.55 ± 0.09 and 0.23 ± 0.06, respectively, which are larger than estimates of h_g^2 in these and other data sets but smaller than twin-based estimates of h^2, consistent with inflation in twin-based estimates due to shared environment or epistasis. We also estimated the heritability of height for each chromosome and found a significant correlation between chromosome length and heritability ($P < 0.003$).

RESULTS

Overview of the method

We consider three approaches to estimating heritability for a phenotype with a narrow-sense heritability of 80%. First, the classic approach to estimating heritability is to divide the phenotypic covariance of related individuals by the fraction of the genome for which they share identity by descent (IBD)[13]. In this instance, the phenotypic covariance of pairs of related individuals will be 0.80 times the fraction of the genome sharing IBD (**Fig. 1a**). The second approach, developed by Yang *et al.*[10], is to estimate the genetic relationship of unrelated individuals over genotyped SNPs and apply a linear mixed model with the genetic relationship matrix to estimate phenotype. To illustrate this approach, we simulated 2 million independent pairs of individuals, regressing their normalized genetic similarity over the product of their normalized phenotypes, which yielded a regression coefficient of 0.79 ± 0.014 (**Fig. 1b**). This Haseman-Elston regression[16] shows how the genetic similarity of unrelated individuals can be used to estimate the heritability of genotyped SNPs (h_g^2). In general, the heritability explained by genotyped SNPs is less than the total narrow-sense heritability (h^2) because phenotypic variation determined by poorly tagged SNPs such as rare variants will not be captured[10].

The approach used in this work is similar to that of Yang *et al.*[10], but, instead of using genotypes to estimate genetic similarity, we use the number of copies of local ancestry in an admixed population. A crucial element of our approach is that the phenotypic variation described by variation in local ancestry (h_γ^2) is a function of all causal variation, not just that tagged by SNPs on a particular genotyping platform. This is because local ancestry tags both common and rare variation. To illustrate this approach, we simulated 4 million unrelated admixed individuals from ancestral populations with genetic distance $F_{STC} = 0.08$ and an equal proportion of ancestry from each ancestral population $\theta = 0.5$ (Online Methods). Applying Haseman-Elston regression to regress the product of normalized phenotypes against the genetic similarity of local ancestry, we observed a regression coefficient of $0.033 \pm 0.007 \approx 2F_{STC}\theta(1-\theta)h^2 = 0.032$, corresponding to $h^2 = 0.83$ (standard error (s.e.) = 0.18) (**Fig. 1c**). The Haseman-Elston regression used in generating these plots is for illustrative purposes (as in Fig. 3 of ref. 10). In practice, we used a mixed-model approach because of its lower standard errors[10].

We first generated a local ancestry–based kinship matrix K_γ, which was constructed similarly to the genotype-based kinship matrix K in previous methods[10] but with local ancestry substituted for genotypes at each SNP. We used a variance components approach to estimate the phenotypic variance explained by variation in local ancestry (σ_γ^2) and the residual phenotypic variance (σ_ε^2) (refs. 10,17). We included the genome-wide ancestry proportion θ and the top five principal components as fixed effects when fitting the mixed model (Online Methods). The heritability explained by local ancestry was given by:

$$h_\gamma^2 = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2}$$

Finally, to estimate h^2, we used the formula $h_\gamma^2 = 2F_{STC}\theta(1-\theta)h^2$, where F_{STC} is a specific measure of weighted allele frequency differences between ancestral populations at causal loci (Online Methods). For dichotomous phenotypes, we applied the same approach but converted the observed scale estimates to a liability scale estimate of heritability using ref. 18 and published disease prevalence in African Americans. In our previous work[11], this conversion was not possible because non-randomly ascertained individuals in multiple relatedness classes (for example, sibling, first cousin and avuncular) were studied, and there is currently no method for accounting for ascertainment in such complex pedigrees. A complete description of the approach, along with an analytical derivation, is given in the Online Methods.

Simulations with simulated genotypes

We first verified the analytical derivations and examined the properties of the approach under a simple simulation framework. We simulated the genotypes and local ancestry of 4,000 unrelated diploid individuals at 1,000 SNPs from a 2-way admixed population with causal variant genetic distance F_{STC} and either normally or uniformly distributed ancestry proportion θ. Each local ancestry segment contained exactly one SNP, and all segments were generated independently. Phenotypes were simulated under an additive model with heritability h^2 in which a proportion r of the 1,000 SNPs were

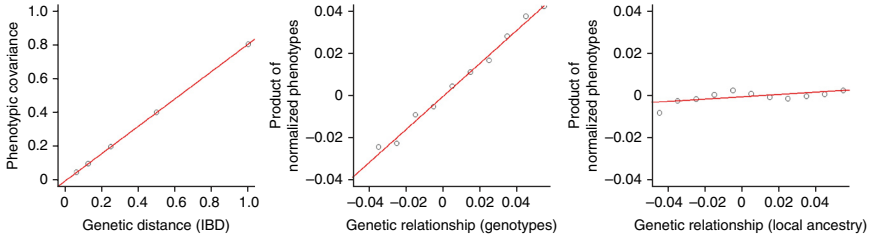

Figure 1 Relationships between genetic distance and phenotype for a trait with heritability of 80%. (**a**) The phenotypic covariance of pairs of individuals at different expected fractions of genomic shared IBD is 0.8 × percent IBD. (**b**) Regression of genetic distance estimated from genetic variation against the product phenotypes normalized to have a mean of 0 and variance of 1 has a coefficient of 0.79 (s.e. = 0.014). (**c**) Regression of genetic distance estimated from local ancestry variation against normalized phenotypes has a coefficient of 0.033 (s.e. = 0.007) ≈ $2F_{STC}\theta(1-\theta)$ $h^2 = 0.032$, corresponding to $h^2 = 0.83$ (s.e. = 0.18).

Table 1 Results of local ancestry–based heritability estimation from simulated genotypes and phenotypes

h^2	F_{ST}	r	$\hat{h}^2$ (s.e.)
0.8	0.30	1.0	0.802 (0.003)
0.8	0.30	0.1	0.802 (0.005)
0.8	0.15	1.0	0.800 (0.005)
0.8	0.15	0.1	0.804 (0.006)

Mean heritability estimates and standard errors are reported from 2,000 simulations for each choice of parameters.

causal (Online Methods). We applied our method to estimate heritability over a range of values of F_{STC}, θ, r and h^2. For each parameter setting, we estimated heritability from 2,000 independent simulated data sets. The results, shown in **Table 1**, demonstrate that our heritability estimates were accurate across a range of parameter settings, confirming our analytical derivation. Results for additional parameter settings are shown in **Supplementary Table 1**.

