Population genetic history and polygenic risk biases in 1000 Genomes populations

Alicia R. Martin[1,2,3], Christopher R. Gignoux[3], Raymond K. Walters[1,2], Genevieve L. Wojcik[3], Simon Gravel[4], Mark J. Daly[1,2], Carlos D. Bustamante[3], Eimear E. Kenny[5]

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114

[2] Medical and Population Genetics, Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, MA USA

[3] Department of Genetics, Stanford University, Stanford, CA 94305, USA

[4] Department of Human Genetics, McGill University, Montreal, Quebec, Canada

[5] Department of Genetics and Genomic Sciences, Mt. Sinai School of Medicine, New York, NY, USA

Corresponding author: eimear.kenny@mssm.edu

Abstract

Background: Genome-wide association studies (GWAS) have largely focused on European descent populations, and the transferability of these findings to diverse populations is dependent on many factors, including selection, genetic divergence, heritability, and phenotype complexity. As medical genomics studies become increasingly large and ethnically diverse, gaining clear insight into population history and genetic diversity from available reference panels is critically important.

Results: We disentangle the population history of the widely-used 1000 Genomes Project reference panel, with an emphasis on underrepresented Hispanic/Latino and African descent populations. By leveraging haplotype sharing, linkage disequilibrium decay, and ancestry deconvolution along chromosomes in admixed populations, we gain insights into ancestral allele frequencies, the origins, rates, and timings of admixture, and sex-biased demography. We make empirical observations to evaluate the impact of population structure in association studies, with conclusions that inform rare variant association in diverse populations, how we use standard GWAS tools, and transferability of findings across populations. Finally, we show through coalescent simulations that inferred polygenic risk scores derived from European GWAS are biased when applied to diverse populations.

Conclusions: Our study provides fine-scale insight into the sampling, genetic origins, divergence, and sex-biased history of admixture in the 1000 Genomes Project populations. We show that the transferability of results from GWAS are dependent on the ancestral diversity of the study cohort as well as the phenotype polygenicity, causal allele frequency divergence, and heritability. This work highlights the need for inclusion of more diverse samples in medical genomics studies to enable broadly applicable disease risk information.

Keywords: population genetics, demographic inference, human evolution, health disparities

Background

The majority of genome-wide association studies (GWAS) have been performed in populations of European descent [1-3]. GWAS have yielded tens of thousands of common genetic variants associated with human medical and evolutionary phenotypes, most of which have been replicated in other ethnic groups [4-7]. However, it has been far harder to identify the role of rare variants. Large-scale sequencing studies have demonstrated that, due to the effects of recent population expansions and purifying selection, rare variants show stronger geographic clustering than common variants [8-10]. For example, African populations contain a larger number of rare variant sites than out-of-Africa populations, and within Europe, there is an increasing gradient of cumulative rare variation on the northern to southern cline [11, 12]. Therefore, rare variants associated with disease may be expected to track with recent population demography and/or be population restricted [9, 13, 14], and require studies incorporating data from hundreds of thousands or millions of people for discovery [15-17]. As the next era of GWAS expand to evaluate the disease-associated role of rare variants, it is not just scientifically imperative to include multi-ethnic populations, but also likely that increasing genetic heterogeneity will be encountered in very large study populaitons. A comprehensive understanding of the genetic diversity and demographic history of multi-ethnic populations is critical for appropriate applications of GWAS, and ultimately for reducing health disparities.

The most recent release of the 1000 Genomes Project (phase 3) provides one of the largest global reference panels of whole genome sequencing data, enabling a broad

survey of human genetic variation [12]. The depth and breadth of diversity queried facilitates a deep understanding of the evolutionary forces (e.g. selection and drift) shaping existing genetic variation in present-day populations that contribute to adaptation and disease [18-24]. Coupling well-documented historical events with demographic inferences from genetic data is important to ensure that methods are well-calibrated. This in turn enables inferences in scenarios where a clear historical record is unavailable, as well as identifying older, previously unknown demographic events [25-27].

Studies of admixed populations have been particularly fruitful in identifying genetic adaptations and risks for diseases that are stratified across diverged ancestral origins [28-36]. These patterns are especially complex in populations from the Americas. Recent admixture in the Americas is extensive, with significant genetic contributions spanning Native American, European, Oceanian, and South and East Asian populations, as well as Western and Central African ancestry from the trans-Atlantic slave trade. Processes shaping structure in these admixed populations include sex-biased migration and admixture, isolation-by-distance, differential drift in mainland versus island populations, and variable admixture timing [9, 37, 38].

Understanding admixture patterns is critical for the inclusion of ancestrally diverse and recently admixed populations (i.e. the majority of populations from the Americas) in GWAS [39]. Standard GWAS strategies approach population structure as a nuiscance factor. A typical step-wise procedure first detects global population structure in each

individual, using principal component analysis (PCA) or other methods [40-43], and often excludes outlier individuals from the analysis and/or corrects for population structure in the statistical model for association. Such strategies may reduce false positives in test statistics, but can also reduce power for association in heterogeneous populations, and are less likely to work for rare variant association [44-47]. Recent methodological advances have leveraged patterns of global and local ancestry for improved association power [32, 48, 49], fine-mapping [50] and genome assembly [51]. At the same time, population genetic studies have demonstrated the presence of fine-scale sub-continental structure in the African, Native American and European components of populations from the Americas [52-54]. If trait-associated variants follow the same patterns of demography, then we expect that modeling sub-continental ancestry may enable their improved detection in admixed populations.

Even when GWAS studies conducted in European populations yield transferable genetic findings across populations, they often explain less phenotypic variation in other populations. While disease loci that are common and at similar frequencies in multiple populations will have shared effects, the phenotypic variance explained for polygenic traits that have undergone significant drift, as is expected across continents, is reduced [55]. The transferability of GWAS results across ancestries is dependent on the similarity of population allele frequencies, genetic architecture of phenotypes, and effects of polygenic selection. For example, a previous study calculated polygenic risk scores for schizophrenia in East Asians and Africans based on GWAS summary statistics derived from a European cohort, and found that prediction accuracy was

reduced by more than 50% in non-European populations [56]. Having arisen more recently (e.g. after out-of-Africa migrations began), rare variants are more ancestry stratified than common variants. Rare clinical variant studies have gained traction recently [57], and have been aided considerably by large-scale diverse databases, e.g. the Exome Aggregation Consortium (ExAC) [58], enabling rapid filtering of variants that are too common to confer highly penetrant disease risk.

Here, we explore the admixture history of the 26 global populations in the 1000 Genomes Project. We focus particularly on admixed populations from the Americas (e.g. African descent and Hispanic/Latino populations). We integrated sequencing and genotyping variants to infer the proportions, timings, and quantity of diverse ancestries contributing to multi-way cross-continental admixture. By combining haplotype sharing, linkage disequilibrium decay, and ancestry patterns across chromosomal segments, we gain insight into ancestral origins and sex-biased demography, which are supported by historical records. To link this work to ongoing efforts to improve study design, disease variant discovery and annotation, we assessed the effect of biased clinical databases and GWAS efforts on inferences into diverse and admixed populations. Finally, we showed through coalescent simulations that inferred polygenic risk scores derived from European GWAS are biased when applied to diverse populations.

Results

The 1000 Genomes project (Phase 3) produced sequence data for 2,504 unrelated individuals from 26 globally diverse populations [12]. The South Asian ‘super population’

(SAS; population abbreviations in Table S1) is new to this phase along with a larger representation of populations from other continental groups, including the Americas. The Barbadians (ACB) and Peruvians (PEL) are also new to Phase 3 and have more African and Native ancestry, respectively, than the other recently admixed African descent and Hispanic/Latino populations in the Americas in the 1000 Genomes Project.

Global ancestry in the 1000 Genomes Project

We first assessed the overall diversity at the global and sub-continental level of the 1000 Genomes populations using a likelihood model via ADMIXTURE [59] in all populations (Figure 1A). This analysis recapitulates population structure observed in 1000 Genomes populations in [12]. At K=5, this analysis separates ancestries from the super populations (AFR, AMR, EAS, EUR, and SAS, population abbreviations in Table S1). These results also highlight sub-continental substructure; for example, there is detectable European (EUR) and East Asian (EAS) ancestry in the SAS populations (population means range from 6.1-15.9% and 0.3-12.2%, respectively), with the highest rates of East Asian ancestry in the Bengalis from Bengladesh (BEB). In contrast, the greatest quantity of European ancestry in the SAS populations is in the Punjabi from Lahore, Pakistan (PJL), which are geographically the closest to Europeans. Ancestral clines have been observed along geographical, caste, and linguistic axes in more densely sampled studies of South Asia [60, 61]. Increasing the model to K=6 there is also an east-west cline among African populations, while at K=7 we observe the north-south cline of European ancestry [13].

