Detection and quantification of inbreeding depression for complex traits from SNP data

Loic Yengo[a,1], Zhihong Zhu[a], Naomi R. Wray[a,b], Bruce S. Weir[c], Jian Yang[a,b], Matthew R. Robinson[a,d], and Peter M. Visscher[a,b,1]

[a]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia; [b]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia; [c]Department of Biostatistics, University of Washington, Seattle, WA 98195; and [d]Department of Computational Biology, University of Lausanne, Lausanne, CH-1015, Switzerland

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved June 21, 2017 (received for review December 22, 2016)

Quantifying the effects of inbreeding is critical to characterizing the genetic architecture of complex traits. This study highlights through theory and simulations the strengths and shortcomings of three SNP-based inbreeding measures commonly used to estimate inbreeding depression (ID). We demonstrate that heterogeneity in linkage disequilibrium (LD) between causal variants and SNPs biases ID estimates, and we develop an approach to correct this bias using LD and minor allele frequency stratified inference (LDMS). We quantified ID in 25 traits measured in ∼140,000 participants of the UK Biobank, using LDMS, and confirmed previously published ID for 4 traits. We find unique evidence of ID for handgrip strength, waist/hip ratio, and visual and auditory acuity (ID between −2.3 and −5.2 phenotypic SDs for complete inbreeding; $P < 0.001$). Our results illustrate that a careful choice of the measure of inbreeding combined with LDMS stratification improves both detection and quantification of ID using SNP data.

inbreeding depression | directional dominance | quantitative genetics | single-nucleotide polymorphism | homozygosity

Mating between close relatives has detrimental consequences on the survival and fertility of resulting offspring (1). This overall reduction of fitness, referred to as inbreeding depression (ID), is observable in a wide range of organisms, including plants (2), animals (3, 4), and humans (5). In humans, major abnormalities are more frequent in children from consanguineous marriages (6) and genes causing rare diseases can be mapped by ascertaining children from such matings (7). To date, although the genetic basis of ID is not completely elucidated, two main hypotheses are proposed to explain this phenomenon: homozygosity for partially recessive deleterious mutations and heterozygous advantage (overdominance) (1, 8). More generally, ID can be estimated for any complex trait, even if the trait is not an obvious component of fitness. For polygenic traits, ID can be detected if there is directional dominance (DD) across loci, which means that the phenotype of individuals who are heterozygous deviates from the average phenotypes of homozygous individuals in a consistent direction. For fitness components, DD is negative; i.e., on average homozygosity reduces fitness.

In practice, ID can be estimated from pedigree studies when the relationships between parents are known (6, 9). However, given the limited number and the small sizes of such studies in humans, contemporary efforts (5, 10) to quantify ID have instead used SNP genotyping platforms to directly estimate inbreeding coefficients (F). SNP data may allow a more accurate evaluation of inbreeding (11), in particular for distant and cryptic inbreeding, and allow inference to be drawn from large population data (10). Conceptually, once a measure of F is derived from SNP data, ID can subsequently be estimated by correlating phenotype with the estimated F.

Genome-wide estimators of F fall in two categories: average homozygosity measures across loci (irrespective of position) and measures of continuous runs of homozygosity (ROH). Using ROH, ID has been reported for diseases (12, 13), height (5), and cognition (10). ROH-based estimates of F (F_{ROH}) have been previously shown to better correlate with the unobserved pedigree-inbreeding coefficient compared with other measures of inbreeding (14, 15), which has made them a gold standard. However, the sampling variance of these estimates is large, and consequently large sample sizes (10) are required to detect ID with F_{ROH} measures. In addition, F_{ROH} estimation depends on arbitrary (although optimized) choices of multiple parameters like the minimum number of SNPs covered by a ROH, the distance between two consecutive ROHs, and the number of heterozygous genotypes allowed in each ROH. Setting ROH length cutoffs ignores the contribution to ID of smaller identity by descent segments due to distant ancestors.

Therefore, quantifying the theoretical properties (bias and variance) of ID estimates derived from F_{ROH} is challenging. These two critical limitations led us to consider two other commonly used measures of inbreeding (3, 15), namely the excess of homozygosity inbreeding coefficient (hereafter denoted F_{HOM}) as estimated in PLINK (16) and the correlation between uniting gametes (hereafter denoted F_{UNI}) previously introduced as $\hat{F}^{III}$ in Yang et al. (17), as potential efficient measures for detecting and quantifying ID. We present the theory underlying unbiased estimation of ID and compare through simulations the performances of these three measures of inbreeding. We then quantify ID in 25 quantitative traits measured in a large dataset of ∼140,000 individuals from the UK Biobank, using an approach that is robust to different assumptions on the distribution of effect sizes, to possible directional effects of minor alleles and to population stratification.

Significance

Inbreeding depression (ID) is the reduction of fitness in offspring of related parents. This phenomenon can be quantified from SNP data through a number of measures of inbreeding. Our study addresses two key questions. How accurate are the different methods to estimate ID? And how and why should investigators choose among the multiple inbreeding measures to detect and quantify ID? Here, we compare the behaviors of ID estimates from three commonly used SNP-based measures of inbreeding and provide both theoretical and empirical arguments to answer these questions. Our work illustrates how to analyze SNP data efficiently to detect and quantify ID, across species and traits.

Author contributions: L.Y., J.Y., M.R.R., and P.M.V. designed research; L.Y., N.R.W., B.S.W., J.Y., M.R.R., and P.M.V. performed research; L.Y., Z.Z., M.R.R., and P.M.V. analyzed data; and L.Y., J.Y., M.R.R., and P.M.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: l.yengodimbou@uq.edu.au or peter.visscher@uq.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1621096114/-/DCSupplemental.

