bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tractor: A framework allowing for improved inclusion of admixed individuals in large-scale association studies.

Elizabeth G. Atkinson[1,2,3,*], Adam X. Maihofer[4], Masahiro Kanai[1,2,3,5,6], Alicia R. Martin[1,2,3], Konrad J. Karczewski[1,2], Marcos L. Santoro[2,7], Jacob C. Ulirsch[1,2,3,8], Yoichiro Kamatani[9], Yukinori Okada[6,10,11], Hilary K. Finucane[1,2,3], Karestan C. Koenen[2,12], Caroline M. Nievergelt[4,†], Mark J. Daly[1,2,3,13,†], Benjamin M. Neale[1,2,†]

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

[2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

[5]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA

[6]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan

[7]Departamento de Psiquiatria, Universidade Federal de São Paulo, São Paulo, Brazil

[8]Program in Biological and Biomedical Science, Harvard Medical School, Boston, MA, USA

[9]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

[10]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan

[11]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan

[12]Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA, USA

[13]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

[†] These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*Corresponding author contact information:

eatkinso@broadinstitute.org

(202) 246-0666

Analytic and Translational Genetics Unit

Richard B. Simches Building, 6th Floor

185 Cambridge Street

Boston, MA 02114

Abstract

Admixed populations are routinely excluded from medical 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 simulations and empirical data focused on admixed African-European individuals. *Tractor* generates ancestry-specific effect size estimates, can boost GWAS power, and improves the resolution of association signals. Using a local ancestry aware regression model, we replicate known hits for blood lipids in admixed populations, discover novel hits missed by standard GWAS procedures, and localize signals closer to putative causal variants.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Admixed groups, whose genomes contain more than one ancestral population such as African American and Hispanic/Latino individuals, make up more than a third of the US populace, and the population is becoming increasingly mixed over time[1]. Many common, heritable, diseases including prostate cancer[2–5], asthma[6–9], and several cardiovascular disorders such as atherosclerosis[10,11] are enriched in admixed populations of the US. However, only a minute proportion of association studies address the genetic architecture of complex traits in such groups[12,13]; admixed individuals are systematically removed from many studies due to the lack of methods and pipelines to effectively account for their ancestry such that population substructure can infiltrate analyses and bias results[14–21]. Large-scale efforts to collect genetic data alongside medically-relevant phenotypes are beginning to focus more on non-Eurasian ethnic groups that contain higher amounts of admixture[22–27], motivating the timely 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 the concerning health disparities that already exist[28–32].

In GWAS, the specific concern regarding including admixed participants is obtaining false positive hits due to alleles being at different frequencies across populations. Most studies currently attempt to control for this by using Principle Components (PCs) in a linear or linear mixed model framework. However, PCs 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 PCs 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. Perhaps the most notable example of this is in multi-ethnic fine-mapping, which can dramatically reduce the variant credible set by leveraging the differing LD structures observed across populations[33–38]. This is particularly helpful in populations of African descent, where LD blocks are the shortest and individuals have nearly a million more variants per person than individuals outside of the continent[39]. We find that with admixed populations we not only can utilize the LD patterns from multiple

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ancestries, but have further disrupted LD blocks within each one, offering a more refined LD landscape with which to localize GWAS signal.

To help ensure that advances in genomic medicine will apply globally, we have developed a scalable framework that allows for the easy incorporation of admixed individuals into psychiatric genomics efforts by using local ancestry inference (LAI). Our framework, distributed as a scalable software package named *Tractor*, generates ancestry dosages at each site from input LAI calls, extracts painted haplotype segments for correction at the genotype level and runs a local ancestry-aware regression model, producing ancestry-specific effect size estimates and *p* values. Through testing in simulations and with empirical data on phenotypes with differing levels of polygenicity, we demonstrate that *Tractor* produces accurate results in admixed cohorts and boosts GWAS power across many genetic contexts. We further demonstrate improvements in association signal localization from the higher resolution of haplotype breakpoints in admixed genomes. These efforts fill a gap in existing resources and will improve our understanding of complex diseases across diverse populations.

The incorporation of local ancestry into variant identification for admixed populations is a concept that has been discussed previously[40–50], particularly with regard to 'admixture mapping,' whereby researchers associate an elevation of a given ancestry at a locus in the genome 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 which are highly stratified across populations, such as asthma and cardiovascular phenotypes[56–61]. We build upon this important work by modeling the local ancestry haplotype dosage for each person at each variant in a way that allows for the generation of ancestry-specific effect size estimates while allowing for differences in minor allele frequency (MAF) across populations without an increased false positive risk.

The statistical model built into *Tractor* for binary phenotypes tests each SNP for an association with the phenotype using the logistic regression model: $\mathrm{logit}(\mathrm{Y}) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots + b_kX_k$

where 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 to X_k are other covariates such as PCs, age, sex, etc. The significance of the risk allele is evaluated with a likelihood ratio test comparing the full model to a model fit without the risk allele, thus

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

allowing estimation of the aggregated effects in the presence of effect size heterogeneity. To further test if a risk allele is ancestry-specific, we evaluate if the difference between b_2 and b_3 is non-zero using a *Z*-test. The model presented is for a 2-way admixed scenario but can be scaled to an arbitrary number of ancestries. We have built pipelines to implement this joint model as well as generate genotype files containing extracted ancestry portions. These tools can be implemented in python and Hail either locally or on the cloud.

