



Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power

Elizabeth G. Atkinson^{1,2,3}✉, Adam X. Maihofer¹, Masahiro Kanai^{1,2,3,5,6}, Alicia R. Martin^{1,2,3}, Konrad J. Karczewski^{1,2}, Marcos L. Santoro^{2,7,8}, Jacob C. Ulirsch^{1,2,3,9}, Yoichiro Kamatani¹⁰, Yukinori Okada^{6,11,12}, Hilary K. Finucane^{1,2,3}, Karestan C. Koenen^{2,13}, Caroline M. Nievergelt^{1,4,15}, Mark J. Daly^{1,2,3,14,15} and Benjamin M. Neale^{1,2,15}

Admixed populations are routinely excluded from genomic studies due to concerns over population structure. Here, we present a statistical framework and software package, Tractor, to facilitate the inclusion of admixed individuals in association studies by leveraging local ancestry. We test Tractor with simulated and empirical two-way admixed African-European cohorts. Tractor generates accurate ancestry-specific effect-size estimates and Pvalues, can boost genome-wide association study (GWAS) power and improves the resolution of association signals. Using a local ancestry-aware regression model, we replicate known hits for blood lipids, discover novel hits missed by standard GWAS and localize signals closer to putative causal variants.

Admixed groups, whose genomes contain more than one ancestral population, make up more than one-third of the US populace, with the population becoming increasingly mixed over time¹. Many common, heritable diseases, including prostate cancer^{2–5}, asthma^{6–9} and certain cardiovascular disorders^{10,11}, are enriched in admixed populations, which include African American and Hispanic/Latino individuals. However, only a minute proportion of association studies address the genetic architecture of complex traits in these groups^{12,13}; admixed individuals are systematically removed from many large-scale collections. This is due in large part to the lack of methods and pipelines to effectively account for their ancestry such that population substructure can infiltrate analyses and bias the results if they are run with standard analysis of homogeneous individuals^{14–21}. Efforts to collect genetic data alongside medically relevant phenotypes are beginning to focus more on diverse groups containing higher amounts of admixture^{22–27}, motivating the development of scalable methods to allow well-calibrated statistical genomic work on these populations. If not addressed, this inability to analyze admixed people will limit the clinical utility of large-scale data-collection efforts for minorities, exacerbating existing health disparities^{28–32}.

In genome-wide association studies (GWASs), the specific concern regarding including admixed participants is false positive hits due to alleles being at different frequencies across populations. Most studies currently attempt to control for this by using

principal components in a linear or linear mixed model framework. However, principal components capture broader admixture fractions, and individuals' local ancestry makeup may differ between case and control cohorts even if their global fractions are identical. Even including principal components as covariates, then, still leaves open the possibility for false positive associations, as well as absorbing power.

Studying diverse populations in gene discovery efforts not only reduces disparities but also benefits genetic analysis for individuals of all ancestries. A notable example of this is multi-ancestry fine-mapping, which can reduce the variant credible set by leveraging populations' differing linkage disequilibrium patterns^{33–38}. This is particularly helpful in populations of African descent, where linkage disequilibrium blocks are the shortest and individuals have nearly one million more variants per person than individuals outside of the continent³⁹. We note that, in admixed populations, we can utilize linkage disequilibrium patterns from multiple ancestries as well as disrupted linkage disequilibrium blocks within each, offering a more refined linkage landscape with which to localize GWAS signals.

Here, we have developed a scalable framework that allows for the incorporation of admixed individuals into large-scale genomics efforts by using local ancestry inference (LAI) (Fig. 1). Our framework, distributed as a software package named Tractor, generates ancestry dosages at each site from LAI calls, extracts painted

¹Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. ²Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ³Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁴Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA. ⁵Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. ⁶Department of Statistical Genetics, Graduate School of Medicine, Osaka University, Suita, Japan.

⁷Departamento de Psiquiatria, Universidade Federal de São Paulo, São Paulo, Brazil. ⁸Departamento de Morfologia e Genética, Universidade Federal de São Paulo, São Paulo, Brazil. ⁹Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. ¹⁰Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan.

¹¹Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. ¹²Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan. ¹³Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ¹⁴Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. ¹⁵These authors contributed equally: Caroline M. Nievergelt, Mark J. Daly, Benjamin M. Neale. ✉e-mail: eatkinso@broadinstitute.org

haplotype segments to correct population structure at the genotype level, and runs a local ancestry-aware regression model, producing ancestry-specific effect-size estimates and *P*values. Admixed individuals can thereby be analyzed in a well-calibrated manner, either in a cohort alone or alongside homogenous groups. Through testing in simulations and with empirical data on continuous and case/control phenotypes with differing levels of polygenicity, we demonstrate that Tractor produces accurate results in admixed cohorts and boosts the GWAS power across many genetic contexts. We further demonstrate improvements in association signal localization from accounting for the ancestral backbone on which alleles fall. These efforts fill a gap in existing resources and will improve our understanding of complex diseases across diverse populations.

Incorporating local ancestry into variant identification for admixed populations has been discussed previously^{40–50}, particularly with regard to admixture mapping, whereby researchers associate an increase of a given ancestry at a locus with increased risk of a disease that is known to be stratified in prevalence across ancestries^{51–55}. Admixture mapping has proven successful in diseases that are highly stratified, such as asthma and cardiovascular phenotypes^{56–61}. However, this strategy cannot be employed for phenotypes that are observed at equivalent rates across groups and can be subject to false positives from local ancestry increases that are unrelated to the phenotype of interest. We build on this important work by modeling the local ancestry dosage for each person at each variant in a way that accounts for differences in minor allele frequency (MAF) across populations, controlling for demography without an increased false positive risk. Tractor thereby generates accurate ancestry-specific summary statistics, which admixture mapping and traditional GWAS both cannot, and can be utilized for phenotypes that are seen at similar or different rates across populations.

The statistical model built into Tractor for binary phenotypes tests each single-nucleotide polymorphism (SNP) for an association with the phenotype using the following logistic regression model:

$$\text{logit}[Y] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \dots + b_kX_k$$

where b values represent effect estimates, X_1 is the number of haplotypes of the index ancestry present at that locus for each individual, X_2 is the number of copies of the risk allele coming from the first ancestry, X_3 is the number of copies coming from the second ancestry and $X_4–X_k$ are other covariates, such as age, sex, the estimate of global ancestry and so on. The significance of the risk allele is evaluated with a likelihood ratio test comparing the full model with a model fit without the risk allele, thus allowing estimation of the aggregated effects in the presence of effect-size heterogeneity. The (two-degrees-of-freedom) model presented here is designed for a two-way admixed scenario but can be readily scaled to an arbitrary number of ancestries with the addition of terms.

Results

LAI has high accuracy for African Americans. We ran LAI using RFMix2, a discriminative approach that estimates local ancestry using conditional random fields parameterized with random forests⁶². RFMix can run on multi-way admixed populations, outperforms other LAI methods for minority populations and leverages the ancestry components in admixed reference panel individuals, which is important when there is a lack of homogenous reference panels, as is often the case for understudied groups^{63,64}. As Tractor relies heavily on LAI calls, we quantified the accuracy of RFMix using a simulated two-way admixed cohort created from African American individuals from the Psychiatric Genomics Consortium Posttraumatic Stress Disorder (PGC-PTSD) working group. LAI was highly accurate in a realistic demographic model for African American individuals^{55–67}, assigning the correct ancestry ~98% of the time (Methods and Supplementary Table 1).

To ensure that Tractor performed well across demographic models, we varied the admixture fractions and pulse timings. Specifically, we varied the pulse of admixture in time to three generations and 20 generations ago and changed the admixture fractions to 30/70% and 50/50% European (EUR) and African (AFR) ancestry, respectively (Extended Data Fig. 1). We also checked the ancestry-specific accuracy in the realistic demographic scenario to assess whether there was a bias in calling dependent on ancestry. Across all demographic models and ancestries, site-wise LAI calls were similarly accurate, with the correct call being obtained ~98% of the time (Supplementary Table 1). While we refer solely to continental-level ancestry here, we appreciate the high level of diversity and admixture within the continents and particularly in Africa. As reference panels for diverse groups grow in size, we will have increased ability to examine ever more geographically refined groups.

Recovering haplotypes disrupted by statistical phasing. While errors in statistical phasing can lead to errors in LAI, we found that iterating between LAI and statistical phasing improved the accuracy of both. Errors in statistical phasing are a major concern^{68,69}, but few methods to recover disrupted haplotypes exist. Phasing errors result in incorrect switches of the chromosome strand on which the haplotype is assigned, which artificially reduces tract lengths and increases the overall tract counts. This phenomenon could also be misinterpreted as recombination events. Taking advantage of the ability to visualize tracts offered by admixed individuals, we are able to consistently correct switch errors and recover disrupted haplotypes, making tract distributions significantly more realistic (Extended Data Fig. 2, Supplementary Table 2 and Supplementary Information).

Evaluating the landscape of GWAS power gains from Tractor. To quantify the potential increase in power from the inclusion of local ancestry, we simulated individuals' likelihoods of being cases as a function of AFR admixture fraction, the risk allele dosage and the ancestral background of each allele. This initial framework can be thought of as modeling a risk allele with a true effect only in the AFR background and/or as modeling a tagging variant having a marginal effect size estimated in only one ancestry. This latter case may arise for multiple reasons due to genomic differences across ancestries, such as a mutation being monomorphic in EUR but variable in AFR (which is probable as individuals from Africa contain almost one million more variants than other populations⁷⁰), differing demographic histories resulting in linkage disequilibrium structure variation that affects causal variant tagging, and MAF differences. Therefore, our model quantifies Tractor's ability to estimate and leverage power from ancestral differences in true effect sizes, as well as for differing marginal effect sizes of tagging SNPs, as may be systematically expected across the genome. Our simulation framework also incorporates the clinically observed phenomenon of disease prevalence differing as a function of ancestry. We then ran association tests and compared the power across the odds ratio spectrum under the traditional GWAS and Tractor models.

Compared with the traditional model, we observed a notable gain in power using the Tractor framework, with comparable improvements across sample sizes and disease prevalences (Fig. 2a). We ran similar simulations, varying the effect-size difference, absolute MAF, MAF difference across ancestries, and admixture fractions. To summarize, Tractor is most powerful when there is heterogeneity in the apparent effect size for a variant across ancestries (Fig. 2 and Extended Data Figs. 3 and 4). As described above, such heterogeneity may be a consequence of a variant truly having different effects in different populations (for example, in the context of gene-environment interactions) or may arise from differences in the indirect association evidence of the variant (that is, the contribution to the estimated effect size from tagging other causal genetic variants).

The biggest power gain within the case of effect-size heterogeneity comes if an allelic effect is present in the smaller-fraction ancestry only (Fig. 2d). For example, using a realistic African American demographic history, if we model an allele with an effect only in the EUR background (~20% of the sample), analyzing the tracts without LAI information will have essentially no power to detect an association due to the higher noise relative to the signal from uninformative tracts. However, Tractor is able to recover the effective sample size and power by analyzing just the effect haplotypes (that is, the EUR segments alone).

We assessed the required level of heterogeneity in effect sizes required for Tractor to benefit from power gains over standard GWAS. We note that the amount of heterogeneity required depends heavily on other parameters, in particular on the relative admixture fraction of the ancestry containing a larger effect. This is because the inflation of variance and bias of the effect estimate will vary based not only on the difference in means, but also the relative sizes of groups. However, all parameters being equal, unless group effects are substantially different, the penalty of additional model complexity can be more severe than the penalty of a biased estimate of effect and higher variance estimate found in a regular model. In our African American simulation framework with equivalent MAF and an effect present in both ancestries but stronger in AFR, Tractor would require an effect difference of more than ~60% to benefit from effect-size heterogeneity (Fig. 2e and Extended Data Fig. 3e,f).

Notably, however, under scenarios where there is no expected benefit from incorporating local ancestry (that is, when all features are equivalent across ancestries), the power loss is minimal, reflecting the penalty for adding an additional model parameter (Extended Data Figs. 3 and 4). In no case does Tractor dramatically underperform compared with the traditional GWAS model. Additionally, Tractor reliably estimates ancestry-specific effects, whereas in a regular GWAS the effect estimate generated would be a weighted average of these. Therefore, even when failing to demonstrate elevated power in variant identification, the Tractor model may provide more power in downstream applications where correct effect estimates are required.

We additionally benchmarked Tractor against asaMap⁵⁰. Instead of local ancestry, asaMap is based on a mixture model, where the mixture components are the phenotype distributions corresponding to the given ancestries of the tested SNP, and the mixture weights are the probabilities of these ancestries. asaMap thereby generates ancestry-specific effect sizes without local ancestry, making it a useful point of comparison. We also tested the impact of variability in LAI accuracy on Tractor, modeling perfect, realistic accuracy (98%) and a lower bound of 90% LAI accuracy. All Tractor models reached higher statistical power than both standard GWAS and asaMap (Extended Data Fig. 5). As expected, decreasing the LAI accuracy reduced the power for Tractor; however, even at the lowest accuracy tested, Tractor outperformed these other methods.

Tractor accurately estimates ancestry-specific effects. To ensure that Tractor generates reliable ancestry-specific effect sizes, we checked the effect size estimated compared with that modeled

across a range of absolute and ancestrally differing effect sizes using the simulation framework described above. Across all genomic models, Tractor accurately estimated the ancestry-specific effect size (Fig. 3).

