Ascertainment bias can create the illusion of genetic health disparities

Michelle S Kim[1], Kane P Patel[1], Andrew K Teng[1], and Ali J Berens[1],

and Joseph Lachance[1]*

[1]School of Biological Sciences, Georgia Institute of Technology

*Corresponding author

Contact information:

Joseph Lachance
950 Atlantic Dr.
Atlanta, GA 30332
joseph.lachance@biology.gatech.edu

Short title: Ascertainment bias and genetic disease risks

Keywords: ascertainment bias, genetic risk scores, genetic epidemiology, genome-wide association studies, global health, health disparities, population genetics

Abstract

Accurate assessment of health disparities requires unbiased knowledge of genetic risks in different populations. Unfortunately, most genetic association studies use genotyping arrays and European samples. Here, we integrate whole genome sequence data, GWAS results, and computer simulations to examine how ascertainment bias causes disease risks to be mis-inferred in non-study populations. We find that genetic disease risks are substantially overestimated for individuals with African ancestry – risk allele frequencies at known disease loci are 1.15% higher on average in Africa. These patterns hold for multiple disease classes (e.g., cancer, gastrointestinal, morphological, and neurological diseases). A contributing factor to this bias is that existing genotyping arrays are enriched for SNPs that have higher frequencies of ancestral alleles in Africa. Computer simulations of GWAS that use samples from bottlenecked non-African populations recapitulate regional differences in allele frequencies at disease susceptibility loci. These differences cause genetic disease risks to be overestimated for individuals with African ancestry and underestimated for individuals with non-African ancestry. We find that the extent of ascertainment bias depends on the genotyping platform used, numbers of cases and controls, demographic history, the proportion of ancestral vs. derived risk alleles, and choice of study population (African GWAS are less biased). Importantly, biases are only moderately reduced if GWAS use whole genome sequences and hundreds of thousands of cases and controls. Our results indicate that caution must be taken when using GWAS results from one population to predict disease risks in another population.

Introduction

In the past decade, over 3,000 genome-wide association studies (GWAS) have successfully identified more than 39,000 genetic variants that are associated with common diseases and other traits [1, 2]. Most GWAS use genotyping arrays to test whether specific risk alleles are more common in cases vs. controls. However, the vast majority of published GWAS have used samples of European ancestry [3, 4], and a looming challenge is to be able to generalize GWAS results across populations [5-10]. Results from GWAS can be combined to generate polygenic risk scores to predict individual risks of disease [11-13]. These polygenic risk scores involve summing the number of risk alleles in each individual's genome to quantify hereditary disease burdens. Further refinement of genetic risk scores involves weighting SNPs by effect size [14]. Additional complications for genetic risk scores include the “missing heritability” problem [15] , which implies that the bulk of causal variants remain undiscovered. Diseases can also have different genetic architectures in different populations [16]. Because of these issues, genetic predictions of disease risk are not always accurate, and it is important to be able to distinguish between situations where genetic risks actually differ between populations and when predictions of genetic health disparities are spurious.

Genetic health disparities can arise when allele frequencies at disease-associated loci differ across populations [14]. These allele frequency differences are magnified for pairs of populations that do not share recent evolutionary history [17, 18]. Population bottlenecks and founder effects have influenced hereditary disease risks in a number of global populations. Many of these effects are disease-specific, such as elevated risks of cystic fibrosis among the Québécois [19] and cardiovascular disease among the descendants of the HMS Bounty mutineers [20]. Evolutionary history also affects whether there are genetic differences in disease risks across populations, including recent natural selection near disease susceptibility loci [21] and whether risk alleles are ancestral (shared with other primates) or derived (due to new mutations) [22, 23]. Although risks of individual diseases can differ across populations, the overall burden of hereditary diseases is expected to be similar across the globe [24]. Systematic departures from this null expectation may arise because many disease alleles are presently unknown.

Even if real differences exist between populations, SNP ascertainment bias can cause genetic disease risks to be misestimated. There are multiple sources of bias in GWAS, including choice of genotyping technology, the ancestry of study participants, and whether sample sizes are large or small [25-28]. Most commercially available genotyping arrays use SNPs that were originally ascertained in European populations, and arrays are enriched for intermediate frequency alleles (i.e., alleles with frequencies that are closer to 50%) [25]. Because of this, allele frequencies at presently known disease loci are not independent of the genotyping technology used to detect genetic associations. As of 2016, the ancestry of 81% of all GWAS samples was European and 14% was Asian [3], and this is likely to cause the set of known disease associations to be enriched for alleles that are polymorphic or intermediate frequency in Europe or Asia, but not Africa. There is also evidence that disease-associated alleles have elevated minor allele frequencies in study populations [5]. Biases in genetic studies parallel what is observed in social science research: most samples are from Western, educated, industrialized, rich and democratic (WEIRD) societies [29, 30]. An additional consideration is that large sample sizes are required to detect associations between SNPs and genetic diseases when risk alleles have small effect sizes or are rare [31]. Because of this, ascertainment bias is more problematic for GWAS that have small numbers of cases and controls.

At present, the extent to which ascertainment bias hinders precision medicine and personal genomics is unknown. To bridge this knowledge gap, we tested empirical data for systematic bias in risk allele frequencies across populations. Extensive computer simulations of GWAS were then used to provide insight into multiple causes of what appears to be a genetic health disparity (including the effects of different genotyping arrays, study designs, mode of inheritance, and evolutionary histories). Here, we focus on the problem of using disease associations discovered in one population to predict disease risks in another population, as opposed to whether GWAS findings can be successfully replicated across multiple populations.

Results

Empirical patterns of genetic risk

Allele frequencies at 3036 disease-associated loci were analyzed for each continental super-population in the 1000 Genomes Project dataset. Contrary to null expectations, the mean frequencies of risk alleles at disease susceptibility loci vary across populations (**Fig. 1A**). Specifically, the overall risk allele frequencies are significantly higher in African populations compared to non-African populations (mean difference: +1.15%, p-value = 0.02129, paired Wilcoxon signed-rank test). However, what appear to be genetic health disparities (elevated risk allele frequencies in Africa) are due to SNP ascertainment bias.

