Genetic drift from the out-of-Africa bottleneck leads to biased estimation of genetic architecture and selection

Bilal Ashraf[1,2] and Daniel John Lawson[1,3]*

[1]: Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Oakfield House, Bristol, BS8 2BN, UK

[2]: Department of Anthropology, Durham Research Methods Centre, Dawson Building, University of Durham, DH13LE, UK

[3]: Department of Statistical Sciences, School of Mathematics, University of Bristol, Fry Building, BS8 1UG, UK

* Corresponding author: dan.lawson@bristol.ac.uk

Abstract

Most complex traits evolved in the ancestors of all modern humans and have been under negative or balancing selection to maintain the distribution of phenotypes observed today. Yet all large studies mapping genomes to complex traits occur in populations that have experienced the Out-of-Africa bottleneck. Does this bottleneck affect the way we characterise complex traits? We demonstrate using the 1000 Genomes dataset and hypothetical complex traits that genetic drift can strongly affect the joint distribution of effect size and SNP frequency. Characterisations that rely on this distribution therefore conflate genetic drift and selection. We provide a model to identify the underlying selection parameter in the presence of drift, and demonstrate that a simple sensitivity analysis may be enough to validate existing characterisations. We conclude that biobanks characterising more worldwide diversity would benefit studies of complex traits.

Introduction

Understanding complex traits is one of the most important questions facing genetics as we progress into the Biobank era. The number of Single Nucleotide Polymorphisms (SNPs) that influence complex traits may vary from tens to thousands in human and non-human species (Goddard et al., 2016; de los Campos et al., 2018). The effect of each SNP on a trait is estimated using Genome Wide Association Studies (GWAS) in the very large biobanks and meta-analyses needed for statistical power. Because of the requirement for large sample sizes, almost everything that we know comes from studies in Eurasia in which these datasets are available; for example the UK Biobank (Bycroft et al., 2018), the China Kadoori Biobank (Chen et al., 2011), the Japanese Biobank (Kanai et al., 2018) and large GWAS consortia (Lee et al., 2018; Visscher et al., 2017). Yet, most selection acting on complex traits occurred primarily in our evolutionary history. How did the out-of-Africa bottleneck (Lipson and Reich, 2017) influence our quantification of complex traits?

Genomic architecture (Timpson et al., 2018) is a key tool for quantifying a complex trait. If a trait is under negative or balancing selection, then SNPs with a large effect are selected against, and reduced in frequency. Genomic (or Genetic) architecture quantifies the relationship between SNP frequency and the effect the SNP has on the trait (Eyre-Walker and Govindaraju, 2010). Many models (Speed et al., 2017; Zeng et al., 2018) contain an explicit parameter that we will denote S that describes this shape, and which is often linked to selection. $S = 0$ means that effect size and SNP frequency are unrelated. $S < 0$ means that rare SNPs have larger effect, and is expected if large effect SNPs are driven to low frequency by negative or balancing selection. Conversely, $S > 0$ implies that common SNPs have a larger effect, and is expected if selection increases the frequency of large effect SNPs via positive selection.

Genetic drift is a well-understood concept in population genetics (Kimura, 1983) and is well understood in a nearly-neutral context (Ohta, 1992) allowing for limited selection. Clearly, the genomic architecture representation as a conditional model describing the effect size, conditional on the SNP frequency, is incomplete. Whilst the allele frequency spectrum is related to selection (Tajima, 1989), a joint model is much more difficult, especially when ascertainment, linkage and other statistical artefacts are accounted for. Figure 1 illustrates how Genetic Drift and Complex Trait Genomic Architecture interact to change the whole SNP-frequency and effect size distribution.