The results also demonstrate the relationship between h^2_γ and the parameters F_{STC}, θ and h^2. For a fixed value of r, phenotypes with a larger h^2 value will have larger genetic effects, resulting in larger h^2_γ. When ancestral populations are genetically distant (larger F_{STC}), variants are more likely to have different frequencies in the ancestral populations, resulting in a concomitant increase in h^2_γ. Increasing the variance of θ results in larger standard errors around the heritability estimates.

Simulations with real genotypes

We made several simplifying assumptions in the above simulations that do not hold in real data sets. These included a single SNP per ancestry block, no genotyping error, no local ancestry inference error, no linkage disequilibrium, a normal or uniform distribution of the ancestry proportion, continuous phenotypes, and the use of identical effect size distributions for common and rare variants in computing F_{STC}. To address these complexities, we took the approach of using real genotypes and simulating phenotypes. We simulated continuous and case-control phenotypes over 5,129 individuals (excluding close relatives) from the CARe cohort (Online Methods). Although phenotypes were generated from SNPs sampled across all genotyped SNPs, we only used local ancestry information from every fifth SNP.

We tried a range of parameters for h^2. Instead of simulating phenotypes under an infinitesimal model, we sampled a proportion of causal variants r. We could not alter the ancestry proportion θ, as this was fixed in the real data set. However, we altered the effect size distribution of SNPs according to their F_{STC} values.

The data did not contain a sufficient number of genotyped variants that were rare in both ancestral populations to simulate rare versus common effects. Instead, we examined SNPs common in both populations ('common') versus SNPs rare in at least one population ('uncommon'). Only common variants were used in constructing the kinship matrix, and uncommon variants would thus only contribute to h^2_g via linkage disequilibrium. The common SNPs had $F_{STC} = 0.15$, whereas the uncommon SNPs had $F_{STC} = 0.25$. We simulated phenotypes using a different proportion of phenotypic variance derived from uncommon variants (α). When α was not 0, the variant and causal variant frequencies in the kinship matrix were different. Simulations involving a large proportion of causal variants not included in the kinship matrix (high α) had h^2_g less than h^2 because the common variants did not completely capture the phenotypic variance driven by the uncommon variants (**Table 2**). The parameter α also determined study-wide F_{STC} according to the formula $F_{STC} = (0.15(1 - \alpha) + 0.25\alpha)$ (Online Methods). The results shown in **Table 2** use the correct value of α, and, hence, the estimates of h^2 are unbiased. However, if we incorrectly assumed that α was 0 when this was not the case, then h^2 would be biased by a factor of $(0.15(1 - \alpha) + 0.25\alpha)/0.15$. We describe this (and other potential sources of bias) in detail in the Discussion.

Setting individuals with the lowest P percent of phenotypes as cases and all others as controls, we generated dichotomous phenotypes with prevalence P. The small number of individuals prevented simulation of case-control ascertainment, which might produce a downward bias in h^2 estimates for low-prevalence diseases in very large studies (see Supplementary Table 9 of ref. 19). Such bias is expected to be small in the prostate cancer data analyzed here because of the high prevalence of prostate cancer and the moderate sample size. For large sample sizes, replacing mixed model–based estimates with Haseman-Elston regression estimates will alleviate the issue of ascertainment bias[20].

The results in **Table 2** also demonstrate that complexities such as genotyping error, linkage disequilibrium or error in local ancestry inference in African Americans do not introduce bias into the heritability estimates when phenotypes are generated under a non-infinitesimal mixture model. This might not be the case for other admixed populations such as Latinos[21].

Application to WHI, CARe and AAPC cohorts

We applied our method to 21,497 African-American individuals from the WHI, CARe and AAPC cohorts over a total of 12 quantitative phenotypes and 1 case-control phenotype (Online Methods). Local ancestry was inferred using the HAPMIX, SABER+ and RFMix methods, which are extremely accurate in African Americans ($r^2 = 0.98$ or greater)[22–24]. For each phenotype, we estimated h^2_g, h^2_γ and, by extension, h^2. For h^2_g and h^2_γ, we used the GCTA software package applied to the genotypes and local ancestry at each SNP, respectively[17]. For the phenotypes measured in both cohorts, we computed the inverse variance–weighted mean and standard error. For each phenotype, we also list previously published estimates of heritability from family-based studies using twins and African-American estimates, where available (h^2_{pub}). The results are shown in **Table 3**, and published African-American estimates are marked for reference. Estimates from European populations may not be directly comparable if the genetic or environmental bases for the phenotype differ substantially.

Several phenotypes, including height, BMI, high-density lipoprotein–cholesterol (HDL) levels, triglyceride levels, prostate cancer and white blood cell (WBC) count (conditioned on ancestry at the Duffy antigen locus *FY*), had h^2 estimates lower than

Table 2 Results of heritability simulations over 5,129 African-American individuals from the CARe cohort

h^2	r	P	α	$\hat{h}^2_g$ (s.e.)	$\hat{h}^2$ (s.e.)
0.8	0.01	NA	0.0	0.797 (0.001)	0.800 (0.004)
0.8	0.001	NA	0.0	0.801 (0.002)	0.793 (0.005)
0.5	0.01	NA	0.0	0.499 (0.001)	0.498 (0.003)
0.5	0.001	NA	0.0	0.499 (0.001)	0.501 (0.004)
0.8	0.01	NA	0.25	0.689 (0.003)	0.802 (0.004)
0.8	0.01	0.2	0.25	0.691 (0.002)	0.782 (0.005)
0.8	0.01	0.5	0.25	0.703 (0.003)	0.800 (0.006)
0.8	0.01	NA	0.50	0.625 (0.002)	0.805 (0.005)
0.8	0.01	0.5	0.50	0.637 (0.003)	0.797 (0.006)
0.8	0.01	NA	1.0	0.473 (0.002)	0.796 (0.005)
0.8	0.01	0.5	1.0	0.498 (0.003)	0.792 (0.007)

Average estimates and standard errors are reported for heritability explained by genotyped SNPs ($\hat{h}^2_g$) and local ancestry–based heritability explained by all SNPs ($\hat{h}^2$) from 2,500 simulations for representative choices of 4 parameters: true heritability (h^2), the proportion of causal variants (r), the prevalence (P) (NA for continuous phenotypes) and the proportion of heritability from uncommon variants (α).