We explored differences in risk allele frequencies by binning each disease-associated locus into one of seven different categories: gastrointestinal (GI) or liver, metabolic, morphological, cancer, neurological, miscellaneous, and cardiovascular disease. As illustrated in **Fig. 1A**, population-level differences in risk allele frequencies persist when GWAS results were binned by disease type. Compared to other populations, African populations have the highest risk allele frequency in five out of seven disease types: metabolic (p-value = 0.005502), morphological (p-value = 0.09494), cancer (p-value = 0.1169), neurological (p-value = 0.0995), and miscellaneous disease (p-value = 0.3865, paired Wilcoxon signed-rank tests). African populations have intermediate frequencies of risk alleles at loci that are associated with GI or liver diseases (p-value = 0.6965), and lower frequencies of risk alleles at loci that are associated with cardiovascular disease (p-value = 0.01404, paired Wilcoxon signed-rank tests). Among non-African populations there was no underlying trend.

Further stratification according to ancestral vs. derived status reveals a clear pattern: disease types that have a larger proportion of ancestral alleles tend to have elevated risk allele frequencies in Africa (**Fig. 1B**). After binning GWAS SNPs by disease category, we find that the differences in the mean frequency of risk alleles between African and non-African populations are highly correlated with the proportion of risk alleles that are ancestral (r^2 = 0.842). This suggests that continental patterns of disease risk may vary for risk alleles that are ancestral vs. derived.

The joint site frequency spectrum (SFS) of risk alleles in African and non-African populations provides empirical evidence of SNP ascertainment bias (**Fig. 2**). In this study, we focused on unfolded allele frequencies, rather than minor allele frequencies. In general, ancestral risk alleles tend to be the major allele and derived risk alleles tend to be the minor allele. This is expected given that derived alleles are, by definition, evolutionarily younger than ancestral alleles. 69.2% of the ancestral risk alleles are found at higher frequency in African populations (below the diagonal), and 64.5% of the derived risk alleles are found at higher frequency in non-African populations (above the diagonal). The null expectation is that equal numbers of alleles would be found on each side of the diagonal. Examining the borders of the joint frequency spectrum between African and non-African populations emphasizes the effects of study populations in GWAS. Many disease-associated alleles are found at extreme allele frequencies in Africa (close to 0 or 1) and at intermediate allele frequencies outside of Africa. This occurs because most GWAS have used non-African samples and statistical power is maximized at intermediate frequencies.

The difference in risk allele frequencies between African and non-African populations is expected to be zero when bias is absent. Conditioning on whether risk alleles are ancestral or derived reveals a striking pattern: ancestral risk alleles are found at much higher frequencies in Africa and derived risk alleles are found at much lower frequencies in Africa (**Fig. 2B**). The mean difference in ancestral risk allele frequencies between African and non-African populations is +9.51%, and the mean difference in derived risk allele frequencies between African and non-African populations is -5.40% (p-value < 2.2×10^{-16} for both comparisons, Wilcoxon signed-rank tests). The overall continental difference in risk allele frequencies of +1.15% arises because 44% of presently known disease-associated SNPs have ancestral risk alleles and 56% of disease-associated SNPs have derived risk alleles.

Because many disease-associations involve imputed SNPs, we tested whether continental differences in risk allele frequencies persist for SNPs that are not on the Affymetrix Genome-Wide Human SNP 6.0 Array. For this set of disease-associated loci, we find that SNPs with ancestral risk alleles have higher allele frequencies in Africa (+8.63% on average) and that SNPs with derived risk alleles have lower allele frequencies in Africa (-4.83% on average). This suggests that biases persist even for imputed SNPs.

Genotyping arrays are biased

One potential source of bias is genotyping platform: GWAS use microarrays with pre-ascertained SNPs. Whole genome sequencing (WGS) data from the 1000 Genomes Project reveals that each population has a similar mean derived allele frequency (**Fig. 3A**). This is expected since all human populations share the same evolutionary distance to chimpanzees. Compared to WGS data, derived allele frequencies are elevated for SNPs on the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Omni 5M microarray. However, commonly used genotyping arrays also exhibit continental patterns of bias: derived allele frequencies in African populations are markedly lower than derived allele frequencies in non-African populations (p-value < $2.2\text{x}10^{-16}$ for both arrays, Wilcoxon signed-rank tests). This bias is due to the fact that genotyping arrays contain SNPs that were ascertained in non-African populations. WGS data have an unbiased SFS with similar numbers of SNPs above and below the diagonal (**Fig. 3B**). By contrast, the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Omni 5M microarray are enriched for SNPs that are above the diagonal, i.e. SNPs with higher derived allele frequencies outside of Africa (**Fig. 3C** and **Fig. 3D**). This pattern mirrors what is seen for empirical GWAS data (**Fig. 2A**), which suggests that genotyping arrays contribute to continental differences in risk allele frequencies.

Simulated GWAS capture the effects of bias

Using computer simulations, we set out to test whether ascertainment bias is sufficient to explain observed patterns at disease-associated loci. Simulations use allele frequency data from the 1000 Genomes Project, knowledge of which SNPs are on genotyping arrays, and GWAS power calculations [31]. Importantly, these simulations do not assume that there were any underlying differences in hereditary disease risks across populations (i.e. simulated differences in risk allele frequencies are due to ascertainment bias). Results from computer simulations are similar to what is observed in empirical data: compared to non-African populations, African populations have elevated frequencies of ancestral risk alleles and reduced frequencies of derived risk alleles (**Fig. 4**). Note that that empirical risk alleles have been discovered in a heterogeneous set of studies. By varying the parameters of GWAS simulations we are able to quantify individual effects of each potential source of ascertainment bias (study population, genotyping technology, sample size, and the dominance of disease alleles).

Choice of study population has a profound effect on the relative frequencies of risk alleles in different populations. Simulated GWAS that use African (AFR) samples yield similar risk allele frequencies across each of the five continental super-populations. However, simulated GWAS that use American (AMR), East Asian (EAS), European (EUR), or South Asian (SAS) samples produce a set of disease-associated loci with elevated frequencies of ancestral risk alleles and reduced frequencies of derived risk alleles in Africa (**Fig. 4A**). The magnitudes of these differences in allele frequencies are comparable to what is observed in empirical GWAS data. Regardless of study population, risk allele frequencies are similar for each non-African population, and this may be due in part to the relatively recent divergence times between these populations. Because statistical power is maximized at intermediate allele frequencies, mean risk allele frequencies in study populations are shifted closer towards 50%. We note that simulated GWAS that use a mixture of samples from different continents (MIX) still produce a set of disease-associated loci with elevated frequencies of ancestral risk alleles and reduced frequencies of derived risk alleles in Africa. Similarly, simulated GWAS that use admixed American (AMR) samples yield biased allele frequencies. Taken together, these results suggest that pooling samples with different ancestries is unlikely to alleviate the problem of SNP ascertainment bias.