Results

Theoretical Determinants of Unbiased Estimation of ID. We assume that the phenotype of interest is a quantitative trait y with genetic component that is underlain by random additive and dominance effects of m independent causal variants. We denote $\mathrm{b} = -\sum_{j=1}^{m} 2p_j(1-p_j)\delta_j$ as the expected depression in y resulting from complete inbreeding, where p_j is the minor allele frequency (MAF) of the jth causal variant and δ_j the expectation of its dominance effect. In the absence of epistasis, fitness-related phenotypes linearly decrease with increasing inbreeding (Eq. S2). This well-established linear relationship naturally implies the use of linear regression methods to estimate ID.

Least-squares estimates of ID obtained with $\mathrm{F_{UNI}}$ converge with increasing sample size toward $\mathrm{b_{UNI}} = \mathrm{cov}[\boldsymbol{y}, \mathrm{F_{UNI}}]/\mathrm{var}[\mathrm{F_{UNI}}]$. When explicitly calculating $\mathrm{cov}[\boldsymbol{y}, \mathrm{F_{UNI}}]$ and $\mathrm{var}[\mathrm{F_{UNI}}]$ with respect to the genotypes and the effect size distributions, we found under classical assumptions (Eq. S4 and Table S1) that $\mathrm{b_{UNI}}$ is unbiased when the average linkage disequilibrium (LD) among observed SNPs equals the weighted (by effect sizes) average LD between causal variants and observed SNPs. Although influenced by the effect size distribution, the consistency of $\mathrm{b_{UNI}}$ toward b is mainly driven by differences in LD between causal variants and observed SNPs. Therefore, a simple condition under which $\mathrm{b_{UNI}}$ is unbiased is if the causal variants are a random subset of the observed SNPs. However, if the causal variants are enriched in high-LD regions of the genome, $\mathrm{b_{UNI}}$ will overestimate the actual inbreeding depression. In contrast, if the causal variants are enriched in a low-LD region like DNAse-I hypersensitive sites or enhancers (18), or if they are enriched among low-frequency variants, $\mathrm{b_{UNI}}$ is expected to underestimate the true effect. This is further illustrated in our first simulation (Fig. 1).

LD heterogeneity between causal variants and SNPs used for inference has been previously shown to determine the consistency of heritability estimates (19–21). We leveraged this estimation problem to propose a strategy to correct the differential LD bias when estimating ID. Following a previous approach (19), we explored how stratifying SNPs according to their LD score (22) and their MAFs before analyses (details given in *Supporting Information*) could correct or at least reduce these biases. We illustrate in our first simulation that LD score and MAF (LDMS) stratification performs well in correcting these biases (Fig. 1).

Similar to that shown with $\mathrm{F_{UNI}}$, we prove that the consistency of ID estimates obtained with $\mathrm{F_{HOM}}$ (hereafter denoted $\mathrm{b_{HOM}}$) is also determined by LD differences between SNPs and causal variants (Eq. S5). However, the bias of $\mathrm{b_{HOM}}$ cannot simply be predicted by the ratio of the mean LD score in causal variants over the mean LD score in SNPs (*Supporting Information*). Nevertheless, our derivations predict that $\mathrm{b_{HOM}}$ behaves similarly to $\mathrm{b_{UNI}}$ with respect to LD differences between causal variants and SNPs. Importantly, we also prove that possible directional effects of minor alleles (DEMA) could confound $\mathrm{b_{HOM}}$ because of the correlation between minor allele counts and $\mathrm{F_{HOM}}$ (*Supporting Information*). Such directional effects could arise as a consequence of directional selection (when the minor allele is also the derived allele) as previously reported in human height (23) or simply because of population stratification (PS).

Simulation Study. The complete description of the simulation study is given in *Supporting Information*.

Influence of differential MAF and LD between causal variants and SNPs. We first considered three scenarios to illustrate the influence of LD and MAF heterogeneity between causal variants and SNPs. In all these scenarios, we assumed no DEMA, i.e., parameter $\mathrm{s} = 0$ in Eq. S3, and that $\mathrm{b} = -3$ phenotypic SDs. Moreover, we assumed the expectation of the dominance effects to be either constant, i.e., $\delta_j = -b/\sum_{j=1}^{m} 2p_j(1-p_j)$, or inversely proportional to the variance of the minor allele count, i.e., $\delta_j = -b/2mp_j(1-p_j)$. The first assumption corresponds to neutral traits, whereas the second one assigns a larger effect to SNPs with lower MAF and therefore corresponds more to traits under directional selection. Unbiasedness is defined below as when the average estimate of ID over multiple simulation replicates does not significantly differ from the value of b used for simulation.