No increase in false positive rate with the Tractor model. We quantified the type I error rate of the Tractor model by simulating a variant with no effect in either ancestry and counting the spurious significant associations identified in a simulated African American population with a 10% disease prevalence given $\alpha=0.05$. Across our tests, we observed no statistically significant difference in false

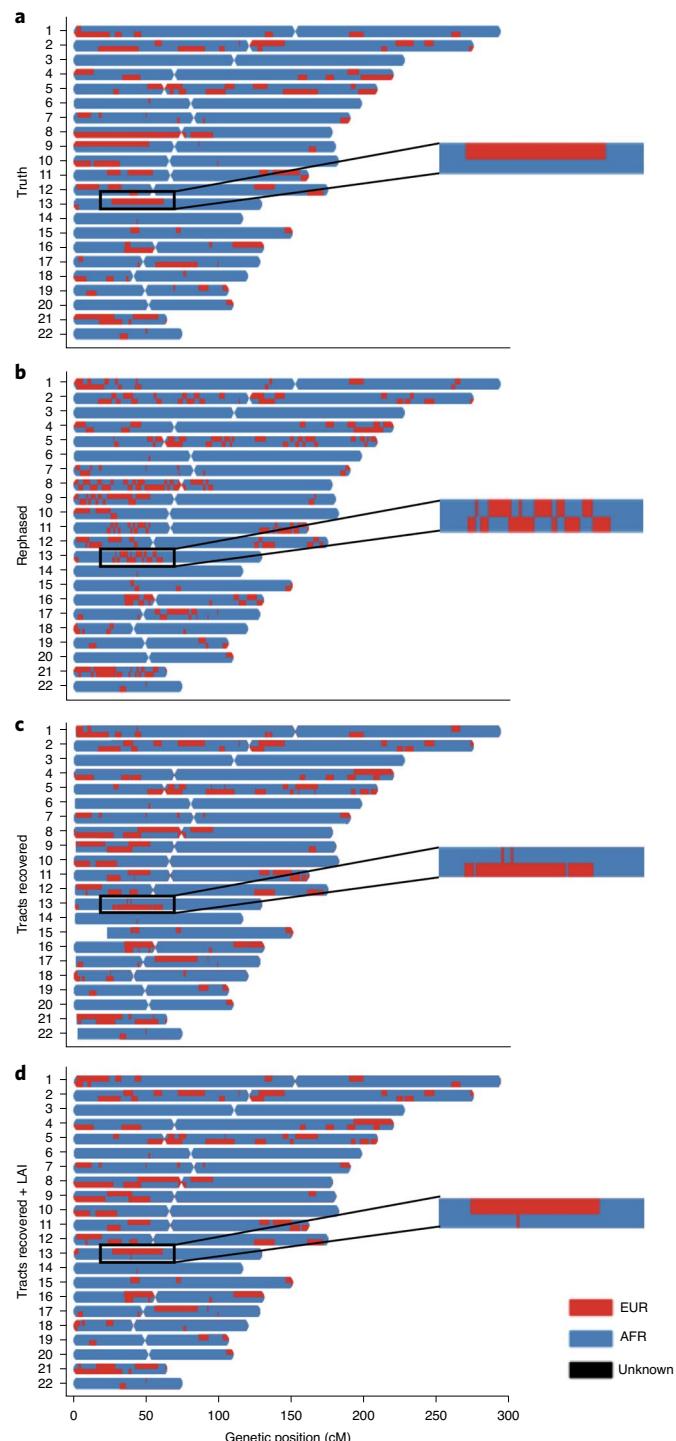


Fig. 1 | Painted karyograms of a simulated African American individual showing local EUR and AFR ancestral tracts across data treatments.

a, Truth results for an example individual in our simulated African American cohort. **b**, Results for the person after statistical phasing. Note the disruption of long haplotypes resulting from phasing switch errors. **c**, Recovery of tracts broken by switch errors in phasing. **d**, Smoothing and further improvement of tracts acquired through an additional round of LAI. The same section of chromosome 13, showing an example tract at higher resolution, is pictured on the right of each panel to highlight tract recovery. Local EUR and AFR ancestral tracts are shown in red and blue, respectively.

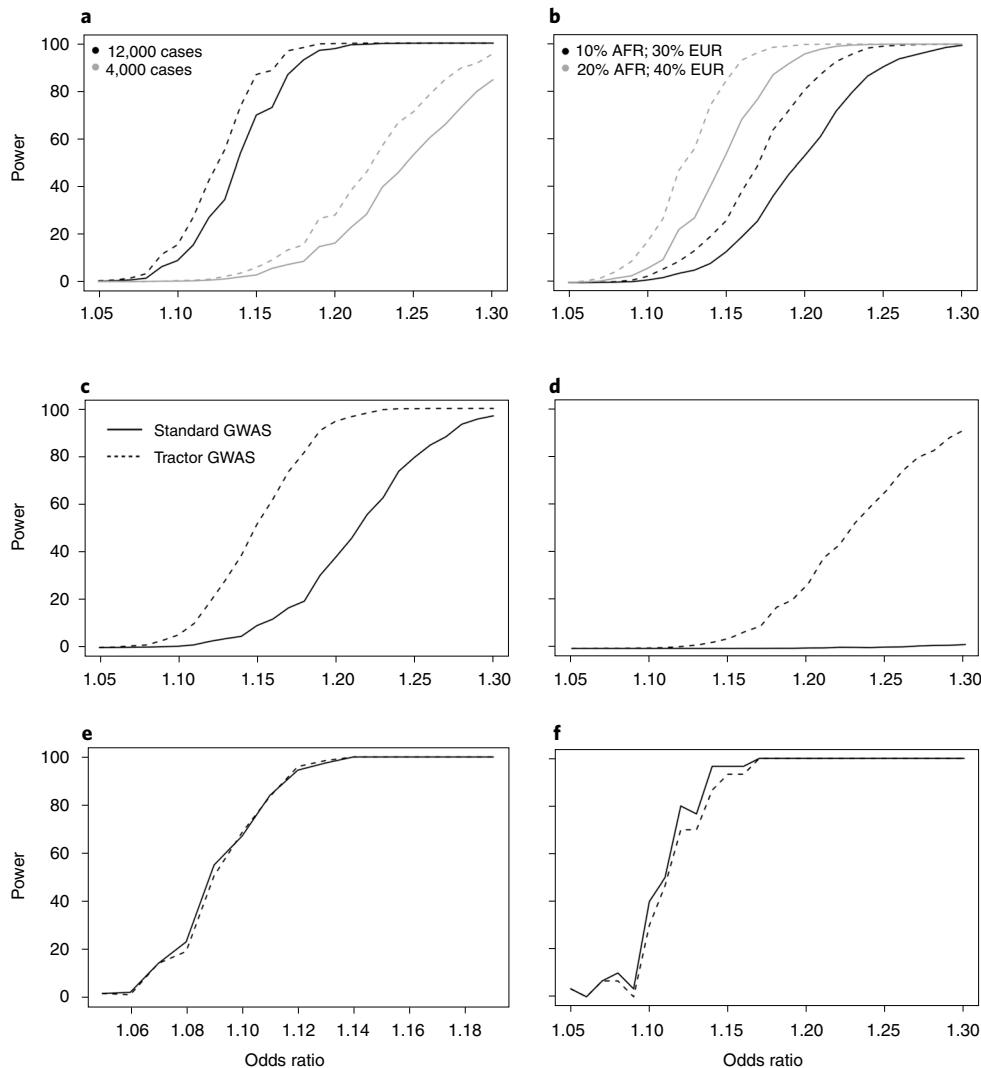


Fig. 2 | GWAS power gains across sample sizes, ancestral MAF differences, admixture proportions and effect-size differences. **a**, There are similar gains in GWAS power when using the Tractor LAI-aware model across sample sizes of 4,000 (gray) and 12,000 cases (black) with two controls. **b**, When there is a MAF difference between ancestries (black = MAF 10% AFR and 30% EUR; gray = MAF 20% AFR and 40% EUR), the boost in power is even more pronounced. Gains vary across the allele frequency spectrum. **c**, Gains become more pronounced when the admixture fractions are modified to 50/50. **d**, Dramatic gains are obtained when the effect is switched to instead only be present on the less common EUR background. **e**, Threshold for heterogeneity in ancestral effect sizes required to observe gains in Tractor power over traditional GWAS assuming 20% MAF in both ancestries and an effect that is stronger in AFR with varying difference to the EUR effect. **f**, There is a small loss in power from incorporating local ancestry into the GWAS model when all parameters are identical across ancestries. In all scenarios shown, dashed lines correspond to the power from the Tractor model incorporating local ancestry, whereas solid lines are for a traditional GWAS model. In all panels, we modeled a 10% disease prevalence. Unless otherwise noted, we used the parameters for a realistic demographic scenario for African American individuals: 80% AFR ancestry, an effect present only in the AFR genetic background, 12,000 cases and 30,000 controls, and 20% MAF.

positive rate between Tractor and the expected type I error rate of 5% (Supplementary Fig. 1). In addition, we calculated the genomic inflation factor, λ_{GC} , of null simulated phenotypes across GWAS permutations and confirmed no inflation using the Tractor GWAS model (Supplementary Fig. 2). Therefore, there does not appear to be an elevation in false positive rates with the Tractor framework, suggesting that the observed power increases result from improved detection of true biological signals.

Tractor replicates known loci and identifies novel hits. To ensure that our Tractor joint-analysis GWAS model also performs well on empirical data, we ran the method on well-characterized blood lipid phenotypes previously demonstrated to have ancestry-specific

effects: total cholesterol and low-density lipoprotein (LDL) cholesterol. We constructed a pseudo-cohort of 4,309 two-way African-European admixed individuals from the UK Biobank with blood panel phenotype data to serve as our sample. To ensure LAI was unbiased across regions of the genome, we examined its genomic distribution. Local ancestry was relatively evenly distributed and was proportional to global fractions, being 97.5% correlated with admixture proportions (Supplementary Fig. 3).

Tractor GWAS replicated known associations for blood lipids in this cohort^{43,71–73}, reaching the standard genome-wide significance level of 5×10^{-8} at previous top associations, including in the genes *PCSK9*, *LDLR* and *APOE* (Fig. 4 and Supplementary Fig. 4). In some cases, Tractor improved the observed top-hit

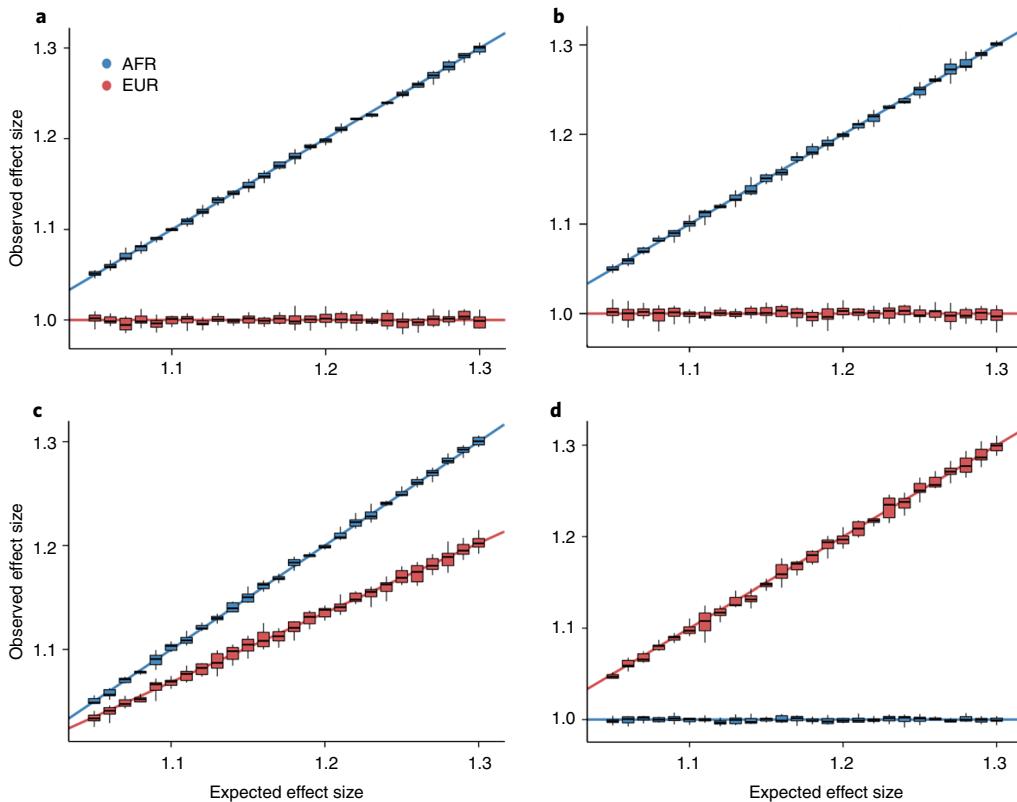


Fig. 3 | Tractor accurately estimates ancestry-specific effect sizes. **a**, The initial simulation framework of an effect only in AFR. **b**, An effect only in AFR with differing MAFs across ancestries, with AFR being 10% and EUR being 30%. **c**, An effect in both ancestries, with a 30% weaker effect modeled for EUR. **d**, Effect only in EUR. The boxplots show the effect size estimated by Tractor compared with that modeled in the simulation across a range of effect sizes. The center is the median, the bounds of the box represent the first to third quantiles and the whiskers are 1.5 \times the interquartile range. The lines indicate the simulated values for each ancestry. Blue represents effects in AFR, while red represents effects in EUR. The models presented all include 1,000 simulation replicates with 12,000 cases and 30,000 controls at 10% disease prevalence in a realistic African American population with an admixture ratio of 80/20 AFR/EUR. Unless otherwise noted, the risk allele MAF was set at 20% in both ancestries.

significance. We describe total cholesterol in detail here (see the Supplementary Information for further description of validation with LDL cholesterol).