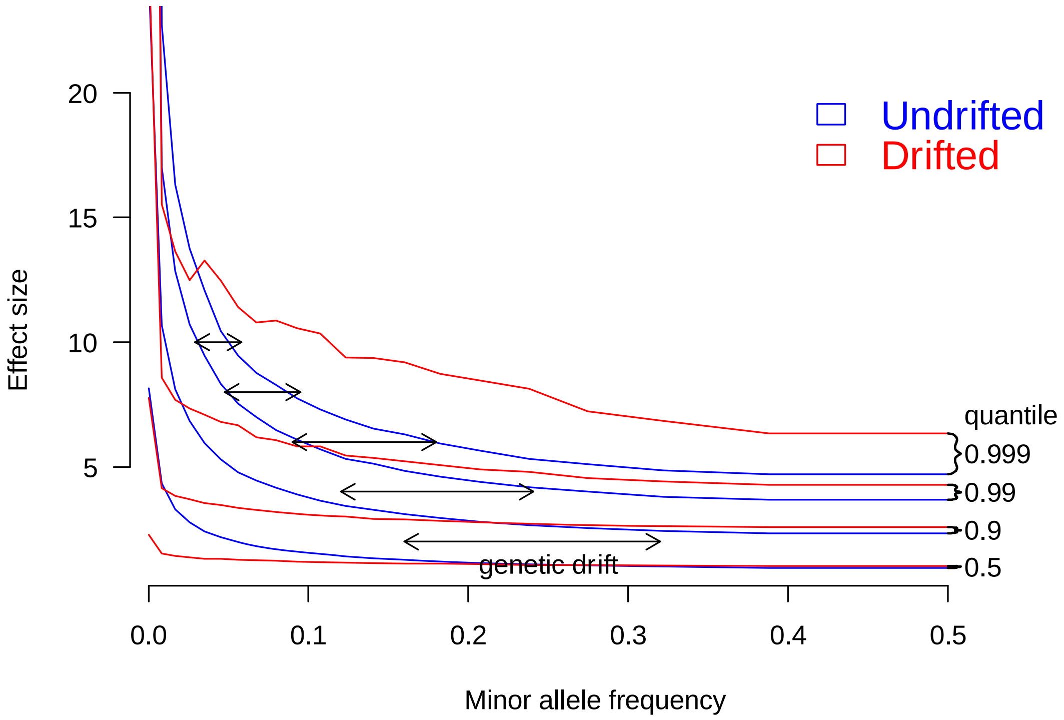


Figure 1 Simulation of Complex Trait Genomic Architecture with Genetic Drift. The Complex Trait has S=-1, meaning that most large effect alleles are very rare. The blue distribution shows quantiles of effect size in the population in which the trait evolved, conditional on frequency. Genetic drift (here, $F_{st} = 0.1$) changes the blue to the red distribution. Drift is larger for common SNPs with modest effect, so most rare SNPs either become a little more common, or go to fixation. The result is a much flatter distribution (e.g. the 0.5, 0.9, 0.99 quantiles) which resembles a smaller magnitude shape parameter S. However, the most extreme SNPs at a given frequency (q=0.999) arrive from lower frequency and hence have much larger effect. Whilst the red distribution cannot be exactly replicated by a different shape parameter S, it can be closely approximated if relatively few SNPs contribute to the complex trait.

We use a simulation approach to examine whether the out-of-Africa bottleneck should change the interpretation of parameters in the genomic architecture of complex traits. We find that unfortunately, Europeans, and any other non-African population have a rather different genomic architecture to the African population in which selection predominantly occurred. As a consequence, S cannot be understood as a direct quantification of selection, and indeed the value obtained depends on many things including any SNP-frequency thresholding performed in quality control. Models of genomic architecture that do not correct for drift are a useful description of the data, but further work is needed for inference about selection.

Results

Genomic architecture is changed by genetic drift

To assess the effect of genetic drift on genetic architecture we need a large sample of individuals from around the world, which is not currently available. To address this we resample data from the 1000 Genomes dataset (1000 Genomes Consortium, 2015) using HAPGEN2 (Su et al., 2011) to create realistic population structure complete with linkage disequilibrium between Africa, Europe, South Asia, East Asia, and America. We then simulate complex traits in the African population using GCTA (Yang et al., 2011) with a specified heritability (using $h^2 = 0.5$ throughout) and SNP frequency relationship S. Recall that $S < 0$ implies "negative selection" on the trait, and therefore high frequency SNPs can only have a small effect on the trait, whilst rare SNPs are permitted to have larger effect sizes.

We then generate genetic variability in each of our populations (Supplementary Figures 1-2), conditional on a SNP frequency threshold. By assuming a constant value for environmental variability, determined to be that required in Africans to give $h^2 = 0.5$ with the specified SNP frequency, we can compute an observed heritability h^2. We also report values computed with GCTB using --bayes S (Zeng et al., 2018).

The resulting heritability for simulated complex traits in African and other populations is given in Figure 2. Both our approach and GCTB agree that heritability in non-Africans is strongly biased by the bottleneck, and that the magnitude of this effect is a function of the simulated value of S. However, we observe that thresholding critically impacts the inferred heritability. If no thresholding is performed, the inferred h^2 is significantly larger than simulated, whilst if thresholding is strict, the inferred h^2 may be smaller.

This is a direct consequence of genetic drift changing allele frequencies independently of SNP effect size (Figure 2). Low frequency SNPs with large effects can become common, leading to an increased genetic variation of the trait (Supplementary Figures 1-2). This is precisely why bottlenecked populations including Ashkenazi Jews (Levy-Lahad et al., 1997), Finns (Cannon et al., 1998) and Icelanders (Lill et al., 2012) are used in GWAS studies for generally rare diseases that are common in those populations.

Environmental variation for real phenotypes varies due to factors including lifestyle, societal organisation, and so on. We report these heritability results to emphasise how important assumptions are in modelling. Of course, it is possible to scale the environmental variation with the genetic variation to ensure a desired heritability. This is merely a case of adding noise to phenotypes and will not affect any other inference.