We assessed population substructure by querying 3-way continental admixture contributions in the six populations from the Americas (AFR/AMR and AMR) using principal components analysis (PCA) and ADMIXTURE (Figure 1B-C). We find widely varying African, European, and Native American ancestry proportions in the ACB, ASW, CLM, MXL, PEL, and PUR populations (Table 1). When compared to the ASW, the ACB have a higher proportion of African ancestry (μ = 0.88, 95% CI = [0.71-0.97] versus μ = 0.76, 95% CI = [0.40-0.90]; p =3.3e-7) and a smaller proportion of European and Native American ancestry. The PEL have more Native American ancestry than the other AMR populations (μ = 0.81, 95% CI = [0.49-1.0] versus μ = 0.28, 95% CI = [0.07, 0.73]; p = 1.4e-95) ascertained in 1000 Genomes, providing a valuable resource of Native American haplotypes reflective of South America.

While there is minimal Native American ancestry (~1%) in most African Americans across the United States, there is a substantial enrichment in several ASW individuals from 1000 Genomes (mean of 3.1%, and 9 samples with >5%, including NA19625, NA19921, NA20299, NA20300, NA20314, NA20316, NA20319, NA20414, and NA20274) [52, 62]. Interestingly, one ASW individual has no African ancestry (NA20314, EUR= 0.40, NAT=0.59) but is the mother of NA20316 in an ASW duo with few Mendelian inconsistencies. The ancestry proportions of NA20316 (EUR=0.30, NAT=0.29, AFR=0.41) suggest that his father mostly likely has ~80% African and ~20% European ancestry, similar to other ASW individuals. We also find evidence of East Asian admixture in several PEL samples (39% in HG01944, 12% in HG02345, 6% in HG0192, 5% in HG01933, and 5% in HG01948). Consistent with the autosomal

evidence, the Y chromosome haplogroup for HG01944 (Q1a-M120) clusters most closely with two KHV samples and other East Asians rather than the Q-L54 subgroup expected in samples from South America [63].

Local ancestry provides insight into recent admixture events in the Americas

The genomes of recently admixed samples from the Americas contain information about both continental and sub-continental origins [31, 37, 64, 65], which we investigated by identifying local ancestry tracts (Methods, Figure S1). As proxy source populations for the recent admixture, we used European and African continental samples from the 1000 Genomes Project as well as Native Americans genotyped previously [66]. Concordance between global ancestry estimates inferred using ADMIXTURE and RFMix was typically very high (Pearson's correlation ≥ 98%, Figure S2). Correlations were lower when there was minimal ancestry from a given source population and/or tracts were very short (<5cM), e.g. NAT ancestry in the ACB (ρ=0.79) and AFR ancestry in the MXL (ρ=0.94). Substantial deviations from ADMIXTURE estimates occurred for 1 ACB individual (HG01880), where considerable South Asian ancestry (31.8%, Figure S2A) was classified in RFMix as European ancestry and in 2 PEL individuals (HG01944 and HG02345, Figure S2E), where considerable East Asian ancestry (38.2% and 12.3%, respectively) was classified in RFMix as EUR and NAT ancestry due to limitations of the 3-way ancestry reference panel.

The distribution of continuous ancestry tracts (i.e. lengths of unbroken segments from one ancestry) provides insight into the time-dependent migration rates from multiple

source populations. We used *Tracts* [67] to infer admixture events from ancestry tracts in the ACB and PEL (Figure S3A-B). The best fit model (log-likelihood=-307.1 compared to -320.3 and -433.0 for next two best models) for the PEL begins ~12 generations ago, which is slightly more recent than for insular and Caribbean mainland populations (Figure S3C). For example, admixture in Colombian and Honduran mainland populations was previously inferred to have begun 14 generations ago, whereas admixture in Cuban, Puerto Rican, Dominican, and Haitian populations began 16-17 generations ago [37]. There is minimal African ancestry (2.9%), some European ancestry (37.6%) and primarily Native ancestry (59.4%) in the first pulse of admixture, followed by a later pulse ~5 generations ago of additional primarily Native ancestry (91.1%). This later pulse of primarily Native ancestry is unique to the PEL compared to other admixed Americas populations [37].

The best fit model for the ACB was an initial pulse of admixture between Europeans and Africans followed by a later pulse of African ancestry (Figure S3D). This model was nearly indistinguishable from the next best fit model: a single pulse of admixture between Europeans and Africans beginning ~8 generations ago (log-likelihood = -276.8 vs -280.2). The best model indicates that admixture in the ACB began ~8 generations ago with the initial pulse containing 87.4% African ancestry and 12.6% European ancestry. The second pulse of African ancestry began ~5 generations ago and had only a minor overall contribution (4.4% of total pulse ancestry), which is consistent with either a later small pulse of African ancestry or movement of populations within the Caribbean. The admixture events we infer in the ACB are more recent than previous ASW and

African American two-pulse models, which estimated that admixture began ~10-11 generations ago [52, 67]. Potential explanations for this small difference include differences in the ages of individual between the two cohorts and the fact that pulse timings indicate the generations that admixture most likely spanned rather than the exact generation during which admixture began [37].

Ancestry-specific PCA

We explored the origin of the sub-continental-level ancestry from recently admixed individuals by masking haplotypes to continental ancestry tracts one at a time based on local ancestry assignments inferred with RFMix. Then, we used a version of PCA modified to handle highly masked data (ancestry-specific PCA) [68]. Example ancestry tracts in a PEL individual subset to AFR, EUR, and NAT components are shown in Figures 2A, 2C, and 2E, respectively. As previously demonstrated, the inferred European tracts in Hispanic/Latino populations most closely resemble southern European IBS and TSI populations with some additional drift [37] (Figure 2D). The European tracts of the PUR show evidence of excess drift compared to the CLM, MXL, and PEL populations, consistent with a founder effect in this island population seen previously [37]. In contrast to the southern European tracts from the Hispanic/Latino populations, the African descent populations in the Americas have European admixture that more closely resembles the northwestern CEU and GBR European populations. The clusters are less distinct, owing to lower overall fractions of European ancestry.

The ability to localize aggregated ancestral genomic tracts enables insights into the mechanistic origins of admixed populations. To disentangle whether the considerable Native American ancestry in the ASW individuals arose from recent admixture with Hispanic/Latino individuals or recent admixture with indigenous Native American populations, we queried the European tracts. We find that the European tracts of all ASW individuals with considerable Native American ancestry are well within the ASW cluster and project closer in Euclidean distance with PC1 and PC2 to northwestern Europe than the European tracts from Hispanic/Latino samples (p=1.15e-3), providing support for the latter hypothesis and providing regional nuance to previous findings [52].

We also investigated the African origin of the admixed African descent populations (ACB and ASW), as well as the Native American origin of the Hispanic/Latino populations (CLM, MXL, PEL, and PUR). The African tracts of ancestry from the AFR/AMR populations project closer to the YRI and ESN of Nigeria than the GWD, MSL, and LWK populations (Figure 2B). This is consistent with slave records and previous genome-wide analyses of African Americans indicating that most sharing occurred in West and West-Central Africa [69, 70]. There are subtle differences between the African origins of the ACB and ASW populations (e.g. p=6.4e-6 for difference in distance from YRI on ancestry-specific PC1 and PC2), likely due either to mild island founder effects in the ACB samples or differences in African source populations for individuals who remained in the Caribbean versus those who were brought to the US. The Native tracts of ancestry from the AMR populations first separate the southernmost PEL populations from the CLM, MXL, and PUR on PC1,

then separate the northernmost MXL from the CLM and PUR on PC2, consistent with a north-south cline of divergence among indigenous Native American ancestry (Figure 2F) [37, 71].

Sex-biased demographic events in the Americas

Sex-biased admixture has previously been shown to be ubiquitous in the Americas, impacting phenotypes strongly correlated with ancestry such as pigmentation [29, 37, 38, 72-75]. We inferred sex-biases in admixture events by separately querying ploidy-adjusted admixture proportions on the X chromosome versus the autosomes, as previously described [72]. We computed 3-way admixture proportions for AMR and AFR/AMR via ADMIXTURE [59]. We consistently find across all six admixed AMR populations that the ratio of European ancestry is depleted on the X chromosome compared to the autosomes, indicating a ubiquitous excess of breeding European males in the Americas, as seen previously [76]. There is a significant depletion of European ancestry on the X chromosome in each of the AMR populations ($p < 1e-4$) and an excess of Native American ancestry ($p<1e-2$, Table 2 and Figure 3B). There is no significant excess/depletion of African ancestry in AMR populations, likely owing to limited African ancestry proportions and sample sizes, as larger and more heterogeneous population surveys have identified an X chromosome excess in Latinos [74]. Across admixed individuals of African descent, European ancestry is significantly depleted on the X chromosome in aggregate across the ACB and ASW populations ($p=1.03E-03$, Figure 3A) as has been seen previously in larger studies [52, 74].