Although genotyping arrays contribute to ascertainment bias, GWAS simulations reveal that biases in risk allele frequencies persist even if whole genome sequences are used. Recall that empirical GWAS data come from heterogeneous set of studies, while simulated results assume a single study design and effect size. Despite this, allele frequency differences between Africa and Europe are similar for real and simulated data (**Table 1**). Disease associations from simulations of European GWAS yield similar results for the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Omni 5M microarray (ancestral risk allele frequencies were 10.7% and 11.0% higher in Africa and derived risk alleles were 8.0% and 8.2% higher in Europe, respectively). Somewhat surprisingly, disparities in allele frequencies also occur for European GWAS simulations that use whole genome sequences (**Fig. 4B**). However, continental differences in allele frequencies were reduced for simulations that used whole genome sequences (ancestral risk allele frequencies were 9.7% higher in Africa and derived risk alleles were 7.2% higher in Europe). The fact that allele frequency differences arise from WGS simulations lends additional support to the claim that biases will persist for imputed SNPs.

Table 1

Data type	Risk allele frequency difference	
	Ancestral	Derived
NHGRI-EBI GWAS Catalog (empirical)	+11.7%	-6.7%
Affymetrix Genome-Wide Human SNP Array 6.0 (simulated)	+10.7%	-8.0%
Illumina Omni 5M microarray (simulated)	+11.0%	-8.2%
Whole genome sequences (simulated)	+9.7%	-7.2%

Table 1. Differences in allele frequencies between African and European populations for different genotyping technologies. Simulation parameters: sample size = 3500 cases and 3500 controls, study population = EUR, p-value threshold = 1x10^{-5}, mode of inheritance = additive, prevalence = 0.1, genotype relative risk = 1.211.

Continental biases in risk allele frequencies occur even if GWAS use large sample sizes. Simulated GWAS with less than 10,000 European cases and controls yield large differences in African and non-African allele frequencies (**Fig. 4B**). This occurs regardless of whether simulations use SNPs from the Affymetrix Genome-Wide Human SNP Array 6.0 or WGS. We find that well-powered studies with hundreds and thousands of cases and controls still results in notable differences in continental allele frequencies. There are diminishing returns for increasing sample sizes if simulated GWAS use genotyping arrays. By contrast, whole genome sequencing of one million cases and controls minimizes the amount of bias. Statistical power is also a function of the p-value threshold used in a GWAS. Holding the default parameter values constant, we find that using a more stringent p-value threshold amplifies risk allele frequency differences; ancestral risk allele frequencies are 12.2% higher in Africa and derived risk allele frequencies that are 8.8% higher in Europe if a p-value threshold of $5x10^{-8}$ is used.

Although we focused on additive effects, continental biases in risk allele frequencies vary for other modes of inheritance. It is easier to detect associations for low frequency dominant alleles, intermediate frequency additive alleles, and high frequency recessive alleles. However, the power to detect a genetic association does not solely depend on minor allele frequency (e.g. disease-causing alleles at 10% and 90% have a different chance of being successfully detected) [31]. Using simulations of European GWAS, we find that African risk allele frequencies are expected to be higher than European risk allele frequencies for dominant models of disease and lower than European risk allele frequencies for recessive models of disease (**Table 2**). These trends occur whether risk alleles are ancestral (dominant: +19.7%, recessive: +2.9%) or derived (recessive: -2.2%, recessive: -17.9%).

Table 2

Mode of inheritance	Risk allele frequency difference	
	Ancestral	Derived
Dominant	+19.7%	-2.2%
Additive	+10.7%	-8.0%
Recessive	+2.9%	-17.9%

Table 2. GWAS simulations reveal that risk allele frequency differences between African and European populations depend upon whether disease alleles are dominant or recessive. Simulation parameters: technology = Affymetrix Genome-Wide Human SNP Array 6.0, study population = EUR, sample size = 3500 cases and 3500 controls, study population = EUR, p-value threshold = $1x10^{-5}$, prevalence = 0.1, genotype relative risk = 1.211.

Discussion

SNP ascertainment bias confounds GWAS results and creates the illusion of genetic health disparities. Specifically, African populations tend to have higher frequencies of ancestral risk alleles and lower frequencies of derived risk alleles at existing GWAS loci. Taking into account the magnitude of these differences and the proportion of ancestral alleles in GWAS results yields risk allele frequencies that are 1.15% higher in Africa. This has important implications with respect to precision medicine and personal genomics: disease risks are likely to be misestimated if GWAS results are naively used to calculate genetic risk scores. Biased predictions of genetic risks are expected to be magnified for individuals of African descent, potentially complicating existing health disparities that are due to socio-cultural factors including access to medical care [32, 33].

Importantly, elevated risk allele frequencies in African populations are the opposite of what one expects to see given what is known about human demographic history. Natural selection is more efficient at purging deleterious variants when population sizes are large[34], and an important difference between African and non-African populations is that the latter have been subjected to multiple bottlenecks and founder effects following the out-of-Africa migration. Because of this, non-African genomes carry an excess load of homozygous deleterious alleles (as identified via GERP scores) [35]. By contrast, geographic patterns at known disease-associated loci differ by continent (**Fig. 1**), and this is due in part to SNP ascertainment bias.