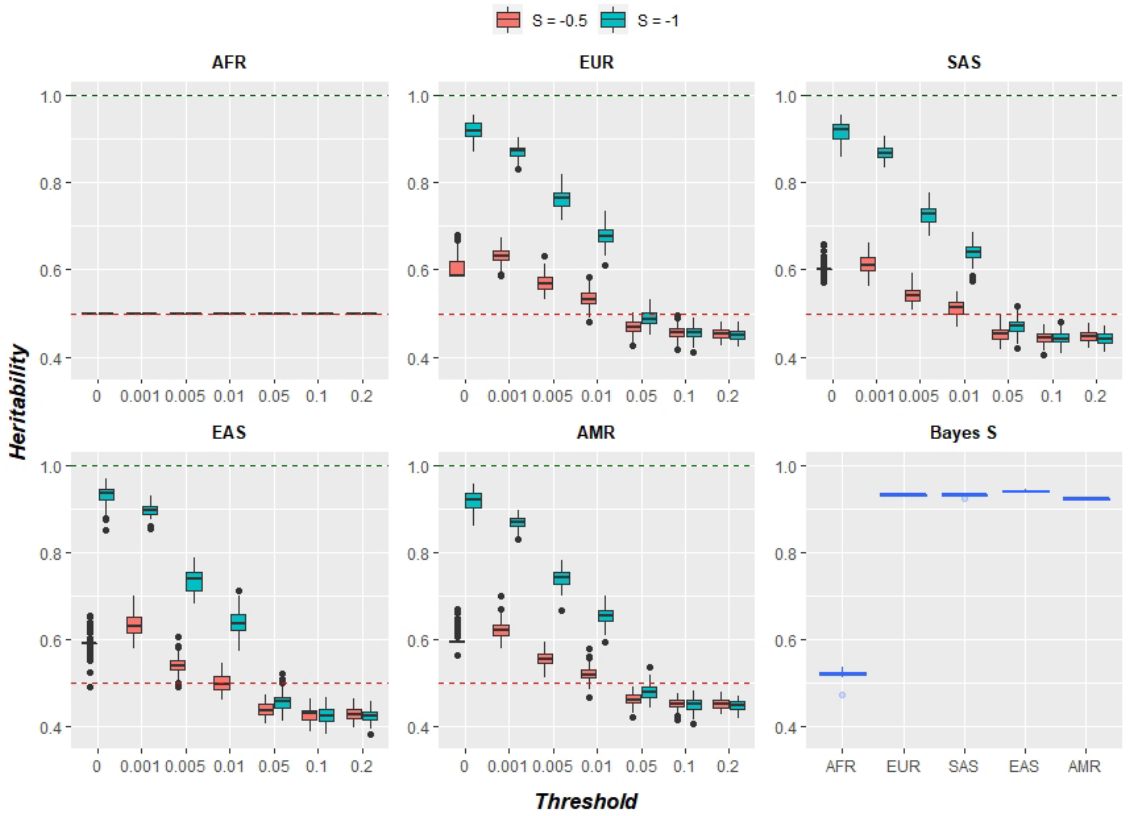


Figure 2: Estimates of heritability when a complex trait is simulated in 1000 Genomes Africans (AFR) with $h^2 = 0.5$ and observed in any other worldwide population, when environmental variability constant across all populations. Each plot shows observed heritability at different thresholds for SNP frequency, for a different population group at S = -0.5 and S = -1. The final plot (Bayes S panel) shows results from GCTB --bayes R (Moser et al., 2015) for S=-1, which agree with our unthresholded estimates .

Inferred selection is affected by genetic drift

We then asked whether the relationship between SNP frequency and effect size has been distorted by genetic drift, by estimating the selection coefficient S. For this we implemented the effect size β_i prior of (Zeng et al., 2018) (see Materials and Methods) conditional on the SNP frequency f_i:

$$\beta_i \sim N\left(0, \sigma_i^2\right),$$

$$\sigma_i^2 = \sigma_\beta^2 [f_i(1 - f_i)]^S.$$

We call this the "simple model" as it does not account for genetic drift. In (Zeng et al., 2018) this is a prior that affects effect size estimates; for our model this is a likelihood for the observed effect size, which we assume given. These would be taken from GWAS or, in simulations are taken from the simulation. This eliminates the estimation error that often dominates genomic architecture studies.

Figure 3 shows that S, like h^2, is biased by genetic drift. If no thresholding is performed, the inferred S is typically of larger magnitude than the true S in all drifted populations. When thresholding is strict, S tends towards the prior mean of 0, due to a lack of variation in the data. There is a transition around minor-allele-frequency of 0.05 where the biases cancel out. However, there is significant variability in the inferred S, due to the random nature of genetic drift and the sensitivity of the inference to the most extreme causal SNPs, since only 10000 are generated for each simulation.

Unlike for heritability, it is not clear how a simulation should be updated to maintain a desired S. The choice of environmental variation does not effect S as it is simply adding different amounts of noise to the phenotype. This is therefore a rather different sort of bias.

Critically however, the choice of thresholding does not affect inference in the population that experienced the selection; in our simulations this is Africa (AFR). In this population, accurate estimates of S are recovered for a range of thresholds (up to MAF 0.1, above which power is lost) which induced considerable bias in every other population. MAF thresholding is therefore a potential sensitivity analysis tool for the interpretation of S.