Scenario 1. In this scenario the causal variants were randomly sampled from the 3,857,369 autosomal SNPs that passed the genotypes quality control (*Supporting Information*). As predicted by our derivations, we observed that $\mathrm{F_{UNI}}$-based estimates of b were unbiased when dominance effects are assumed inversely proportional to the variances of allele counts, whereas an overestimation of ~14% of b was observed when dominance effects are assumed constant (Fig. 1*A*). This overestimation is explained by the fact that assuming a constant dominance effect, regardless

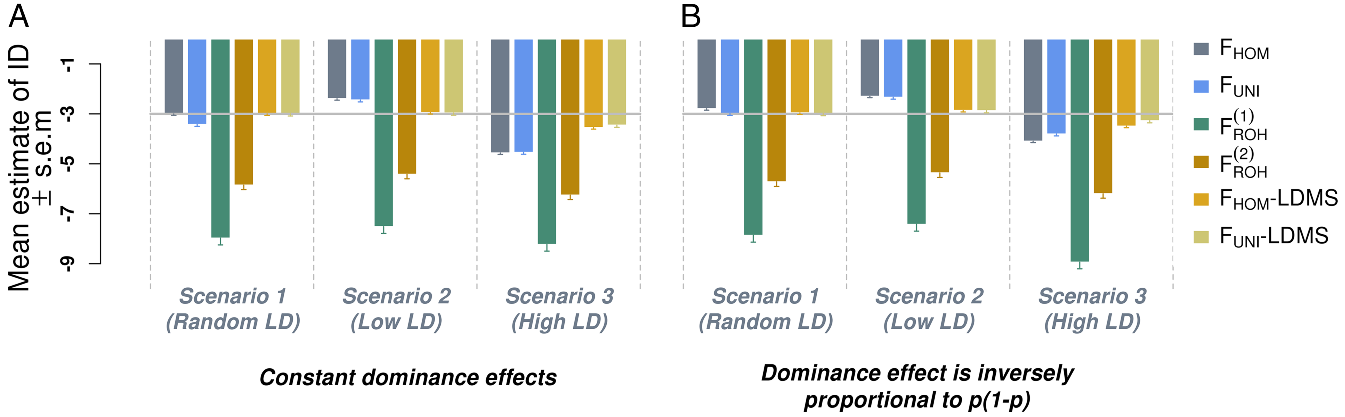

Fig. 1. Averaged estimates of inbreeding depression (ID) from 1,000 simulated datasets. Datasets were simulated assuming a true ID parameter b = −3 (horizontal gray line) phenotypic SD for complete inbreeding. In scenario 1 the $m = 1{,}000$ causal variants were randomly sampled from all observed SNPs, whereas in scenarios 2 and 3 they were respectively sampled from low- and high-LD regions of the genome. In *A* the expectation of the dominance effects (δ_j for the *j*th causal variant) is constant (neutral model) whereas in panel *B* δ_j is inversely proportional to the variance of the minor allele count at each causal variant. $\mathrm{F_{HOM}}$, excess homozygosity inbreeding measure; $\mathrm{F_{ROH}}$, runs of homozygosity-based inbreeding measures; $\mathrm{F_{UNI}}$, measure of inbreeding based on correlation between uniting gametes; LDMS, LD and minor allele frequency stratified inference; SEM, SE of the mean.

of allele frequencies, creates an apparent MAF and LD heterogeneity between SNPs and causal variants by relatively up-weighting common SNPs compared with rarer SNPs (Eq. **S3**). We observed that LDMS stratification, which accounts for that heterogeneity, completely corrected this upward bias as presented in Fig. 1*A*. In addition, we found that F$_{\text{HOM}}$ produced unbiased estimates of b when dominance effects are assumed constant as for a neutral trait (Fig. 1*A*), but was biased downward (−7% of b) when dominance effects are inversely proportional to the variances of allele counts (Fig. 1*B*). This downward bias can be explained using the same reasoning presented above because in that case assuming dominance effects inversely proportional to the variances of allele counts relatively up-weights rarer SNPs compared with common ones. This downward bias could similarly be corrected using LDMS stratification. We also found that estimates of b obtained with ROH-based measures of inbreeding were strongly biased: +162% of b using the definition of ROH from Joshi et al. (10) [hereafter denoted $F^{(1)}_{\text{ROH}}$] and +91% of b using an alternative definition from Gazal et al. (15) or Howrigan et al. (24) [hereafter denoted as $F^{(2)}_{\text{ROH}}$]. The main difference between those two definitions of ROH is that $F^{(2)}_{\text{ROH}}$ requires LD pruning of the SNPs before calling the ROHs, whereas $F^{(1)}_{\text{ROH}}$ explicitly imposes a constraint on the ROH lengths (here >1.5 Mb). This result highlights that LD pruning improves ID estimation using ROH-based inbreeding measures but still remains insufficient to produce unbiased estimates. Indeed, using more stringent LD pruning thresholds did not change our conclusion (Fig. S1). Overall, we found that LDMS stratified estimates for F$_{\text{UNI}}$ and F$_{\text{HOM}}$ were unbiased in all cases (Fig. 1 *A* and *B*), which emphasizes that this strategy can be safely used even when causal variants are perfectly tagged by SNPs.

On average over 1,000 simulation replicates we found that F$_{\text{HOM}}$-associated estimates had smaller standard errors (SE) compared with F$_{\text{UNI}}$ or F$_{\text{ROH}}$ ($F^{(1)}_{\text{ROH}}$ and $F^{(2)}_{\text{ROH}}$) (Fig. S2 *A* and *B*). F$_{\text{HOM}}$ consequently yielded the largest statistical power whereas F$_{\text{UNI}}$ was second best with a power on average 10% below that of F$_{\text{HOM}}$. On the other hand, because of their large SEs, $F^{(1)}_{\text{ROH}}$ and $F^{(2)}_{\text{ROH}}$ yielded the smallest statistical power to detect ID. Finally, we found that LDMS stratified estimates had ∼13% larger SEs compared with nonstratified estimates. This increase in SE corresponds on average over all inbreeding measures to an ∼8% loss of statistical power and is explained by the larger underlying effective dimensionality (4 LD score strata × 6 MAF strata = 24 parameters actually estimated; *Supporting Information*) of the LDMS approach compared with the nonstratified inference.