Our model also identified additional hits in these admixed individuals that were missed by standard GWAS (Fig. 4). For example, we identified an association present only on the AFR background on chromosome 1 (rs12740374; $P = 3.46 \times 10^{-8}$). This locus has previously been shown to affect blood lipid levels, metabolic syndrome and coronary heart disease risk in independent African American cohorts^{71,74–79} and was determined to be the causal variant for affecting LDL cholesterol in a multi-ancestry fine-mapping study⁸⁰. Had we not deconvolved ancestral tracts for our GWAS, we would have missed this site with a demonstrated effect in the phenotype and population of interest.

We additionally identified a novel peak on chromosome 15 that only reached significance in the AFR tracts. The lead SNP (rs12594517; $P = 1.915 \times 10^{-8}$) lies in an intergenic area and is uncharacterized. The closest gene neighboring it is *MEIS2*, lying ~70 kilobases (kb) upstream, followed by *C15orf41*. While the precise role and mechanism this locus plays in affecting blood lipids remains unclear, *MEIS2* has previously been found to be associated with body mass index and waist circumference, and *C15orf41* was a significant hit in a previous GWAS of cholesterol^{81,82}. Although further follow-up is needed to clarify any direct relationship to total cholesterol, this association highlights the utility of Tractor to identify signals that would be undetectable in admixed cohorts without accounting for local ancestry.

Tractor is also able to refine the location of signals closer to the causal variant than is possible using standard GWAS procedures. Total cholesterol was previously mapped to an intron in the gene *DOCK6* in African American cohorts⁷¹, a finding we replicated at the suggestive significance threshold with standard GWAS. Tractor identified a lead *DOCK6* SNP 20 kb downstream in the AFR tracts and a meta-analysis of hits from deconvolved AFR and EUR tracts. This new lead SNP (rs2278426) spans *DOCK6* as well as *ANGPTL8*, where it is a missense mutation ([NC_000019.9:g.11350488C>T](#)) that is predicted to be possibly damaging and deleterious by PolyPhen and SIFT, respectively^{83,84} (see Discussion and Fig. 5). *ANGPTL8*, also known as lipasin, has been shown to regulate plasma lipid levels in mice by inhibiting the enzyme lipoprotein lipase^{85–88}. In humans, *ANGPTL8* levels correlate with metabolic phenotypes including type 2 diabetes and obesity^{89–92}, and HDL cholesterol levels across diverse populations have been demonstrated to better correlate with *ANGPTL8* than *DOCK6* (ref. ⁹³). Altogether, *ANGPTL8* appears to be a more promising candidate, highlighting how leveraging diverse populations allows for improved identification of risk variants. Notably, asaMap improved localization compared with standard GWAS, but was unable to identify the putative causal variant (Supplementary Fig. 5).

To assess whether our ability to identify rs2278426 was due to a true or marginal effect-size difference driven by MAF or linkage disequilibrium differences across ancestries, we attempted validating its association with fine-mapping in 345,235 white, British UK Biobank individuals and 135,808 Japanese individuals from

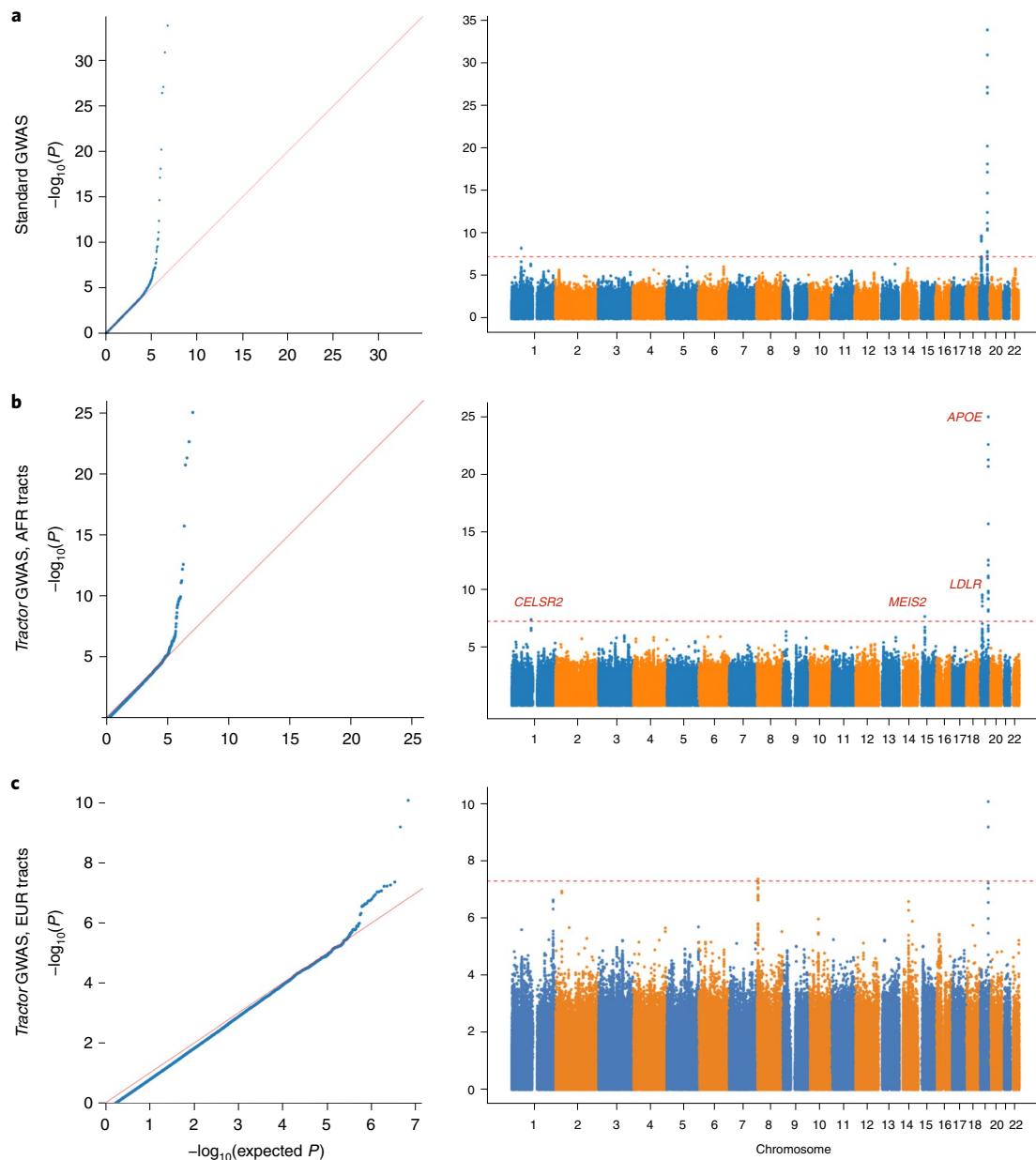


Fig. 4 | Tractor GWAS replicates established hits for total cholesterol in individuals of admixed African-European ancestry and identifies new ancestry-specific loci. **a–c,** Q–Q and Manhattan plots for total cholesterol using the standard GWAS model (**a**) compared with Tractor joint-analysis results for the AFR (**b**) and EUR backgrounds (**c**). The traditional genome-wide significance threshold of 5×10^{-8} is shown as a red dashed line.

BioBank Japan⁹⁴. rs2278426 was successfully fine-mapped to a 95% credible set in both populations, with a maximum posterior inclusion probability of 0.993 in BioBank Japan. This variant occurs at 26% frequency in gnomAD East Asian individuals, 18% in Africans and 4% in non-Finnish Europeans⁹⁵. We estimated the frequencies in our admixed empirical cohort to be ~22% in the AFR tracts and ~4% in the EUR tracts. To assess the role of linkage disequilibrium on localization improvement, we created within-sample linkage disequilibrium heatmaps for this *DOCK6* region of interest from the full dataset, AFR tracts and EUR tracts, finding broadly similar patterns (Supplementary Fig. 6). These trends suggest that localization improvement was driven by higher power in non-European populations to identify the causal variant rather than linkage disequilibrium differences directly. Although below the traditional genome-wide significance level, this locus highlights the improved ability to localize GWAS signals thanks to leveraging additional

ancestral breakpoints in admixed genomes (see Supplementary Fig. 7 for a demonstration of signal localization for LDL cholesterol in the canonical gene *PCSK9*).

Discussion

Despite recent advances in understanding the genetics of complex diseases, major limitations remain in our knowledge of the architecture of such disorders in minority and admixed populations. Here, we present an analytic framework and gene discovery method distributed as a scalable software package named Tractor, which allows admixed samples to be appropriately analyzed alone or alongside homogenous cohorts in statistical genomics efforts. We tested our framework in a simulation model designed to emulate African American cohorts. We then applied it to empirical data from admixed individuals of the UK Biobank of African descent. We observed a gain in power to detect risk loci across sample sizes,

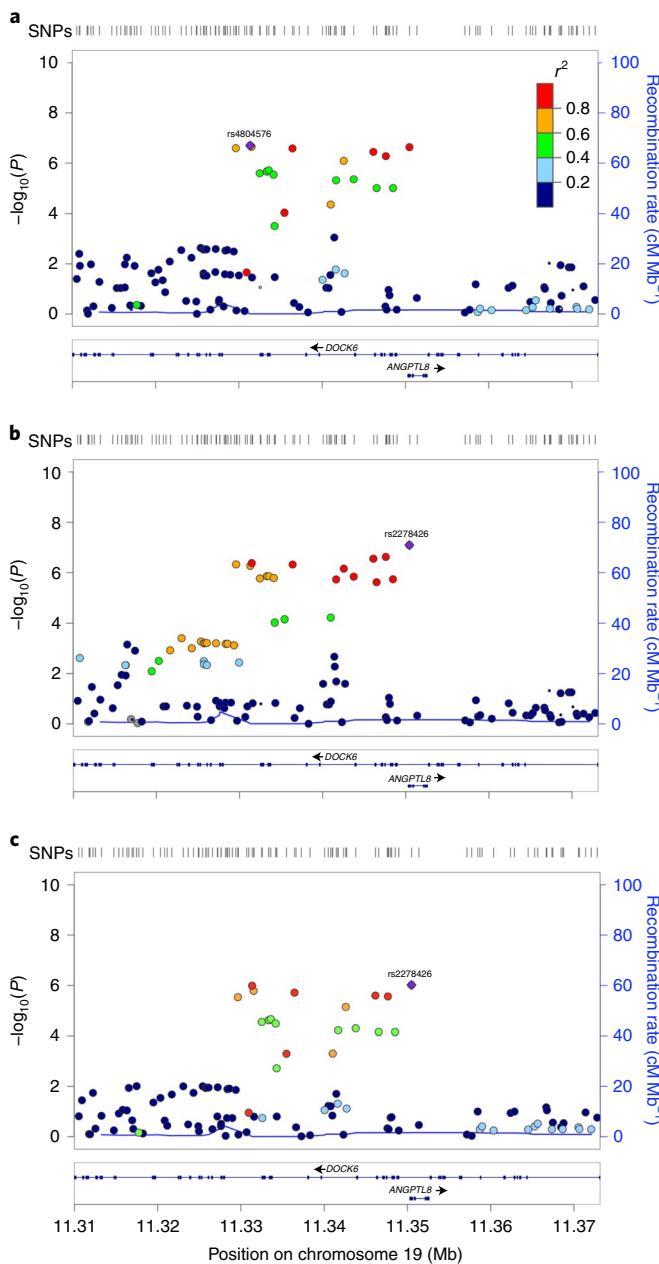


Fig. 5 | Tractor better localizes a top hit for total cholesterol. **a–c**, Runs on UK Biobank admixed individuals with the standard GWAS model (**a**), AFR-specific GWAS with Tractor (**b**) and a meta-analysis of GWAS runs on deconvolved EUR and AFR tracts (**c**). Both Tractor runs pinpoint a lead SNP ~20 kb downstream of the intronic standard GWAS top hit in *DOCK6*, spanning a better-candidate gene, *ANGPTL8*. No significant signal was seen in the EUR segments. In all plots, the point size is proportional to the number of samples included for that test, and the color indicates r^2 to the named lead SNP. The recombination rate is shown as a blue line, generated from the AFR superpopulation of the 1000 Genomes Project in **b** and from the EUR superpopulation in **a** and **c**.

demographic models and disease prevalences using Tractor, particularly when true or marginal effect sizes were heterogeneous across populations. Our approach incorporates a local ancestry-aware GWAS method that generates ancestry-specific estimates, which admixture mapping and standard GWAS cannot, and which can be extremely helpful in downstream efforts such as constructing

genetic risk scores for understudied populations. Tractor also gives increased precision in localizing GWAS signals closer to the causal variant in recently admixed groups. This reduces the credible set of SNPs and aids in the prioritization of variants for subsequent functional testing.

The Tractor pipeline requires several inputs; most importantly, accurate local ancestry calls. Users should ensure good LAI performance in their target cohorts. A major determinant of accurate LAI calls is a comprehensive and well-matched reference panel^{47,96}. Existing reference panels are more plentiful for Eurasian populations than for other groups, underscoring the need to expand sequencing efforts in diverse global populations. An additional consideration is to ensure high-quality imputation if Tractor is to be run on imputed genotype array data. A final consideration is to ensure consistent phenotyping across ancestry groups, as is standard in multi-ancestry GWAS. We note that we have thoroughly tested Tractor here in the two-way admixture scenario reflecting African American demographic history. The analytic infrastructure can currently also run on three-way admixed populations, and our statistical model can scale to an arbitrary number of ancestries. Future work will test the power and optimize the code for multi-way admixed models.