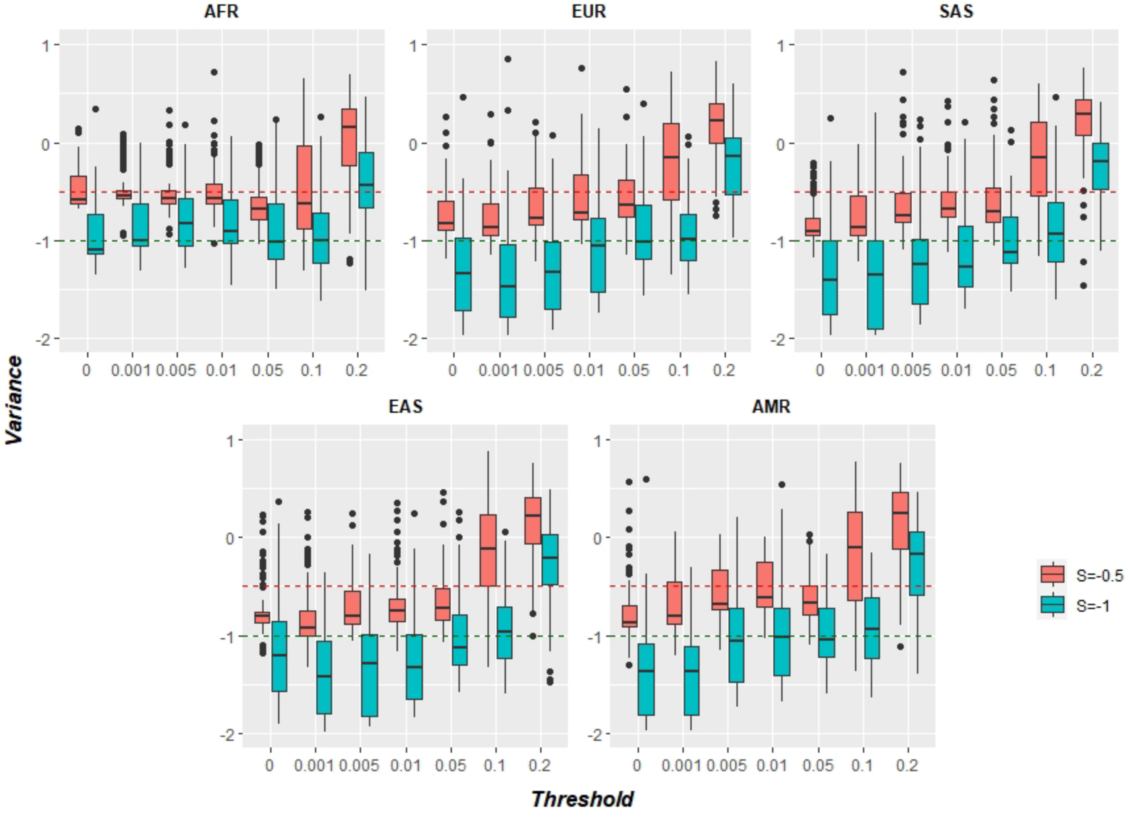


Figure 3 Inferred architecture parameter S with different thresholds for all 1000 Genomes population groups, using a simulated S = 0.5 and S = 1. The complex trait was simulated in Africans and inferred in the specified population using the "Simple model". See Methods for details.

Separating drift and selection

Bias in heritability and S are both natural consequences of genetic drift. To model genetic drift and hence recover the pre-drift values (see Materials and Methods) we allow for genetic drift in a "drift model", in which the drift process is represented using the Balding-Nichols model (Balding and Nichols, 1995).

We then applied the model to a simplified simulated dataset (see Methods) as shown in Figure 4a-b. The bias experienced by ignoring population structure is very large and grows with the genetic drift parameter F_{st}. For a given F_{st}, S is biased by a constant factor; i.e. the observed S is around half what it should be when F_{st} is around 0.1. When we perform inference with our "drift" model, we are able to accurately recover the true S that relates to selection in the pre-drifted population. We confirmed that these results apply also to the 1000 Genomes datasets (Figure 4c-d) discussed above, in which F_{st} is fixed and genetic drift has not occurred under our inference model.