Scenarios 2 and 3. For the two other scenarios we used 1,358,699 SNPs within exons, introns, 3'-UTRs, 5'-UTRs, and promoter regions ±500 bp (*SI Materials and Methods, URLs*). SNPs within these five genomic (sets of) regions have distinct MAF and LD distributions as shown in Figs. S3 and S4. In scenario 2, we sampled the causal variants among 542,379 intronic SNPs with MAF <5% whereas in scenario 3 causal variants were sampled among 28,341 SNPs within exons, 3'-UTRs, and 5'-UTRs. Our theoretical derivations predict that F$_{\text{UNI}}$ and F$_{\text{HOM}}$ would underestimate the true ID in scenario 2 because causal variants in that scenario had on average lower LD scores (Fig. S3). Accordingly, we found over 1,000 simulation replicates an underestimation of ∼19% of the true ID for F$_{\text{UNI}}$ and F$_{\text{HOM}}$ (Fig. 1). These downward biases could be reduced below 1% of b using LDMS stratification (Fig. 1 *A* and *B*) and were not significantly different from 0 ($P > 0.5$). In scenario 3 causal variants had on average larger LD scores and MAF (Figs. S3 and S4). We therefore expected an overestimation of ID estimates in that scenario according to our theoretical derivations. This predicted upward bias was indeed more noticeable in our simulations (∼40% on average over all inbreeding measures) compared with scenario 2. Still, using LDMS stratification we were able to reduce these biases down to <15% of b on average over all inbreeding measures (Fig. 1) and more specifically <10% of b for F$_{\text{UNI}}$. Overall, LDMS stratification using F$_{\text{UNI}}$ yielded the smallest biases compared with all other strategies.

Influence of DEMA. Let α_j denote the expectation of the additive effect of the minor allele at the *j*th causal variant. We define $s = \sum_{j=1}^{m} 2p_j\alpha_j$ as an overall measure of DEMA. Under this assumption we prove that estimates of ID obtained with F$_{\text{HOM}}$ are confounded because of the correlation between F$_{\text{HOM}}$ and minor allele counts (*Supporting Information*). We illustrate and quantify here that confounding bias. The parameters of this simulation are similar to scenario 1 of the first simulation, except that data are now simulated assuming no ID, i.e., b = 0 and using s = 10. With s = 10, DEMA (which contributes to the additive genetic variance) accounts in our simulations for ∼0.3% of the total phenotypic variance (*Supporting Information*). We considered two alternatives for the expectation of the additive effects α_j: (*i*) $\alpha_j = s/\sum_{j=1}^{m} 2p_j$ is constant and (*ii*) $\alpha_j = s/2mp_j$ is inversely proportional to the MAF p_j. Under both alternatives, we found that estimates of ID obtained with F$_{\text{HOM}}$ were severely biased (between −2 and −3 phenotypic SDs whereas the true value is 0; Fig. 2*A*) unlike those derived from other inbreeding measures. Whether this bias can be corrected is a difficult question in practice. Indeed, under the simplistic scenario considered here where causal variants are well tagged, adjusting for a genetic score summing all minor alleles should completely correct this bias. However, in more realistic situations where causal variants are only partially tagged, this would remain insufficient. In contrast, theory shows that F$_{\text{UNI}}$ is orthogonal to minor allele counts and consequently would not be influenced by such directional effects if these exist. This result has motivated our decision to restrict real data analyses to F$_{\text{UNI}}$ only even if F$_{\text{HOM}}$ had better statistical power in our previous simulations.

DEMA could also arise as a consequence of PS. To illustrate that second aspect, we performed another simulation with settings similar to scenario 1 of the first simulation (with b = 0 and s = 0) but now include the effect of 10 genotypic principal components (PC) as a proxy for PS (*Supporting information*). When varying the contribution of PS to the phenotypic variance from 0 to 5%, we observed that PS had a larger influence on estimates derived from F$_{\text{HOM}}$ and F$_{\text{ROH}}$. These observations are consistent with our theoretical results (at least for F$_{\text{HOM}}$) as they directly derive from the correlation between F$_{\text{HOM}}$ and minor allele counts, the weighted sum of which constitutes the PC. The biases of F$_{\text{HOM}}$- and F$_{\text{ROH}}$-associated ID estimates were on average ∼1 phenotypic SD (Fig. 2*B*) when PS explained 5% of the phenotypic variance but were significant only for F$_{\text{HOM}}$ (Fig. S5*B*). On the contrary, the biases observed when using F$_{\text{UNI}}$ never exceeded 0.5 phenotypic SD (Fig. 2*B*) and were not statistically different from 0 (Fig. S5*B*). In conclusion, orthogonality between F$_{\text{UNI}}$ and minor allele counts (25) (*Supporting Information*) guarantees that confounding by DEMA or PS is negligible for ID estimates obtained with this measure.

To summarize this section, we show in our simulations that biases in ID estimates induced by MAF and LD heterogeneity between SNPs and causal variants can be corrected using LDMS stratification. Moreover, we show on average that LDMS correction performs better when applied to F$_{\text{UNI}}$ and that F$_{\text{UNI}}$-based ID estimates are robust to DEMA and PS. Overall, F$_{\text{UNI}}$ offers the best trade-off between statistical efficiency and unbiasedness in the situations covered in this simulation study. We therefore recommend its use and focus hereafter our analyses on real data to F$_{\text{UNI}}$.