We evaluated the landscape of when Tractor does and does not add GWAS discovery power using simulated data modeled after African American cohorts (Fig. 2 and Extended Data Figs. 3–5). Modifying several interacting parameters (absolute and difference in effect size, absolute and difference in MAF, sample size, disease prevalence and admixture proportions), power gains from Tractor are generally most dramatic when there is a large effect-size difference between ancestries. Gains from observed effect-size differences are heightened when coupled with highly differing MAF across groups. The most extreme examples of such a case are when there is an effect only on one haplotype background or an allele only present in one ancestry. Another interacting feature is the overall admixture proportions and whether the stronger effect allele is on the more frequent or more rare ancestry background. Tractor power gains are most dramatic when a variant with an effect in only one ancestry falls on the ancestry that is less common in the dataset. In such an instance in a standard GWAS setting, noise from the uninformative majority haplotypes would result in extremely low power to detect the locus (Fig. 2d). Power can be recovered, however, by analyzing genotypes on ancestry-specific haplotypes, thus controlling for population structure as well as identifying risk variants that would otherwise be undetectable.

Conversely, we find that it is generally not necessary to include local ancestry in a GWAS model when there is no difference in the estimated effect size between ancestry groups. Notably, we are referring here to the detection of both marginal as well as true effects. By a true effect, we mean an effect at a causal mutation directly impacting the phenotype. By a marginal effect, we imply the estimated effect at a variant tagging a causal variant; these marginal effects will represent the majority of results from a standard GWAS. Differences in indirect association can come from features such as ancestry-specific variation or different patterns of linkage disequilibrium. Both true and marginal effect sizes are useful in post-GWAS efforts such as the construction of genetic risk scores and heritability estimation. There is evidence suggesting that in most cases (with some notable exceptions^{8,71,97}), the effects of causal variants are similar across ancestries^{33,38,98–103}. However, the marginal effect sizes of tag SNPs are routinely expected to differ across ancestries due to differences in ancestral MAF and the linkage disequilibrium patterns resulting from each ancestral population's demographic history^{104,105}. Therefore, the most powerful use case for Tractor—when there are effect-size differences across ancestries—should impact many more sites across the genome than just variants that have a true biological difference in effect across ancestry groups. Said

differently, we clarify that Tractor benefits from power gains to detect the marginal beta in addition to the rarer case of variants with true effects only on one haplotypic backbone.

Tractor is also expected to benefit from increased power when functionally important (and probably rare) alleles only present in one population are missed by genotyping or imputation. In such situations, the common alleles in linkage disequilibrium, despite being shared across populations, would be associated with the phenotype as a function of which haplotypic background they are found on and thus would have a haplotype-specific effect. Another relevant scenario is the presence of linkage disequilibrium in regions where there are ancestry-specific markers intermingled with shared ones. This would affect univariate scan results such that considering the haplotypic background on which alleles fall would particularly aid in localizing signals through improved marginal beta estimates, even with consistent causal effects in both ancestries. To expand on this idea, assuming a shared causal variant across ancestries, even in genomic regions with all variants present in both groups, one could hone in on the causal variant by minimizing the difference between betas across ancestries. If one assumes the beta at the causal variant is the same in both populations, then optimizing for reduced beta difference across groups using ancestry-specific effect-size estimates can help in fine-mapping signals.

To summarize, the primary benefits we observed from Tractor are power gains from leveraging ancestral genomic differences, accurate estimation of ancestry-specific effect sizes, and improved GWAS signal resolution. At admixed cohort sample sizes of the scale we tested (4,000–12,000 cases), ten principal components adequately controlled for false positives using standard GWAS procedures. With increasing sample sizes, however, Tractor's fine-scale correction of population structure is expected to better control false positive associations than principal components alone, in addition to improving GWAS signals.

In empirical data, Tractor was able to replicate established GWAS hits, discover new ones and aid in the localization of GWAS signals. We replicated known hits for blood lipids in ~4,300 admixed African-European individuals from the UK Biobank and demonstrated an improved ability to identify putative causal SNPs fine-mapped in another diverse collection, BioBank Japan⁹⁴ (Figs. 4 and 5 and Supplementary Figs. 4 and 7). We additionally identified novel hits using our local ancestry-aware method that would have been missed using standard GWAS.

Portions of the Tractor pipeline may be helpful for working with admixed cohorts in use cases beyond GWAS. For example, recovering long-range haplotypes and correcting for population structure is key in evolutionary genomic studies for analyses such as genome-wide scans of selection^{106–108}. Within medical genetics, accounting for the ancestral background of alleles will be valuable in studies of rarer variants that are more population specific^{109,110}.

In summary, Tractor allows users to account for genotype-level ancestry in a precise manner, enabling the well-calibrated inclusion of admixed individuals in large-scale gene discovery efforts. This approach provides a number of benefits over traditional GWAS, including the production of accurate ancestry-specific summary statistics, improved localization of GWAS signals, and power boosts in many genetic contexts. This infrastructure is designed as a series of steps to be flexible and easily ported into other statistical genomics activities. Tractor advances the existing methodologies for studying the genetics of complex disorders in admixed populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of

data and code availability are available at <https://doi.org/10.1038/s41588-020-00766-y>.

Received: 15 May 2020; Accepted: 15 December 2020;
Published online: 18 January 2021

References

- Parker, K., Menasce Horowitz, J., Morin, R. & Lopez, M. H. *Multiracial in America: Proud, Diverse and Growing in Numbers* (Pew Research Center, 2015); <https://www.pewsocialtrends.org/2015/06/11/multiracial-in-america/>
- Bhardwaj, A. et al. Racial disparities in prostate cancer: a molecular perspective. *Front. Biosci.* **22**, 772–782 (2017).
- Grizzel, W. E. et al. Self-identified African Americans and prostate cancer risk: West African genetic ancestry is associated with prostate cancer diagnosis and with higher Gleason sum on biopsy. *Cancer Med.* **8**, 6915–6922 (2019).
- Duggan, M. A., Anderson, W. F., Altekruze, S., Penberthy, L. & Sherman, M. E. The Surveillance, Epidemiology, and End Results (SEER) program and pathology: toward strengthening the critical relationship. *Am. J. Surg. Pathol.* **40**, e94–e102 (2016).
- Freedman, M. L. et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA* **103**, 14068–14073 (2006).
- Bateman, E. D. et al. Global strategy for asthma management and prevention: GINA executive summary. *Eur. Respir. J.* **31**, 143–178 (2008).
- Daya, M. & Barnes, K. C. African American ancestry contribution to asthma and atopic dermatitis. *Ann. Allergy Asthma Immunol.* **122**, 456–462 (2019).
- Wyss, A. B. et al. Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat. Commun.* **9**, 2976 (2018).
- Demenaïs, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–50 (2018).
- Benetos, A. & Aviv, A. Ancestry, telomere length, and atherosclerosis risk. *Circ. Cardiovasc. Genet.* **10**, e001718 (2017).
- Mozaffarian, D. et al. Heart disease and stroke statistics—2015 update. *Circulation* **131**, e29–e322 (2015).
- Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
- Popejoy, A. B. & Fullerton, S. M. Genomics is falling. *Nature* **538**, 161–164 (2016).
- Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet.* **14**, e1007309 (2018).
- Huang, H. et al. Bootstrat: population informed bootstrapping for rare variant tests. Preprint at *bioRxiv* <https://doi.org/10.1101/068999> (2016).
- Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
- Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
- Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
- Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *Am. J. Hum. Genet.* **101**, 218–226 (2017).
- Walters, R. K. et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).
- Martin, E. R. et al. Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet. Epidemiol.* **42**, 214–229 (2018).
- Stevenson, A. et al. Neuropsychiatric genetics of African populations—psychosis (NeuroGAP—Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. *BMJ Open* **9**, e025469 (2019).
- The H3Africa Consortium., Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
- Kowalski, M. H. et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
- The Precision Medicine Initiative Cohort Program—Building a Research Foundation for 21st Century Medicine*. Precision Medicine Initiative Working Group Report to the Advisory Committee to the Director, NIH (Precision Medicine Initiative Working Group, 2015).
- Logue, M. W. et al. The Psychiatric Genomics Consortium Posttraumatic Stress Disorder Workgroup: posttraumatic stress disorder enters the age of large-scale genomic collaboration. *Neuropsychopharmacology* **40**, 2287–2297 (2015).

27. Bien, S. A. et al. The future of genomic studies must be globally representative: perspectives from PAGE. *Ann. Rev. Genom. Hum. Genet.* **20**, 181–200 (2019).
28. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
29. Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
30. Hero, J. O., Zaslavsky, A. M. & Blendon, R. J. The United States leads other nations in differences by income in perceptions of health and health care. *Health Aff.* **36**, 1032–1040 (2017).
31. Williams, D. R., Priest, N. & Anderson, N. B. Understanding associations among race, socioeconomic status, and health: patterns and prospects. *Health Psychol.* **35**, 407–411 (2016).
32. *2016 National Healthcare Quality and Disparities Report* (Agency for Healthcare Research and Quality, 2017).
33. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).
34. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111–R119 (2015).
35. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
36. Wu, Y. et al. Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet.* **9**, e1003379 (2013).
37. Van de Bunt, M. et al. Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS Genet.* **11**, e1005535 (2015).
38. Mahajan, A. et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
39. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
40. Zhang, J. & Stram, D. O. The role of local ancestry adjustment in association studies using admixed populations. *Genet. Epidemiol.* **38**, 502–515 (2014).
41. Lachance, J. et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).
42. Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I. & London, S. J. Joint testing of genotype and ancestry association in admixed families. *Genet. Epidemiol.* **34**, 783–791 (2010).
43. Coram, M. A. et al. Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *Am. J. Hum. Genet.* **92**, 904–916 (2013).
44. Aschard, H., Gusev, A., Brown, R. & Pasaniuc, B. Leveraging local ancestry to detect gene–gene interactions in genome-wide data. *BMC Genet.* **16**, 124 (2015).
45. Zaitlen, N., Pas, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).
46. Pasaniuc, B. et al. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7**, e1001371 (2011).
47. Pasaniuc, B. et al. Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* **29**, 1407–1415 (2013).
48. Chimusa, E. R. et al. Genome-wide association study of ancestry-specific TB risk in the South African coloured population. *Hum. Mol. Genet.* **23**, 796–809 (2014).
49. Smith, E. N. et al. Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol. Psychiatry* **14**, 755–763 (2009).
50. Skotte, L., Jørsboe, E., Korneliussen, T. S., Moltke, I. & Albrechtsen, A. Ancestry-specific association mapping in admixed populations. *Genet. Epidemiol.* **43**, 506–521 (2019).
51. Shriner, D. Overview of admixture mapping. *Curr. Protoc. Hum. Genet.* **94**, 1.23.1–1.23.8 (2013).
52. Chen, M. et al. Admixture mapping analysis in the context of GWAS with GAW18 data. *BMC Proc.* **8**, S3 (2014).
53. Chen, W. et al. A generalized sequential Bonferroni procedure for GWAS in admixed populations incorporating admixture mapping information into association tests. *Hum. Hered.* **79**, 80–92 (2015).
54. Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. & McKeigue, P. M. Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* **74**, 965–978 (2004).
55. Patterson, N. et al. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
56. Spear, M. L. et al. A genome-wide association and admixture mapping study of bronchodilator drug response in African Americans with asthma. *Pharmacogenomics J.* **19**, 249–259 (2019).
57. Gignoux, C. R. et al. An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos. *J. Allergy Clin. Immunol.* **143**, 957–969 (2019).
58. Shetty, P. B. et al. Variants for HDL-C, LDL-C, and triglycerides identified from admixture mapping and fine-mapping analysis in African American families. *Circ. Cardiovasc. Genet.* **8**, 106–113 (2015).
59. Shetty, P. B. et al. Variants in CXADR and F2RL1 are associated with blood pressure and obesity in African-Americans in regions identified through admixture mapping. *J. Hypertens.* **30**, 1970–1976 (2012).
60. Reiner, A. P. et al. Genome-wide association and population genetic analysis of c-reactive protein in African American and Hispanic American women. *Am. J. Hum. Genet.* **91**, 502–512 (2012).
61. Florez, J. C. et al. in *Racial Identities, Genetic Ancestry, and Health in South America: Argentina, Brazil, Colombia, and Uruguay* (eds Gibbon, S. et al.) 137–153 (Palgrave Macmillan, 2011).
62. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
63. Geza, E. et al. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Brief. Bioinform.* **20**, 1709–1724 (2019).
64. Schubert, R., Andaleon, A. & Wheeler, H. E. Comparing local ancestry inference models in populations of two- and three-way admixture. *PeerJ* **8**, e10090 (2020).
65. Tishkoff, S. A. et al. The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
66. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* **108**, 11983–11988 (2011).
67. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
68. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, e1007308 (2018).
69. Andrés, A. M. et al. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epidemiol.* **31**, 659–671 (2007).
70. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
71. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
72. Musunuru, K. & Kathiresan, S. Genetics of common, complex coronary artery disease. *Cell* **177**, 132–145 (2019).
73. Rotimi, C. N. et al. The genomic landscape of African populations in health and disease. *Hum. Mol. Genet.* **26**, R225–R236 (2017).
74. Superko, H. R., Momary, K. M. & Li, Y. Statins personalized. *Med. Clin. North Am.* **96**, 123–139 (2012).
75. Fu, J. et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8**, e1002431 (2012).
76. Avery, C. L. et al. A phenomics-based strategy identifies loci on *APOC1*, *BRAP*, and *PLCG1* associated with metabolic syndrome phenotype domains. *PLoS Genet.* **7**, e1002322 (2011).
77. Lettre, G. et al. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe project. *PLoS Genet.* **7**, e1001300 (2011).
78. Talmud, P. J. et al. Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip. *Am. J. Hum. Genet.* **85**, 628–642 (2009).
79. Sandhu, M. S. et al. LDL-cholesterol concentrations: a genome-wide association study. *Lancet* **371**, 483–491 (2008).
80. Sanna, S. et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.* **7**, e1002198 (2011).
81. Fox, C. S. et al. Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. *BMC Med. Genet.* **8**, S18 (2007).
82. Kathiresan, S. et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* **8**, S17 (2007).
83. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit 7.20 (2013).
84. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