Table 3 Heritability estimates for phenotypes from 21,497 African Americans in the WHI, CARe and AAPC cohorts

Heritability explained by genotyped SNPs (h^2_g) in WHI and CARe

Phenotype	WHI $\hat{h}^2_g$ (s.e.)	CARe $\hat{h}^2_g$ (s.e.)	Meta $\hat{h}^2_g$ (s.e.)[a]	$h^2_{g,\,pub}$[b]
Height	0.461 (0.058)	0.378 (0.029)	0.395 (0.026)	0.45 (ref. 10)
BMI	0.198 (0.055)	0.078 (0.065)	0.148 (0.042)	0.14 (ref. 38)
log(HDL)	0.316 (0.057)	0.224 (0.066)	0.277 (0.043)	0.12 (ref. 38)
LDL	0.294 (0.056)	0.156 (0.067)	0.238 (0.043)	0.10 (ref. 11)
WBC	0.725 (0.051)	0.848 (0.091)	0.755 (0.044)	0.2 (ref. 11)
WBC|FY	0.188 (0.054)	0.167 (0.097)	0.183 (0.047)	NA
log(TG)	0.226 (0.056)	NA	NA	NA
Glucose	0.16 (0.063)	NA	NA	0.10 (ref. 38)
log(insulin)	0.086 (0.051)	NA	NA	0.09 (ref. 38)
QT interval	0.251 (0.098)	NA	NA	NA
log(CRP)	0.295 (0.056)	NA	NA	NA
DBP	0.148 (0.053)	0.170 (0.066)	0.157 (0.041)	NA
SBP	0.162 (0.054)	0.189 (0.066)	0.173 (0.042)	0.24 (ref. 38)

Total narrow-sense heritability ($\hat{h}^2$) as derived from $\hat{h}^2_\gamma$ in WHI and CARe

Phenotype	WHI $\hat{h}^2$ (s.e.)	CARe $\hat{h}^2_g$ (s.e.)	Meta $\hat{h}^2_g$ (s.e.)[a]	$h^2_{g,\,pub}$[b]
Height	0.611 (0.135)	0.503 (0.120)	0.550 (0.090)	0.77 (ref. 39)*
BMI	0.252 (0.097)	0.208 (0.085)	0.227 (0.064)	0.47 (ref. 39)*
log(HDL)	0.418 (0.117)	0.470 (0.146)	0.438 (0.091)	0.52 (ref. 39)*
LDL	0.395 (0.116)	0.333 (0.140)	0.370 (0.089)	0.53 (ref. 39)*
WBC	3.267 (0.322)	3.703 (0.447)	3.415 (0.261)	0.48 (ref. 40)*
WBC|FY	0.172 (0.084)	0.247 (0.166)	0.187 (0.075)	NA
log(TG)	0.225 (0.094)	NA	NA	0.40 (ref. 39)*
Glucose	0.104 (0.087)	NA	NA	0.29 (ref. 41)*
log(insulin)	0.105 (0.077)	NA	NA	0.28 (ref. 41)*
QT Interval	0.336 (0.164)	NA	NA	0.41 (ref. 42)*
log(CRP)	0.542 (0.139)	NA	NA	0.56 (ref. 43)*
DBP	0.179 (0.088)	0.238 (0.119)	0.200 (0.071)	0.13 (ref. 39)*
SBP	0.187 (0.092)	0.233 (0.117)	0.205 (0.072)	0.17 (ref. 39)*

Complete AAPC results

Phenotype	AAPC $\hat{h}^2_g$ (s.e.)	$h^2_{g,\,pub}$[b]	AAPC $\hat{h}^2$ (s.e.)	h^2_{pub}[c]
Prostate cancer	0.182 (0.040)	NA	0.328 (0.093)	0.58 (ref. 44)
PCI8q24	0.174 (0.040)	NA	0.315 (0.092)	NA

BMI, body mass index; log(HDL), log-transformed high-density lipoprotein–cholesterol levels; LDL, low-density lipoprotein–cholesterol levels; WBC, white blood cell count; log(TG), log-transformed triglyceride levels; log(insulin), log-transformed insulin levels; log(CRP), log-transformed C-reactive protein levels; DBP, diastolic blood pressure; SBP, systolic blood pressure; PC, prostate cancer; NA, not available.

[a]Estimates from inverse variance–weighted meta-analysis for phenotypes analyzed in both the WHI and CARe data sets. [b]Previously published estimate of $\hat{h}^2_g$. [c]Previously published estimate of h^2 from family-based studies. Published heritability studies of African Americans are indicated with an asterisk.

family-based estimates. These differences could be due to phenotype-specific effects of epistasis, gene environment interaction and/or shared environmental factors that can inflate family-based estimates[12,13]. In our recent work using an extended genealogy inclusive of more distantly related individuals, we also found height and BMI estimates lower than previous heritability estimates, providing further evidence of inflation[11]. The lower estimates could also reflect a difference in the heritability between African Americans and the previous study populations. There were no statistically significant differences in h^2 estimates between the cohorts.

Yang *et al.* proposed an adjustment to account for the incomplete coverage of genotyping platforms[10]. We applied this approach in the CARe data (**Supplementary Table 2**) and observed an increase in h^2_g of less than 1% in all phenotypes. We included the genome-wide ancestry proportion as a fixed effect in our mixed model. If an environmental factor exists that affects phenotype and is correlated with ancestry, our heritability estimates would discount this environmental effect, leading to higher estimates of heritability. Specifically, our method would remove the variance of the environmental factor that could be explained by ancestry from the environmental component (σ^2_e) in the denominator of the heritability estimate (Online Methods).

$$h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e}$$

Differences between our heritability estimates and those of previous studies could also be due to differences between the value of F_{STC} we used in this study and the true value of F_{STC} for the phenotype in question. Considering recent evidence that rare variants are unlikely to contribute to a large proportion of phenotypic variation[25,26], we computed an F_{STC} of 0.182 over the common variants (minor allele frequency (MAF) > 5%) in African Americans. However, this estimate dropped to 0.165 for low-frequency variants (MAF < 5%) and 0.054 for rare variants (MAF < 1%). Estimates of heritability assuming a rare variants–only phenotype model would be more than three times as large as from a common variants–only phenotype model. Therefore, if rare variants contribute substantially to phenotypic variation or if balancing or negative selection constrains the genetic distance at causal variants, then our estimates of heritability would be biased downward.

Positive selection acting at causal variants could induce such a bias in F_{STC}, and we included WBC count as a positive control for this type of bias. A SNP in the *FY* gene (*DARC*, Duffy null allele) is highly differentiated between the CEU (Utah residents of Northern and Western European ancestry) and YRI (Yoruba from Ibadan, Nigeria) populations, likely owing to its protective effect against *Plasmodium vivax* malaria[27]. It is also a SNP of large effect size for WBC count[28]. Therefore, the average F_{STC} at causal variants for WBC count was much higher than the 0.182 value estimated from common variants (Online Methods). The h^2 estimate of WBC count was 3.42 owing to the effect of this positive selective pressure. Ancestry at the *FY* locus accounts for ~20% of the phenotypic variation in WBC count[28]. By including ancestry at *FY* as a fixed effect (WBC|FY), we obtained an h^2 estimate of 0.19, which is lower than the published estimate of 0.48.