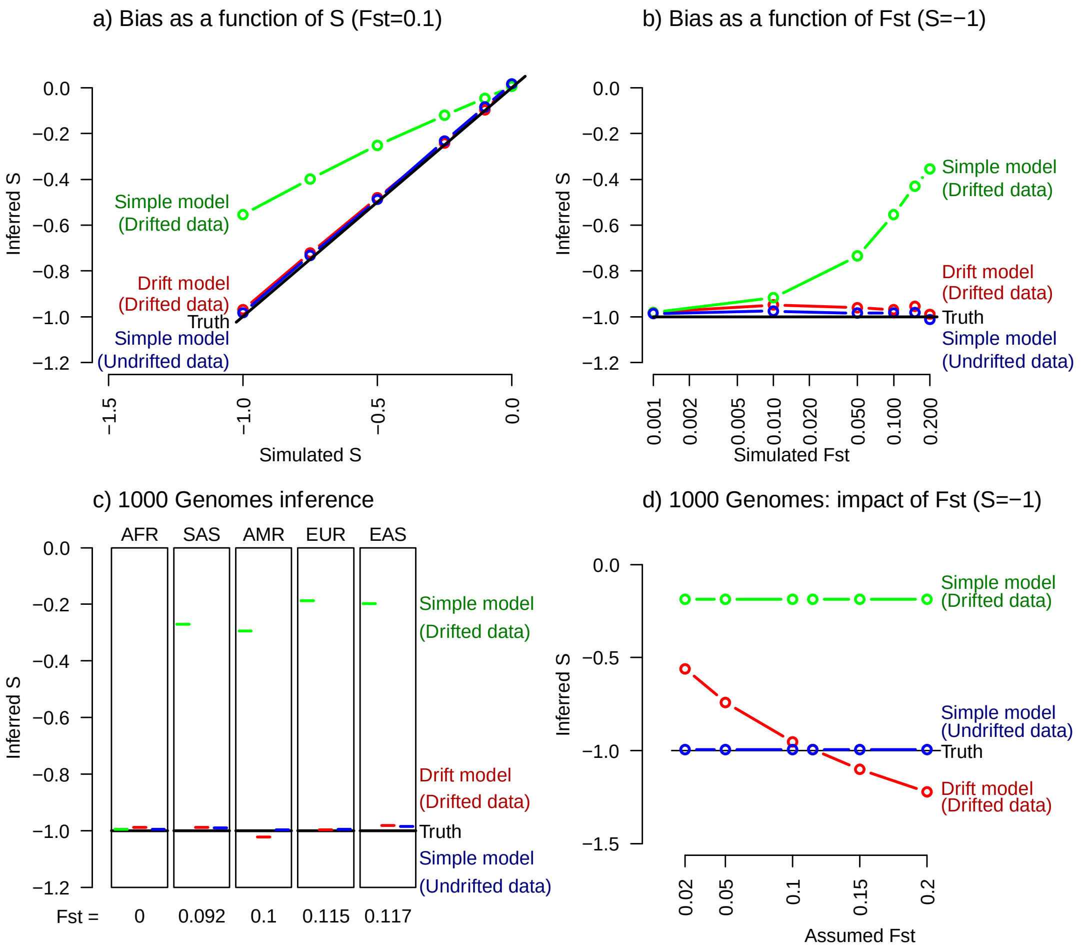


Figure 4: Drift aware inference of genetic architecture removes bias. a-b) Simulation of genetically drifted genetic data with a specific F_{st} and a phenotype with specific S leads to biased inference using the standard "Simple model", which is corrected by our "Drift model" (see Methods). a) The bias is linear in S, and b) grows rapidly with F_{st}. c-d) When the genetic data uses the 1000 Genomes haplotype structure with phenotypes generated in Africa, the bias is strong in all other populations, but is still corrected by our model. c) When F_{st} is well-estimated (see Methods) the inference is approximately unbiased. d) The estimate is sensitive to the estimate of F_{st}. (Plots show median and 90% credible sets for inference.)

Discussion

Selection occurred on most complex traits in the evolution of modern humans; that is, most selection will have acted on the African population prior to the out-of-Africa event that led to the peopling of Eurasia and beyond. This bottleneck led to considerable genetic drift in all non-Africans.

We demonstrated that a simple sensitivity analysis, that of performing inference at a range of minor-allele frequencies, can identify whether genetic drift has an influence on the inferences made on a particular complex trait. We then showed that correcting for genetic drift was possible and desirable, and provided a Bayesian inference algorithm for this. Whilst our implementation lacks the SNP selection component of established tools, our model can be directly used by performing SNP selection within other software, or software could be updated to allow more appropriate models. S is always a valid summary of a specific genomic architecture, but to link S to selection it is essential that sensitivity analysis or further modelling supports this interpretation.

Our model uses relatively little information and is not likely to reconstruct true allele frequencies from the past; it instead learns ancestral SNP frequencies that make the Complex Trait effect size distribution most plausible. It also does not implement inference of F_{st}, as it would be inconsistent to infer F_{st} on a trait-by-trait basis for the same SNP set. However, it is the case that F_{st} varies considerably between SNP sets and the F_{st} we observed across populations was low, which may be due to the relatively high frequency imposed on this during SNP selection.

Genome-Wide, F_{st} between Africans and Eurasians is high at ~ 0.2 (1000 Genomes Consortium, 2015); within Eurasians is moderate (~ 0.1 between Europe/China) and small within ancestry groups (~ 0.01 between North and South Europe). Yet the appropriate F_{st} from the ancestor of all humans is not completely clear. Diversity within Africa is extremely high (again $\sim 0.2 - 0.3$) (Henn et al., 2011). As larger datasets within Africa become available, we will need to establish whether selection has continued to operate effectively on complex traits, leading to unbiased estimates from these populations. If not, it may still be inappropriate to use a specific modern African population as a proxy for the ancestral population of modern humans. Despite this, African individuals who have not experienced the bottleneck will be essential in establishing the true genomic architecture of complex traits, as drift modelling alone will have limited power to infer the original SNP frequencies.

On Complex Traits whose variation is dominated by relatively few SNPs, it will be hard to separate genetic drift and selection. This leads to two independent avenues of further research. The first is to increase diversity of large-scale population studies and especially African ancestry, to access the genetic diversity that was lost in the Out-of-Africa bottleneck. The second is to develop multi-ethnic models of genomic architecture to account for population structure.

Materials and Methods

Data sets

The 1000 Genomes Project

We use the 1000 Genomes Project data for simulation analyses. The latest release phase 3, containing 84.4 million variants for 2504 individuals. Population groups in this data are African, European, South Asian, East Asian and American (1000 Genomes Consortium, 2015).