Performance of standard GWAS tools in admixed populations

We recapitulate results showing that there is less variation across the genome in out-of-Africa versus African populations, but that GWAS variants are more polymorphic in European and Hispanic/Latino populations (Figure S5A-B), reflecting a strong bias towards European study participants [1-3, 12]. We reassessed this bias on Affymetrix 6.0 sites, which were used in local ancestry calling to incorporate a Native American reference panel [66] for admixed samples from the Americas. The average minor allele frequency in a population is an indicator of the amount of diversity captured by a technology. We use a normalized measure of the minor allele frequency to obtain a background coverage of each population, as done previously (e.g. Figure S4 from [12]). The array background closely resembles the genome-wide background with an additional slight European bias (Figure S4A-B). Minor allele frequencies on the Affymetrix 6.0 array as well as on this array specifically at GWAS sites are higher in populations sharing European ancestry, including the AMR, EUR, and SAS super populations (Figure S4C-D). We further stratified these analyses in admixed AMR and AFR/AMR populations by local ancestry, computing minor allele frequency on different ancestry tracts (Figure S5C-D). We show that GWAS variants are at increased allele frequencies specifically on European ancestry tracts, and that there is a dearth of GWAS variation on both African and Native American tracts (Figure S5D).

We also assessed imputation accuracy across the 3-way admixed populations from the Americas (CLM, MXL, PEL, PUR) for two arrays: the Illumina OmniExpress and the Affymetrix Axiom World Array LAT. Imputation accuracy was estimated as the

correlation (r^2) between the original genotypes and the imputed dosages. For both array designs, imputation accuracy across all minor allele frequency (MAF) bins was highest for populations with the largest proportion of European ancestry (PUR) and lowest for populations with the largest proportion of Native American ancestry (PEL, Figure S6A-B). We also stratified imputation accuracy by local ancestry tract diplotype within the Americas. Consistently, tracts with at least one Native American ancestry tract had lower imputation accuracy when compared to tracts with only European and/or African ancestry (Figure S7A-B).

Transferability of GWAS findings across populations

To quantify the transferability of European-biased genetic studies to other populations, we next used published GWAS summary statistics to infer polygenic risk scores across populations for well-studied traits, including height [15], waist-hip ratio [77], schizophrenia [16], type II diabetes [78, 79], and asthma [80] (Figure 4A-D, Figure S8, Methods). Most of these summary statistics are derived from studies with primarily European cohorts, although GWAS of type II diabetes have been performed in both European-specific cohorts as well as across multi-ethnic cohorts. We identify clear directional inconsistencies in these score inferences. For example, although the height summary statistics show the expected cline of southern/northern cline of increasing European height (FIN, CEU, and GBR populations have significantly higher polygenic risk scores than IBS and TSI, p=1.5e-75, Figure S8A), polygenic scores for height across super populations show biased predictions. For example, the African populations sampled are genetically predicted to be considerably shorter than all Europeans and

minimally taller than East Asians (Figure 4A), which contradicts empirical observations (with the exception of pygmy populations) [81, 82]. Additionally, polygenic risk scores for schizophrenia, which is at a similar prevalence across populations where it has been well-studied [83] and shares significant genetic risk across populations [84], shows considerably decreased scores in Africans compared to all other populations (Figure 4B). Lastly, the relative order of polygenic risk scores computed for type II diabetes across populations differs depending on whether the summary statistics are derived from a European-specific (Figure 4C) or multi-ethnic (Figure 4D) cohort.

Detecting biases in polygenic risk score estimates

We performed coalescent simulations to determine how GWAS signals discovered in one ancestral case/control cohort (i.e. 'single-ancestry' GWAS) would impact polygenic risk score estimates in other populations (for details, see Methods). Breifly, we simulated variants according to a previously published demographic model inferred from Africans, East Asians, and Europeans [9]. We specified "causal" alleles and effect sizes randomly, such that each causal variant has evolved neutrally and has a mean effect of zero with the standard deviation equal to the global heritability divided by number of causal variants. We then computed the true polygenic risk for each individual as the product of the effect sizes and genotypes, then standardized the scores across all individuals. We calculated the total liability as the sum of the genetic and random environmental contributions, then identified 10,000 European cases with the most extreme liabilities and 10,000 other European controls. We then computed Fisher's exact tests with this European case-control cohort, then quantified inferred polygenic

risk scores as the sum of the product of genotypes and log odds ratios for 10,000 samples per population not included in the GWAS.

In our simulations, most variants are rare and population-specific; “causal” variants are sampled from the global site frequency spectrum, resulting in subtle differences in true polygenic risk across populations (Figure S9). We mirrored standard practices for performing a GWAS and computing polygenic risk scores (see above and Methods). We find that the correlation between true and inferred polygenic risk is generally low (Figure 4E-G), consistent with limited variance explained by polygenic risk scores from GWAS of these cohort sizes for height (e.g. ~10% of variance explained for a cohort of size 183,727 [85]) and schizophrenia (e.g. ~7% variance explained for a cohort of size 36,989 cases and 113,075 controls, [16]). Low correlations in our simulations are most likely because common tag variants are a poor proxy for rare causal variants. As expected, correlations between true and inferred risk within populations are typically highest in the European population (i.e. the population in which variants were discovered, Figure 4E-G). Across all populations, the mean Spearman correlations between true and inferred polygenic risk increase with increasing heritability while the standard deviation of these correlations significantly decreases ($p=0.05$); however, there is considerable within-population heterogeneity resulting in high variation in scores across all populations. We find that in these neutral simulations, a polygenic risk score bias in essentially any direction is possible even when choosing the exact same causal variants, heritability, and varying only effect size (i.e. inferred polygenic risk in Europeans can be higher, lower, or intermediate compared to true risk relative to East

Asians or Africans, Figure S9). These observed biases and inconsistent predictive power of polygenic risk scores across populations highlight the importance of exercising caution when inferring polygenic risk in admixed and substructured populations, population-centering polygenic risk scores in cross-population analyses [86], and the potential for increased health disparities emerging from the application of polygenic risk scores ascertained in a single-ancestry cohort.

Discussion

To date, GWAS have been performed opportunistically in primarily single-ancestry European cohorts, and an open question remains about their biomedical relevance for disease associations in other ancestries. As studies gain power by increasing sample sizes, effect size estimates become more precise and novel associations at lower frequencies are feasible. However, rare variants are largely population-private, and their effects are unlikely to transfer to new populations. Because linkage disequilibrium decay and allele frequencies vary across ancestries, effect size estimates from diverse cohorts are typically more precise than from single-ancestry cohorts (and often tempered) [4], and the resolution of causal variant fine-mapping is considerably improved [79]. Across a range of genetic architectures, diverse cohorts provide the opportunity to reduce false positives. At the Mendelian end of the spectrum, for example, disentangling risk variants with incomplete penetrance from benign false positives and localizing functional effects in genes is much more feasible with large diverse population cohorts than possible with single-ancestry analyses [87, 88]. Multiple false positive reports of pathogenic variants causing hypertrophic cardiomyopathy, a relatively simple trait, have been returned to

patients of African descent or unspecified ancestry that would have been prevented if even a small number of African American samples were included in control cohorts [89]. At the highly complex end of the polygenicity spectrum, we and others have shown that the utility of polygenic risk inferences and the heritable phenotypic variance explained in diverse populations is improved with more diverse cohorts [84, 90].

Standard single-ancestry GWAS typically apply linear mixed model approaches with primarily European-descent cohorts [1-3] and incorporate principal components as covariates to control for confounding from population structure. A key concern when including multiple diverse populations in a GWAS is that there is increasing likelihood of identifying false positive variants associated with disease that are driven by allele frequency differences across ancestries. However, previous studies have analyzed association data for diverse ancestries and replicated findings across ethnicities, assuaging these concerns [6, 79, 91]. In this study, we show that long tracts of ancestrally diverse populations present in admixed samples from the Americas are easily and accurately detected. Querying population substructure within these tracts recapitulates expected trends, e.g. European ancestry in African Americans primarily descends from northern Europeans in contrast to European ancestry from Hispanic/Latinos, which primarily descends from southern Europeans, as seen previously [52]. Additionally, population substructure follows a north-south cline in the Native component of Hispanic/Latinos, and the African component of admixed African descent populations in the Americas most closely resembles reference populations from Nigeria. Admixture mapping has been successful at large sample sizes for identifying

ancestry-specific genetic risk factors for disease [92]. Given the level of accuracy and sub-continental-resolution attained with local ancestry tracts in admixed populations, we emphasize the utility of a unified framework to jointly analyze genetic associations with local ancestry simultaneously [48].

The transferability of GWAS is aided by the inclusion of diverse populations [93]. We have shown that European discovery biases in GWAS are recapitulated in local ancestry tracts in admixed samples. We have quantified GWAS study biases in ancestral populations and shown that GWAS variants are at lower frequency specifically within African and Native tracts and higher frequency in European tracts in admixed American populations. Imputation accuracy is also stratified across diverged ancestries, including across local ancestries in admixed populations. With decreased imputation accuracy especially on Native American tracts, there is decreased power for potential ancestry-specific associations. This differentially limits conclusions for GWAS in admixed population in a two-pronged manner: the ability to capture variation and the power to estimate associations.