85. Zhang, R. The ANGPTL3-4-8 model, a molecular mechanism for triglyceride trafficking. *Open Biol.* **6**, 150272 (2016).
86. Fu, Z., Abou-Samra, A. B. & Zhang, R. A lipasin/Angptl8 monoclonal antibody lowers mouse serum triglycerides involving increased postprandial activity of the cardiac lipoprotein lipase. *Sci. Rep.* **5**, 18502 (2015).
87. Zhang, R. Lipasin, a novel nutritionally-regulated liver-enriched factor that regulates serum triglyceride levels. *Biochem. Biophys. Res. Commun.* **424**, 786–792 (2012).
88. Siddiqua, A. et al. Visualizing the regulatory role of Angiopoietin-like protein 8 (ANGPTL8) in glucose and lipid metabolic pathways. *Genomics* **109**, 408–418 (2017).
89. Yamada, H. et al. Circulating betatrophin is elevated in patients with type 1 and type 2 diabetes. *Endocr. J.* **62**, 417–421 (2015).
90. Espes, D., Martinell, M. & Carlsson, P.-O. Increased circulating betatrophin concentrations in patients with type 2 diabetes. *Int. J. Endocrinol.* **2014**, 323407 (2014).
91. Hu, H. et al. Increased circulating levels of betatrophin in newly diagnosed type 2 diabetic patients. *Diabetes Care* **37**, 2718–2722 (2014).
92. Fu, Z. et al. Elevated circulating lipasin/betatrophin in human type 2 diabetes and obesity. *Sci. Rep.* **4**, 5013 (2015).
93. Cannon, M. E. et al. Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at the ANGPTL8 HDL-C GWAS Locus. *G3* **7**, 3217–3227 (2017).
94. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
95. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
96. Lin, M. et al. Population-specific reference panels are crucial for genetic analyses: an example of the CREBRF locus in Native Hawaiians. *Hum. Mol. Genet.* **29**, 2275–2284 (2020).
97. Ntzani, E. E., Liberopoulos, G., Manolio, T. A. & Ioannidis, J. P. A. Consistency of genome-wide associations across major ancestral groups. *Hum. Genet.* **131**, 1057–1071 (2012).
98. Marigorta, U. M. & Navarro, A. High *trans*-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* **9**, e1003566 (2013).
99. Waters, K. et al. Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.* **6**, e1001078 (2010).
100. Lam, M. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).
101. Liu, J. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
102. Carlson, C. S. et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661 (2013).
103. Kuchenbaecker, K. et al. The transferability of lipid loci across African, Asian and European cohorts. *Nat. Commun.* **10**, 4330 (2019).
104. Mägi, R. et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
105. Wegmann, D. et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.* **43**, 847–853 (2011).
106. Atkinson, E. G. et al. No evidence for recent selection at FOXP2 among diverse human populations. *Cell* **174**, 1424–1435.e15 (2018).
107. Deng, L., Ruiz-Linares, A., Xu, S. & Wang, S. Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci. Rep.* **6**, 21766 (2016).
108. Jin, W. et al. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* **22**, 519–527 (2012).
109. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
110. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Quality control and LAI pipeline. The core feature of the Tractor framework relies on accounting for fine-scale population structure as informed by local ancestry (that is, ancestral chromosome painting). Tractor then uses this information to: (1) correct for individuals' ancestral dosage at all variant sites; (2) recover long-range tracts in admixed individuals; and (3) extract the tracts and ancestry dosage counts from each ancestry component for use in ancestry-specific association tests. We have tested and built this framework around LAI calls from RFMix2 (ref. ⁶²) and have built an automated pipeline (<https://github.com/eatkinson/Post-QC>) to perform post-genotyping quality control, data harmonization, phasing and LA inference to consistently prepare the data for downstream analysis (Supplementary Information)^{11,112}.

In all tests, we ran RFMix with one expectation–maximization iteration and a window size of 0.2 cM with the HapMap combined recombination map¹¹³ to inform switch locations. The -n 5 flag (terminal node size for random forest trees) was included to account for an unequal number of reference individuals per reference population. We used the --reanalyze-reference flag, which recalculates admixture in the reference samples for improved ability to distinguish ancestries. This is especially important when the reference samples are themselves admixed. As a reference panel, we used AFR and EUR individuals from the 1,000 Genomes reference panel. Painted karyogram plots were produced using a modified version of publicly available code (https://github.com/armartin/ancestry_pipeline). We optimized this pipeline under the two-way admixed African American demographic model. Tractor additionally supports three-way admixture calls with an expanded set of scripts.

LAI accuracy. We validated that LAI was performing well in the African American use case. To do this, we generated a truth dataset by simulating individuals with known phase and local ancestry from empirical data. Our simulation reference panel consisted of haplotypes from homogenous PGC-PTSD individuals who had $\geq 95\%$ EUR or AFR ancestry as inferred by SNPweights¹⁴. We simulated admixture between these reference individuals with admix-simu¹⁵ using a realistic African American demographic model of one pulse of admixture nine generations ago with 84% contribution from Africa and 16% from Europe. The resultant population mixes among itself until the present day, copying haplotypes from the previous generation with break points informed by the recombination map. This retains the linkage disequilibrium structure and genetic variation present in real genomic data and ensures that the truth dataset resembles cohort data as closely as possible. We then ran LAI and calculated the LAI accuracy as how often the ancestry call was correct in the simulated truth data. The overall proportion of the genome estimated by RFMix in the realistic scenarios was within the range of expectations given the simulation model of 16% EUR ancestry and 84% AFR ancestry (15.1 and 84.9%, respectively).

Correcting switch errors using local ancestry. Despite LAI calling ancestry dosage accurately, frequent chromosomal switches were visible in painted karyograms, which we determined were due to phasing errors. It is important to retain complete tracts, as spurious breakpoints will reduce the accuracy of haplotype-based tests. Tractor detects and fixes phase switches, which we define as a swap of ancestry across a chromosome within a 1-cM window at a region with heterozygous ancestry dosage, using local ancestry. This level of information is accessible due to the admixture; no local ancestry switches would be visible if a cohort were homogenous.

To ensure that correcting phase switch errors improved the results compared with the truth expectations, we modeled the expected distributions of the EUR tract within African American individuals using a Poisson process with a rate equal to the number of generations ago when the pulse of admixture occurred. The waiting time until a recombination event disrupts a tract is expected to follow this distribution, with a slight shortening of tracts proportional to the percent admixture due to the inability to visualize tract switches that occur across regions of the same ancestry. In this way, we could quantify the probability of observing the observed number of tracts after a particular data treatment given the truth expectations (see Supplementary Information, Supplementary Table 2 and Extended Data Figs. 1 and 2).

GWAS power simulations incorporating local ancestry. We assessed the improvements in GWAS power from using Tractor through simulations. We formulated our simulation framework on the suggestions of Skotte et al.⁵⁰. Power calculations were based on a simulation framework that initially models an African American population assuming a biallelic disease risk allele with a 20% overall MAF and an additive effect in the AFR genetic background but not in the EUR genetic background. Specifically, the overall admixture proportions were drawn from a beta distribution with shape parameters 7.76 and 2.17—the fitted parameters to this distribution for AFR ancestry proportions observed in the PGC-PTSD Freeze 2 African American cohorts. The genotype of each copy of the allele was drawn from a binomial distribution with the probability of having the minor allele set to the MAF. We simulated a disease phenotype with individuals' risk drawn from a binomial distribution assuming a 10% disease prevalence. The risk of developing the phenotype was modified on a log-additive scale according

to the admixture proportions and the presence of the minor allele on an AFR background using a logit model. In this model, the probability of disease was set to $-2.19 + \log[\text{allelic risk effect size}] \times \text{the number of copies of the minor allele coming from an AFR ancestral background} + 0.5 \times \text{AFR admixture proportion}$. -2.19 was chosen as it represents a 10% probability of disease given no AFR admixture or copies of the minor allele from either ancestral background. The 0.5 value in $0.5 \times \text{AFR admixture proportion}$ was set to induce stratification in the simulated population, as is observed in empirical data. In other words, all of our simulations modeled increasing disease prevalence with admixture fractions, reflective of clinical observation. With this simulation design, individuals with higher AFR ancestry proportions are more likely to be cases whereas those with higher EUR ancestry proportions are more likely to be controls. Subjects' disease statuses were then drawn from a binomial distribution with the probability parameterized to their individual disease risk according to the logit model. Cases and controls were sampled at random from the simulated population at a 2.5:1 control-to-case ratio—the approximate ratio of controls to cases in PGC-PTSD Freeze 2.

Under each simulation, we fit three logistic regression models of disease status that included: (M1) admixture only; (M2) the number of copies of the risk allele only; and (M3) admixture + the number of copies of the risk allele on a EUR background + the number of copies on an AFR background. M1 serves as a null comparison to evaluate the significance of including the SNP as a predictor. M2 is equivalent to standard GWAS procedures. M3 models our Tractor method. In all cases, we included the first ten principal components to account for population structure, as is common practice in current GWASs on admixed cohorts. The significance of M2 and M3 was evaluated by likelihood ratio tests comparing them with M1. For each 100 simulations at a given effect size and sample size, for both M2 and M3, we estimated power as the proportion of time for which the likelihood ratio test was significant ($P < 5 \times 10^{-8}$). We performed 1,000 rounds of simulation with this model at each level of allelic effect size, with an odds ratio of 1.05–1.30 and case sample sizes of $n = 4,000$ and 12,000. It should be noted that the two-way Tractor model is by nature a two-degrees-of-freedom test. Adding additional ancestries/terms into the model could reduce power gains due to the additional penalty.

We additionally benchmarked Tractor against asaMap under our primary models, assessing the effect of differing sample sizes, MAF differences, admixture fractions and relative ancestry-specific effects (Supplementary Fig. 5). The specific asaMap model test that we assessed in these runs is equivalent to the M1 versus M5 comparison described in Skotte et al.⁵⁰; namely, an effect in either population. A comparison of results from different asaMap models for the DOCK6 region of interest can be seen in Supplementary Fig. 5.

We also built LAI uncertainty into the Tractor model in these comparisons to quantify the effect of LAI accuracy on Tractor power. For this, we modeled perfect, realistic accuracy (98%) and a lower tolerable bound of 90% LAI accuracy (Extended Data Fig. 5). In all cases, 1,000 replicates were run under a model of 12,000 cases, 30,000 controls, 10% disease prevalence and a realistic 80/20 AFR-to-EUR admixture ratio using ten principal components as covariates.

Characterizing the landscape of Tractor power gains. To evaluate Tractor power gains, we ran similar sets of simulations, varying effect-size differences across ancestries, MAF differences, admixture fractions and disease prevalence (Fig. 2 and Extended Data Figs. 3 and 4). Additional details of these runs can be found in the Supplementary Information.

Varying effect size across populations. To examine the effect of modifying the effect sizes, we introduced an effect on EUR haplotypes as well, rather than just on AFR. All of these simulations assumed 80% admixture, 10% disease prevalence and 20% MAF in both groups. We modeled cases across the odds ratio spectrum where there was an effect of equal size in both ancestries, a 30% larger effect size on the EUR background, an effect size 30% larger on the AFR haplotype and an effect size only in the EUR.

Varying absolute MAF. Next, we fixed all other parameters and modified the absolute MAF of the simulated risk allele, with the relative difference in MAF between ancestries remaining constant. We changed our MAF from 20% to 10 and 40% under both the models of an effect only in the AFR background and with matching effect sizes between EUR and AFR.

MAF differences between groups. To see whether having a difference in the MAF between the two ancestral groups affected GWAS power, we varied the MAF in the EUR background to be 10, 20 and 30% while keeping the AFR MAF set to 20%.

False positive rate. We quantified the false positive rate by simulating a variant with no effect and counting significant associations identified in a simulated realistic African American population given $\alpha = 0.05$ across a range of absolute and relative differences in ancestral MAF. In all cases, 5,000 replicates were run, individuals were 80/20 admixed AFR/EUR, the disease prevalence was set at 10% and the sample size was modeled as 12,000 cases and 30,000 controls.

Selection of two-way admixed African–European individuals. To select individuals with two-way admixture with European and West African ancestry

(Supplementary Fig. 3), we took a two-pronged approach. First, we combined genetic reference data from the 1000 Genomes Project and Human Genome Diversity Panel¹¹⁶, then we harmonized metadata according to consistent continental ancestries. We then ran principal component analysis on unrelated individuals from the reference dataset. To partition individuals in the UK Biobank based on their continental ancestry, we used the principal component loadings from the reference dataset to project UK Biobank individuals into the same principal component space. We trained a random forest classifier given continental ancestry metadata based on the top six principal components from the reference training data. We applied this random forest to the projected UK Biobank principal component analysis data and assigned AFR ancestries if the random forest probability was >50%, otherwise individuals were dropped from further analysis. To restrict to only two-way admixed West African–European ancestry individuals, we restricted to individuals with at least 12.5% EUR ancestry and at least 10% AFR ancestry and who did not deviate more than one standard deviation from the AFR–EUR cline. This resulted in approximately 4,300 individuals per blood lipid trait. Global ancestry fraction and ancestral allele frequency estimates were obtained from running ADMIXTURE¹¹⁷ with $k=2$ (which was the best-fit k value to this dataset based on fivefold cross-validation) on these individuals with 1000 Genomes Project EUR and AFR superpopulation individuals as reference data. To ensure that there were no major areas of the genome where LAI was skewing from the expected global fractions, we also assessed the cumulative local ancestry calls across the genome for the UK Biobank admixed subset.