1000 Genomes data (genome wide) were pruned based on linkage disequilibrium. Variant pruning was done using PLINK 1.9 (Purcell et al., 2007) with command LD --indep-pairwise 200 10 0.07. After pruning 354,443 SNPs were retained. These SNPs were further passed to HAPGEN2 (Su et al., 2011) to simulate 10,000 individuals from each population. Ultimately, the data set for analysis was 10,000 number of individuals, 354,443 SNPs for five population groups were available for further analysis.

Complex trait simulation

We generate a random complex trait by selecting 20 sets of 10,000 random SNPs (causal SNP list), using African SNP frequencies and use GCTA (Yang et al., 2011) with $h^2 = 0.5$ to generate effect sizes with varied values of S. To generate phenotypes in other populations, we compute the environmental variation that was present in the African population given a particular SNP effect size threshold. We then apply the same threshold to the non-African population (which may retain different SNPs above the threshold) as was applied within the African population and generate phenotypes by adding the genetic and environmental variability. Narrow sense Heritability was calculated as: Vg/(Vg + Ve) (Falconer, 1996).

We repeated this process 100 times and considered each threshold for the same 100 random complex traits.

Bayesian model for Genomic Architecture with drift

We created a novel MCMC algorithm in Stan (Carpenter et al., 2017) (mc-stan.org) using the Rstan interface.

Model 0 is the baseline model which is an implementation of the BayesS model in which there are no SNPs that do not affect the trait, because we know which these are. Model 0 can be written for each SNP $i = 1..L$ for the observed frequency f_i and observed effect size β_i:

$$S \sim U(-2,2),$$

$$\sigma_\beta \sim U(0,2),$$

$$\beta_i \sim N\left(0, \sigma_\beta^2 [f_i(1-f_i)]^S\right).$$

The "drift model" is an extension accounting for genetic drift. It follows Model 0, except that we simulate the complex trait in a "pre-drifted population". SNP frequencies in this population is $p_i$which generates the "drifted data" frequency f_i using the Baldings-Nichols model (Balding and Nichols, 1995) to represent drift using the "Fixation Index" F_{st}, treated as known. This leads to:

$$f_i \sim Beta\left(p_i \frac{(1-F_{st}\ )}{F_{st}}, (1-p_i)\frac{(1-F_{st}\ )}{F_{st}}\right),$$

$$\beta_i \sim N\left(0, \sigma_\beta^2 [p_i(1-p_i)]^S\right).$$

Here, Normal distributions are specified via (mean, variance) and the Beta distribution is specified as $Beta(\alpha, \beta)$ defined in terms of shape and scale parameters with expectation $\alpha/(\alpha+\beta)$. Therefore f_i has expectation $\mathbb{E}(f_i) = p_i$, and variance $\mathrm{Var}(f_i) = F_{st}\ p_i(1-p_i)$.

When F_{st} is known (Figure 4a-b) this is provided to the model. When F_{st} is unknown, we estimate it on our dataset using plink1.9 (www.cog-genomics.org/plink/1.9/) (Chang et al., 2015) using "--fst –within", providing only the individuals belonging to the two populations being compared.

For 1000 Genomes inference, only SNPs with frequency >0.01 in Africans are considered. All SNPs are considered without thresholding in the drifted population, except for those that have reached fixation which are omitted as they have zero probability under the likelihood.

Simulation model for Figure 4a-b

We created a simulation model that could characterise our model rapidly without going through the 1000 Genomes data, hence providing a faster simulation that could generate a range of simulated F_{st}values. We choose a value of S and F_{st} and then simulate data from the "drift model" with a specified L (=10000 throughout). We also threshold minor-allele frequency to 0.01, i.e. in the inference model, any frequency less than 0.01 is treated as 0.01.

Acknowledgements

DJL and BA were funded by the Wellcome Trust under grant number WT104125MA.

Code availability

The code necessary to replicate the results presented here are given at https://github.com/danjlawson/genomicarchitecture.

References

1000 Genomes Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

Balding, D.J., Nichols, R.A., 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96, 3–12. https://doi.org/10.1007/BF01441146

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., Marchini, J., 2018. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209. https://doi.org/10.1038/s41586-018-0579-z

Cannon, T.D., Kaprio, J., Lönnqvist, J., Huttunen, M., Koskenvuo, M., 1998. The Genetic Epidemiology of Schizophrenia in a Finnish Twin Cohort: A Population-Based Modeling Study. Arch. Gen. Psychiatry 55, 67–74. https://doi.org/10.1001/archpsyc.55.1.67

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A Probabilistic Programming Language. J. Stat. Softw. 76, 1–32. https://doi.org/10.18637/jss.v076.i01

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4. https://doi.org/10.1186/s13742-015-0047-8

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., 2011. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int. J. Epidemiol. 40, 1652–1666. https://doi.org/10.1093/ije/dyr120

de los Campos, G., Vazquez, A.I., Hsu, S., Lello, L., 2018. Complex-Trait Prediction in the Era of Big Data. Trends Genet. TIG 34, 746–754. https://doi.org/10.1016/j.tig.2018.07.004

Eyre-Walker, A., Govindaraju, D.R., 2010. Genetic Architecture of a Complex Trait and Its Implications for Fitness and Genome-Wide Association Studies. Proc. Natl. Acad. Sci. U. S. A. 107, 1752–1756.

Falconer, D.S., 1996. Introduction to quantitative genetics. Prentice Hall, Harlow, England.