Results

LAI has high accuracy for African Americans

We ran LAI using the program RFmix, a discriminative approach which estimates local ancestry by using conditional random fields parameterized with random forests[62]. RFmix can run on multi-way admixture populations, outperforms other local ancestry inference methods for minority populations, and leverages the ancestry components in admixed reference panel individuals, highly important when there is a lack of homogenous reference panels – often the case for understudied groups[63,64]. As *Tractor* relies heavily on LAI calls, we first ran simulations to ensure that RFmix called local ancestry accurately. LAI was highly accurate in a realistic demographic model for African American (AA) individuals (one pulse of admixture 9 generations ago with 84% contribution of haplotypes from Africa (AFR) and 16% from Europe (EUR); see *Methods*), assigning the correct ancestry ~98% of the time (Table S1). To ensure that *Tractor* performed well across demographic models, we varied demographic parameters including admixture fractions and pulse timings. Specifically, we varied the pulse of admixture in time to 3 generations and 20 generations ago and changed the admixture fractions to 30/70% and 50/50% EUR and AFR ancestry, respectively (**Figure** S1). We also checked the ancestry-specific accuracy in the realistic demographic scenario to assess if 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 (Table S1). 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 more geographically refined groups and deconvolve ever more specific haplotypes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recovery of long-range 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[65,66], but few methods to recover disrupted haplotypes exist. Taking advantage of the unique ability to visualize tracts offered by admixed individuals, we additionally improve long-range haplotype resolution by correcting chromosome strand switch errors from phasing, which we find to be common in admixed cohorts. We demonstrate that using local ancestry information, we are able to consistently correct switch errors and recover disrupted haplotypes, making tract distributions look significantly more realistic (**Figures** 1, 2).

To replicate standard analytical procedures employed on cohort data, we statistically phased our truth dataset using SHAPEIT2[67] software with a balanced reference panel composed of EUR and AFR continental individuals from the 1000 Genomes Project[39]. We then examined the distribution and lengths of the EUR tracts. Analyzing the less common ancestry tracts allows for more precise quantification of tract counts because it is less likely that recombination will mask their phase switch errors. The probability of obtaining the observed number of tracts after phasing (131 after phasing vs 42 in the truth dataset) given the input demographic model was extremely unlikely, p=5.0x10^{-26}. After correcting phase switch errors, the likelihood of the tract distribution improved, albeit still with significantly more switches than in the truth data (p=2.7x10^{-11}, 96 tracts) – approximately half the excess tracts without phase error correction. After then implementing one additional round of LAI on the corrected genotype files, the number of excess tracts was further reduced (p=0.009, 62 tracts). Thus, our procedure for correcting phase switch errors successfully recovers long-range haplotypes and better approximates the true tract length distributions (**Figure** 2).

To ensure that phase switch error correction performs well across different population histories, we ran simulations to assess how closely the tract length distributions approximated the truth for a range of demographic models. Specifically, we checked performance varying the timing of admixture pulses (including 3, 9, and 20 generations ago) and the admixture proportions in the simulation (70/30, and 50/50 AFR/EUR, respectively). Under all scenarios, our tract recovery procedure improves strand flips in painted karyograms

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**Figures** 1, S1) and decreases the significance of the difference between the observed versus true tract length distributions (**Figure** 2, Table S2). We note that running an additional round of LAI after recovering haplotypes produced the most accurate tract length distributions. This is due to the improved ability of the model to recognize ancestry switch points once more complete haplotypes have been recovered, resulting in smoothing over previous short miscalls.

Evaluating the landscape of GWAS power gains from Tractor

We simulated individuals' likelihoods of being cases as a function of AFR admixture fraction, the ancestral haplotype of each copy of the risk allele, and the risk allele dosage (See *Methods, Supplementary Information).* This framework is equivalent to modifying the marginal effect sizes due to a tag SNP for a shared causative mutation being monomorphic in EUR but variable in AFR, which is plausible as individuals from Africa contain almost a million more variants than other populations[68]. This 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 model and under our model. Compared to the traditional model, there is a significant gain in power using the *Tractor* framework with similar improvements across sample sizes and disease prevalences (**Figure** 3). Power increases further when there is a difference in MAF across ancestries.

We ran similar sets of simulations varying the paraments of the effect size difference, the absolute MAF, MAF difference across ancestries, and admixture fractions (**Figures** 3, S2, S3). The biggest power gain comes if an allelic effect is present in the smaller fraction ancestry only. For example, in a realistic AA demographic model, EUR ancestry makes up only ~20% of the sample. If we model an allele with an effect only active in the EUR background (**Figure** 3D), analyzing the tracts together without LAI information will have essentially no power to detect an association due to the higher noise relative to signal from uninformative tracts. However, *Tractor* is able to recover the effective sample size and power that one would have had if analyzing just the effect haplotypes, i.e. the EUR segments alone.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The scenario where *Tractor* is most powerful is when there is heterogeneity in the apparent effect size for the same variant across ancestries. Such heterogeneity in effect sizes may be a consequence of the same variant having different effects in different populations (e.g., in the context of gene-environment interactions) or may arise from differences in the indirect association evidence of the variant (i.e., the contribution to the estimated effect size from tagging other causal genetic variants). Differences in indirect association can come from ancestry-specific variation or from different patterns of linkage disequilibrium. Moreover, under the converse scenarios where there is predicted to be no benefit, the power loss is minimal—we lose a degree of freedom, resulting in a less precise error estimate for each SNP effect (Figure S2,3). In no case does *Tractor* dramatically underperform compared to the traditional GWAS model. See *Supplementary Information* for additional simulations and detail about power results.