Software implementations. We developed separate scripts to deconvolve ancestry tracts and calculate haplotype dosages, correct phase switch errors and run a Tractor GWAS to obtain ancestry-specific effect-size estimates and P values. Pre-GWAS steps are available as independent Python scripts to allow for maximum flexibility. To implement the joint modeling GWAS approach with the novel linear regression model described here, we built a scalable pipeline in Hail¹¹⁸, which can be implemented locally or on the Google Cloud Platform¹¹⁹. Descriptions of the steps and an example Jupyter notebook¹²⁰ demonstrating analytical steps and visualization of the results of the Tractor joint-analysis GWAS are freely available on GitHub (<https://github.com/eatkinson/Tractor>). The Hail implementation of Tractor GWAS generally runs on the order of minutes depending on the sample size and variant density. Alternatively, users can run a separate/meta-analysis GWAS version of Tractor (see Supplementary Information). This pipeline requires the initial processing steps to optionally correct phase switch errors and deconvolve ancestry tracts into their own VCF files. Next, GWAS can be run for the deconvolved files containing different ancestral components with the user's preferred GWAS software, such as PLINK¹²¹. In this implementation, a standard GWAS model can be run on each ancestral component separately using the ancestry-specific VCF output by Tractor, which contains fully or partially missing data including only haplotypes from the ancestry in question. Results from the different ancestry runs could then be meta-analyzed to increase the power by incorporating summary statistics from both populations, although we recommend preferentially using the joint-analysis method described in this manuscript to avoid any potential bias from combining multiple ancestral portions of the genome of the same individuals. This implementation is also compatible in large-scale collections where there are large numbers of homogenous individuals (for example, many Europeans) but too limited a number of admixed individuals to be run in a GWAS alone. The EUR sections of the admixed cohorts could be analyzed alongside the homogenous European cohorts, making better use of the admixed samples even if other ancestry portions are not utilized, and increasing the effective sample size.

Empirical test of Tractor on blood lipid phenotypes. To ensure that Tractor replicated well-established associations, we ran standard GWAS, the Tractor joint-analysis model and inverse variance-weighted fixed-effects meta-analysis of summary statistics from EUR and AFR deconvolved tracts on ~4,300 admixed African–European individuals from the UK Biobank on the biomarker blood lipid traits of total cholesterol and LDL cholesterol. We included covariates capturing global ancestry (principal components or global ancestry fraction), age, sex and blood dilution factor in all runs. All simulation results for both Tractor and standard GWAS include the first ten principal components to account for population stratification, as is currently standard practice in studies of admixed cohorts. To confirm good performance regardless of the global ancestry estimation strategy, we assessed meta-analysis performance using both metrics to capture global ancestry (namely, principal components versus the AFR fraction as determined by ADMIXTURE), finding that these did not result in substantive differences. In the empirical joint-analysis framework, we used the measure of global AFR ancestry fraction to more directly capture global ancestry and avoid any potential collinearity with local ancestry from principal components. We generated Q–Q plots alongside each trait and compared the inflation of test statistics in each GWAS case by looking at the genomic inflation factor, λ_{GC} . We then compared the results with those obtained from the same individuals using a standard GWAS approach. No individuals overlapped between the previous study of interest (Natarajan et al.⁷¹) and the UK Biobank individuals included here. As expected, near-identical results were obtained from the meta- and

joint-approaches. Gene visualizations were produced with LocusZoom¹²², while Manhattan and Q–Q plots were produced with bokeh in Hail¹²³.

We further investigated the pattern of linkage disequilibrium in the *DOCK6* region of interest to assess how much linkage disequilibrium differences across ancestries aided in localization improvements with Tractor. For this, we generated within-sample linkage disequilibrium heatmaps for the region of interest (Supplementary Fig. 6) in the full dataset, as well as within just the AFR and EUR tracts. Squared intervariant allele count correlations were generated with PLINK¹²¹ for the region 11,325,000–11,360,000 of chromosome 19. Linkage disequilibrium heatmaps were created with the LDheatmap package¹²⁴ in R.

Statistical fine-mapping of top hits in independent cohorts. We conducted GWAS and statistical fine-mapping in two additional large-scale cohorts: 345,235 white British individuals from UK Biobank and 135,808 Japanese individuals from BioBank Japan. For the UK Biobank white British, we used previously conducted fine-mapping results for total cholesterol (<https://www.finucanelab.org/data>). Briefly, we computed association statistics for the variants with INFO > 0.8, MAF > 0.01% (except for rare coding variants with MAC > 0) and HWE P value > 1×10^{-10} using BOLT-LMM¹²⁵, with covariates including the top ten principal components, sex, age, age², sex × age, sex × age² and blood dilution factor. We used FINEMAP version 1.3.1 (refs. ^{126,127}) and susieR version 0.8.1.0521 (ref. ¹²⁸) for fine-mapping using the GWAS summary statistics and in-sample dosage linkage disequilibrium matrices computed by LDstore version 2.0b. We defined regions based on 3-megabase window surrounding lead variants and merged them if they overlapped. The maximum number of causal variants in a region was specified as ten. For BioBank Japan, we additionally conducted fine-mapping using the same pipeline as we did for UK Biobank. The GWAS summary statistics of total cholesterol were computed for the variants with $R^2 > 0.7$ and MAF > 0.01% using BOLT-LMM with the covariates including the top ten principal components, sex, age, age², sex × age, sex × age² and disease status (affected versus non-affected) for the 47 target diseases in BioBank Japan. The details of genotyping and imputation were described extensively previously^{94,129}.

Assessment of the empirical P value threshold. To evaluate the appropriate P value threshold for Tractor associations, we estimated ancestry-specific empirical null P value distributions via permutation. Although the genome-wide significance threshold ($P < 5 \times 10^{-8}$) is widely adopted in the current literature, previous work has shown that different ancestry groups have different numbers of independent variants³⁹. Here, we permuted a null continuous phenotype 1,000 times using the same admixed African–European individuals from UK Biobank as in the Tractor cholesterol GWAS to assess the correct P value threshold for the admixed individuals in this study. We measured the minimum P values of associations (P_{min}) for each ancestry and derived an ancestry-specific empirical genome-wide significance threshold as the fifth percentile ($\alpha = 0.05$) of P_{min} across permutations, as previously described¹³⁰. We calculated this percentile using the Harrell–Davis distribution-free quantile estimator¹³¹ and calculated the 95% confidence interval via bootstrapping. Based on the permutation results (Supplementary Fig. 2), a study-wide significance threshold at a conservative level of $P = 1 \times 10^{-8}$ for both AFR- and EUR-specific associations was deemed appropriate for admixed African–European cohorts, consistent with the presence of additional recombination breakpoints present in such admixed populations. In addition, we calculated the genomic inflation factor, λ_{GC} , of null phenotypes across permutations and confirmed no inflation using the Tractor GWAS model (Supplementary Table 3).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All summary statistics described here for total and LDL cholesterol in ~4,300 admixed UK Biobank individuals can be found at https://github.com/eatkinson/Tractor_ms_results and have been uploaded to the GWAS catalog under accession numbers GCST90012868–GCST90012873 (<https://www.ebi.ac.uk/gwas/deposition/bodyofwork/GCP000093>). The UK Biobank raw data can be obtained through a data access application available at <https://www.ukbiobank.ac.uk>. PGC-PTSD data can be obtained through a data access application at <https://pgc-ptsd.com/data-samples/access-data/>. BioBank Japan summary statistics are available at <http://jenger.riken.jp/en/>. The 1000 Genomes reference panel is available at <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>. The Human Genome Diversity Project dataset is available at <https://www.internationalgenome.org/data-portal/data-collection/hgdp>.

Code availability

All code is freely available. The automated quality control pipeline to prepare datasets for Tractor and run LAI is located at <https://github.com/eatkinson/Post-QC>. We freely provide Tractor code in Python and Hail, as well as examples of implementation in Jupyter notebook at <https://github.com/eatkinson/Tractor> alongside a detailed wiki. Specific scripts used to produce the simulated data

and results are additionally freely provided at https://github.com/eatkinson/Tractor_ms_results.

References

111. Van Rossum G. & Drake, F. L. Jr. *Python Reference Manual* (Centrum Wiskunde en Informatica, 1995).
112. GNU Project, Free Software Foundation. Bash (3.2.48) [Unix shell program] (2007).
113. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
114. Chen, C. Y. et al. Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).
115. Williams, A. admix-simu: program to simulate admixture between multiple populations. *Zenodo* <https://doi.org/10.5281/ZENODO.45517> (2016).
116. Cann, H. M. et al. A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
117. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
118. The Hail Team. Hail <https://github.com/hail-is/hail> (2008).
119. Google Compute Engine launches, expanding Google's cloud offerings. *Google Cloud Platform Blog* <https://cloudplatform.googleblog.com/2012/06/google-compute-engine-launches.html> (2012).
120. Kluyver, T. et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (eds. Loizides, F. & Schmidt, B.) 87–90 (IOS Press, 2016).
121. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
122. Pruijm, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
123. Bokeh: Python Library for Interactive Visualization (Bokeh Development Team, 2020); <https://bokeh.org/citation/>
124. Shin, J.-H., Blay, S., McNeney, B. & Graham, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide oolymorphisms. *J. Stat. Softw.* **16**, 1–10 (2006).
125. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
126. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
127. Benner, C., Havulinna, A., Salomaa, V., Ripatti, S. & Pirinen, M. Refining fine-mapping: effect sizes and regional heritability. Preprint at *bioRxiv* <https://doi.org/10.1101/318618> (2018).
128. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
129. Akiyama, M. et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 4393 (2019).
130. Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.* **61**, 861–866 (2016).
131. Harrell, F. E. & Davis, C. E. A new distribution-free quantile estimator. *Biometrika* **69**, 635–640 (1982).

Acknowledgements

We thank the PGC-PTSD working group, P. Natarajan, S. Gagliano Talun and many other scientists within and beyond Boston for their intellectual contributions to this work. This project was supported by the National Institute of Mental Health (K01 MH121659 and T32 MH017119 to E.G.A., K99MH117229 to A.R.M., R37 MH107649 to B.M.N., and 2R01MH106595 to C.M.N. and K.C.K.). M.K. was supported by a Nakajima Foundation Fellowship and the Masason Foundation. M.L.S. was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (#2018/09328-2). The BioBank Japan Project was supported by the Tailor-Made Medical Treatment Program of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Agency for Medical Research and Development (AMED). This research has been conducted using the UK Biobank Resource under application number 31063.

Author contributions

E.G.A. designed and implemented the pipeline, ran the analyses and drafted the manuscript. A.X.M. designed and ran the analyses. M.K. designed and ran the analyses with the aid of J.C.U., Y.K., Y.O. and H.K.F. A.R.M. contributed code and aided in writing the manuscript. K.J.K. and M.L.S. aided in code implementation. K.C.K., C.M.N., B.M.N. and M.J.D. supervised and advised on the project. All authors reviewed and approved the final draft.

Competing interests

M.J.D. is a founder of Maze Therapeutics. A.R.M. serves as a consultant for 23andMe and is a member of the Precise.ly Scientific Advisory Board. B.M.N. is a member of the Deep Genomics Scientific Advisory Board and serves as a consultant for the CAMP4 Therapeutics Corporation, Takeda Pharmaceutical and Biogen. The remaining authors declare no competing interests.

Additional information

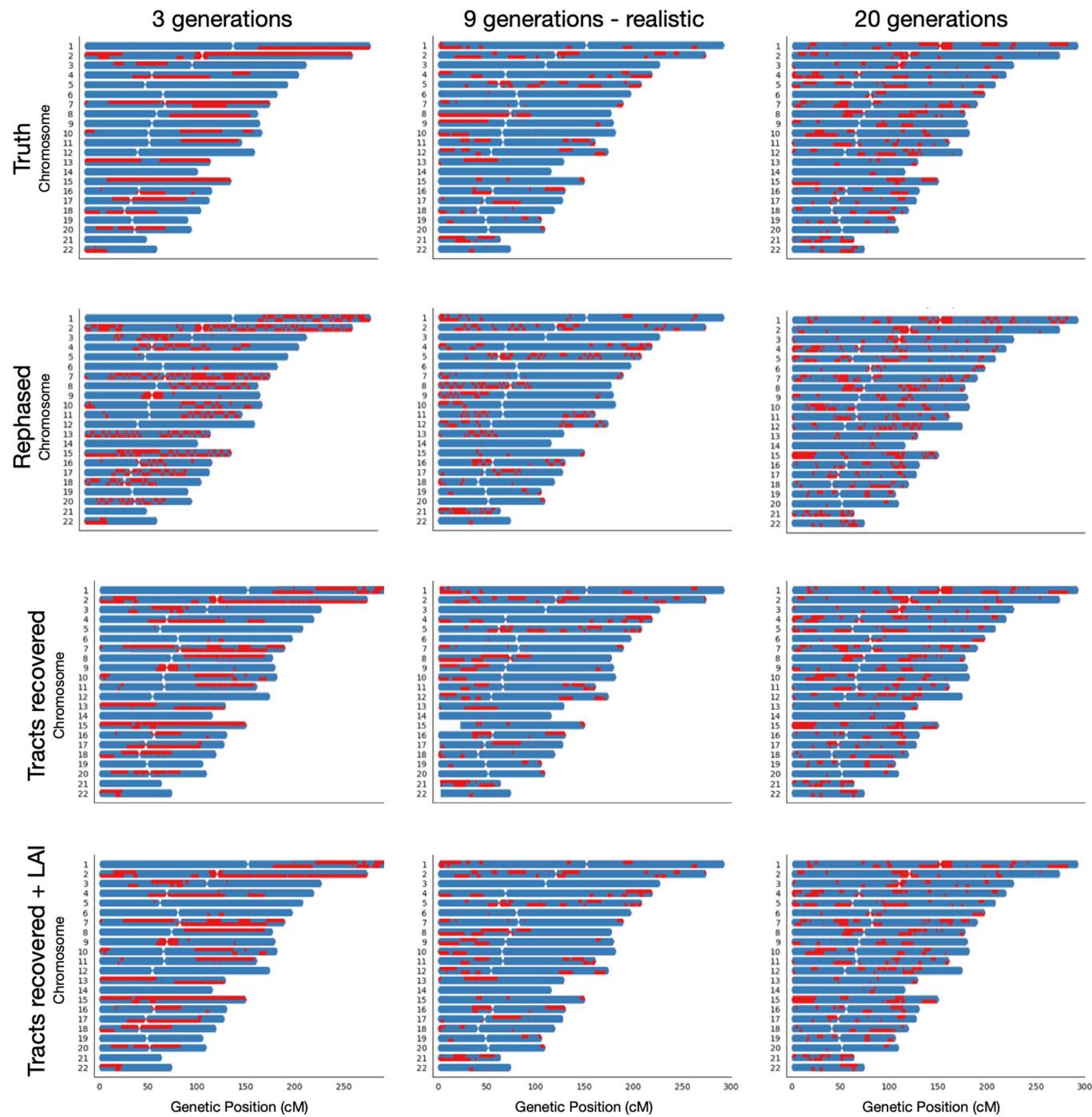
Extended data is available for this paper at <https://doi.org/10.1038/s41588-020-00766-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-020-00766-y>.