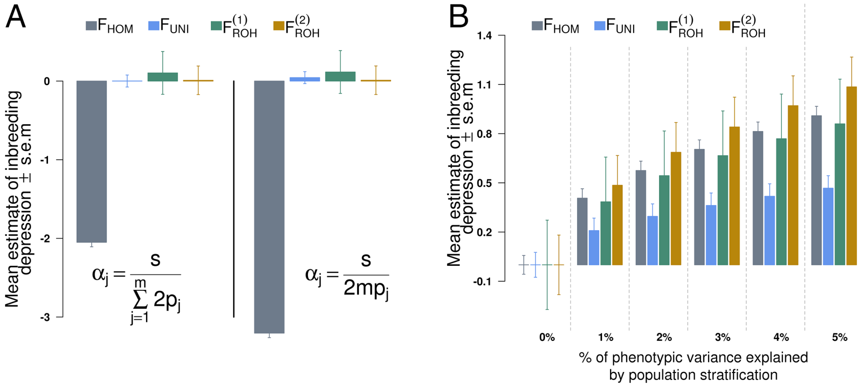

Fig. 2. Averaged estimates of ID from 1,000 simulated datasets. In *A* datasets were simulated assuming no ID, i.e., $b = 0$ and nonnegative expectation for the additive effects (i.e., $\alpha_j > 0$, for the *j*th causal variant). In *B* datasets were simulated assuming no ID ($b = 0$) and no directional effect of minor alleles ($s = 0$) but including the contribution of the first 10 genotypic PCs to model the effect of population stratification. F_{HOM}, excess homozygosity inbreeding measure; F_{ROH}, runs of homozygosity-based inbreeding measures; F_{UNI}, measure of inbreeding based on correlation between uniting gametes; SEM, SE of the mean.

Analysis of UK Biobank Data. We quantified ID in 25 quantitative traits measured in ∼140,000 (Table 1) participants from the UK Biobank (*Supporting Information*). These traits can be grouped into three categories including physical measures (standing height, weight, body mass index, waist and hip circumferences, waist/hip ratio, bone mineral density, body fat percentage, hand grip strength, systolic blood pressure, diastolic blood pressure, heart pulse rate, peak expiratory flow, visual acuity measured on log minimum angle of resolution (MAR) scale, and auditory acuity measured as the speech reception threshold), cognitive traits and educational attainment (fluid intelligence score, mean time to identify matches, maximum number of digits remembered, and age when completed full education), and sex-specific reproductive traits (number of children fathered, number of live births, age at menarche, and age at menopause) (Table S2). Some of these traits like standing height, peak expiratory flow (strongly correlated with forced expiratory volume in 1 s: $r = 0.79$, $P < 10^{-10}$), educational attainment, and cognitive ability were previously reported to be associated with inbreeding (10). Beyond quantifying the effect of inbreeding on these traits, we also aimed to evaluate whether differential LD and MAF distribution in causal variants influenced classical least-squares estimates and if so to correct these biases using our LDMS inference. The analyses were performed using linear regression adjusted for age, sex (for traits not specific to males or females), recruitment center, Townsend deprivation index (26) as a proxy for socioeconomical status, and the first 10 genotypic PCs. The last three adjustments were considered to account for geographical and socio-economic structures in the UK population, which we found to correlate with levels of inbreeding (Table S3).

After Bonferroni correction ($P < 0.05/25$ traits $= 2 \times 10^{-3}$), we detected significant ID in eight traits (Table 1), using LDMS stratified inference based on F_{UNI}. Ranked by decreasing magnitude of depression, these traits are auditory acuity (AA), fluid intelligence score (FIS), visual acuity (VA), peak expiratory flow (PEF), number of children fathered (NCF), hand grip strength (HGS), mean time to correctly identify matches (MTCIM), and waist/hip ratio (WHR). This analysis included 16,781 related individuals (estimated to be first, second, or third degree) and 19 participants with extreme inbreeding ($F_{UNI} > 0.15$). As a sensitivity analysis we reran all analyses without related and inbreeding outliers, which reduced the number of traits passing the

Table 1. Statistically significant estimates of inbreeding depression for eight quantitative traits measured in the UK Biobank.

		F_{UNI}			F_{UNI}(LDMS)			
Traits	N	Estimate	SE	*P* value	Estimate	SE	*P* value	P_{HET}
PEF	117,575	−4.1 (−4.1)	0.62	4.57×10^{-11}	−4.12 (−4.19)	0.69	2.25×10^{-9}	0.941 (0.81)
	(103,781)		(0.74)	(3.48×10^{-8})		(0.84)	(6.35×10^{-7})	
AA, speech	43,175	5.23 (4.44)	1.04	5.55×10^{-7}	5.34 (4.39)	1.16	4.6×10^{-6}	0.828 (0.94)
reception threshold	(38,449)		(1.26)	(4.45×10^{-4})		(1.45)	(2.51×10^{-3})	
FIS	45,043	−3.9 (−3.14)	0.97	5.36×10^{-5}	−4.72 (−4.32)	1.06	8.52×10^{-6}	0.061 (0.07)
	(40,089)		(1.23)	(1.08×10^{-2})		(1.42)	(2.25×10^{-3})	
HGS, average of left	139,623	−2.36 (−2.72)	0.54	1.38×10^{-5}	−2.43 (−3.31)	0.6	4.64×10^{-5}	0.771 (0.1)
and right hands	(122,950)		(0.68)	(5.79×10^{-5})		(0.76)	(1.51×10^{-5})	
NCF	51,494	−3.09 (−3.69)	0.96	1.27×10^{-3}	−4.01 (−4.58)	1.07	1.79×10^{-4}	0.039 (0.14)
	(45,483)		(1.16)	(1.44×10^{-3})		(1.32)	(5.17×10^{-4})	
MTCIM	138,902	1.94 (1.7)	0.54	3.66×10^{-4}	2.05 (2.1)	0.6	6.14×10^{-4}	0.647 (0.26)
	(122,334)		(0.68)	(1.21×10^{-2})		(0.77)	(6.24×10^{-3})	
WHR	140,295	2.35 (3.03)	0.53	1.12×10^{-5}	1.97 (2.89)	0.6	7.85×10^{-4}	0.121 (0.69)
	(123,540)		(0.67)	(6.18×10^{-6})		(0.76)	(1.37×10^{-4})	
VA, log	29,616	4.04 (5.77)	1.20	7.42×10^{-4}	4.39 (6.68)	1.32	9.04×10^{-4}	0.518 (0.26)
MAR scale	(26,596)		(1.51)	(1.37×10^{-4})		(1.73)	(1.14×10^{-4})	