No increase in false positive rate with the Tractor model

We quantified the false positive rate of the *Tractor* model by simulating a variant with no effect and counting the spurious significant associations identified in a simulated realistic AA population given $\boldsymbol{\alpha} = 0.05$. Across our tests at various MAFs and between-ancestry MAF differences, we observe no clear difference in false positive rate between *Tractor* and traditional GWAS (**Figure** S4). In addition, we calculated the genomic inflation factor, λ_{GC}, of null phenotypes across GWAS permutations and confirmed no significant inflation using the *Tractor* GWAS model (**Figure** S5b). Therefore, there does not appear to be an elevation in false positive rates with the *Tractor* framework, suggesting that the observed power increases are from improved detection of true biological signal.

Tractor replicates known associations and identifies hits for blood lipids in admixed individuals missed by standard GWAS

To ensure that our *Tractor* joint-analysis GWAS model also performs well on empirical data, we ran the method for three well characterized blood lipid phenotypes which have been demonstrated to have ancestry-specific effects: Total Cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), and low-density

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lipoprotein cholesterol (LDL-C). We constructed a pseudo-cohort of 4309 two-way African-European admixed individuals from the UK Biobank (UKBB) with biomarker phenotype data to serve as our sample (**Figure** S4). *Tractor* GWAS replicated previously implicated associations for blood lipids in this cohort[45,69–71], reaching the standard genome-wide significance level of $5x10^{-8}$ at previous top associations, including in genes *PCSK9, LDLR*, and *APOE* (**Figure** 4). In some cases, *Tractor* improved the observed top hit significance.

Our LAI-incorporating model was also able to identify hits in these admixed UKBB individuals that standard GWAS was not when using the traditional genome-wide significance threshold (**Figure** 4). For example, we identify a hit missed by standard GWAS on the same dataset that is present only on the AFR background on chromosome 1 (rs12740374, p=$3.46x10^{-8}$). This locus has previously been shown to affect blood lipid levels, metabolic syndrome, and coronary heart disease risk in independent AA cohorts[69,72–77], and was determined to be the causal variant for affecting LDL-C in a multi-ethnic fine-mapping study[78]. Had we not deconvolved ancestral tracts for our GWAS, we would have missed this site with a demonstrated effect on our phenotype and population of interest.

We additionally identify a novel peak on chr15 that only reached significance in the AFR tracts in this UKBB cohort. The lead SNP (rs12594517, p= 1.915x10-8) lies in an intergenic area and is uncharacterized. The closest gene neighboring it is *MEIS2,* lying ~70kb 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 to body mass index and waist circumference and *C15orf41* was a significant hit in a previous GWAS of cholesterol [79,80]. Though further follow-up is needed to clarify any direct relationship to TC, 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 GWAS signals to closer to the causal variant than is possible using standard GWAS procedures. TC has previously been mapped to the gene *DOCK6* in AA cohorts[69], a finding we replicate for the suggestive GWAS threshold with standard GWAS on UKBB admixed individuals in the same intronic area as previously found. However, when we run the *Tractor* model, we identify a lead *DOCK6* SNP 20kb downstream in the AFR samples, as well as in a meta-analysis of hits from

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

deconvolved AFR and EUR tracts. This new lead SNP (rs2278426) is a missense mutation spanning both *DOCK6* as well as *ANGPTL8*, a gene which may play a key role in blood lipid regulation (see *Discussion,* **Figure** 5). To assess whether our improved ability to localize to this variant was due to a true effect size differences between the EUR and AFR or to a marginal effect size difference driven by MAF or LD differences across the ancestries, we further attempted validating its association by fine-mapping TC in other large-scale populations: 345,235 white British individuals from UKBB and 135,808 Japanese individuals from Biobank Japan ([81], Ulirsch, JC., Kanai, M. et al., in prep., Kanai, M. et al., in prep). rs2278426 was successfully fine-mapped to a 95% credible set in both populations, with maximum posterior inclusion probability of 0.993 in Biobank Japan. This variant is at 26% frequency in the gnomAD [82] East Asian ancestry individuals, 18% in African, and 4% frequency in the non-Finish Europeans. These frequency patterns suggest higher power in non-European population to localize a causal variant compared to Europeans. Though below the traditional genome-wide significance level in our sample of ~4300 individuals, this locus highlights the improved ability to localize GWAS signal thanks to leveraging additional breakpoints in admixed genomes.