The effects of different study populations are asymmetric. For example, if a GWAS uses European samples, allele frequencies at disease-associated loci will be similar across non-African populations and different for Africa (**Fig. 4**). By contrast, risk allele frequencies from African GWAS are relatively similar across all global populations. Because successful detection of a SNP-disease association requires that a causal locus is polymorphic in the study population [31, 36], bottlenecks and founder effects can contribute to the illusion of genetic health disparities. Consider a disease-causing allele that is initially found at the same frequency in two populations (i.e. prior to the divergence of these populations). Over time, genetic drift causes allele frequencies at this locus to change in each daughter population

(**Fig. 5A**). Importantly, non-African populations have experienced a history of population bottlenecks, including a drastic reduction in population size during the out-of-Africa migration [37], and there is a greater chance that non-African populations will have allele frequencies that are either 0 or 1. Note that derived alleles tend to be low frequency and ancestral alleles tend to be high frequency [22]. African GWAS result in minimal bias and non-African GWAS result in an excess of ancestral risk alleles with elevated allele frequencies in Africa and an excess of derived risk alleles with elevated allele frequencies outside of Africa (compare **Fig. 5B** and **Fig. 5C**). Biases in genetic predictions of disease risk depend upon historical population sizes and divergence times.

Although whole genome sequencing can identify many genetic variants that are missing from genotyping arrays, many of these variants are rare and population-specific (**Fig. 3B**). Because of this, disease-associations that use WGS data need not generalize well to other populations. We find that continental biases in risk allele frequencies persist even if GWAS use whole genome sequences and hundreds of thousands of cases and controls (**Fig, 4B**). This has important implications for genetic risk score calculations: estimates of disease risk depend upon the population(s) in which disease-associations were originally discovered, regardless of whether WGS data were used.

Going forward, there are multiple ways to extend the benefits of precision medicine and personal genomics to a wide range of global populations. One option is to replicate every existing GWAS in as many populations as possible. However, this option has limited feasibility: even if sufficient funds and epidemiological resources are available, it is not always possible to obtain large sample sizes for each population. Instead, genetic risk scores can correct for SNP ascertainment bias. This requires understanding how risk allele frequencies differ between populations (as shown here), and leveraging linkage disequilibrium information to infer the effect sizes of risk alleles in non-study populations [38, 39]. Only by understanding the effects of SNP ascertainment bias can accurate predictive models of genetic disease risks be built.

Methods

Population genetic data

Allele frequencies were obtained for each of the five continental super-populations of the 1000 Genomes Project: Africa (AFR), Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS) [17]. Phase 3 data were used. These frequencies were used to generate risk allele frequencies and derived allele frequencies at disease-associated loci from the NHGRI-EBI GWAS Catalog and simulated datasets. Ancestral and derived states in phase 3 1000 Genomes Project VCF files were used (these ancestral states were inferred via the EPO pipeline from Ensembl). We found that derived allele frequencies were elevated for large chunks of chromosome 8, which is indicative of misidentified ancestral states. To compensate for this, we masked SNPs found in the chr8: 89,00,000-146,364,022 region (hg19). Individuals in phase 3 of 1000 Genomes Project were genotyped using WGS. Allele frequencies of SNPs on the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Omni 5M microarray were found by merging data from the 1000 Genomes Project with lists of SNP ids obtained from the Affymetrix and Illumina websites.

Identification of disease-associated variants

Using the NHGRI-EBI GWAS Catalog [1], Berens and colleagues generated a curated set of 3180 disease-associated loci [40]. This involved filtering out SNPs that were not associated with a disease, eliminating SNPs lacking risk allele or odds ratio information, and LD-pruning. Here, we further constrained the set of disease-associated loci from [40] by requiring knowledge of whether risk alleles are ancestral or derived. After excluding 144 SNPs with unknown ancestral states, we were left with a focal set of 3036 disease-associated loci. We classified these 3036 disease-associated loci into seven non-overlapping categories: gastrointestinal/liver, metabolic, morphological, cancer, neurological, miscellaneous, and cardiovascular. Wilcoxon signed-rank tests were used to compare disease allele frequencies between African and non-African populations.

GWAS simulations

Computer simulations were used to test whether SNP ascertainment bias alone can produce what appears to be genetic health disparities. The goal here was to generate simulated datasets comparable to the set of 3036 disease-associated loci from the NHGRI-EBI GWAS Catalog. These simulations assume that the underlying risks of disease are the same across the globe. Two general types of simulations were run: simulations with ancestral risk alleles and simulations with derived risk alleles. Simulations involved randomly drawing a ***test SNP*** from a list of known genetic variants ascertained via WGS or found on commercial genotyping arrays. Conditioning on whether risk alleles are ancestral or derived, the risk allele frequency of the ***test SNP*** was found in the study population. We then used a Perl script based on the GAS/CaTS power calculator [31] to determine the probability of detecting a successful genetic association at the ***test SNP***. The GAS power calculator leverages information about the number of cases and controls, p-value threshold, disease model, prevalence, disease allele frequency, and genotype relevant risk (http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/). For each ***test SNP***, we generated a uniformly distributed random number between 0 and 1. The ***test SNP*** was retained if the random number was less than the power to successfully detect a genetic association, and the ***test SNP*** was rejected if the random number was greater than the probability of detection. This process was repeated until a set of 3036 successful disease associations were detected. At each of these 3036 SNPs, we obtained simulated risk allele frequencies for five super-populations in the 1000 Genomes Project dataset (AFR, AMR, EAS, EUR, SAS). Our default parameters were as follows: genotyping technology = Affymetrix Genome-Wide Human SNP Array 6.0, study population = Europe (EUR), sample size = 3500 cases and 3500 controls, genetic model = additive, p-value threshold = 10^{-5}, prevalence = 0.1, and genotype relative risk = 1.211. These parameter values were chosen to be representative of the empirical data found in the NHGRI-EBI GWAS Catalog.

Our default model was modified to test which aspects of SNP ascertainment bias contribute the most to continental differences in risk allele frequencies. This involved varying the following simulation parameters: genotyping technology, sample size, mode of inheritance, and the p-value threshold required for association detection. The effects of different genotyping technologies were simulated by drawing random SNPs from either the Affymetrix Genome-Wide Human SNP Array 6.0, the Illumina Omni 5M microarray, or WGS data from the 1000 Genomes Project. To examine the effects of different study populations, simulated risk allele frequencies were chosen from one of five different populations (AFR, AMR, EAS, EUR, or SAS) or from an equal mixture of all five populations (MIX). The effects of different sample sizes were simulated by varying the number of cases and controls from 3 to 6 on a $\log_{10}$ scale at intervals of 0.1 (i.e. between 1,000 and 1,000,000 cases and controls). Three genetic modes of inheritance were simulated: dominant, additive, and recessive. Two different p-value thresholds were simulated: $1\text{x}10^{-5}$ and $5\text{x}10^{-8}$.