Effect sizes and standard errors are expressed in phenotypic SD of the trait. Results presented in parentheses were obtained after removing 16,781 related individuals and 19 extreme cases of inbreeding ($F_{UNI} > 0.15$). N: sample size in the analysis. P_{HET} is the *P* value from the LDMS heterogeneity test comparing nonstratified and LDMS-stratified estimates.

significance threshold to five. The three traits dropped in this secondary analysis were AA ($P = 2.51 \times 10^{-3}$ in secondary analysis), FIS ($P = 2.25 \times 10^{-3}$ in secondary analysis), and MTCIM ($P = 6.24 \times 10^{-3}$ in secondary analysis). To test whether the differences in ID estimates between full and reduced analyses are significant, we used a jackknife procedure to compare the observed differences with differences generated when randomly excluding 16,800 participants. Over 1,000 resampling events, we found that the observed differences in ID estimates for the eight traits highlighted above were not significantly different from those obtained when excluding random subsets (empirical $P > 0.14$; Fig. S6). We consequently believe that the drop of significance between those two analyses is mainly explained by the reduced statistical power and not by confounding. In addition, we explored how much of ID could be captured at genome-wide significant (GWS) SNPs. We therefore selected trait-specific GWS SNPs from the genome-wide association studies (GWAS) catalog (*SI Materials and Methods, URLs*) and assessed inbreeding depression for the same traits using F_{UNI} at those loci. We could not, however, detect any significant association with the traits analyzed in our study. Even for height, for which ∼700 common GWSs are now reported (27), the estimate of inbreeding depression at GWS was only -0.08 SD for complete inbreeding ($P = 0.072$).

We observed for all traits that ID estimates derived from F_{ROH} were systematically larger than those obtained with F_{UNI} (Table S3). As expected, their SEs were also larger. In particular, only four and six traits (of eight detected with F_{UNI}) passed the Bonferroni threshold when using $F_{ROH}^{(1)}$ and $F_{ROH}^{(2)}$, respectively. On the other hand, ID estimates obtained with F_{HOM} were systematically smaller than those obtained using F_{UNI} (Table S3), with an average over the eight traits significantly associated with F_{UNI}, $b_{HOM} \approx 0.64 \times b_{UNI}$. The latter observation would be expected if the traits analyzed are under directional selection as observed in our simulations when rarer variants were assumed to have larger effects.

We observed for most traits that LDMS stratified and nonstratified F_{UNI} estimates were similar (Table 1), suggesting weak differential LD and MAF distributions in SNPs tagging causal variants. Nonetheless, a marginally significant (LDMS heterogeneity test $P < 0.05$; Table 1 and Fig. S7) difference could be observed in NCF for which the LDMS ID estimate was ∼1 SE larger than the nonstratified one (Table 1). This also translated into an improvement of the association P value from 1.27×10^{-3} to 1.79×10^{-4} (Table 1). We subsequently assessed which component(s) in the LDMS stratification contributed the most to NCF (Fig. S8). We therefore fitted a first multivariate regression model adjusted for four inbreeding coefficients specific to each LD score strata component and then another multivariate regression model adjusted for six inbreeding coefficients specific to each MAF stratum. We chose to fit two different models (for MAF and LD separately) instead of one including 24 covariates to minimize the effects of colinearity between inbreeding measures in each LDMS stratum. We found a nominally significant contribution of SNPs with minor alleles <5% ($b_{FUNI} = -4.01$ phenotypic SD; $P = 0.01$) but no significant enrichment in LD strata despite a large contribution of the second-lowest LD score stratum ($b_{FUNI} = -1.62$ phenotypic SD; $P = 0.12$). According to our derivations, this enrichment of ID in lower-frequency SNPs and more generally in low-LD regions explains why nonstratified analyses produced smaller estimates compared with the LDMS approach. These results imply a disproportionate contribution of low-frequency SNPs to ID in NCF.

Discussion

We comprehensively quantified the behavior of ID estimators based on three commonly used measures of inbreeding. Our study illustrated some of the shortcomings of the most commonly used ROH-based estimates of ID, which not only are biased but also have large SE (approximately three times larger compared with F_{UNI}). This, along with the arbitrary choices underlying the definition of ROHs, leads us to recommend the use of F_{UNI} over F_{ROH}. Overall, our results suggest that F_{UNI}-based ID estimates are robust to different assumptions about the distribution of effect sizes, to possible directional effects of minor alleles, and also to population stratification. The contribution of population stratification reported in this study needs, however, to be put in perspective as our simulations and real data analyses were based upon a relatively homogeneous population within the United Kingdom. For stronger population stratification (e.g., between European and Asian populations) F_{UNI}-based ID estimates can also be biased. This somewhat extreme situation, which would in general be handled as part of quality control, is discussed in *Supporting Information* (Table S4 and Dataset S1).