Discussion

Despite the 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 analytical framework and statistical gene discovery method distributed as a scalable software package named *Tractor*, which allows admixed samples to be appropriately included alongside homogenous ones in a well calibrated manner in statistical genomics efforts. We test our framework in a simulation model designed to emulate real AA cohorts. We also apply it to empirical data from admixed African-descent individuals of the UKBB. We observe a gain in power to detect risk loci across sample sizes, demographic models, and disease prevalences using the *Tractor* framework, particularly when effect sizes are heterogeneous across populations. Our approach incorporates a local-ancestry aware GWAS method that can extend the traditional GWAS model. *Tractor* generates ancestry specific *p* values and effect size estimates, which admixture mapping cannot, and which can be extremely helpful in post-GWAS efforts such as

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

256 constructing genetic risk scores for understudied populations. We demonstrate that our framework also gives
257 increased precision in localizing GWAS signal by leveraging the disrupted LD blocks visible with ancestral
258 chromosome painting in recently admixed groups. This reduces the credible set of SNPs and aids in the
259 prioritization of variants for subsequent functional testing.

260 The *Tractor* pipeline requires several inputs, most importantly accurate local ancestry calls. Users
261 should ensure good LAI performance in their target cohorts. A major determinant of accurate LAI calls is a
262 comprehensive and well-matched reference panel[49,83]. Of relevance is that reference panels are more plentiful
263 for Eurasian populations than for other groups, underscoring the need to expand sequencing efforts in more
264 global populations. To ensure LAI was unbiased across regions of the genome in our GWAS, we examined the
265 distribution of local ancestry across the genome. Local ancestry inference appears relatively evenly distributed
266 across the genome and proportional to global admixture proportions (**Figures** 1, S6). However, we caution that
267 calls around centromeres and at the ends of chromosomes are most likely to include error, as these genomic
268 regions do not have anchor points on one edge. We similarly recommend that LAI ideally to be conducted on
269 whole genome sequencing data to avoid the introduction of biases. We also note that we have thoroughly
270 tested *Tractor* here in the two-way admixture model that reflects AA demographic history. The analytic
271 infrastructure, however, can currently also run on a three-way admixed model and our statistical model can
272 scale to an arbitrary number of ancestries. Future work will test power and optimize the code in a variety of
273 multi-way admixed demographic scenarios. A final consideration is to ensure consistent phenotyping across
274 ancestry groups, as is standard in multi-ethnic GWAS.

275 We thoroughly evaluated the landscape of when *Tractor* does and does not add power to association
276 studies in simulated data modeled after AA cohorts of the PGC-PTSD (**Figure**s 3, S2, S3). In situations where
277 there are differences across ancestries, *Tractor* recovers the power that would be lost from analyzing
278 populations together. In particular, power gains are greatest when there is an effect size difference at a locus
279 between ancestries coupled with differing MAF. For example, our simulated case of a 20% MAF difference for
280 an allele with an effect only in the AFR genetic background would allow for identification of risk variants with an
281 odds ratio ~0.1 smaller than would be possible with traditional GWAS. This allows for detection of additional

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

loci that would have been undetected without modeling local ancestry. *Tractor* can also boost power when there is an effect only on one haplotype background or an allele only present in one ancestry, again most dramatically when that ancestry is less frequent in the dataset. In such an instance in a standard GWAS setting, the signal would be dramatically reduced due to noise from the uninformative majority haplotypes, resulting in extremely low power to detect the locus (**Figure** 3D). Power can be recovered, however, by deconvolving local ancestry and 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 not generally necessary to include local ancestry in a GWAS model when there is no effect size difference between groups. Notably, we are referring here to detection of marginal effect sizes as well as true effect sizes. There is evidence suggesting that in most cases (with some notable exceptions[8,69,84]), the true effect sizes of causative variants are likely to be equal across ancestries[33,38,85–90]. However, the marginal effect size of a tag SNP might routinely be different across ancestries due to differences in ancestral MAF and the LD patterns resulting from each ancestral population's demographic history[91,92]. We underscore that power gains appear to be from true biological signal rather than false positives, as we quantified the *Tractor* false positive rate to be no higher than standard GWAS (Figures S4, S5).

We would like to highlight that *Tractor* benefits from power gains to detect the marginal beta (pertaining to, for example, an allele which is only present in one population and tags a nearby causal variant), in addition to the rarer case of variants with true effects only on one haplotypic backbone. Our framework therefore will improve power at substantially more locations across the genome than only at sites which have ancestry-specific causal effect differences. *Tractor* also benefits from increased power in cases where functionally important (and likely rare) alleles only present in one population are missed by genotyping or imputation. In such situations the common alleles in LD, despite being shared across populations, would be associated to the phenotype as a function of which haplotypic background they are found on, and thus would have a haplotype-specific effect. Another relevant scenario to consider would be the presence of LD in regions where there are ancestry-specific markers intermingled with shared ones. This would affect the univariate scan results such

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that considering the haplotypic background on which alleles fall would particularly aid in localizing signal through improved marginal beta estimates, even when the causal effect is the same in both ancestries.