We performed a sensitivity analysis to assess whether this type of bias is likely to be problematic. Because strong positive selection is unusual[29], we considered a single locus under positive selection. We estimated bias as a function of F_{STC} at the locus and the variance explained by the locus. The results showed that only for extreme values of both locus F_{STC} and heritability will there be significant bias in heritability due to positive selection (**Supplementary Table 3**). As an example, we considered the 8q24 locus in prostate cancer, which contains causal SNPs that are highly differentiated between African and European ancestors, producing an admixture mapping peak[30]. However, because this locus explains less than 2% of the heritability of prostate cancer, even exceedingly strong population differentiation at the locus would not substantially bias our overall results.

Table 4 Number of individuals for each phenotype in the CARe and WHI data sets

Phenotype	WHI	CARe
Height	8,109	5,024
BMI	8,153	5,026
log(HDL)	8,014	4,928
LDL	7,979	4,794
WBC	8,035	3,367
WBCIFY	8,035	3,367
log(TG)	8,015	NA
Glucose	6,826	NA
log(insulin)	7,749	NA
QT Interval	4,143	NA
log(CRP)	8,014	NA
DBP	8,153	5,030
SBP	8,153	5,029

The AAPC data set contained 4,207 prostate cancer cases and 4,008 controls.

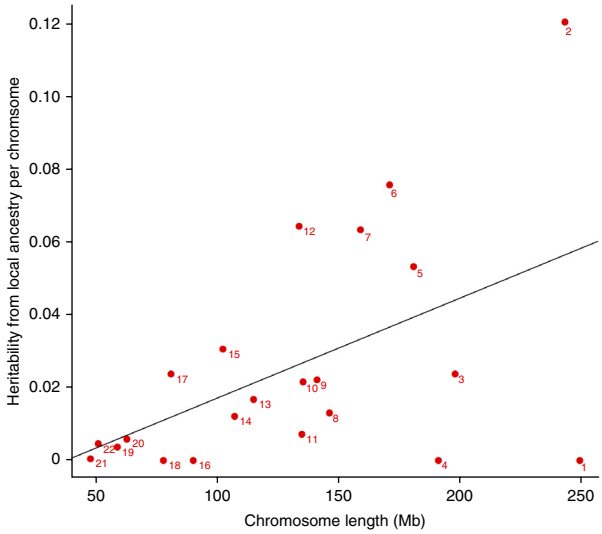

Figure 2 Estimated heritability of height for each chromosome in the CARe data set. The number adjacent to each point represents the chromosome number. We plot the regression line of h^2 per chromosome regressed on chromosome length. We find a strong correlation between chromosome length and height heritability (Pearson correlation = 0.513, weighted P value = 0.0028).

Partitioning heritability across the genome

Our method is also capable of estimating the total narrow-sense heritability attributable to a particular genomic region. This is accomplished by constructing the kinship matrix using just the ancestry segments in the region of interest and applying the variance component model to the phenotype of interest using the region-specific kinship matrix (Online Methods). We partitioned the heritability for each of the phenotypes from the CARe data set across each of the chromosomes[31]. We applied weighted linear regression to determine the relationship between heritability and chromosome length (Online Methods). The results for height are presented in **Figure 2**, and the full results are provided in **Supplementary Table 4**. We found a strong correlation between chromosome length and heritability of height (Pearson corelation r = 0.513, weighted P value = 0.0028). Log-transformed HDL levels, BMI and systolic blood pressure (SBP) also produced significant results (weighted P value < 0.03, 0.02 and 0.02, respectively). Other phenotypes had standard errors too large to yield meaningful results. To address this, we averaged the heritability from each chromosome across all phenotypes (using WBC|FY instead of WBC), observing a significant correlation between chromosome length and mean chromosomal heritability (Pearson corelation r = 0.686, weighted P value < 0.0002).

DISCUSSION

We developed a method for estimating narrow-sense heritability from unrelated individuals by leveraging the two ancestral genomes in recently admixed populations, such as African Americans. We used a population genetic approach to derive the relationship between heritability and variation in local ancestry in admixed populations. Theory and simulations confirm that, under an infinitesimal phenotypic model, our approach produces unbiased estimates of heritability. Because the individuals are distantly related, our approach will not produce heritability estimates inflated by epistasis, gene environment interactions or shared environmental effects.

Our method is also able to partition total narrow-sense heritability (h^2) along genomic segments such as chromosomes, as we have shown by application to the phenotypes in the CARe data set. This feature is distinct from recent work that instead partitioned the heritability explained by genotyped SNPs (h^2_g) across chromosomes[31–33]. Although a previous method also partitioned h^2 along chromosomes[34,35], it relied on the use of siblings, leading to very large standard errors, and was limited by the coarseness of shared IBD segments (which extend for tens of megabases). Our approach is limited by the coarseness of local ancestry segments (which extend for megabases) and thus cannot be applied at the level of individual genes.

In this study, we applied our method to an African-American population. Application to more complex admixed populations such as Latinos will have to account for reduced accuracy in local ancestry inference[21] to avoid downward bias. Restricting to two ancestry categories (for example, Native American versus non–Native American ancestry)[36] is one approach to handle multi-way admixture, but it may be possible to extend our derivation to multi-way admixture. There is evidence that African Americans have a small proportion of admixture from Native American populations (0.5%)[24], but this very small proportion is unlikely to significantly change our results. Substantial errors in the assumed population genetic structure would perturb the F_{STC} and θ values, and resulting h^2 estimates would be biased in proportion to these errors. Application to sex chromosomes can be adapted from the approach taken in ref. 31, but these analyses must be performed separately because of differences in the admixture proportion of European ancestry on autosomes and sex chromosomes.

In our previous work, we found that heritability estimates from related individuals followed a pattern consistent with biases due to shared environment[11]. In this work, we found that a linear additive model, implicitly including both rare and common variants, typically explained less phenotypic variation than was predicted in family-based studies. These new estimates of narrow-sense heritability are less susceptible to bias and provide additional evidence that family-based estimates are inflated. Unlike in ref. 11, we were able to obtain estimates for both quantitative and case-control traits. We also found that chip-based additive models explained less phenotypic variation than our estimates. In the phenotypes common to CARe and WHI for which meta-analysis was performed, the averages of these estimates were 24.7% and 31.1%, respectively. Rare variants and poorly tagged common variants are the most likely explanation for the difference between these two estimates. We discuss other possible explanations below.

Our method does produce biased estimates when model assumptions are violated. Specifically, if the genetic distance we estimated

over common variants (0.182; Online Methods) differs from the distribution over causal variants, our method can result in either inflated or deflated estimates. If selection acted on the causal variants, their F_{STC} values could be higher or lower, depending on the direction of selection. In the case of positive selection in one of the ancestral groups but not the other, the true value of F_{STC} would be larger than our genome-wide estimate and our h^2 estimate would therefore be inflated. For example, estimates for WBC count were larger than h^2_{pub} owing to strong selective pressure at the Duffy locus[27,37]. However, strong positive selection is believed to be rare in recent human evolution[29]. If a large proportion of phenotypic variance is due to rare variants, then incorrect estimates of F_{STC} might induce bias. However, previous reports suggest that rare variation explains a small proportion of total heritability[25,26].