Goddard, M.E., Kemper, K.E., MacLeod, I.M., Chamberlain, A.J., Hayes, B.J., 2016. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proc. R. Soc. B Biol. Sci. 283. https://doi.org/10.1098/rspb.2016.0569

Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A.A., Underhill, P.A., Comas, D., Kidd, K.K., Norman, P.J., Parham, P., Bustamante, C.D., Mountain, J.L., Feldman, M.W., 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. 108, 5154–5162. https://doi.org/10.1073/pnas.1017511108

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., Kubo, M., Okada, Y., Kamatani, Y., 2018. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. 50, 390–400. https://doi.org/10.1038/s41588-018-0047-6

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Linnér, R.K., Fontana, M.A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P.N., Walters, R.K., Willoughby, E.A., Yengo, L., Alver, M., Bao, Y., Clark, D.W., Day, F.R., Furlotte, N.A., Joshi, P.K., Kemper, K.E., Kleinman, A., Langenberg, C., Mägi, R., Trampush, J.W., Verma, S.S., Wu, Y., Lam, M., Zhao, J.H., Zheng, Z., Boardman, J.D., Campbell, H., Freese, J., Harris, K.M., Hayward, C., Herd, P., Kumari, M., Lencz, T., Luan, J., Malhotra, A.K., Metspalu, A., Milani, L., Ong, K.K., Perry, J.R.B., Porteous, D.J., Ritchie, M.D., Smart, M.C., Smith, B.H., Tung, J.Y., Wareham, N.J., Wilson, J.F., Beauchamp, J.P., Conley, D.C., Esko, T., Lehrer, S.F., Magnusson, P.K.E., Oskarsson, S., Pers, T.H., Robinson, M.R., Thom, K., Watson, C., Chabris, C.F., Meyer, M.N., Laibson, D.I., Yang, J., Johannesson, M., Koellinger, P.D., Turley, P., Visscher, P.M., Benjamin, D.J., Cesarini, D., 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. 50, 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., Shohat, M., Weber, B.L., Beller, U., Lahad, A., Halle, D., 1997. Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. Am. J. Hum. Genet. 60, 1059–1067.

Lill, C.M., Roehr, J.T., McQueen, M.B., Kavvoura, F.K., Bagade, S., Schjeide, B.-M.M., Schjeide, L.M., Meissner, E., Zauft, U., Allen, N.C., Liu, T., Schilling, M., Anderson, K.J., Beecham, G., Berg, D., Biernacka, J.M., Brice, A., DeStefano, A.L., Do, C.B., Eriksson, N., Factor, S.A., Farrer, M.J., Foroud, T., Gasser, T., Hamza, T., Hardy, J.A., Heutink, P., Hill-Burns, E.M., Klein, C., Latourelle, J.C., Maraganore, D.M., Martin, E.R., Martinez, M., Myers, R.H., Nalls, M.A., Pankratz, N., Payami, H., Satake, W., Scott, W.K., Sharma, M., Singleton, A.B., Stefansson, K., Toda, T., Tung, J.Y., Vance, J., Wood, N.W., Zabetian, C.P., 23andMe, T.G.E. of P.D. (GEO-P.C., Consortium (IPDGC), T.I.P.D.G., Consortium, T.P.D.G., Consortium 2 (WTCCC2), T.W.T.C.C., Young, P., Tanzi, R.E., Khoury, M.J., Zipp, F., Lehrach, H., Ioannidis, J.P.A., Bertram, L., 2012. Comprehensive Research Synopsis and Systematic Meta-Analyses in Parkinson's Disease Genetics: The PDGene Database. PLOS Genet. 8, e1002548. https://doi.org/10.1371/journal.pgen.1002548

Lipson, M., Reich, D., 2017. A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. Mol. Biol. Evol. 34, 889–902. https://doi.org/10.1093/molbev/msw293

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., Visscher, P.M., 2015. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLOS Genet. 11, e1004969. https://doi.org/10.1371/journal.pgen.1004969

Ohta, T., 1992. The Nearly Neutral Theory of Molecular Evolution. Annu. Rev. Ecol. Syst. 23, 263–286.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575.

Speed, D., Cai, N., Consortium, the U., Johnson, M.R., Nejentsev, S., Balding, D.J., 2017. Reevaluation of SNP heritability in complex human traits. Nat. Genet. 49, 986–992. https://doi.org/10.1038/ng.3865

Su, Z., Marchini, J., Donnelly, P., 2011. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics 27, 2304–2305. https://doi.org/10.1093/bioinformatics/btr341