This study also highlights that differential LD distribution between causal variants and SNPs could bias ID estimates. As previously reported for genomic-relatedness-based restricted maximum-likelihood (GREML) (28) heritability estimates (19–21), we demonstrate through simulations that an LDMS approach successfully corrects these biases when estimating ID. More generally, the flexibility of the LDMS approach in terms of numbers and types of MAF/LD strata allows adaptation to any effect size distribution. Indeed, all SNP-based inbreeding measures are defined upon an underlying assumption on the distribution of dominance effects (e.g., when assuming constant dominance effects, the underlying inbreeding measure is F_{HOM}), which, when not verified, creates biases in ID estimates even when causal variants are randomly distributed among observed SNPs. We showed in our simulations using two distributions of dominance effects that such biases are explained by MAF and LD heterogeneity between causal variants and SNPs and therefore can be corrected using the LDMS approach. This result is important as it guarantees an unbiased estimation of ID regardless of the distribution of dominance effects.

Beyond methodological considerations, we confirmed in this study known associations between increased inbreeding and reduced lung function (PEF), cognitive ability (FIS and MTCIM), and fertility (NCF), which were previously reported, however, using different proxy traits (10, 29–31). We also replicated the association between inbreeding and decreased height ($\mathbf{b}_{UNI}$ – LDMS: -1.71 phenotypic SD for complete inbreeding; $P = 0.003$) even though it was below the Bonferroni threshold. We did not replicate the association with educational attainment (EA) measured, as the "age when completed full education." However, when measuring EA as whether or not participants went to college or university as in Joshi et al. (10), we found a strong although nominally significant negative association (odds ratio of 0.04, $P = 0.02$).

We report evidence of ID for HGS, VA, AA and WHR. Although HGS, VA, and AA seem obvious fitness-related traits, these results still require replication. The association with WHR is particularly interesting as it illustrates on real data that one may benefit from using a less variable measure of inbreeding. Joshi et al. (10) reported a positive effect of inbreeding on WHR using F_{ROH}; however, even with ∼200,000 individuals the effect did not reach statistical significance ($P = 0.09$). Although heterogeneity between the cohorts involved in that meta-analysis may explain that apparent lack of statistical power, our theoretical and simulation results predict that if the authors had used a different metric [F_{UNI} instead of $F_{ROH}^{(1)}$ to reduce the SE] combined with a LD and MAF stratified inference, then such an effect would have been easily detected. We have not considered in this study the detection and estimation of nonlinear effects of inbreeding because of a lack of statistical power to detect such effects. Nonlinearity is predicted by theory in the presence of epistasis involving dominance (32) and was implied by

Szpiech et al. (33), who showed that long runs of homozygosity were enriched for coding variants that are predicted to be deleterious. In conclusion, we have demonstrated that LD and MAF stratified inference based on F_{UNI} as a measure of inbreeding minimizes bias relative to the other ID estimation strategies compared in this study. As illustrated here on real data, we believe that our approach will lead to more discoveries in forthcoming and larger studies.

Materials and Methods

Statistical Models and Notations. We consider the following model:

$$y_i = \mu + \sum_{j=1}^{m} a_j x_{ij} + \sum_{j=1}^{m} d_j H_{ij} + \varepsilon_i, \qquad [1]$$

For individual i, y_i is the observed value of the phenotype of interest, and x_{ij} is the minor allele count at the jth causal SNP ($x_{ij} \in \{0, 1, 2\}$). We denote p_j the minor allele frequency of the jth causal SNP, $H_{ij} = x_{ij}(2 - x_{ij})$ is the indicator of heterozygosity, and ε_i is a residual term capturing nongenetic effects on the observed phenotype. The additive and dominance effect sizes of the minor allele at the jth causal SNP are respectively denoted a_j and d_j. We assume independence between the m causal variants, between the genotypes and the effect sizes, and between the genetic and the nongenetic effects. Finally, we assume the effect sizes to be random and such that $E[a_j] = \alpha_j$ and $E[d_j] = \delta_j$.

Measures of Inbreeding. We studied three measures of inbreeding. All these measures of inbreeding require individual SNP genotypes and can be used in the absence of any pedigree information. The first inbreeding measure is the excess of homozygosity measure defined here as

$$F_{HOM} = 1 - \frac{\sum_{k=1}^{p} x_k(2 - x_k)}{\sum_{k=1}^{p} 2p_k(1 - p_k)},$$

where x_k is the minor allele count of SNP k, p_k is the minor allele frequency, and p is the number of genotyped or imputed SNPs available. F_{HOM} is implemented in PLINK2 (command: –het).

The second measure (F_{UNI}) is based on the correlation between uniting gametes. This measure was defined in Yang et al. (17) as

$$F_{UNI} = \frac{1}{p}\sum_{k=1}^{p} \frac{x_k^2 - (1 + 2p_k)\,x_k + 2p_k^2}{2p_k(1 - p_k)}.$$

F_{UNI} is implemented in PLINK2 and GCTA software (command: –ibc).