The application of our approach to two large cohorts of African Americans showed a difference between previously published family-based estimates of the heritability of height and BMI and our estimates. This disparity suggests that there is substantial contribution from non-additive genetic effects or shared environmental effects that differ between monozygotic and dizygotic twins. The future application of our method to large-scale studies of African Americans will provide a mechanism of estimating the total narrow-sense heritability of phenotypes as well as determining the genetic architecture of complex phenotypes.

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 US National Institutes of Health grants R01 HG006399, R01 GM073059, 1K25HL121295-01A1 and R21 ES020754. The WHI program is funded by the National Heart, Lung, and Blood Institute, US National Institutes of Health, US Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C and HHSN271201100004C.

AUTHOR CONTRIBUTIONS

N.Z., B.P., S.S., G.B., A.G., B.J.V., C.H., J.G.W., C.K., D.S., A.P.R., H.T. and A.L.P. designed experiments. N.Z., J.Z., T.Y., A.T., S.P., H.T. and A.L.P. performed experiments. N.Z., S.S., C.H., J.G.W., C.K., D.S., A.P.R., H.T. and A.L.P. wrote the text. T.L.A., S.I.B., W.J.B., S.C., N.F., P.J.G., J.H., A.J.M.H., A.H., S.A.I., W.I., R.A.K., E.A.K., L.A.L., B.N., N.P., D.R., B.A.R., J.L.S., V.L.S., S.S.S., E.A.W., J.S.W. and J.X. provided data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Wray, N.R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
2. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
3. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Hum. Genet.* **131**, 1655–1664 (2012).
4. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
5. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
6. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).
7. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
8. Hindorff, 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).
9. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2011).
10. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
11. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
12. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA* **109**, 1193–1198 (2012).
13. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).
14. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
15. Bhatia, G. *et al.* Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet.* **89**, 368–381 (2011).
16. Sham, P.C. & Purcell, S. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* **68**, 1527–1532 (2001).
17. 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).
18. 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).
19. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
20. Golan, D. & Rosset, S. Narrowing the gap on heritability of common disease by direct estimation in case-control GWAS. http://arxiv.org/abs/1305.5363 (2013).
21. Pasaniuc, B. *et al.* Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* **29**, 1407–1415 (2013).
22. Price, A.L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
23. Johnson, N.A. *et al.* Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet.* **7**, e1002410 (2011).
24. Maples, B.K., Gravel, S., Kenny, E.E. & Bustamante, C.D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
25. 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).
26. Morrison, A.C. *et al.* Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat. Genet.* **45**, 899–901 (2013).
27. Hamblin, M.T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66**, 1669–1679 (2000).
28. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).
29. Hernandez, R.D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
30. Freedman, M.L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. USA* **103**, 14068–14073 (2006).
31. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
32. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
33. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
34. 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).
35. 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).
36. Price, A.L. *et al.* A genomewide admixture map for Latino populations. *Am. J. Hum. Genet.* **80**, 1024–1036 (2007).
37. Nalls, M.A. *et al.* Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am. J. Hum. Genet.* **82**, 81–87 (2008).
38. Vattikuti, S., Guo, J. & Chow, C.C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* **8**, e1002637 (2012).
39. Wilson, J.G. *et al.* Study design for genetic analysis in the Jackson Heart Study. *Ethn. Dis.* **15**, S6-30–S6-37 (2005).
40. Reiner, A.P. *et al.* Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet.* **7**, e1002108 (2011).
41. Freedman, B.I. *et al.* Genome-wide scans for heritability of fasting serum insulin and glucose concentrations in hypertensive families. *Diabetologia* **48**, 661–668 (2005).
42. Akylbekova, E.L. *et al.* Clinical correlates and heritability of QT interval duration in blacks: the Jackson Heart Study. *Circ Arrhythm Electrophysiol* **2**, 427–432 (2009).
43. Fox, E.R. *et al.* Epidemiology, heritability, and genetic linkage of C-reactive protein in African Americans (from the Jackson Heart Study). *Am. J. Cardiol.* **102**, 835–841 (2008).
44. Hjelmborg, J.B. *et al.* The heritability of prostate cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiol. Biomarkers Prev.* doi:10.1158/1055-9965.EPI-13-0568 (8 May 2014).

[1]Department of Medicine, University of California, San Francisco, San Francisco, California, USA. [2]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, California, USA. [3]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [6]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [7]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [8]Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. [9]Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA. [10]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [11]Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [12]International Epidemiology Institute, Rockville, Maryland, USA. [13]Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, USA. [14]SWOG Statistical Center, Seattle, Washington, USA. [15]Department of Preventive Medicine, Stony Brook University, Stony Brook, New York, USA. [16]Chronic Disease Research Centre, University of the West Indies, Bridgetown, Barbados. [17]Faculty of Medical Sciences, University of the West Indies, Bridgetown, Barbados. [18]Ministry of Health, Bridgetown, Barbados. [19]Cancer Prevention Institute of California, Fremont, California, USA. [20]Division of Epidemiology, Stanford University School of Medicine, Stanford, California, USA. [21]James Buchanan Brady Urological Institute, Johns Hopkins Hospital and Medical Institutions, Baltimore, Maryland, USA. [22]Department of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA. [23]Glickman Urologic and Kidney Institute, Cleveland Clinic, Cleveland, Ohio, USA. [24]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [25]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts, USA. [26]Department of Public Health Sciences, Henry Ford Hospital, Detroit, Michigan, USA. [27]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [28]Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA. [29]Department of Epidemiology, Division of Cancer Prevention and Population Sciences, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [30]Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA. [31]Center for Cancer Genomics, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA. [32]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA. [33]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, USA. [34]Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA. [35]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [36]These authors contributed equally to this work. Correspondence should be addressed to N.Z. (noah.zaitlen@ucsf.edu), H.T. (huatang@stanford.edu) or A.L.P. (aprice@hsph.harvard.edu).

ONLINE METHODS

Given a set of M admixed individuals with two ancestral populations (P_0 and P_1), let the local ancestry for individual i at SNP s, $\gamma_{i,s} \in \{0,1,2\}$, be the number of alleles inherited from a P_1 ancestor. We use a mixed-model approach to estimate h^2_γ, the contribution of variation in local ancestry to phenotypic variation for the phenotype $\mathbf{Y} = y_1, y_2, \ldots, y_M$. We first generate a local ancestry–based kinship matrix K_γ, which is constructed similarly to the genotype-based kinship matrix K but with local ancestry substituted for genotypes at each SNP. We then find the parameters σ^2_γ and σ^2_ε that maximize the likelihood of the mixed model $Y \sim N(0, K_\gamma\sigma^2_\gamma + I\sigma^2_\varepsilon)$. The heritability explained by local ancestry is given by h^2_γ. Finally, we use the formula $h^2_\gamma = h^2 2F_{\mathrm{STC}}\theta(1-\theta)$ to estimate h^2.