Data access

Global allele frequencies are publicly available from the 1000 Genomes Project website: http://www.internationalgenome.org/data. Disease associations are publicly available from the NHGRI-EBI GWAS Catalog: https://www.ebi.ac.uk/gwas/. R and Perl scripts used in GWAS simulations are available upon request.

Acknowledgements

We thank A. Martin, U. Martinez-Marigorta, and M. Quiver for helpful discussions during the writing of this paper. This work was supported by NIH/NCI grant U01CA184374 and start-up funds from Georgia Institute of Technology.

Disclosure declaration

All authors declare that they have no competing interests.

References

1. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901. doi: 10.1093/nar/gkw1133. PubMed PMID: 27899670; PubMed Central PMCID: PMCPMC5210590.

2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106(23):9362-7. doi: 10.1073/pnas.0903103106. PubMed PMID: 19474294; PubMed Central PMCID: PMCPMC2687147.

3. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4. doi: 10.1038/538161a. PubMed PMID: 27734877; PubMed Central PMCID: PMCPMC5089703.

4. Manolio TA. In Retrospect: A decade of shared genomic associations. Nature. 2017;546(7658):360-1. doi: 10.1038/546360a. PubMed PMID: 28617469.

5. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. The American Journal of Human Genetics. 2017;100(4):635-49. doi: 10.1016/j.ajhg.2017.03.004.

6. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475(7355):163-5. doi: 10.1038/475163a. PubMed PMID: 21753830; PubMed Central PMCID: PMCPMC3708540.

7. Marigorta UM, Navarro A. High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet. 2013;9(6):e1003566. doi:

10.1371/journal.pgen.1003566. PubMed PMID: 23785302; PubMed Central PMCID: PMCPMC3681663.

8. Palmer C, Pe'er I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLoS Genet. 2017;13(7):e1006916. doi: 10.1371/journal.pgen.1006916. PubMed PMID: 28715421.

9. Shriner D. Mixed Ancestry and Disease Risk Transferability. Current Genetic Medicine Reports. 2015;3(4):151-7.

10. Coram MA, Fang H, Candille SI, Assimes TL, Tang H. Leveraging Multi-Ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. Am J Hum Genet. 2017. doi: 10.1016/j.ajhg.2017.06.015. PubMed PMID: 28757202.

11. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17(7):392-406. doi: 10.1038/nrg.2016.27. PubMed PMID: 27140283.

12. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52. doi: 10.1038/nature08185. PubMed PMID: 19571811; PubMed Central PMCID: PMCPMC3912837.

13. Shi J, Park JH, Duan J, Berndt ST, Moy W, Yu K, et al. Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. PLoS Genet. 2016;12(12):e1006493. doi: 10.1371/journal.pgen.1006493. PubMed PMID: 28036406; PubMed Central PMCID: PMCPMC5201242.

14. Corona E, Chen R, Sikora M, Morgan AA, Patel CJ, Ramesh A, et al. Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. PLoS Genet. 2013;9(5):e1003447. doi: 10.1371/journal.pgen.1003447. PubMed PMID: 23717210; PubMed Central PMCID: PMCPMC3662561 consultants to Personalis, Inc. RC is now employed by Personalis, Inc.

15. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53. doi: 10.1038/nature08494. PubMed PMID: 19812666; PubMed Central PMCID: PMCPMC2831613.

16. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210-7. doi: 10.1016/j.cell.2010.03.032. PubMed PMID: 20403315.

17. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi: 10.1038/nature15393. PubMed PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.

18. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-4. doi: 10.1126/science.1153717. PubMed PMID: 18292342.

19. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, et al. Population history and its impact on medical genetics in Quebec. Clin Genet. 2005;68(4):287-301. doi: 10.1111/j.1399-0004.2005.00497.x. PubMed PMID: 16143014.

20. Macgregor S, Bellis C, Lea RA, Cox H, Dyer T, Blangero J, et al. Legacy of mutiny on the Bounty: founder effect and admixture on Norfolk Island. Eur J Hum Genet. 2010;18(1):67-72. doi: 10.1038/ejhg.2009.111. PubMed PMID: 19584896; PubMed Central PMCID: PMCPMC2987173.

21. Huang M, Graham BE, Zhang G, Harder R, Kodaman N, Moore JH, et al. Evolutionary triangulation: informing genetic association studies with evolutionary evidence. BioData Min. 2016;9:12. doi: 10.1186/s13040-016-0091-7. PubMed PMID: 27042214; PubMed Central PMCID: PMCPMC4818851.

22. Lachance J. Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. BMC Med Genomics. 2010;3:57. doi: 10.1186/1755-8794-3-57. PubMed PMID: 21143973; PubMed Central PMCID: PMCPMC3017004.

23. Di Rienzo A, Hudson RR. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet. 2005;21(11):596-601. doi: 10.1016/j.tig.2005.08.007. PubMed PMID: 16153740.

24. Lohmueller KE. The distribution of deleterious genetic variation in human populations. Curr Opin Genet Dev. 2014;29:139-46. doi: 10.1016/j.gde.2014.09.005. PubMed PMID: 25461617.

25. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays : news and reviews in molecular, cellular and developmental biology. 2013;35(9):780-6. Epub 2013/07/10. doi: 10.1002/bies.201300014. PubMed PMID: 23836388; PubMed Central PMCID: PMCPMC3849385.

26. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356-69. doi: 10.1038/nrg2344. PubMed PMID: 18398418.

27. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15(5):335-46. doi: 10.1038/nrg3706. PubMed PMID: 24739678.

28. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005;15(11):1496-502. doi: 10.1101/gr.4107905. PubMed PMID: 16251459; PubMed Central PMCID: PMCPMC1310637.

29. Jones D. A WEIRD View of Human Nature Skews Psychologists' Studies. Science. 2010;328(5986):1627-. doi: 10.1126/science.328.5986.1627.

30. Henrich J, Heine SJ, Norenzayan A. Most people are not WEIRD. Nature. 2010;466(7302):29-.

31. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38(2):209-13. doi: 10.1038/ng1706. PubMed PMID: 16415888.