Correspondence and requests for materials should be addressed to E.G.A.

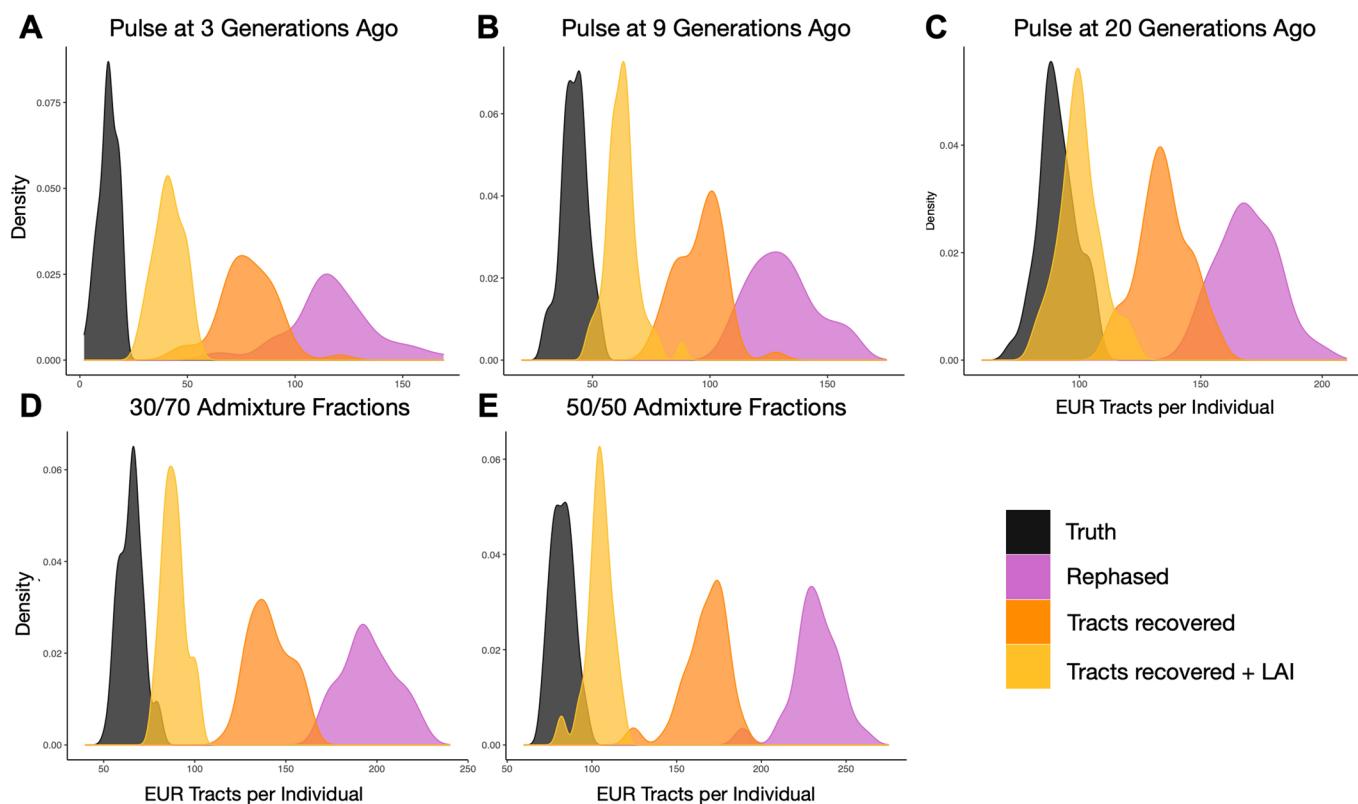
Peer review information *Nature Genetics* thanks Michelle Daya and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

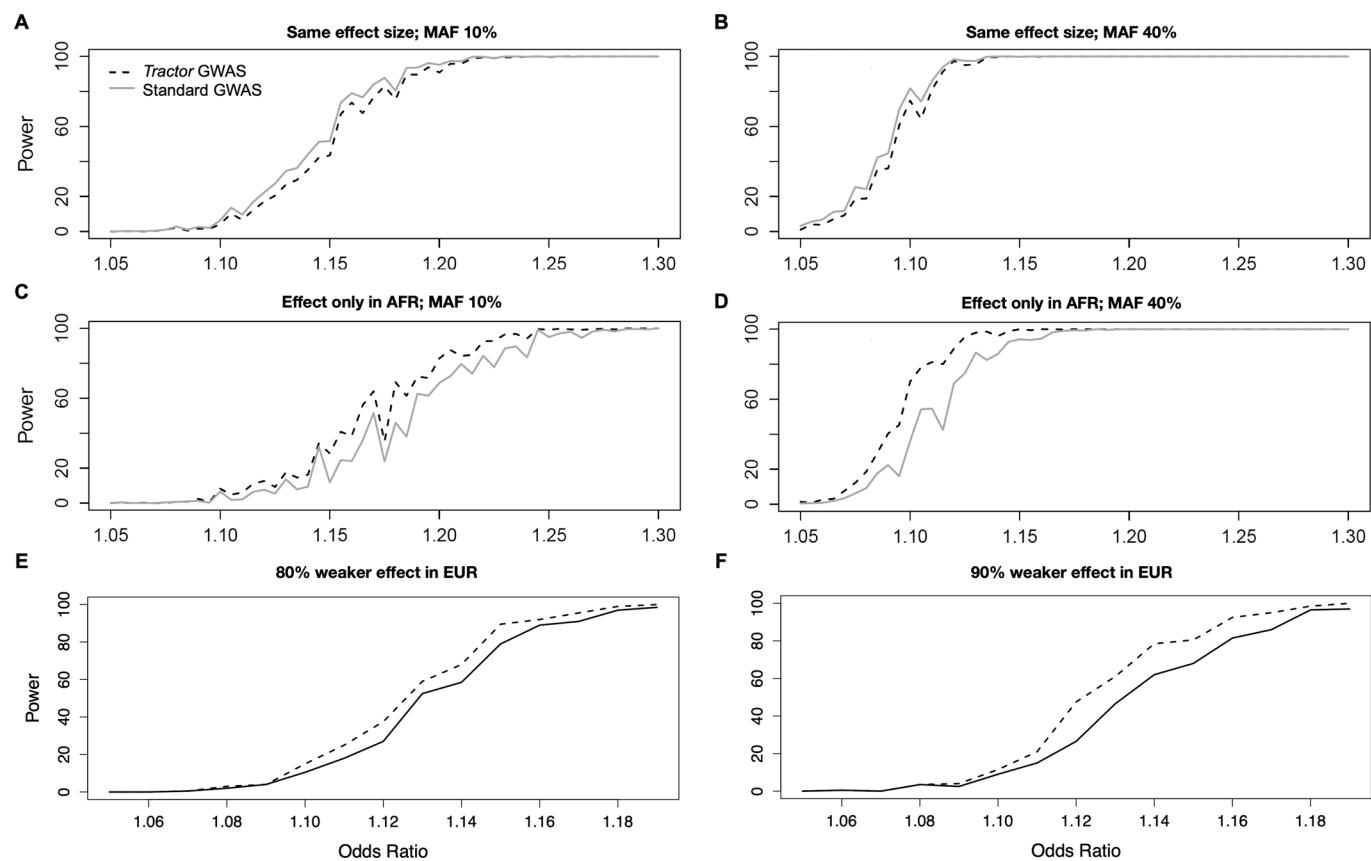


Extended Data Fig. 1 | Painted karyograms of a simulated AA individual showing EUR (red) and AFR (blue) ancestral tracts across demographic models.

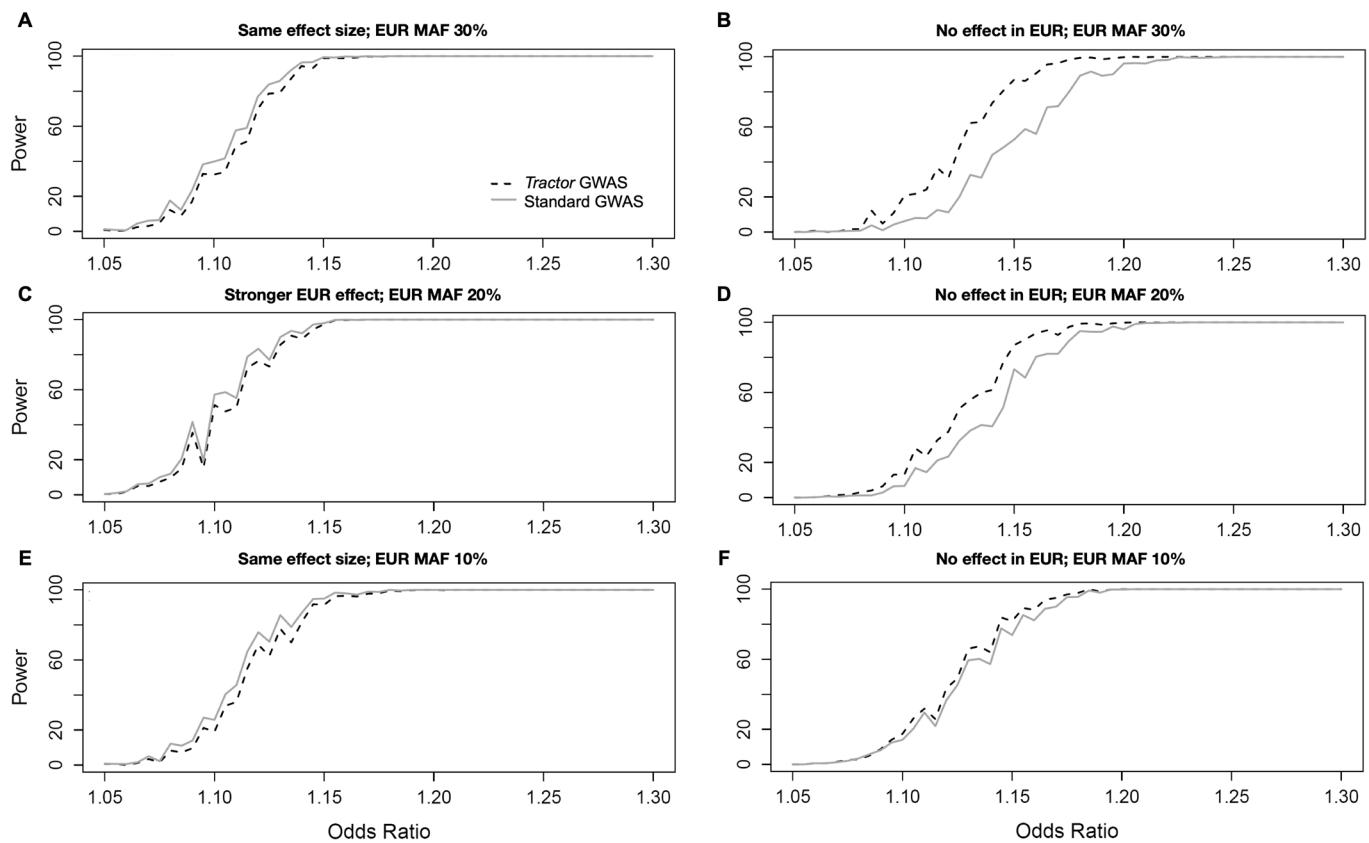
The first column shows the results for the demographic model of one pulse of admixture 3 generations ago, the middle column shows the realistic model of one pulse 9 generations ago, and the right column shows a pulse 20 generations ago. In all cases the model involved 84% AFR ancestry and 16% EUR. The rows show the results from treatments of the data across steps of the *Tractor* pipeline. The top row shows the truth results from our simulations. Painted karyograms after statistical phasing of this truth cohort is shown in the second row. The third row illustrates the recovery of tracts broken by switch errors in phasing obtained by unkinking. The bottom row shows the smoothing and further improvement of tracts acquired through an additional round of LAI.