Definition of h^2. Heritability is the ratio of genetic variance to the sum of genetic and environmental variance:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e}$$

In this case, we are defining these elements with respect to an admixed population. For a given phenotype, both σ^2_g and σ^2_e can vary between the ancestral European and African populations. For example, σ^2_g will vary with ancestry if the MAF at causal variants is systematically larger in one of the two populations. It is also possible for ancestry to be associated with environmental factors. In this case, by conditioning on genome-wide ancestry, our method will remove the environmental variance that can be explained by ancestry and estimate the heritability of the component of phenotype that cannot be predicted by genome-wide ancestry, thereby increasing the heritability estimate.

Estimation of h^2_γ. We apply a variance components approach to determine the phenotypic variance described by local ancestry h^2_γ, using θ as a fixed effect to prevent confounding from environmental factors associated with ancestry. This method is equivalent to recent methods used to determine the phenotypic variance described by genotyped SNPs (h^2_g) but replaces genotypes with inferred local ancestry[10].

Derivation of the relationship between h^2 and h^2_γ. Let i denote (diploid) individuals and s denote index SNPs. Individual i is assigned global ancestry proportion θ_i from some distribution $F()$ with mean $E[\theta_i] = \theta$ and variance σ^2_θ. Given θ_i, individual i is assigned maternal and paternal local ancestries $\gamma_{i,s,M}$ and $\gamma_{i,s,P}$ at each SNP s (0 or 1 copy of European ancestry) from the Bernoulli distribution $\mathrm{Ber}(\theta_i)$. Given the local ancestries $\gamma_{i,s,M}$ and $\gamma_{i,s,P}$ and the allele frequencies $p_{s,0}$ and $p_{s,1}$ at SNP s in populations P_0 and P_1, individuals are assigned maternal genotypes of $g_{i,s,M} = \gamma_{i,s,M}Z_{i,s,1} + (1 - \gamma_{i,s,M})Z_{i,s,0}$, where $Z_{i,s,0} \sim \mathrm{Ber}(p_{s,0})$ and $Z_{i,s,1} \sim \mathrm{Ber}(p_{s,1})$, and paternal genotypes are assigned in a similar manner. The diploid genotype is $g_{i,s} = g_{i,s,P} + g_{i,s,M}$ (0, 1 or 2), and the diploid local ancestry is $\gamma_{i,s} = \gamma_{i,s,P} + \gamma_{i,s,M}$ (0, 1 or 2).

We define $E[g_{i,s}]$ as $\mu_{g,s}$ and $\mathrm{var}[g_{i,s}]$ as $\sigma^2_{g,s}$, and we define the normalized genotype as:

$$\bar{g}_{i,s} = \frac{g_{i,s} - \mu_{g,s}}{\sigma_{g,s}}$$

where:

$$\mu_{g,s} = 2(\mu p_{s,1} + (1-\mu)p_{s,0}) \quad (1)$$

$$\sigma^2_{g,s} = 2[\mu(1-\mu)(p_{s,1} - p_{s,0})^2 + (\mu p_{s,1}(1-p_{s,1}) + (1-\mu)p_{s,0}(1-p_{s,0})] \quad (2)$$

Similarly, we define $E[\gamma_{i,s}]$ as μ_γ and $\mathrm{var}[\gamma_{i,s}]$ as σ^2_γ, and we define the normalized local ancestry at each locus as:

$$\bar{\gamma}_{i,s} = \frac{\gamma_{i,s} - \mu_\gamma}{\sigma_\gamma}$$

where:

$$\mu_\gamma = 2\theta \quad (3)$$

$$\sigma^2_\gamma = 2\theta(1-\theta) \quad (4)$$

Although equation (4) might not be strictly true (for example, in a population where all individuals have one European parent and one African parent), it is approximately true for African Americans[22]. Furthermore, σ^2_γ can be estimated empirically, and we do so in this work. We model the phenotype of individual i as:

$$y_i = \sum_s \beta_s \bar{g}_{i,s} + \varepsilon_i \quad (5)$$

where $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, $\mathrm{var}[y_i] = 1$, $E[y_i] = 0$, the effect size of SNP s is β_s and $h^2 = \sum_s \beta^2_s$. By substitution and algebra, we get:

$$\begin{aligned}
\bar{g}_{i,s} &= \frac{g_{i,s} - \mu_{g,s}}{\sigma_{g,s}} = \frac{1}{\sigma_{g,s}}(\sigma_{g,s,M} + \sigma_{g,s,P} - 2\mu p_{s,1} - 2(1-\mu)p_{s,0}) \\
&= \frac{1}{\sigma_{g,s}}(\gamma_{i,s,M} - \theta)(Z_{i,s,1} - Z_{i,s,0}) + \frac{1}{\sigma_{g,s}}[\theta(Z_{i,s,1} - p_{s,1}) + (1-\theta)(Z_{i,s,0} - p_{s,0})] \\
&\quad + \frac{1}{\sigma_{g,s}}(\gamma_{i,s,P} - \theta)(Z_{i,s,1} - Z_{i,s,0}) + \frac{1}{\sigma_{g,s}}[\theta(Z_{i,s,1} - p_{s,1}) + (1-\theta)(Z_{i,s,0} - p_{s,0})] \\
&= \frac{\sigma_\gamma}{\sigma_{g,s}}\bar{\gamma}_{i,s}(Z_{i,s,1} - Z_{i,s,0}) + \frac{2}{\sigma_{g,s}}[\theta(Z_{i,s,1} - p_{s,1}) + (1-\theta)(Z_{i,s,0} - p_{s,0})] \quad (6)
\end{aligned}$$

Plugging into equation (5), we get:

$$\begin{aligned}
y_i &= \sum_s \beta_s \bar{g}_{i,s} + \varepsilon_i \\
&= \sum_s \beta_s \frac{\sigma_\gamma}{\sigma_{g,s}}\bar{\gamma}_{i,s}(Z_{i,s,1} - Z_{i,s,0}) \\
&\quad + \sum_s \beta_s \frac{2}{\sigma_{g,s}}\left[\theta(Z_{i,s,1} - p_{s,1}) + (1-\theta)(Z_{i,s,0} - p_{s,0})\right] + \varepsilon_i \\
&= \sum_s \beta_s \frac{\sigma_\gamma}{\sigma_{g,s}}\bar{\gamma}_{i,s}(Z_{i,s,1} - Z_{i,s,0}) + \delta_i \quad (7)
\end{aligned}$$