32. Braveman P, Egerter S, Williams DR. The social determinants of health: coming of age. Annu Rev Public Health. 2011;32:381-98. doi: 10.1146/annurev-publhealth-031210-101218. PubMed PMID: 21091195.

33. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016;375(7):655-65. doi: 10.1056/NEJMsa1507092. PubMed PMID: 27532831; PubMed Central PMCID: PMCPMC5292722.

34. Ohta T. The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics. 1992;23(1):263-86.

35. Henn BM, Botigue LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad

Sci U S A. 2016;113(4):E440-9. doi: 10.1073/pnas.1510805112. PubMed PMID: 26712023; PubMed Central PMCID: PMCPMC4743782.

36. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22. doi: 10.1016/j.ajhg.2017.06.005. PubMed PMID: 28686856; PubMed Central PMCID: PMCPMC5501872.

37. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005;102(44):15942-7. doi: 10.1073/pnas.0507611102. PubMed PMID: 16243969; PubMed Central PMCID: PMCPMC1276087.

38. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97(4):576-92. doi: 10.1016/j.ajhg.2015.09.001. PubMed PMID: 26430803; PubMed Central PMCID: PMCPMC4596916.

39. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nat Rev Genet. 2010;11(5):356-66. doi: 10.1038/nrg2760. PubMed PMID: 20395969; PubMed Central PMCID: PMCPMC3079573.

40. Berens AJ, Cooper TL, Lachance J. The genomic health of ancient hominins. Human Biology. 2017;89(1):5-17.

Figures

Figure 1

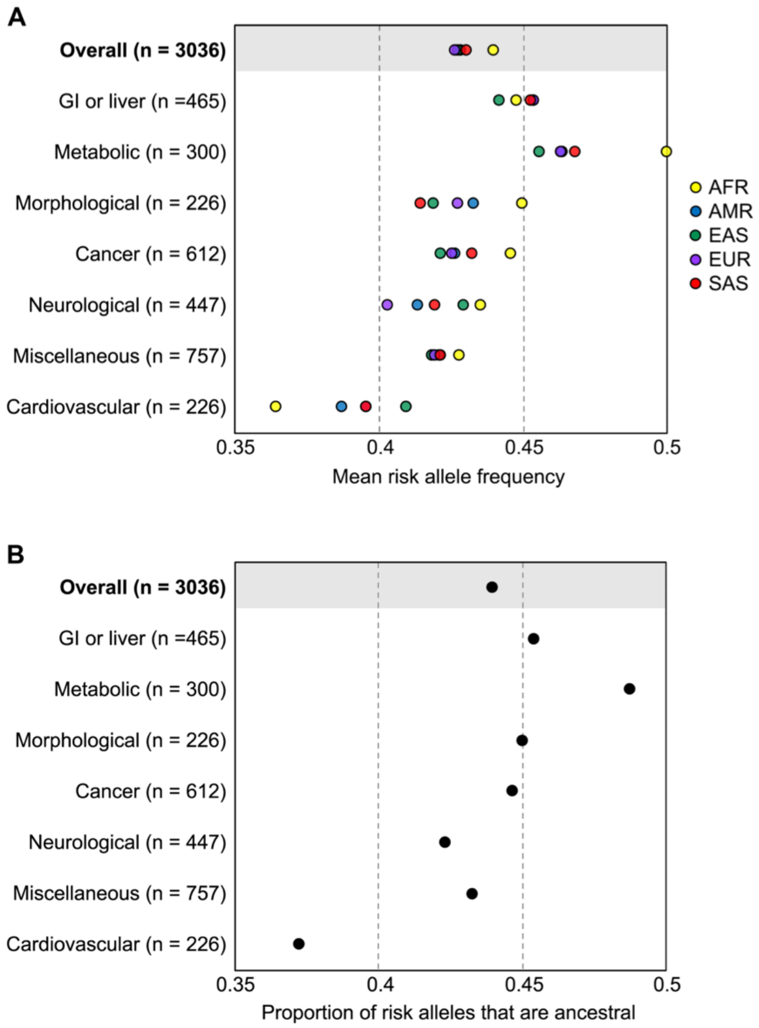

Fig 1. Risk allele frequencies at published GWAS loci suggest the possibility of genetic health disparities. (A) Risk allele frequencies at published GWAS loci vary by population (1000 Genomes Project data shown). For most disease classes, risk allele frequencies are elevated in African populations. Significant allele frequency differences between African and non-African populations are indicated by * (p-values < 0.05, paired Wilcoxon rank sum tests). (B) Proportion of disease-associated SNPs where the risk allele is ancestral, as opposed to derived.