The last measure is defined as the proportion of the genome within ROH. More specifically, the inbreeding measure F_{ROH} was calculated as the cumulated length (in base pairs) of one individual's genome within ROHs divided by 3×10^9 (the approximate length of the autosomal genome in base pairs). We used two definitions of ROHs corresponding to those proposed in Joshi et al. (10) (definition 1) and Gazal et al. (15) (definition 2). Inbreeding measures calculated using definition 1 and definition 2 are respectively denoted as $F_{ROH}^{(1)}$ and $F_{ROH}^{(2)}$. Both definitions used SNPs with MAF $> 5\%$ as previously reported in Joshi et al. (10) and Gazal et al. (15) (*Supporting Information*).

UK Biobank Data. We used baseline data from 152,729 men and women who were genotyped in the first phase of genotyping of the UK Biobank (34). To ensure ancestry homogeneity, we selected individuals who reported to be "British," "Irish," "white," or of "any other white background" and whose coordinates on the first genetic PC were below 0 (Fig. S9). In total, we included 140,720 participants in this analysis. The Northwest Multicentre Research Ethics Committee (MREC) approved the study and all participants in the UK Biobank study provided written informed consent. The first steps of the quality control have been previously described (*SI Materials and Methods, URLs*). Phasing and imputation were performed using SHAPEIT and IMPUTE2 (*SI Materials and Methods, URLs*), respectively, as previously described (35). After imputation, we selected 9,493,148 autosomal SNPs with imputation quality $r^2 > 0.3$, MAF $> 1\%$, and Hardy–Weinberg equilibrium test P value $> 10^{-6}$. Imputed SNPs were then called to the genotypes having the largest posterior probability. Finally, we removed redundancy by LD pruning SNPs with a squared genotype correlation $r^2 > 0.9$. In total we used 3,857,369 SNPs in this analysis.

ACKNOWLEDGMENTS. This research was supported by the Australian Research Council (DP130102666 and DP160103860), the Australian National Health and Medical Research Council (1078037 and 1107258), the National Institute of Health (NIH Grants R01AG042568 and P01GM099568), and the Sylvia and Charles Viertel Charitable Foundation. This research has been conducted using the UK Biobank Resource under Project 12505.

1. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nat Rev Genet* 10:783–796.
2. Huang X, et al. (2015) Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat Commun* 6:6258.
3. Huisman J, Kruuk LEB, Ellis PA, Clutton-Brock T, Pemberton JM (2016) Inbreeding depression across the lifespan in a wild mammal population. *Proc Natl Acad Sci USA* 113:3585–3590.
4. Pemberton JM, Ellis PE, Pilkington JG, Bérénos C (2017) Inbreeding depression by environment interactions in a free-living mammal population. *Heredity* 118:64–77.
5. McQuillan R, et al. (2012) Evidence of inbreeding depression on human height. *PLoS Genet* 8:e1002655.
6. Bittles AH, Neel JV (1994) The costs of human inbreeding and their implications for variations at the DNA level. *Nat Genet* 8:117–121.
7. Najmabadi H, et al. (2011) Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* 478:57–63.
8. Charlesworth B, Charlesworth D (1999) The genetic basis of inbreeding depression. *Genet Res* 74:329–340.
9. Morton NE, Crow JF, Muller HJ (1956) An estimate of the mutational damage in man from data on consanguineous marriages. *Proc Natl Acad Sci USA* 42:855–863.
10. Joshi PK, et al. (2015) Directional dominance on stature and cognition in diverse human populations. *Nature* 523:459–462.
11. Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity* 115:63–72.
12. Keller MC, et al. (2012) Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet* 8:e1002656.
13. Lencz T, et al. (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci USA* 104:19942–19947.
14. Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189:237–249.
15. Gazal S, et al. (2014) Inbreeding coefficient estimation with dense SNP data: Comparison of strategies and application to HapMap III. *Hum Hered* 77:49–62.
16. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
17. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82.
18. Gusev A, et al. (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 95:535–552.
19. Yang J, et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47:1114–1120.
20. Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* 91:1011–1021.
21. Lee SH, et al. (2013) Estimation of SNP heritability from dense genotype data. *Am J Hum Genet* 93:1151–1155.
22. Bulik-Sullivan BK, et al. (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47:291–295.
23. Robinson MR, et al. (2015) Population genetic differentiation of height and body mass index across Europe. *Nat Genet* 47:1357–1362.
24. Howrigan DP, et al. (2016) Genome-wide autozygosity is associated with lower general cognitive ability. *Mol Psychiatry* 21:837–843.
25. Zhu Z, et al. (2015) Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* 96:377–385.
26. Mackenbach JP (1988) Health and deprivation. Inequality and the North. *Health Policy* 10:207.
27. Wood AR, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 46:1173–1186.
28. Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42:565–569.
29. Robert A, Toupance B, Tremblay M, Heyer E (2009) Impact of inbreeding on fertility in a pre-industrial population. *Eur J Hum Genet* 17:673–681.
30. Bittles AH, Grant JC, Sullivan SG, Hussain R (2002) Does inbreeding lead to decreased human fertility? *Ann Hum Biol* 29:111–130.
31. Woodley MA (2009) Inbreeding depression and IQ in a study of 72 countries. *Intelligence* 37:268–276.
32. Lynch M (1991) The genetic interpretation of inbreeding depression and outbreeding depression. *Evolution* 45:622–629.
33. Szpiech Z, et al. (2013) Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* 93:90–102.
34. Allen N, et al. (2012) UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol* 1:123–126.
35. O'Connell J, et al. (2016) Haplotype estimation for biobank-scale data sets. *Nat Genet* 48:817–820.