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

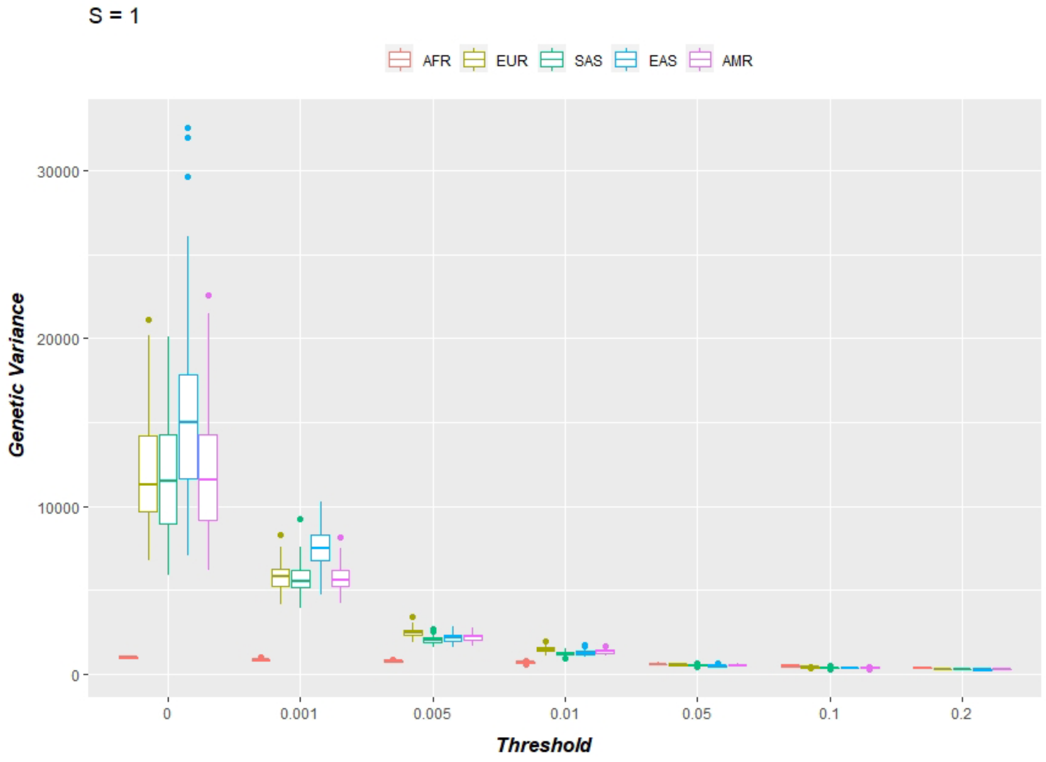
Timpson, N.J., Greenwood, C.M.T., Soranzo, N., Lawson, D.J., Richards, J.B., 2018. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat. Rev. Genet. 19, 110–124. https://doi.org/10.1038/nrg.2017.101

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J., 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

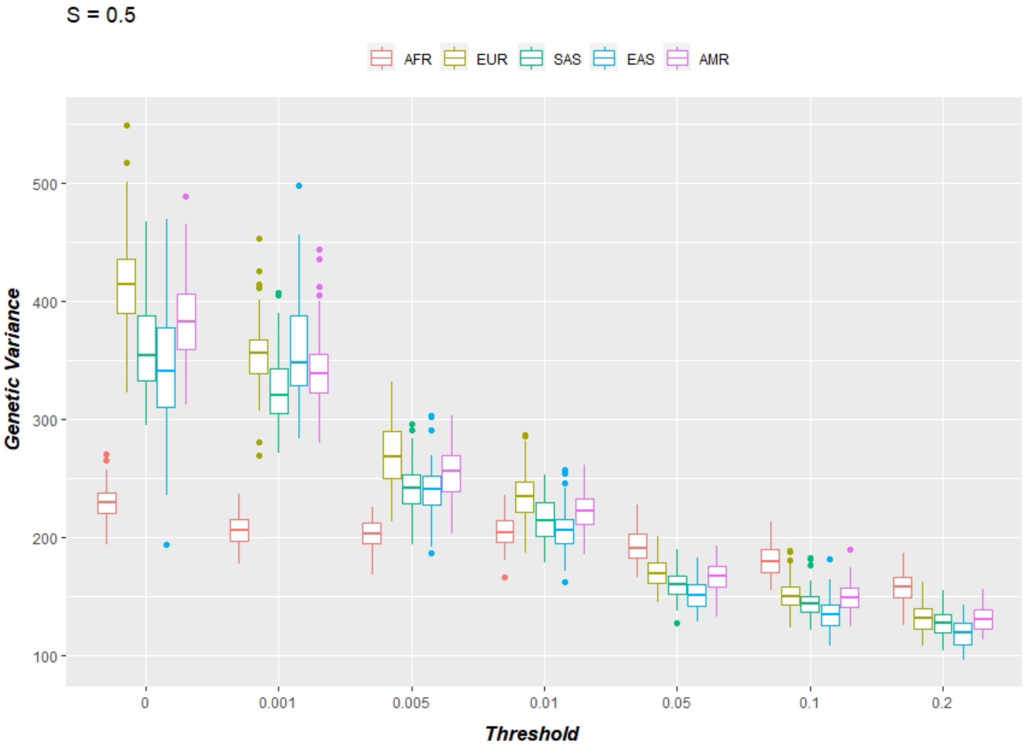
Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am. J. Hum. Genet. 88, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zeng, J., Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., Powell, J.E., Montgomery, G.W., Metspalu, A., Esko, T., Gibson, G., Wray, N.R., Visscher, P.M., Yang, J., 2018. Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. 50, 746–753. https://doi.org/10.1038/s41588-018-0101-4

Supplementary Figures



Supplementary Figure 1. Estimates of genetic variance with different thresholds for all population groups at S = 1.



Supplementary Figure 2 Estimates of genetic variance with different thresholds for all population groups at S = 0.5.