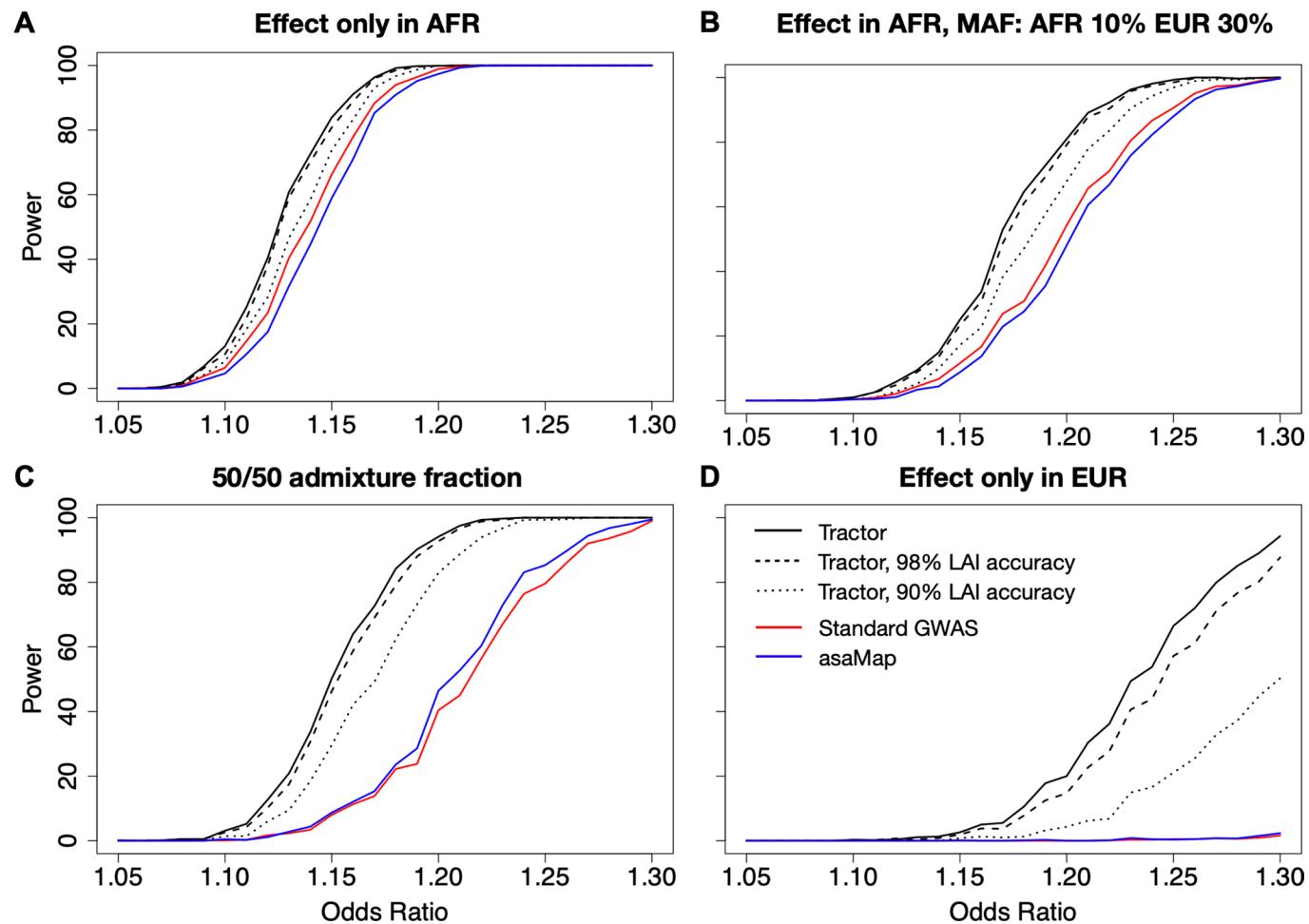
Extended Data Fig. 2 | Tractor recovers disrupted tracts, improving tract distributions. The top row (A-C) shows the improvements to the distributions of the number of discrete EUR tracts observed in simulated AA individuals under demographic models of 1 pulse of admixture at 3, 9 (realistic for AA population history) and 20 generations ago. The bottom row (D,E) shows the results from different initial admixture fractions, of 70% and 50% AFR, respectively, at 9 generations since admixture. These can be compared to the inferred realistic demographic model shown in B. In all panels, the simulated truth dataset is shown in black, after statistical phasing in purple, immediately after tract recovery procedures is in orange, and after one additional round of LAI after tract recovery in yellow.



Extended Data Fig. 3 | The contribution of absolute MAF and effect size to Tractor power. All cases assume an 80/20 AFR/EUR admixture ratio, 10% disease prevalence, 12k cases/30k controls with an effect only in the AFR genetic background. In all panels, the solid line uses a traditional GWAS model while the dashed line is our LAI-incorporating Tractor model. (A,B): Equal effect in EUR and AFR with shifted absolute MAF. (C,D): effect only in AFR background. (A,C): MAF is set to 10% in both AFR and EUR. (B,D): MAF is set to 40% in both AFR and EUR. Panels E and F illustrate the heterogeneity in effect sizes required to observe gains in Tractor power over traditional GWAS assuming 20% MAF in both ancestries and an effect that is stronger in AFR with varying difference to the EUR effect.



Extended Data Fig. 4 | The interaction of between-ancestry MAF differences and effect sizes on Tractor power. In all cases, the grey solid line uses a traditional GWAS model while the black dashed line is our LAI-incorporating model, admixture proportions are 80/20 AFR/EUR, disease prevalence is 10%, and the AFR MAF is fixed at 20%. **A** and **E** model the same effect size between EUR and AFR while varying the EUR MAF. **B,D,F** model the case when there is no effect in the EUR background while varying EUR MAF. **C** models an effect size difference of 30% with the effect being stronger in the EUR background. For comparison, Fig. 2f shows the same effect at matched 20% MAF.



Extended Data Fig. 5 | The impact of LAI accuracy on Tractor's performance as compared to standard GWAS and asaMap. We modeled perfect accuracy, realistic accuracy as derived from simulations of our AA demographic model (98%), and a lower bound of 90% LAI accuracy. Black lines all indicate Tractor runs: the solid black line is Tractor's performance with perfect LAI accuracy, the dashed line is at 98% accuracy, and the dotted line is at 90% accuracy. The red line represents the power obtained from standard GWAS, and the blue line for the asaMap model for the ancestry in which the effect was modeled (AFR for A,B, and C, and EUR for D). In all cases we included 10 PCs as covariates and 1000 replicates were run.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No new data was collected for this manuscript; all data used here has been previously published/described.

Data analysis

Novel code developed for this manuscript is freely available on github. The automated QC pipeline to prepare datasets for Tractor and run LAI is located at <https://github.com/eatkinson/Post-QC>. We freely provide Tractor code in python and Hail, as well as examples of implementation in a Jupyter notebook at <https://github.com/eatkinson/Tractor> alongside a detailed wiki. Specific scripts used to produce the simulated data and results is additionally freely provided at https://github.com/eatkinson/Tractor_ms_results.

We used the following software during the analyses described in this manuscript: linux for x86_64 architecture systems, python version 2.7, R version 3.6.1 and package LDheatmap, Hail version 0.2, Bokeh v2.2.3, Jupyter notebook 6.0, plink 1.9, SHAPEIT2, admix-simu (<https://github.com/williamslab/admix-simu>), RFmix version 2 (<https://github.com/slowkoni/rfmix>), FINEMAP v1.3.116, susieR v0.8.1.0521, LDstore v2.0b, BOLT-LMM.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about availability of data

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All summary statistics described here for Total and LDL Cholesterol in ~4300 admixed UK Biobank individuals can be found at https://github.com/eatkinson/Tractor_ms_results and have been uploaded to the GWAS catalog under accession numbers GCST90012868-GCST90012873. The UK Biobank raw data can be obtained through a Data Access Application, available at <https://www.ukbiobank.ac.uk>. PGC-PTSD data can be obtained through a data access application at <https://pgc-ptsd.com/data-samples/access-data/>. Biobank Japan summary statistics are available at <http://jenger.riken.jp/en/>. The Thousand Genomes reference panel is available at: <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>. The Human Genome Diversity Project dataset is available at <https://www.internationalgenome.org/data-portal/data-collection/hgdp>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

To ensure that our Tractor joint-analysis GWAS model performs well on empirical data, we ran the method for three well-characterized blood lipid phenotypes, and in this paper highlight the results for Total Cholesterol (TC). As our TC cohort we used 4309 two-way African-European admixed individuals from the UK Biobank (UKBB) who had this biomarker data available.

To select individuals with 2-way admixture with European and West African ancestry, we took a two-pronged approach. First, we combined genetic reference data from the 1000 Genomes Project and Human Genome Diversity Panel and ran PCA on unrelated individuals from this reference dataset. To partition individuals in the UKBB based on their continental ancestry, we used the PC loadings from the reference dataset to project UKBB individuals into the same PC space. We trained a random forest classifier given continental ancestry meta-data based on the top 6 PCs from the reference training data. We applied this random forest to the projected UKBB PCA data to assign continental ancestries.

For those individuals classified by their genetic data to have AFR ancestry with probability >0.5, we then combined the 1000 Genomes and Human Genome Diversity Panel reference data with genetic data from the African Genome Variation Project as well as these UKBB individuals. To restrict to only two-way admixed West African-European ancestry individuals, we restricted to individuals with at least 12.5% European ancestry, at least 10% African ancestry, and who did not deviate more than 1 standard deviation from the AFR-EUR cline (Figure S6A, B). This resulted in approximately 4300 individuals per blood lipid trait.

Data exclusions

We excluded UKBB individuals who were not 2-way European-African admixed, as determined using the procedures described above. This was done to create a pseudo-cohort of individuals to be comparable in composition and size to anticipated user cohorts comprising African American individuals.

Replication

We confirmed ancestry-specific top hits produced for TC by Tractor are robust by running blood panel phenotypes in two additional cohorts: another non-European collection, Biobank Japan (BBJ), as well as a different subset of UKBB individuals of white, British ancestry. Specifically, we conducted GWAS and statistical fine-mapping in these two additional large-scale cohorts, comprising 345,235 white British individuals from UKBB and 135,808 Japanese individuals from BBJ. For the UKBB white British, we used previously conducted fine-mapping results for TC (<https://www.finucanelab.org/data>). Briefly, we computed association statistics for the variants with INFO > 0.8, MAF > 0.01% (except for rare coding variants with MAC > 0), and HWE p-value > 1e-10 using BOLT-LMM124 with covariates including the top 20 PCs, sex, age, age2, sex * age, sex * age2, and blood dilution factor. We used FINEMAP v1.3.1125,126 and susieR v0.8.1.0521127 for fine-mapping using the GWAS summary statistics and in-sample dosage LD matrices computed by LDstore v2.0b. We defined regions based on 3 Mb window surrounding lead variants and merged them if overlapped. The maximum number of causal variants in a region was specified as 10. For BBJ, we additionally conducted fine-mapping using the same pipeline as we did for UKBB. The GWAS summary statistics of TC was computed for the variants with Rsq > 0.7 and MAF > 0.01% using BOLT-LMM with the covariates including top 20 PCs, sex, age, age2, sex * age, sex * age2, and disease status (affected versus non-affected) for the 47 target diseases in the BBJ. Top hits fine-mapped in biobank Japan and UKBB British individuals were consistent with the hits prioritized by Tractor.

Randomization

In our empirical GWAS runs, we included covariates capturing global ancestry fraction (AFR fraction or PCs1-10), age, sex, and blood dilution factor in all runs.

Blinding

Blinding is not relevant to this study, as we did not generate new data. The bulk of this manuscript describes a novel method, which we test with simulated data. In our empirical tests, we ran the quantitative traits Total Cholesterol and LDL Cholesterol, which had been previously collected in a biobank setting (UKBB).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study <input checked="" type="checkbox"/> Antibodies <input checked="" type="checkbox"/> Eukaryotic cell lines <input checked="" type="checkbox"/> Palaeontology and archaeology <input checked="" type="checkbox"/> Animals and other organisms <input type="checkbox"/> Human research participants <input checked="" type="checkbox"/> Clinical data <input checked="" type="checkbox"/> Dual use research of concern
-----	---

Methods

n/a	Involved in the study <input checked="" type="checkbox"/> ChIP-seq <input checked="" type="checkbox"/> Flow cytometry <input checked="" type="checkbox"/> MRI-based neuroimaging
-----	---

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

The individuals included in the empirical analyses in this manuscript consist of ~4300 2-way admixed European-African participants in the UK Biobank who had blood lipid biomarker data released. The process by which we determined ancestry and who to include in our analyses from the larger collection is described in detail above in the section entitled "Sample size". Additional details about how the UK Biobank recruited subjects is found immediately below. To control for any confounding effect of age or sex, we included these as covariates in our GWAS. UKBB individuals range in age between the ages of 40 and 69 and are approximately 54% female.

Recruitment

Following the success of an initial pilot study in 2005-2006, the main stage of recruitment for the UK Biobank resource began in 2007, with the goal of recruiting 500,000 individuals between the ages of 40 and 69. The restriction to this cross-section of the population was due to the primary aim of the study; to improve the prevention, diagnosis and treatment of serious illnesses that typically onset later in life; including diabetes, cancer, arthritis, heart disease, stroke, and dementia. To that end, individuals from across the UK were contacted by post to participate in the study, with names, addresses, and dates of birth provided by the UK National Health Service (NHS). The 500,000 recruitment goal was reached in July of 2010, and recruitment ended shortly after. Focusing on voluntary recruitment of an older subset of the UK population sent by mail led to a preponderance of healthy and wealthy Brits of primarily European ancestry. This means that this cohort is not a perfect representation of the UK population due to some selection bias.

Ethics oversight

The empirical data used in these analyses comes from the UK Biobank, a large-scale open database with hundreds of thousands of individuals' genotype data paired to electronic health records and survey data. Researchers can gain access to the UK Biobank by writing a proposal for a research project, which then is reviewed for approval. The UK biobank is more thoroughly described on their website (<https://www.ukbiobank.ac.uk/>).

The Research Ethics Committee reference for the UK Biobank is 11/NW/0382.

Our approved proposal to use UK Biobank data is UK Biobank application #31063 "Methodological extensions to estimate genetic heritability and shared risk factors for phenotypes of the UK Biobank".

Note that full information on the approval of the study protocol must also be provided in the manuscript.