As GWAS scale to sample sizes on the order of hundreds of thousands to millions, genetic risk prediction accuracy at the individual level improves [94]. However, we show that the utility of polygenic risk scores computed using GWAS summary statistics are dependent on genetic similarity to the discovery cohort. Polygenic risk computed from a single-ancestry study cohort can be biased in essentially any direction across diverse populations simply as a result of drift, limiting their interpretability. Directional selection

is expected to bias polygenic risk inferences even more. This study demonstrates the utility of disentangling ancestry tracts in recently admixed populations for inferring recent demographic history and identifying ancestry-stratified analytical biases to date; we also motivate the need to include more ancestrally diverse cohorts in GWAS to ensure that health disparities arising from precision medicine initiatives in genetic risk prediction do not become pervasive in individuals of admixed and non-European descent.

Competing interests

CDB is a SAB member of Liberty Biosecurity, Personalis, Inc., 23andMe, Ancestry.com, IdentifyGenomics, LLC, Etalon, Inc., and is a founder of CDB Consulting, LTD. All other authors declare that they have no competing interests.

Author contributions

ARM, CRG, RKW, CDB, MJD, EEK conceived of and designed the experiments. ARM and GLW performed the data analysis. SG, CDB, and MJD contributed analysis tools/materials. ARM wrote the manuscript with comments from CRG, RKW, GLW, MJD, SG, and EEK. All authors read and approved the final manuscript.

Acknowledgments

We thank Suyash Shringarpure, Brian Maples, and Andres Moreno-Estrada for helpful discussions about sex-biased admixture, local ancestry calling considerations, and the historical context of admixture in Hispanic/Latino populations. We thank Verneri Antilla

for providing GWAS summary statistics. We thank Jerome Kelleher for several conversations about msprime features enabling case/control association studies as well as for providing an example script. This work was supported by funds from several grants; the National Human Genome Research Institute under award numbers U01HG009080 (EEK, CDB, CRG), U01HG007419 (CDB, CRG, GLW), U01HG007417 (EEK), U01HG005208 (MJD), T32HG000044 (CRG) and R01GM083606 (CDB), and the National Institute of General Medical Sciences under award number T32GM007790 (ARM) at the National Institute of Health; the Directorate of Mathematical and Physical Sciences award 1201234 (SG, CDB) at the National Science Foundation; the Canadian Insitutes of Health Research through the Canada Research Chair program and operating grant MOP-136855 (SG), and a Sloan Research Fellowship (to SG).

Tables

Table 1 – Three-way admixture proportions between recently admixed populations in the Americas. Values are computed at K=3 on common autosomal SNPs using ADMIXTURE with mean percentages ± standard deviations.

	AFR	EUR	NAT
ACB	88.0% (7.7%)	11.7% (7.3%)	0.3% (1.1%)
ASW	75.6% (13.8%)	21.3% (9.1%)	3.1% (9.2%)
CLM	7.8% (13.8%)	66.6% (12.8%)	25.7% (9.3%)
MXL	4.3% (2.2%)	48.7% (18.6%)	47.0% (19.1%)
PEL	2.5% (5.4%)	20.2% (12.0%)	77.3% (14.2%)
PUR	13.9% (5.4%)	73.2% (10.0%)	12.9% (3.6%)

Table 2 – Comparison of mean ancestry proportions and ratio on chromosome X versus autosomes across populations. Per Lind et al, proportion X in a population = (fraction

male + 2*fraction female) / 1.5, and proportion autosome in a population = fraction male + fraction female. P-values are from two-sided t-tests on individual ancestries (comparisons are not independent as ancestry proportions must sum to one).

	Ancestry	ACB	ASW	CLM	MXL	PEL	PUR
Relative X/autosome % change	AFR	4.01	0.83	-2.02	-20.32	50.75	12.69
	EUR	-41.73	-17.41	-20.20	-26.60	-41.51	-14.51
	NAT	558.04	87.41	52.70	28.49	9.37	66.89
p-value	AFR	8.9e-2	7.7e-1	9.8e-1	6.8e-2	3.5e-1	4.1e-1
	EUR	1.0e-3	8.9e-2	1.4e-7	7.9e-4	4.5e-6	1.5e-7
	NAT	7.2e-9	1.1e-1	4.0e-9	3.9e-4	1.3e-3	1.4e-10

Figure captions

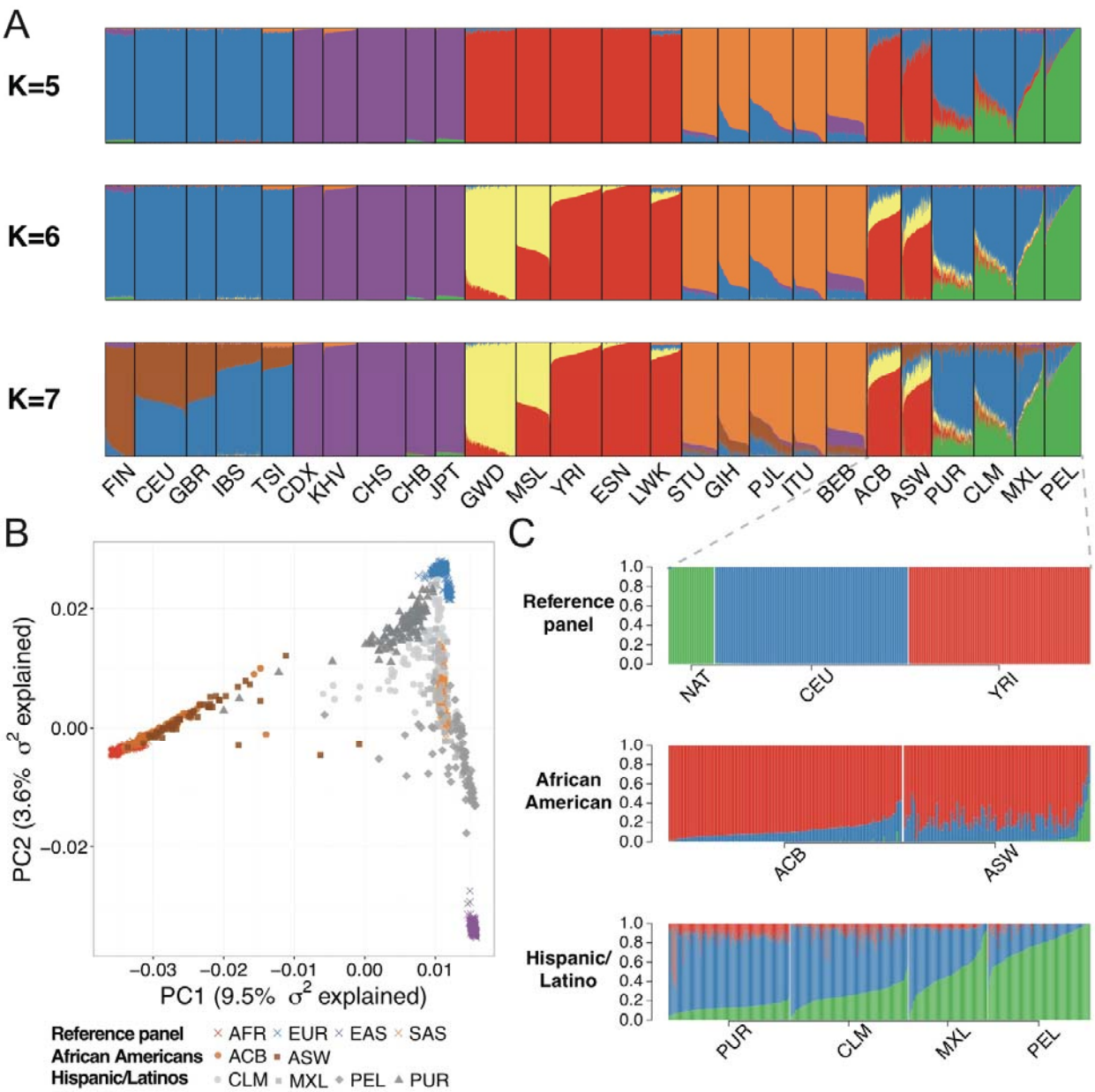

Figure 1 – Global diversity in 1000 Genomes data. A) ADMIXTURE analysis at K=5, K=7, and K=8. K=8 has the lowest 10-fold cross-validation error of K=3-12. B) Principal components analysis of all samples showing the relative homogeneity of AFR, EUR, EAS, and SAS continental groups and continental mixture of admixed samples from the Americas (ACB, ASW, CLM, MXL, PEL, and PUR). C) ADMIXTURE analysis at K=3

focusing on admixed Americas samples, with the NAT [66], CEU, and YRI as reference populations.