Note that δ_i does not depend on local ancestry, which allows us to compute the heritability due to local ancestry as:

$$\begin{aligned}
h^2_\gamma &\equiv \mathrm{var}[E[y_i \,|\, \gamma_{i,1}, \ldots, \gamma_{i,N}]] \\
&= \mathrm{var}\left[\sum_s \beta_s \frac{\sigma_\gamma}{\sigma_{g,s}}\bar{\gamma}_{i,s}(p_{s,1} - p_{s,0})\right] \approx \sum_s \left[\beta_s \frac{\sigma_\gamma}{\sigma_{g,s}}(p_{s,1} - p_{s,0})\right]^2 \\
&= 2\theta(1-\theta)\sum_s \left[\beta_s \frac{1}{\sigma_{g,s}}(p_{s,1} - p_{s,0})\right]^2 \quad (8)
\end{aligned}$$

We define F_{STC} as a measure of the genetic distance between ancestral populations weighted by the square of effect size β_s:

$$F_{\mathrm{STC}} = \sum_s \frac{\beta^2_s}{h^2}\frac{(p_{s,1} - p_{s,0})^2}{\sigma^2_{g,s}} \quad (9)$$

This results in a final relationship:

$$h^2_\gamma = 2\theta(1-\theta)h^2 F_{\mathrm{STC}} \quad (10)$$

In practice, we do not know the effect size of every SNP and must make simplifying assumptions about their distribution to estimate F_{STC}. First, consider a simple phenotypic model in which the genotypic effect size β_s is independent of $p_{s,0}$ and $p_{s,1}$. Then:

$$h^2 = \sum_s \beta^2_s \approx NE[\beta^2_1] \quad (11)$$

where N is the number of SNPs. Then, equation (8) becomes:

$$h^2_\gamma = 2\theta(1-\theta)NE[\beta_1^2]E\left[\frac{(p_{s,1}-p_{s,0})^2}{\sigma^2_{g,s}}\right]$$

$$= h^2 2\theta(1-\theta)E\left[\frac{(p_{s,1}-p_{s,0})^2}{\sigma^2_{g,s}}\right] \quad (12)$$

The F_{STC} in equation (12) is a genome-wide measure of the genetic difference between the ancestral populations. This is related to the classic parameter F_{ST} when all variants are causal (i.e., in the infinitesimal model).

$$F_{ST} = E\left[\frac{(p_{s,1}-p_{s,0})^2}{\sigma^2_{g,s}}\right] \quad (13)$$

Now consider a more complex model in which the effect size of SNPs can fall into one of L classes such that the effect size distribution is a function of class L. These classes could be, for example, rare and common variants (used in this work). We defined the genetic distance between ancestral populations within each class as F_{STL} and the phenotypic variance explained by SNPs in this class as h^2_L. Again, substituting into equation (8), we get:

$$h^2_\gamma = 2\theta(1-\theta)\frac{h^2}{h^2}\sum_L\sum_{s\in L}\left[\beta_s^2\frac{(p_{s,1}-p_{s,0})^2}{\sigma^2_{g,s}}\right]$$

$$= 2\theta(1-\theta)h^2\sum_L\frac{h^2_L}{h^2}F_{STL} \quad (14)$$

Therefore:

$$F_{STC} = \sum_L\frac{h^2_L}{h^2}F_{STL}$$

representing a weighted measure of genetic distance in each class.

To obtain an estimate of h^2, we must estimate θ, F_{STC} and h^2_γ. The parameter θ is estimated from local ancestry inference. The parameter F_{STC} is estimated from an assumptions about the variance explained by SNPs in each genotypic class combined with external reference panels[45,46].

Definition and estimation of F_{STC}. As in the equations above, we are defining F_{ST} to be the weighted average (across all SNPs s) of the ratios calculated as:

$$\frac{(p_{s,1}-p_{s,0})^2}{\sigma^2_{g,s}}$$

Although this definition is similar to standard versions of F_{ST}, a ratio of averages is recommended instead when the goal is to draw population genetic inferences[47]. If the distribution of SNP effect sizes is not a function of F_{ST}, then this would be the appropriate definition for our heritability estimation approach. However, recent work has shown that rare variants are unlikely to contribute to a large proportion of phenotypic variation[25,48]. As has been reported previously[47], the average-of-ratios estimate will shrink when many rare variants are included in the estimate. Such an effect is reflected in the 1000 Genomes Projects–based estimate of F_{ST} of 0.07, which used an average of ratios[49]. Therefore, F_{ST} will produce a biased estimate of heritability because the variance explained by rare variants is different from the variance explained by common variants. To account for this difference, we defined a parameter F_{STC}, which is a weighted measure of the genetic distance between ancestral populations (equation (9)).

In practice, we defined F_{STC} as the average F_{ST} within each class L of SNPs (F_{STL}), weighted by the proportion of phenotypic variance explained by that class:

$$F_{STC} = \sum_L\frac{h^2_L}{h^2}F_{STL} \quad (15)$$

Consider a situation in which L contains two classes, rare and common SNPs, with F_{ST} values of 0.054 and 0.182, respectively. If rare variants explain 10% of the heritability and common variants explain 90% of the heritability, then F_{STC} = 0.1692. We estimated F_{STC} over the HapMap 3 (ref. 37) data set using the CEU and YRI populations as proxies for the ancestral populations of African Americans, with an admixture proportion of 18.3% European ancestry and assuming three different distributions of causal variant frequencies. We estimated a value of 0.182 assuming causal variant MAF > 5% (which we used in this work), 0.165 assuming MAF < 5% and 0.054 assuming MAF < 1%.

Simulations with simulated genotypes. To examine the properties of our approach, we first applied our method to data generated under a simple simulation framework for generating the genotypes, local ancestries and phenotypes of individuals from an admixed population. Allele frequencies p_{A1}, p_{A2}, …, p_{AN} of N SNPs from an ancestral population were drawn uniformly from [0.1, 0.9]. Allele frequencies of SNPs from P_0 were drawn from a beta distribution with parameters $p_{AS}(1-F_{STC})/F_{STC}$ and $(1-p_{AS})(1-F_{STC})/F_{STC}$ for each SNP s, and allele frequencies were similarly drawn for SNPs from P_1. The F_{STC} parameter determines the genetic distance between the two populations. The global proportion of P_0 ancestry θ_1, θ_2, …, θ_M for each of M individuals was drawn either uniformly from [0.4, 0.6], from the normal distribution $N(0.5, 0.1)$ or fixed at 0.5. Local ancestry for individual i at SNP s ($\gamma_{i,s}$) was generated by two draws from a binomial distribution with parameter θ_s. The genotypes for individual i at SNP s ($g_{i,s}$) were then generated by drawing from the binomial distributions with allele frequencies specified by the local ancestry for that individual at that SNP. That is, if the individual had two copies of an ancestry segment from P_0 at SNP s, then two draws from a binomial distribution with parameter $p_{0,s}$ were used. To create a phenotype, we first selected Nr causal variants where r is the proportion of causal variants. Effect sizes were drawn from the normal distribution $N(0, h^2/(Nr))$, and the genetic element of the phenotype was generated by taking the inner product of the causal variants, normalized to have mean = 0 and variance = 1, and the effect sizes for the variants. Normally distributed random noise was added such that the total heritability in the population was h^2.