Tractor was also able to replicate established GWAS hits, discover new ones, and aid in the localization of GWAS signal with empirical data. We replicated known hits for well-characterized blood lipid phenotypes when testing *Tractor* on a dataset consisting of ~4300 2-way admixed African-European individuals from the UKBB (**Figure** 4). We further demonstrate an improved ability to localize GWAS signal to putative causal SNPs previously identified in another diverse collection, Biobank Japan[81] (**Figure** 5). Specifically, previous analyses of blood lipid phenotypes in an admixed AA cohort had pinpointed the TC top hit to lie within the gene *DOCK6*[69], nearby the lead SNP for this region in standard GWAS on the admixed UKBB individuals (rs4804576). The *Tractor* AFR ancestry, as well as results from a meta-analyses of summary statistics from AFR and EUR deconvolved genotype files, identified a different top association ~20kb downstream (rs2278426) which additionally lies over the *ANGPTL8* gene on the positive strand while spanning an intronic area of *DOCK6* on the minus strand. *ANGPTL8,* also known as lipasin and betatrophin, has been shown to regulate plasma lipid levels in mice by inhibiting the enzyme lipoprotein lipase[93–96]. In humans, ANGPTL8 levels correlate with metabolic phenotypes including type 2 diabetes and obesity[97–100] and HDL-C expression levels across diverse ancestry groups have been demonstrated to better correlate with than *DOCK6*[101]. Together these make *ANGPTL8* a more promising candidate gene than *DOCK6*, which has no clear tie to blood lipid phenotypes. Intriguingly, the *Tractor* lead SNP, rs2278426, is a missense mutation in *ANGPTL8* (p.Arg59Trp) that is predicted to be possibly damaging and deleterious by polyphen and SIFT, respectively[102,103]. This variant is at 18% frequency in gnomAD[82] AFR individuals but at 4% frequency in the non-Finnish Europeans. These frequency differences highlight how leveraging different diverse populations allows for the improved identification of risk variants as well as how employing multi-ethnic mapping methods aids in the resolution of association signals.

The *Tractor* infrastructure released here may be helpful in multiple statistical genetics use cases beyond GWAS. For example, correcting for population structure should be a key early step in evolutionary genomic studies on admixed populations running analyses such as genome-wide scans of selection to avoid

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bias in selection statistics[104–106]. Within medical genetics, accounting for the ancestral background on which an allele appears will be key in admixed populations, particularly in studies of rarer variants which are more population specific[107,108]. For example, because AFR and EUR haplotypes have different rates of background variation[68], controlling for the local ancestral background may help pinpoint the differences between cases and controls in burden testing that would previously have been overwhelmed by uninformative markers.

In sum, *Tractor* allows users to account for genotype-level ancestry in a precise manner, allowing for 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 ancestry-specific effect size estimates a *p* values, improved localization of GWAS signal, and power boosts in genetic scenarios such as when there are effect size or MAF differences across ancestries. This infrastructure is designed as a series of steps to be flexible and easily ported into other statistical genomics activities. We freely provide *Tractor* code in python and Hail[109], a scalable cloud-compatible framework, as well as examples of implementation in a Jupyter notebook[110]. *Tractor* advances the existing methodologies for studying the genetics of complex disorders in admixed populations.

Online Methods

QC and LAI Pipeline

The core feature of the *Tractor* framework relies on accounting for fine-scale population structure as informed by local ancestry (i.e. ancestral chromosome painting). *Tractor* then uses this information to (i) correct for individuals' ancestral dosage at all variant sites, (ii) recover long-range tracts in admixed individuals; and (iii) 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 RFmix versions 1 and 2[62], and have built an automated pipeline (https://github.com/eatkinson/Post-QC) to perform all necessary post-genotyping QC, data harmonization, phasing, and LA inference to consistently prepare the data for downstream analysis. The main code is in bash, subscripts are written in python (See *Supplementary Information* for additional details).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In all tests shown here, we ran RFmix_v2 with 1 EM iteration and a window size of 0.2 cM. We used the HapMap b37 recombination map[111] 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 additionally used the --reanalyze-reference flag, which recalculates admixture in the reference samples themselves for improved ability to distinguish ancestries. This is especially important when the reference samples are themselves admixed. As a reference panel for our 2-way admixed simulated African-European cohorts, we used relevant populations of the 1000G reference panel given *a priori* knowledge of AA's demographic history[112–114] consisting of 108 YRI and 99 CEU. Painted karyogram plots were produced using a modified version of publicly available code (https://github.com/armartin/ancestry_pipeline). We have optimized this pipeline under the two-way admixed AA demographic scenario. *Tractor* additionally supports 3-way admixture calls with an expanded set of scripts (also at https://github.com/eatkinson/Tractor). In all cases we recommend conducting tests of LAI accuracy to ensure reliability, as accurate LAI calls are required for good performance.