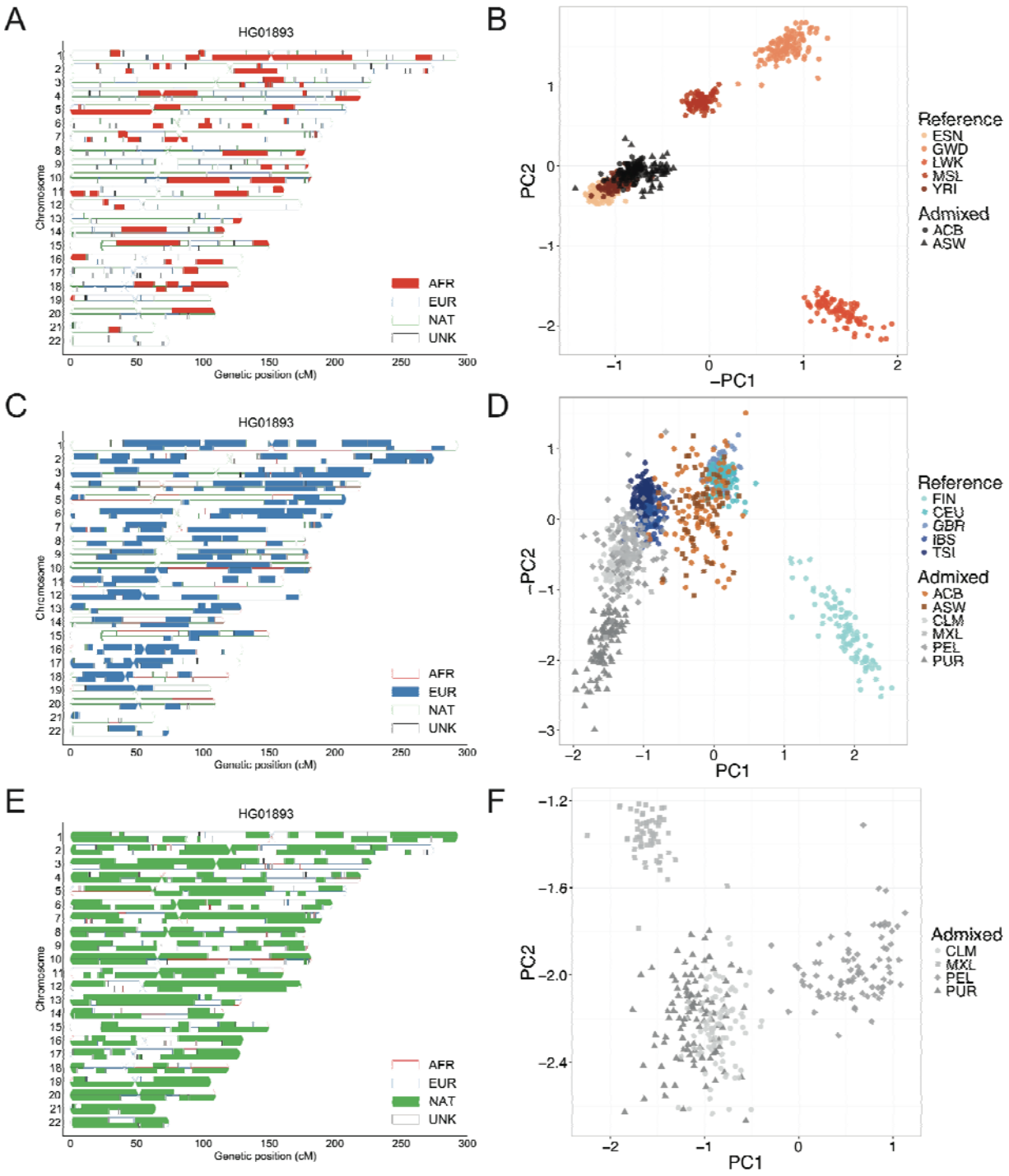

Figure 2 – Sub-continental origins of African, European, and Native American components of recently admixed Americas populations. A,C,E) Local ancestry

karyograms for representative PEL individual HG01893 with A) African, C) European, and E) Native American components shown. B,D,F) Ancestry-specific PCA applied to admixed haploid genomes as well as ancestrally homogeneous continental reference populations from 1000 Genomes (where possible) for B) African tracts, D) European tracts, and F) Native American tracts. A small number of admixed samples that constituted major outliers from the ancestry-specific PCA analysis were removed, including B) 1 ASW sample (NA20314) and D) 8 samples, including 3 ACB, 2 ASW, 1 PEL, and 2 PUR samples.

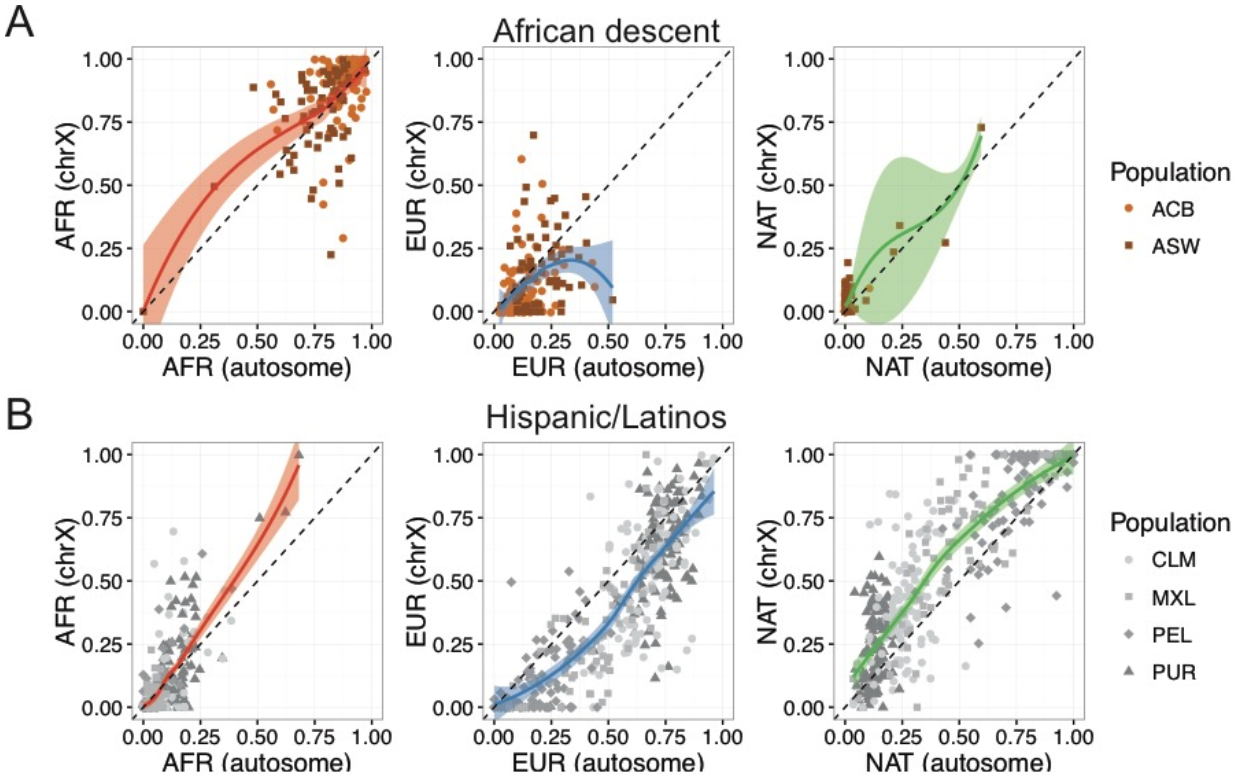

Figure 3 – Comparison of ploidy-adjusted ADMIXTURE ancestry estimates obtained on the autosomes and X chromosome at K=3 with CEU, YRI, and NAT [66] reference samples. 700,093 SNPs on the autosomes and 10,503 SNPs on the X chromosome

were used to infer ancestry proportions. A) African descent and B) Hispanic/Latino samples.