Figure 2

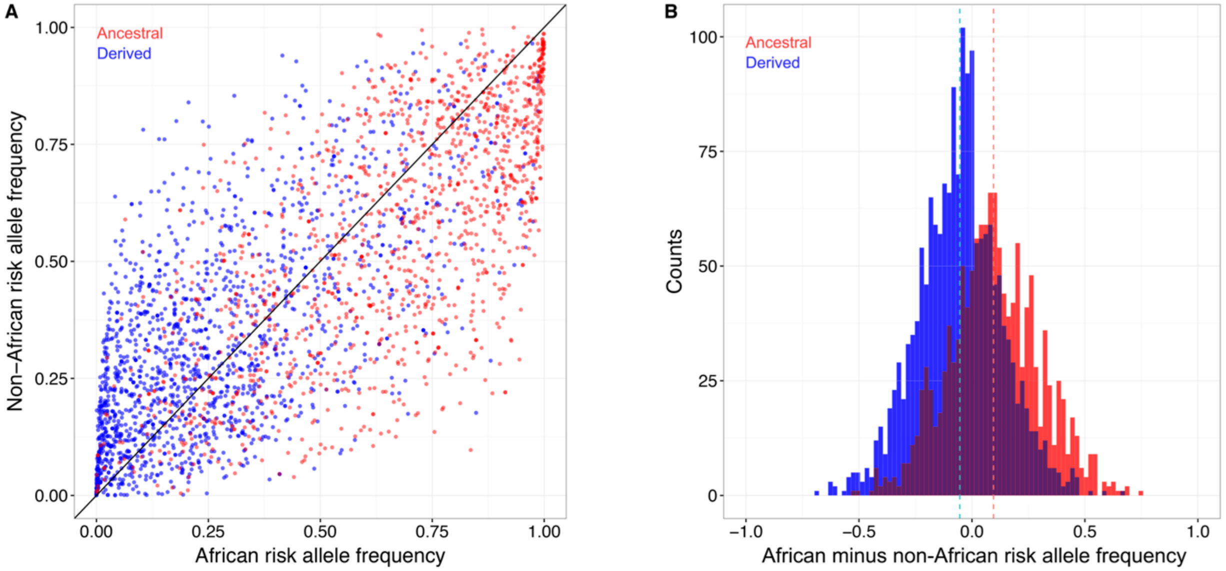

Fig 2. Continental patterns of allele frequencies at disease susceptibility loci depend on whether risk alleles are ancestral or derived. (A) Joint site frequency spectrum of published GWAS loci. Disease susceptibility loci with ancestral risk alleles are labelled red and loci with derived risk alleles are labelled blue. (B) Histogram of the difference in risk allele frequencies between African and non-African populations. The mean allele frequency of ancestral risk alleles is higher in Africa (+9.51%) and the mean risk allele frequency of derived risks is lower in Africa (-5.40%). Disease susceptibility loci with ancestral risk alleles are labelled red and loci with derived risk alleles are labelled blue. Overlap in the histogram is labelled purple, and dashed lines indicate mean values.

Figure 3

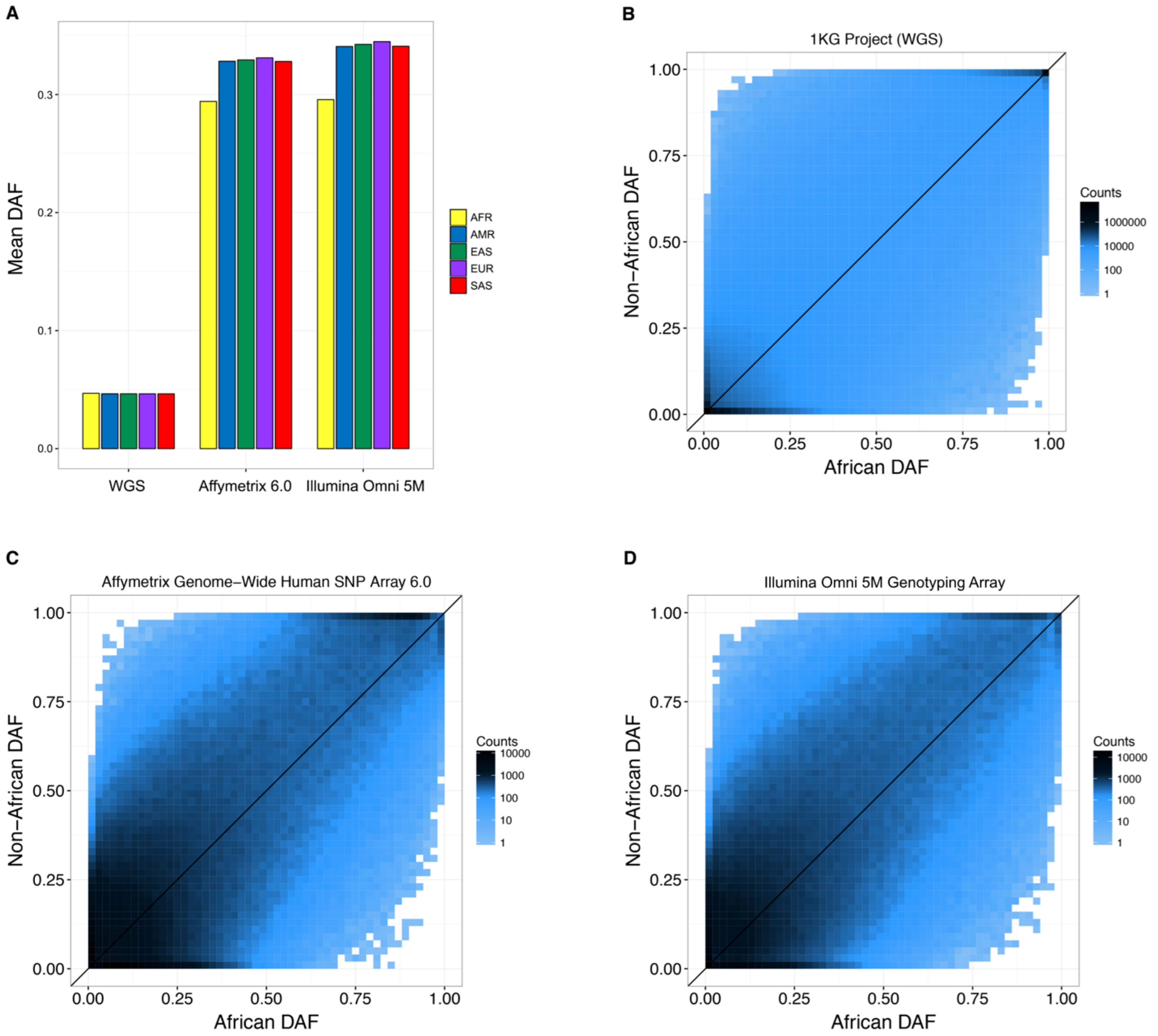

Fig 3. Genotyping arrays bias allele frequencies in African populations. (A) Comparisons show the mean derived allele frequencies (DAF) of five populations from the 1000 Genomes Project. Genotyping arrays are enriched for intermediate frequency derived alleles in non-African populations. (B) Joint site frequency spectrum of whole genome sequence (WGS) data. Non-African and African data from the 1000 Genomes Project are shown. Shading indicates counts of SNPs. (C) Joint site frequency spectrum of ascertained SNPs on the Affymetrix Genome-Wide Human SNP Array 6.0. (D) Joint site frequency spectrum of ascertained SNPs on the Illumina Omni 5M microarray.