LAI Accuracy

We validated that LAI was performing well in the AA use case. To do this, we generated a truth dataset by simulating individuals with known phase and LA from empirical data. Our simulation reference panel consisted of haplotypes from homogenous PGC-PTSD individuals who had ≥95% EUR or AFR ancestry as inferred by SNPweights[115]. We simulated admixture between these reference individuals with admix-simu[116] using a realistic demographic scenario for the AA population [113,114] of 1 pulse of admixture 9 generations ago with 84% contribution from Africa and 16% from Europe. The resultant population mixes amongst itself until the present day, copying haplotypes from the previous generation with break points informed by the HapMap combined recombination map[111]. This retains the LD structure and genetic variation present in real genomic data and ensures that the truth dataset resembles cohort data as closely as possible. We then called LA with the 1000 Genomes[68] AFR and EUR superpopulations as our reference panel, and calculated LAI accuracy as how often the ancestry call was correct in the simulated truth data.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

385

386 ***Correcting Switch Errors from Statistical Phasing using Local Ancestry***

387 Despite LAI calling ancestry dosage accurately, frequent chromosomal switches were visible in painted
388 karyograms (**Figure** 1), which we determined were due to phasing errors. It is important to retain complete
389 tracts, as spurious breakpoints will reduce the accuracy haplotype-based test. *Tractor* detects and fixes phase
390 switches using the most likely ancestry assignment of subpopulations as determined by a conditional random
391 field from RFmix. We define phase switches as a swap of ancestry across a chromosome within a 1 cM
392 window at a region with heterozygous ancestry dosage.

393 To ensure that correcting phase switch errors improved results compared to the truth expectations for
394 the input demographic scenario, we modeled the expected distributions of EUR tract lengths within AA
395 individuals using a Poisson process with rate=9, the number of generations ago when the pulse of admixture
396 occurred (**Figure** 2). The waiting time until a recombination event disrupts a tract is expected to follow this
397 distribution, with a slight shortening of tracts proportional to the percent admixture due to the inability to
398 visualize tract switches that occur across regions of the same ancestry. The overall proportion of the genome
399 in the realistic scenarios was within range of expectations given the simulation model of 16% European, 84%
400 African ancestry (15.1% and 84.9%, respectively).

401

402 ***GWAS power simulations incorporating local ancestry***

403 We assessed the improvements in GWAS power from using *Tractor* through simulations. We
404 formulated our simulation framework on the suggestions of Skotte et al (2019)[117]. Power calculations were
405 based on a simulation framework that initially models an AA population assuming a bi-allelic disease risk allele
406 with a 20% overall MAF and an additive effect in the AFR genetic background but not in the EUR. Specifically,
407 the overall admixture proportions were drawn from a beta distribution with shape parameters 7.76 and 2.17,
408 the fitted parameters to this distribution for AFR ancestry proportions observed in the PGC-PTSD Freeze 2 AA
409 cohorts. The genotype of each copy of the allele was drawn from a binomial distribution with the probability of
410 having the minor allele set to the MAF. We simulated a disease phenotype with individuals' risk drawn from a

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

binomial distribution assuming a 10% disease prevalence. 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 of allelic risk effect size*number of copies of the minor allele coming an AFR ancestral background + 0.5*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*AFR Admixture was set in order 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 status was 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) number of copies of the risk allele only, and M3) admixture + number of copies of the risk allele on a EUR background + number of copies on an AFR background. M1 serves as a null comparison to evaluate the significance of including the SNP as a predictor. The significance of M2 and M3 are evaluated by likelihood ratio tests comparing them to 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 the time that the likelihood ratio test was significant ($p < 5e-8$). We performed 100 rounds of simulation with this model at each level of allelic effect size ranging from Odds Ratio (OR) 1.05 to 1.3 and case sample size *N*=4000 and 12000.

Characterizing the landscape of Tractor power across genomic and disease contexts

To evaluate *Tractor* power gains, we ran similar sets of simulations varying effect size differences across ancestries, MAF differences, admixture fractions, and disease prevalence (**Figure**s 3, S2, S3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

436 *Varying effect size across populations*: To examine the effect of modifying the effect sizes, we introduced
437 an effect on EUR haplotypes as well, rather than just on AFR. All these simulations assumed 80% admixture,
438 10% disease prevalence, and 20% MAF in both groups. We modeled cases across the OR spectrum where
439 there was an effect of equal size in both ancestries, a 30% larger effect size on the EUR background, an effect
440 size 30% larger on the AFR haplotype, and an effect size only in the EUR.
441 *Varying absolute MAF:* We next fixed all other parameters and modified the absolute MAF of the
442 simulated risk allele, with the relative difference in MAF between ancestries remaining constant. We changed
443 our MAF from 20% to 10% and 40% under both the models of an effect only in the AFR background and with
444 matching effect sizes between EUR and AFR.
445 *MAF differences between groups*: To see if having a difference in the MAF between the two ancestral
446 groups affected GWAS power, we varied the MAF in the EUR background to be 10, 20, and 30% while
447 keeping the AFR MAF set to 20%.
448 *False positive rate*: We quantified the false positive rate by simulating a variant with no effect and
449 counting significant associations identified in a simulated realistic AA population given $\alpha = 0.05$.
450
451 ***Selection of two-way admixed African-European empirical individuals***
452 To select individuals with 2-way admixture with European and West African ancestry, we took a two-
453 pronged approach. First, we combined genetic reference data from the 1000 Genomes Project[39] and Human
454 Genome Diversity Panel[118], then harmonized meta-data according to consistent continental ancestries. We
455 then ran PCA on unrelated individuals from the reference dataset. To partition individuals in the UKBB based
456 on their continental ancestry, we used the PC loadings from the reference dataset to project UK Biobank
457 individuals into the same PC space. We trained a random forest classifier given continental ancestry meta-data
458 based on the top 6 PCs from the reference training data. We applied this random forest to the projected UK
459 Biobank PCA data and assigned AFR ancestries if the random forest probability was >50%, otherwise
460 individuals were dropped from further analysis.
461 For those individuals classified by their genetic data to have AFR ancestry, we then combined the 1000

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Global ancestry fraction estimates were obtained from running ADMIXTURE[119] with *k*=2 (which was the best fit *k* value to this dataset based on 5-fold cross-validation) on these individuals with 1000 Genomes Project[39] EUR and AFR superpopulation individuals as reference data (**Figure** S6C). To ensure there were no major areas of the genome where local ancestry inference was skewing significantly from the expected global fractions, we also assessed the cumulative local ancestry calls across the genome for the UKBB admixed subset (**Figure** S6D).