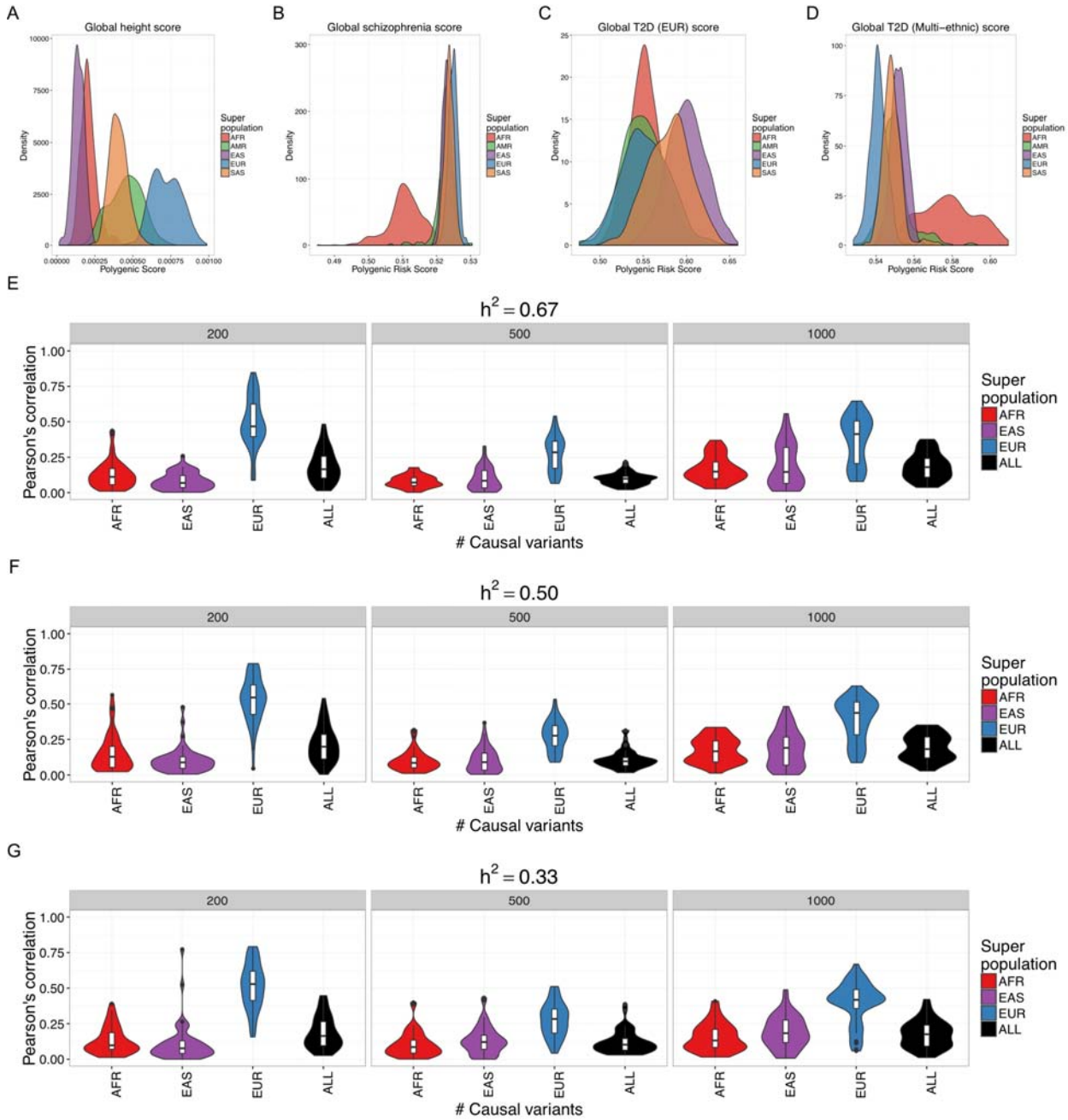

Figure 4 – Biased genetic discoveries influence disease risk inferences. A-D) Inferred polygenic risk scores across individuals colored by population for: A) height based on summary statistics from [15]. B) schizophrenia based on summary statistics from [16].

C) type II diabetes summary statistics derived from a European cohort from [78]. D) type II diabetes summary statistics derived from a multiethnic cohort from [79]. E-G) Coalescent simulation results for true and inferred polygenic risk scores derived from GWAS summary statistics with 10,000 European cases and 10,000 European controls. Violin plots show Pearson's correlation across 50 iterations per parameter set between true and inferred polygenic risk scores across differing genetic architectures, including *m*=200, 500, and 1,000 causal variants. The "ALL" population correlations were performed on population mean-centered true and inferred polygenic risk scores. Heritabilities of each panel are: E) h^2=0.67, F) h^2=0.5, and G) h^2=0.33.

Availability of data and materials

Local ancestry calls are available at the following ftp site: https://personal.broadinstitute.org/armartin/tgp_admixture/. Scripts used to process data from this manuscript are available here: https://github.com/armartin/ancestry_pipeline/. We used SHAPEIT2 (v2.r778) was used for phasing. RFMix v1.5.4 was used for local ancestry calls. PCAmask was used for ancestry-specific PCA and is available here: https://sites.google.com/site/pcamask/dowload. Tracts, used for demographic inference, is available here: https://github.com/sgravel/tracts. Msprime used for coalescent simulations is available here: https://github.com/jeromekelleher/msprime.

Methods

Ancestry deconvolution

We used the phased haplotypes from the 1000 Genomes consortium available here: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/shapeit2_scaffolds/wgs_gt_scaffolds/. We phased the Native American reference haplotypes using SHAPEIT2 (v2.r778) [95]. We merged the haplotypes using scripts made publicly available (https://github.com/armartin/ancestry_pipeline). These combined phased haplotypes were used as input to the PopPhased version of RFMix v1.5.4 [65] with the following flags: -w 0.2, -e 1, -n 5, --use-reference-panels-in-EM, --forward-backward. The node size of 5 was selected to reduce bias resulting from unbalanced reference panel sizes (AFR panel N=504, EUR panel N=503, and NAT panel N=43). We used the default minimum window size of 0.2 cM to enable model comparisons with previously inferred models using *Tracts*. We used 1 EM iteration to improve the local ancestry calls without substantially increasing computational complexity. We used the reference panel in the EM to take better advantage of the Native American ancestry tracts from the Hispanic/Latinos in the EM given the small NAT reference panel. We set the LWK, MSL, GWD, YRI, and ESN as reference African populations, the CEU, GBR, FIN, IBS, and TSI as reference European populations, and the samples from [66] with inferred > 0.99 Native ancestry as reference Native American populations, as in [96].

Ancestry-specific PCA

We performed ancestry-specific PCA, as described in [37] (PCAmask software available here: https://sites.google.com/site/pcamask/dowload). The resulting matrix is not necessarily orthogonalized, so we subsequently performed singular value

decomposition in python 2.7 using numpy. There were a small number of major outliers, as seen previously [37]. There was one outlier (ASW individual NA20314) when analyzing the African tracts, which was expected as this individual has no African ancestry. There were 8 outliers (PUR HG00731, PUR HG00732, ACB HG01880, ACB HG01882, PEL HG01944, ACB HG02497, ASW NA20320, ASW NA20321) when analyzing the European tracts. Some of these outliers had minimal European ancestry, had South or East Asian ancestry misspecified as European ancestry resulting from a 3-way ancestry reference panel, or were unexpected outliers. As described in the PCAmask manual, a handful of major outliers sometimes occur, and AS-PCA is an iterative procedure. We therefore removed the major outliers for each sub-continental analysis and orthogonalized the matrix again.

Tracts

The RFMix output was collapsed into haploid bed files, and an “UNK” or unknown ancestry was assigned where the posterior probability of a given ancestry was < 0.90. These collapsed haploid tracts were used to infer admixture timings, quantities, and proportions for the ACB and PEL (new to phase 3) using *Tracts [67]*. Because the ACB have a very small proportion of Native American ancestry, we fit three 2-way models of admixture, including one model of single- and two models of double-pulse admixture events, using *Tracts*. In both of the double-pulse admixture models, the model includes an early mixture of African and European ancestry followed by another later pulse of either European or African ancestry. We randomized starting parameters and fit each model 100 times and compared the log-likehoods of the model fits. The single-pulse

and double-pulse model with a second wave of African admixture provided the best fits and reached similar log-likelihoods, with the latter showing a slight improvement in fit.

We next assessed the fit of 9 different models in *Tracts* for the PEL [67], including several two-pulse and three-pulse models. Ordering the populations as NAT, EUR, and AFR, we tested the following models: ppp_ppp, ppp_pxp, ppp_xxp, ppx_xxp, ppx_xxp_ppx, ppx_xxp_pxx, ppx_xxp_pxp, ppx_xxp_xpx, and ppx_xxp_xxp, where the order of each letter corresponds with the order of population given above, an underscore indicates a distinct migration event with the first event corresponding with the most generations before present, p corresponding with a pulse of the ordered ancestries, and x corresponding with no input from the ordered ancestries. We tested all 9 models preliminarily 3 times, and for all models that converged and were within the top 3 models, we subsequently fit each model with 100 starting parameters randomizations.

Imputation accuracy

Imputation accuracy was calculated using a leave-one-out internal validation approach. Two array designs were compared for this analysis: Illumina OmniExpress and Affymetrix Axiom World Array LAT. Sites from these array designs were subset from chromosome 9 of the 1000 Genomes Project Phase 3 release for admixed populations. After fixing these sites, each individual was imputed using the rest of the dataset as a reference panel.

Overall imputation accuracy was binned by minor allele frequency (0.5-1%, 1-2%, 2-3%, 3-4%, 4-5%, 5-10%, 10-20%, 20-30%, 30-40%, 40-50%) comparing the genotyped true alleles to the imputed dosages. A second round of analyses stratified the imputation by local ancestry diplotype, which was estimated as described earlier. Within each ancestral diplotype (AFR_AFR, AFR_NAT, AFR_EUR, EUR_EUR, EUR_NAT, NAT_NAT), imputation accuracy was again estimated within MAF bins.

Empirical polygenic risk score inferences

To compute polygenic risk scores in the 1000 Genomes samples using summary statistics from previous GWAS, we first filtered to biallelic SNPs and removed ambiguous AT/GC SNPs from the integrated 1000 Genome callset. To take LD into account when multiple significant p-value associations are in the same region in a GWAS, we used a greedy algorithm that orders SNPs by p-value, then selectively removes SNPs within close proximity and LD in ascending p-value order (i.e. starting with the most significant SNP). To perform the greedy p-value pruning to get relatively independent associations, we used the --clump option in plink [97] for all variants with a MAF ≥ 0.01. As a population cohort with similar LD patterns to the study sets, we used European 1000 Genomes samples (CEU, GBR, FIN, IBS, and TSI). To compute the polygenic risk scores, we considered all SNPs with p-values ≤ 1e-2 in the GWAS study, a window size of 250 kb, and an R^2 threshold of 0.5 in Europeans to group SNPs. After obtaining the top clumped signals, we computed scores using the --score flag in plink.