Figure 4

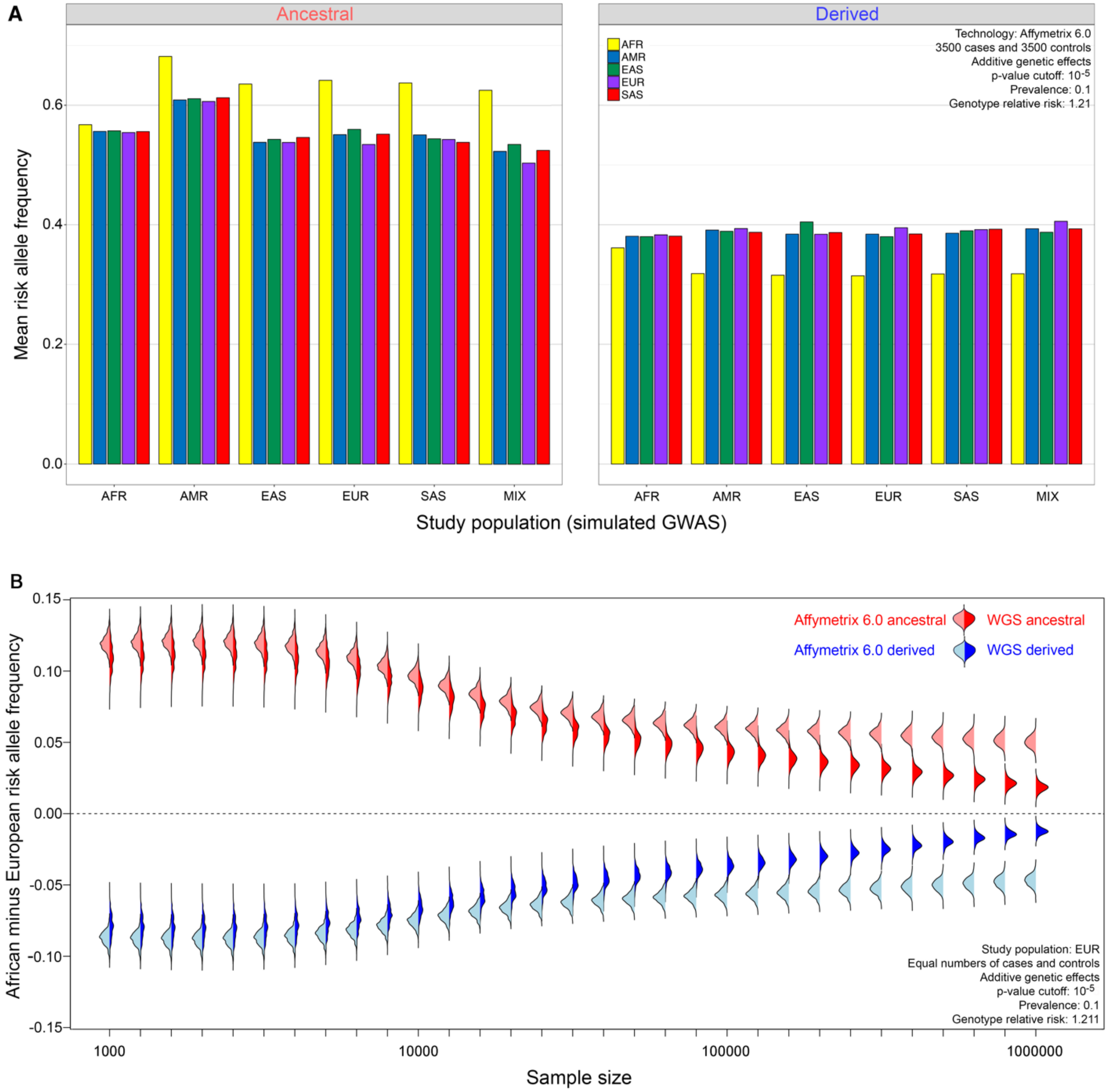

Fig 4. ***GWAS simulations replicate empirical patterns.*** (A) Simulations of non-African GWAS yield elevated frequencies of ancestral risk alleles and reduced frequencies of derived risk alleles in African populations. (B) Larger sample sizes reduce bias for simulated GWAS. Beanplots show the results of 1000 simulations per set of parameter values. Each simulation run involved generating a set of 3036 disease associations. Sample sizes indicate the numbers of cases, and numbers of controls are set equal to the numbers of cases.

Figure 5

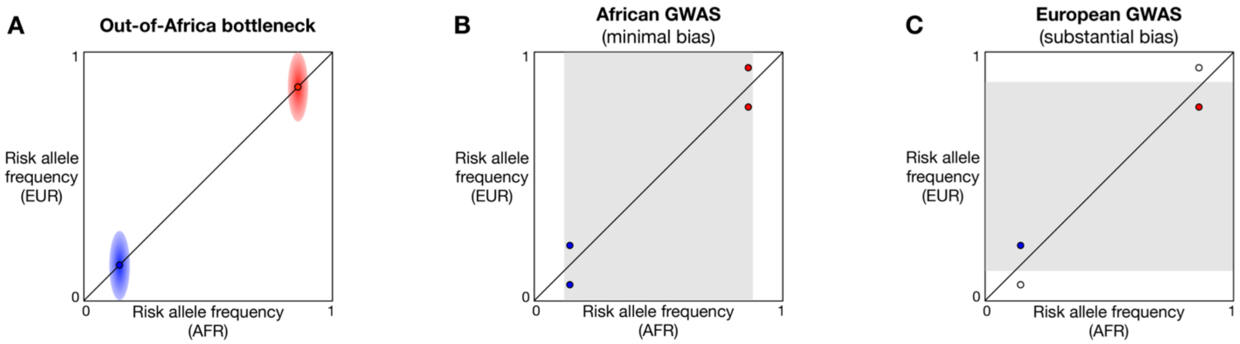
A
Out-of-Africa bottleneck
Risk allele frequency (EUR)
Risk allele frequency (AFR)
0
1
0
1
B
African GWAS (minimal bias)
Risk allele frequency (EUR)
Risk allele frequency (AFR)
0
1
0
1
C
European GWAS (substantial bias)
Risk allele frequency (EUR)
Risk allele frequency (AFR)
0
1
0
1

Fig 5. Synergistic effects between population bottlenecks and choice of study population. Ancestral risk alleles are labelled red and derived risk alleles are labelled blue. Detectable associations in GWAS are indicated by filled circles and gray shading. Undetectable associations are indicated by open circles. (A) Prior to divergence of allele frequencies are the same in different population (along the diagonal). The out-of-Africa bottleneck causes allele frequencies to drift farther in non-African populations than African populations. (B) African GWAS result in minimal bias. (C) Non-African GWAS result in biased frequencies of risk alleles.