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. We separated steps to allow for maximum flexibility when using *Tractor*. To implement the joint modeling GWAS approach with the novel linear regression model described here, we have built a scalable pipeline in Hail[109] which can be implemented locally or on the Google Cloud Platform[120]. Descriptions of the steps and an example Jupyter notebook [110] demonstrating analytical steps and visualization of results of the *Tractor* joint-analysis GWAS are freely available on github (https://github.com/eatkinson/Tractor).

An alternative pipeline designed for use across environments where Hail may not be as readily implemented involves running the separate/meta-analysis GWAS version of *Tractor*. 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 plink[121]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 power by incorporating summary statistics from both populations, though 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 in European-African admixed UKBB individuals

To ensure that *Tractor* replicated well-established associations, we ran standard GWAS, the *Tractor* joint-analysis model, and a meta-analysis of summary statistics from EUR and AFR deconvolved tracts on ~4300 admixed African-European individuals from the UKBB on the biomarker blood lipid traits of Total Cholesterol (TC), high-density lipoprotein cholesterol (HDLC), and low-density lipoprotein cholesterol (LDLC). We included covariates capturing global ancestry, age, sex, and blood dilution factor in all runs. We assessed meta-analysis performance using different metrics to capture global ancestry, namely the first 20 PCs versus the AFR fraction as determined by ADMIXTURE, which did not have substantive differences. In the 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 PCs. We generated QQ 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 results to those obtained from the same individuals using a standard GWAS approach. No individuals overlap between the previous study of interest (Natarajan et al. 2018, which includes diverse TOPMed[24] individuals) and the UKBB individuals included here. As expected, near-identical results were obtained from the meta- and joint-approaches. Gene visualizations were produced with LocusZoom[122], Manhattan and QQ plots with bokeh[123].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

514

515 ***Statistical fine-mapping of top hits in independent cohorts***

516 We conducted GWAS and statistical fine-mapping in two additional large-scale cohorts, 345,235 white
517 British individuals from UKBB and 135,808 Japanese individuals from BBJ. For the UKBB white British, we
518 used previously conducted fine-mapping results for TC (https://www.finucanelab.org/data). Briefly, we
519 computed association statistics for the variants with INFO > 0.8, MAF > 0.01% (except for rare coding variants
520 with MAC > 0), and HWE p-value > 1e-10 using BOLT-LMM[124] with covariates including the top 20 PCs, sex,
521 age, age^2, sex * age, sex * age^2, and blood dilution factor. We used FINEMAP v1.3.1[125,126] and susieR
522 v0.8.1.0521[127] for fine-mapping using the GWAS summary statistics and in-sample dosage LD matrices
523 computed by LDstore v2.0b. We defined regions based on 3 Mb window surrounding lead variants and
524 merged them if overlapped. The maximum number of causal variants in a region was specified as 10. For BBJ,
525 we additionally conducted fine-mapping using the same pipeline as we did for UKBB. The GWAS summary
526 statistics of TC was computed for the variants with Rsq > 0.7 and MAF > 0.01% using BOLT-LMM with the
527 covariates including top 20 PCs, sex, age, age^2, sex * age, sex * age^2, and disease status (affected versus
528 non-affected) for the 47 target diseases in the BBJ. The details about genotyping and imputation was
529 extensively described previously[81,128].

530

531 ***Assessment of the correct empirical p value for admixed individuals***

532 To evaluate the appropriate *p* value threshold for *Tractor* associations, we estimated ancestry-specific
533 empirical null *p* value distributions via permutation. Although the genome-wide significance threshold ($p < 5 \times$
534 10^{-8}) is widely adopted in the current literature, previous work has shown that different ancestry groups have
535 different numbers of independent variants[68]. Here, we permuted a null continuous phenotype 1,000 times
536 using the same admixed African-European individuals from UKBB as in the *Tractor* cholesterol GWAS to
537 assess the correct *p* value threshold for the admixed individuals in this study. We measured the minimum *p*
538 values of associations (p_{min}) for each ancestry and derived an ancestry-specific empirical genome-wide
539 significance threshold as the fifth percentile ($\alpha = 0.05$) of p_{min} across permutations as previously described[129].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We calculated this percentile using the Harrell–Davis distribution-free quantile estimator[130] and calculated the 95% confidence interval via bootstrapping. Based on the permutation results (**Figure** S5a), we defined a study-wide significance threshold at a conservative level of $p = 1 \times 10^{-8}$ for both AFR- and EUR-specific associations. In addition, we calculated the genomic inflation factor, λ_{GC}, of null phenotypes across permutations and confirmed no significant inflation using the *Tractor* GWAS model (**Figure** S5b).