Polygenic risk score simulations

We simulated genotypes in a coalescent framework with msprime v1.3 [98] for chromosome 20 incorporating a recombination map of GRCh37 and an assumed mutation rate of 2e-8 mutations / (base pair * generation). We used a demographic model previously inferred using 1000 Genomes sequencing data [9] to simulate individuals that reflect European, East Asian, and African population histories. To imitate a GWAS with European sample bias and evaluate polygenic risk scores in other populations, we simulated 200,000 European individuals, 200,000 East Asian, and 200,000 African individuals. Next, we assigned "true" causal effect sizes to *m* evenly spaced alleles. Specifically, we randomly assigned effect sizes as:

$$\beta \sim N(0, \frac{h^2}{m})$$

where the normal distribution is specified by the mean and standard deviation (as in python's numpy package). For all other non-causal sites, the effect size is zero. We then define *X* as:

$$X = \sum_{i=1}^{m} g_i \beta_i$$

where the g_i are the genotype states (i.e. 0, 1, or 2). To handle varying allele frequencies, potential weak LD between causal sites, ensure a neutral model with random true polygenic risks with respect to allele frequencies, and to obtain the total desired variance, we normalize *X* as:

$$Z_X = \frac{X - \mu_X}{\sigma_X}$$

We then compute the true polygenic risk score, as:

$$G = \sqrt{h^2} * Z_X$$

such that the total variance of the scores is h^2. We also simulated environmental noise and standardize to ensure equal variance between normalized genetic and environmental effects before, defining the environmental effect *E* as:

$$\epsilon = N(0, 1 - h^2)$$

$$Z_\epsilon = \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon}$$

$$E = \sqrt{1 - h^2} * Z_\epsilon$$

such that the total variance of the environmental effect is $1 - h^2$. We then define the total liability as:

$$T = \sqrt{h^2} * Z_X + \sqrt{1 - h^2} * Z_\epsilon$$
$$= G + E$$

We assigned 10,000 European individuals at the most extreme end of the liability threshold "case" status assuming a prevalence of 5%. We randomly assigned 10,000 different European individuals "control" status. We ran a GWAS with these 10,000 European cases and 10,000 European controls, computing Fisher's exact test for all sites with MAF > 0.01. As before for empirical polygenic risk score calculations from real GWAS summary statistics, we clumped these SNPs into LD blocks for all sites with $p \leq 1e\text{-}2$, $R^2 \geq 0.5$ in Europeans, and within a window size of 250 kb. We used these SNPs to compute inferred polygenic risk scores as before, summing the product of the log odds and genotype for the true polygenic risk in a cohort of 10,000 simulated European, African, and East Asian individuals (all not included in the simulated GWAS). We compared the true versus inferred polygenic risk scores for these individuals across varying complexities (*m* = 200, 500, 1000) and heritabilites (h^2_g = 0.33, 0.50, 0.67).

References

1. Need AC, and Goldstein DB: **Next generation disparities in human genomics: concerns and remedies.** *Trends in genetics : TIG* 2009**, 25**(11):489-94.

2. Bustamante CD, Francisco M, and Burchard EG: **Genomics for the world.** *Nature* 2011**, 475**(7355):163-165.

3. Petrovski S, and Goldstein DB: **Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine.** *Genome Biology* 2016**, 17**(1):157.

4. Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, et al: **Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study.** *PLoS Biol* 2013**, 11**(9):e1001661.

5. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al: **Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations.** *Nature genetics* 2011**, 43**(9):887-92.

6. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, et al: **Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups.** *PLoS Genet* 2010**, 6**(8).

7. 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.** *Proceedings of the National Academy of Sciences* 2009**, 106**(23):9362-9367.

8. Mathieson I, and McVean G: **Differential confounding of rare and common variants in spatially structured populations.** *Nature Genetics* 2012**, 44**(3):243-246.

9. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al: **Demographic history and rare allele sharing among human populations.** *Proceedings of the National Academy of Sciences of the United States of America* 2011**, 108**(29):11983-8.

10. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al: **The UK10K project identifies rare variants in health and disease.** *Nature* 2015.

11. Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean PS, Verzilli C, et al: **An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people.** *Science* 2012**, 337**(6090):100-104.

12. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al: **A global reference for human genetic variation.** *Nature* 2015**, 526**(7571):68-74.

13. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al: **Genes mirror geography within Europe.** *Nature* 2008**, 456**(7218):98-101.

14. Do R, Kathiresan S, and Abecasis GR: **Exome sequencing and complex disease: Practical aspects of rare variant association studies.** *Human Molecular Genetics* 2012**, 21**(R1):1-9.

15. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al: **Defining the role of common variation in the genomic and biological architecture of adult human height.** *Nature genetics* 2014**, 46**(11):1173-86.

16. Schizophrenia Working Group of the Psychiatric Genomics Consortium: **Biological insights from 108 schizophrenia-associated genetic loci.** *Nature* 2014**, 511**(7510):421-7.

17. Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, and Tenesa A: **Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank.** *Nat Genet* 2016.

18. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes.** *Science (New York, N.Y.)* 2012**, 337**(6090):64-9.

19. Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al: **A composite of multiple signals distinguishes causal variants in regions of positive selection.** *Science (New York, N.Y.)* 2010**, 327**(5967):883-6.

20. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al: **A systematic survey of loss-of-function variants in human protein-coding genes.** *Science (New York, N.Y.)* 2012**, 335**(6070):823-8.

21. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et alD: **Distance from sub-Saharan Africa predicts mutational load in diverse human genomes.** *Proceedings of the National Academy of Sciences of the United States of America* 2016**, 113**(4):E440-9.

22. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, et al: **Proportionally more deleterious genetic variation in European than in African populations.** *Nature* 2008**, 451**(7181):994-7.

23. Fu W, Gittelman RM, Bamshad MJ, and Akey JM: **Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations.** *The American Journal of Human Genetics* 2014**, 95**(4):421-436.

24. Simons YB, Turchin MC, Pritchard JK, and Sella G: **The deleterious mutation load is insensitive to recent population history.** *Nature genetics* 2014**, 46**(3):220-4.

25. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh P, Govindaraj P, et al: **Genetic evidence for recent population mixture in India.** *American journal of human genetics* 2013**, 93**(3):422-38.

26. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al: **Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History.** *Science (New York, N.Y.)* 2012**, 374.**

27. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al: **A draft sequence of the Neandertal genome.** *Science (New York, N.Y.)* 2010**, 328**(5979):710-22.

28. Stokowski RP, Pant PVK, Dadd T, Fereday A, Hinds DA, Jarman C, et al: **A genomewide association study of skin pigmentation in a South Asian population.** *American journal of human genetics* 2007**, 81**(6):1119-32.

29. Marcheco-Teruel B, Parra EJ, Fuentes-Smith E, Salas A, Buttenschøn HN, Demontis D, et al: **Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers.** *PLoS genetics* 2014**, 10**(7):e1004488.

30. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque J, et al: **A genome-wide association scan in admixed Latin Americans**

identifies loci influencing facial and scalp hair features. *Nature Communications* 2016**, 7**:10815.

31. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al: **Sensitive detection of chromosomal segments of distinct ancestry in admixed populations.** *PLoS genetics* 2009**, 5**(6):e1000519.

32. Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WHL, et al: **Enhanced statistical tests for GWAS in admixed populations: Assessment using african americans from CARe and a breast cancer consortium.** *PLoS Genetics* 2011**, 7**(4).

33. Fejerman L, Chen GK, Eng C, Huntsman S, Hu D, Williams A, et al: **Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas.** *Human molecular genetics* 2012**, 21**(8):1907-17.

34. Fejerman L, Ahmadiyeh N, Hu D, Huntsman S, Beckman KB, Caswell JL, et al: **Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25.** *Nat Commun* 2014**, 5**:5260.

35. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al: **Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men.** *Proceedings of the National Academy of Sciences of the United States of America* 2006**, 103**(38):14068-14073.

36. Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, Pollack S, et al: **Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection.** *American journal of human genetics* 2011**, 89**(3):368-81.

37. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et al: **Reconstructing the Population Genetic History of the Caribbean.** *PLoS Genetics* 2013**, 9**(11):e1003925.

38. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, et al: **Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations.** *Proceedings of the National Academy of Sciences of the United States of America* 2010**, 107 Suppl** :8954-61.

39. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, Acuña-Alonzo V, et al: **The genetics of Mexico recapitulates Native American substructure and affects biomedical traits.** *Science (New York, N.Y.)* 2014**, 344**(6189):1280-5.

40. Pritchard JK, Stephens M, and Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000**, 155**(2):945-959.

41. Tang H, Peng J, Wang P, and Risch NJ: **Estimation of individual admixture: Analytical and study design considerations.** *Genetic Epidemiology* 2005**, 28**(4):289-301.