Simulations with real genotypes. We split the genotypes from 5,129 distantly related CARe individuals into 2 groups. The common group contained the SNPs with MAF > 5% in both the CEU and YRI populations. The uncommon group contained all other SNPs (i.e., those with MAF < 5% in either or both the CEU and YRI populations). The genotype kinship matrix K was constructed over the common SNPs, and the local ancestry kinship matrix K_γ was constructed using the local ancestry called at every fifth common SNP.

We simulated a phenotype by first selecting a proportion r of causal variants at random from the common and uncommon SNPs, leaving N_c common causal SNPs and N_n uncommon causal SNPs. We then selected a fraction of the phenotypic variance α explained by the uncommon SNPs. At $\alpha = 0.0$, uncommon variants had no effect and the genetic basis of the phenotype was entirely determined by common variants. We then chose effect sizes for each common and uncommon SNP by drawing from the normal distributions $N(0, (1-\alpha)h^2/(N_c))$ and $N(0, (\alpha)h^2/(N_n))$, respectively. The genetic element of the phenotype was generated by taking the inner product of the causal variants, normalized to have mean = 0 and variance = 1 in the admixed population, and the effect sizes for the variants. Normally distributed random noise was added such that the total heritability in the population was h^2. F_{STC} for the common and uncommon SNPs was 0.15 and 0.25, respectively. The study F_{STC} used to estimate heritability was the weighted mean $0.15(1-\alpha) + 0.25\alpha$, as described in the derivation above. Setting individuals with the lowest P percent of phenotypes as cases and all others as controls, we generated dichotomous phenotypes with prevalence P.

Data set approvals. The CARe project has been approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) of the Massachusetts Institute of Technology and by the institutional review board of each of the nine parent cohorts. CARe data were obtained from the database of Genotypes and Phenotypes (dbGaP).

The WHI project has been approved by the Human Subjects Committees at the WHI Clinical Coordinating Center (FHCRC) and at the 40 WHI Field Centers.

doi:10.1038/ng.3139

The AAPC project has been approved by the Institutional Review Board of the University of Southern California.

The studies contributing data to the CARe, SHARe and AAPC projects have each been approved by their local review boards. Sample sizes for the phenotypes in each data set are given in **Table 4**.

CARe data set. Affymetrix 6.0 genotyping and quality control filtering of African-American samples from the CARe cardiovascular consortium was performed as described previously[50]. After quality control filtering for each of the ARIC, CARDIA, CFS, JHS and MESA cohorts and subsequent merging, 8,367 samples and 770,390 SNPs remained. To limit relatedness among samples, we restricted all analyses to a subset of 5,129 samples in which all pairs had genome-wide relatedness of 0.05 or less and had between 5% and 45% European ancestry. We performed local ancestry inference using HAPMIX software with the CEU and YRI HapMap populations as reference ancestral populations. We examined seven phenotypes from the CARe cohort: height, BMI, log-transformed HDL concentration, LDL concentration, WBC count, DBP and SBP. For each phenotype, we included age, sex, study center, proportion of European ancestry and the top five principal components as fixed effects. A detailed description of the phenotypes can be found in ref. 51.

WHI data set. Affymetrix 6.0 genotyping and quality control filtering of African-American samples from the WHI SHARe cohort was performed as described previously[52]. The data set includes extensive phenotypic and genotypic data on 12,008 African-American and Hispanic women aged 50–79 years enrolled in one or more components of the WHI program. We included only African-American samples, and, to limit relatedness among samples, we restricted all analyses to a subset of 8,153 samples in which all pairs had genome-wide relatedness of 0.05 or less. We performed local ancestry inference using SABER+ software[23] with the CEU and YRI HapMap populations as reference ancestral populations. We examined 11 phenotypes from the WHI cohort: height, BMI, log-transformed HDL concentration, LDL concentration, WBC count, log-transformed triglyceride concentration, glucose concentration, log-transformed insulin concentration, QT-interval duration, CRP concentration, DBP and SBP. For each phenotype, we included age and proportion of European ancestry as fixed effects. A detailed description of the phenotypes can be found in ref. 52.

African-American prostate cancer data set (AAPC). Illumina Human1M-Duov3_B genotyping and quality control filtering of African-American samples from AAPC for a total of 11 participating studies was performed as described previously[53–55]. The cleaned data set includes 9,641 African-American subjects and 1,001,899 autosomal SNPs. To limit relatedness among samples, we restricted all analyses to a subset of 8,215 samples in which all pairs had genome-wide relatedness of 0.05 or less. We performed local ancestry inference using RFMix[24] with the CEU and YRI HapMap populations as reference ancestral populations. We examined prostate cancer outcome for each subject. There were 4,207 cases and 4,008 controls after quality control. Because of the admixture signal at the 8q24 locus[54], we also estimated heritability by removing 8q24 markers from the SNPs used to estimate kinship (PC|8q24). For each phenotype, we included age and the top ten principal components as fixed effects. For conversion to the liability scale, we used a prevalence of 5% (ref. 54).

Partitioning heritability across the genome. To estimate the heritability for a particular genomic segment, we compute the genetic relatedness matrix as defined in Yang *et al.*[10], replacing genotypic information with local ancestry calls and restricting to only the SNPs contained in the region of interest. Given a partitioning of segments along the genome (in our case, 22 segments), it is possible to fit them individually or jointly. We attempted both approaches but found that the joint fit resulted in numerical instability in the optimization algorithm that prevented convergence. Thus, all results reported for the single-chromosome analyses are from individual and not joint estimates.

We performed both weighted and standard linear regression to assess the relationship between the heritability explained by a chromosome and the length of the chromosome. The weighted version accounts for differences in the number of SNPs contained on longer and shorter chromosomes, and the weighting factor we used was the length of the chromosome in centimorgans.

45. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
46. Pennisi, E. Genomics. 1000 Genomes Project gives new map of genetic diversity. *Science* **330**, 574–575 (2010).
47. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
48. 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).
49. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
50. Lettre, G. *et al.* Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7**, e1001300 (2011).
51. Pasaniuc, B. *et al.* Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7**, e1001371 (2011).
52. Franceschini, N. *et al.* Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. *Am. J. Hum. Genet.* **93**, 545–554 (2013).
53. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
54. Haiman, C.A. *et al.* Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. *PLoS Genet.* **7**, e1001387 (2011).
55. Olama, A.A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46**, 1101–1109 (2014).