42. Alexander DH, Novembre J, and Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome research* 2009**, 19**(9):1655-64.

43. Price AL, Zaitlen NA, Reich D, and Patterson N: **New approaches to population stratification in genome-wide association studies.** *Nature reviews. Genetics* 2010**, 11**(7):459-63.

44. Mathieson I, and Mcvean G: **Demography and the Age of Rare Variants.** 2014**, 10**(8).

45. O'Connor TD, Fu W, Mychaleckyj JC, Logsdon B, Auer P, Carlson CS, et al: **Rare variation facilitates inferences of fine-scale population structure in humans.** *Mol Biol Evol* 2015**, 32**(3):653-60.

46. Babron MC, de Tayrac M, Rutledge DN, Zeggini E, and Génin E: **Rare and low frequency variant stratification in the UK population: description and impact on association tests.** *PLoS One* 2012**, 7**(10):e46519.

47. Bhatia G, Gusev A, Loh P-R, Finucane HK, Vilhjalmsson BJ, Ripke S, et al: **Subtle stratification confounds estimates of heritability from rare variants.** *bioRxiv* 2016:048181.

48. Szulc P, Bogdan M, Frommlet F, and Tang H: **Joint Genotype-and Ancestry-based Genome-wide Association Studies in Admixed Populations.** *bioRxiv* 2016:062554.

49. Conomos M, Reiner A, Weir B, and Thornton T: **Model-free Estimation of Recent Genetic Relatedness.** *The American Journal of Human Genetics* 2016**, 98**(1):127-148.

50. Zaitlen N, Paşaniuc B, Gur T, Ziv E, and Halperin E: **Leveraging genetic variability across populations for the identification of causal variants.** *Am J Hum Genet* 2010**, 86**(1):23-33.

51. Genovese G, Handsaker RE, Li H, Kenny EE, and McCarroll SA: **Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes.** *Am J Hum Genet* 2013**, 93**(3):411-21.

52. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al: **The Great Migration and African-American Genomic Diversity.** *PLoS genetics* 2016**, 12**(5):e1006059.

53. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al: **Reconstructing Native American population history.** *Nature* 2012**, 488**(7411):370-4.

54. Ruiz-Linares A, Adhikari K, Acuña-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, et al: **Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals.** *PLoS Genet* 2014**, 10**(9):e1004572.

55. Wu Y, Waite LL, Jackson AU, Sheu WH, Buyske S, Absher D, et al: **Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained.** *PLoS Genet* 2013**, 9**(3):e1003379.

56. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al: **Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.** *Am J Hum Genet* 2015**, 97**(4):576-92.

57. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al: **Searching for missing heritability: designing rare variant association studies.** *Proceedings of the National Academy of Sciences of the United States of America* 2014**, 111**(4):E455-64.

58. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al: **Analysis of protein-coding genetic variation in 60,706 humans.** *Nature* 2016**, 536**(7616):285-291.

59. Shringarpure SS, Bustamante CD, Lange KL, and Alexander DH: **Efficient analysis of large datasets and sex bias with ADMIXTURE.** *bioarXiv* 2016**, 1**(1):1-10.

60. Basu A, Sarkar-Roy N, and Majumder PP: **Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure.** *Proc Natl Acad Sci U S A* 2016**, 113**(6):1594-9.

61. Reich D, Thangaraj K, Patterson N, Price AL, and Singh L: **Reconstructing Indian population history.** *Nature* 2009**, 461**(7263):489-494.

62. Mimno D, Blei DM, and Engelhardt BE: **Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure.** *Proc Natl Acad Sci U S A* 2015**, 112**(26):E3441-50.

63. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al: **Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences.** *Nature Genetics* 2016.

64. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al: **Fast and accurate inference of local ancestry in Latino populations.** *Bioinformatics* 2012**, 28**(10):1359-67.

65. Maples BK, Gravel S, Kenny EE, and Bustamante CD: **RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.** *American journal of human genetics* 2013**, 93**(2):278-288.

66. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, et al: **A genomewide admixture mapping panel for Hispanic/Latino populations.** *American journal of human genetics* 2007**, 80**(6):1171-1178.

67. Gravel S: **Population genetics models of local ancestry.** *Genetics* 2012**, 191**(2):607-19.

68. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et al: **Reconstructing the Population Genetic History of the Caribbean.** *PLoS Genetics* 2013**, 9**(11):e1003925.

69. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al: **The genetic structure and history of Africans and African Americans.** *Science (New York, N.Y.)* 2009**, 324**(5930):1035-44.

70. Zakharia F, Basu A, Absher D, Assimes TL, Go AS, Hlatky MA, et al: **Characterizing the admixed African ancestry of African Americans.** *Genome biology* 2009**, 10**(12):R141.

71. Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, et al: **Reconstructing Native American migrations from whole-genome and whole-exome data.** *PLoS genetics* 2013**, 9**(12):e1004023.

72. Lind JM, Hutcheson-Dilks HB, Williams SM, Moore JH, Essex M, Ruiz-Pesini E, et al: **Elevated male European and female African contributions to the genomes of African American individuals.** *Human Genetics* 2007**, 120**(5):713-722.

73. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al: **Genome-wide patterns of population structure and admixture in West Africans and African Americans.** *Proceedings of the National Academy of Sciences of the United States of America* 2010**, 107**(2):786-91.

74. Bryc K, Durand EY, Macpherson JM, Reich D, and Mountain JL: **The genetic ancestry of african americans, latinos, and european Americans across the United States.** *American Journal of Human Genetics* 2015**, 96**(1):37-53.

75. Beleza S, Campos J, Lopes J, Araújo II, Hoppfer Almada A, Correia e Silva A, et al: **The Admixture Structure and Genetic Variation of the Archipelago of Cape Verde and Its Implications for Admixture Mapping Studies.** *PLoS ONE* 2012**, 7**(11):1-12.

76. McHugh C, Thornton TA, and Brown L: **Detecting Heterogeneity in Population Structure Across the Genome in Admixed Populations.** *bioRxiv* 2015:031831.

77. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, et al: **New genetic loci link adipose and insulin biology to body fat distribution.** *Nature* 2015**, 518**(7538):187-96.

78. Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Mägi R, Reschen ME, et al: **Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci.** *Nat Genet* 2015**, 47**(12):1415-25.

79. Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, et al: **Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility.** *Nat Genet* 2014**, 46**(3):234-44.

80. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al: **A large-scale, consortium-based genomewide association study of asthma.** *New England Journal of Medicine* 2010**, 363**(13):1211-1221.

81. N'Diaye A, Chen GK, Palmer CD, Ge B, Tayo B, Mathias RA, et al: **Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry.** *PLoS Genet* 2011**, 7**(10):e1002298.

82. Gustafsson A, and Lindenfors P: **Human size evolution: no evolutionary allometric relationship between male and female stature.** *J Hum Evol* 2004**, 47**(4):253-66.

83. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al: **Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010.** *The Lancet* 2013**, 382**(9904):1575-1586.

84. De Candia TR, Lee SH, Yang J, Browning BL, Gejman PV, Levinson DF, et al: **Additive genetic variation in schizophrenia risk is shared by populations of African and European descent.** *American Journal of Human Genetics* 2013**, 93**(3):463-470.

85. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al: **Hundreds of variants clustered in genomic loci and biological pathways affect human height.** *Nature* 2010**, 467**(7317):832-838.

86. Marquez-Luna C, Price AL, and Consortium ST2D: **Multi-ethnic polygenic risk scores improve risk prediction in diverse populations.** *bioRxiv* 2016:051458.

87. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al: **Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples.** *bioRxiv* 2016:041111.

89. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al: **Genetic Misdiagnoses and the Potential for Health Disparities.** *New England Journal of Medicine* 2016**, 375**(7):655-665.

90. Li YR, and Keating BJ: **Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations.** *Genome Med* 2014**, 6**(10):91.

91. Dumitrescu L, Carty CL, Taylor K, Schumacher FR, Hindorff LA, Ambite JL, et al: **Genetic Determinants of Lipid Traits in Diverse Populations from the Population Architecture using Genomics and Epidemiology (PAGE) Study.** *PLoS Genetics* 2011**, 7**(6):e1002138.

92. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al: **Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men.** *Proceedings of the National Academy of Sciences* 2006**, 103**(38):14068-14073.

93. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, and Boehnke M: **Genome-wide association studies in diverse populations.** *Nature reviews. Genetics* 2010**, 11**(5):356-366.

94. Dudbridge F: **Power and predictive accuracy of polygenic risk scores.** *PLoS Genet* 2013**, 9**(3):e1003348.

95. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al: **A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness.** *PLoS Genetics* 2014**, 10**(4):e1004234.

96. 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012**, 135**(V):0-9.

97. Purcell S, Neale B, Toddbrown K, Thomas L, Ferreira M, Bender D, et al: **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.** *The American Journal of Human Genetics* 2007**, 81**(3):559-575.

98. Kelleher J, Etheridge AM, and McVean G: **Efficient coalescent simulation and genealogical analysis for large sample sizes.** *PLoS Comput Biol* 2016**,** **12**(5):e1004842.