Acknowledgements

We thank Pradeep Natarajan, Sarah Gagliano Taliun, 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.; 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 Fundacao de Amparo a Pesquisa do Estado de Sao 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), the Japan Agency for Medical Research and Development (AMED).

Author Contributions

E.G.A. designed and implemented the pipeline, ran analyses, and drafted the primary manuscript. A.X.M. designed and ran analyses. M.K. designed and ran 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.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 statement

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

565 serves as a consultant for the Camp4 Therapeutics Corporation, Takeda Pharmaceutical and Biogen. The
566 remaining authors declare no competing interests.
567
568 **Code availability**
569 All code is freely available. *Tractor* scripts, in both cloud executable Hail format and python formats, can be
570 found at https://github.com/eatkinson/Tractor. The automated QC pipeline to prepare datasets for *Tractor* and
571 run LAI is located at https://github.com/eatkinson/Post-QC.
572

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

573 **References Cited**

574

575 1. Parker, K., Morin, R., Juliana Menasce Horowitz & Rohal, M. *Multiracial in America: Proud, Diverse and*
576 *Growing in Numbers*. (2015).
577 2. Bhardwaj, A. *et al.* Racial disparities in prostate cancer a molecular perspective. *Front. Biosci.* **22**, 4515
578 (2017).
579 3. Grizzle, W. E. *et al.* Self-Identified African Americans and prostate cancer risk: West African genetic
580 ancestry is associated with prostate cancer diagnosis and with higher Gleason sum on biopsy. *Cancer*
581 *Med.* **8**, 6915–6922 (2019).
582 4. Duggan, M. A., Anderson, W. F., Altekruse, S., Penberthy, L. & Sherman, M. E. The Surveillance,
583 Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical
584 Relationship. *Am. J. Surg. Pathol.* **40**, e94–e102 (2016).
585 5. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-
586 American men. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14068–14073 (2006).
587 6. Bateman, E. D. *et al.* Global strategy for asthma management and prevention: GINA executive
588 summary. *Eur. Respir. J.* **31**, 143–78 (2008).
589 7. Daya, M. & Barnes, K. C. African American ancestry contribution to asthma and atopic dermatitis. *Ann.*
590 *Allergy. Asthma Immunol.* **122**, 456–462 (2019).
591 8. Wyss, A. B. *et al.* Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for
592 pulmonary function. *Nat. Commun.* **9**, 2976 (2018).
593 9. Demenais, F. *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with
594 immune-cell enhancer marks. *Nat. Genet.* **50**, 42–50 (2018).
595 10. Benetos, A. & Aviv, A. Ancestry, Telomere Length, and Atherosclerosis Risk. *Circ. Cardiovasc. Genet.*
596 **10**, (2017).
597 11. Mozaffarian, D. *et al.* Heart Disease and Stroke Statistics—2015 Update. *Circulation* **131**, (2015).
598 12. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

599 26–31 (2019).
600 13. Popejoy, Alice B., Fullerton, S. M. Genomics is falling. *Nature* **538**, 161–164 (2016).
601 14. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and
602 mixed models. *PLOS Genet.* **14**, e1007309 (2018).
603 15. Huang, H. *et al.* Bootstrat: Population Informed Bootstrapping for Rare Variant Tests. *bioRxiv* 068999
604 (2016). doi:10.1101/068999
605 16. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in
606 genome-wide association studies. *Elife* **8**, (2019).
607 17. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK biobank. *Elife* **8**, (2019).
608 18. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science (80-.).* **265**, 2037–2048
609 (1994).
610 19. Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging Multi-ethnic Evidence for
611 Risk Assessment of Quantitative Traits in Minority Populations. *Am. J. Hum. Genet.* **101**, 218–226
612 (2017).
613 20. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic
614 underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).
615 21. Martin, E. R. *et al.* Properties of global- and local-ancestry adjustments in genetic association tests in
616 admixed populations. *Genet. Epidemiol.* **42**, 214–229 (2018).
617 22. Stevenson, A. *et al.* Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-
618 Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. *BMJ*
619 *Open* **9**, e025469 (2019).
620 23. Consortium, T. H. Enabling the genomic revolution in Africa. *Science* **344**, 1346–8 (2014).
621 24. TOPMed Whole Genome Sequencing Project. Freeze 5b, Phases 1 and 2. (2018).
622 doi:10.1155/2013/865181
623 25. Precision Medicine Initiative (PMI) Working Group. The precision medicine initiative cohort program –
624 building a research foundation for 21st century medicine. *Precis. Med. Initiat. Work. Gr. Rep. to Advis.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comm. to Dir. NIH **Sept 17**, 1–108 (2015).

26. 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**, 2287–97 (2015).

27. Bien, S. A. *et al.* The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE. (2019). doi:10.1146/annurev-genom-091416

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. *Heal. Psychol.* **35**, 407–411 (2016).

32. Agency for Healthcare Research & Quality. *2016 National Healthcare Quality and Disparities Report*. (2017).

33. Li, Y. R. & Keating, B. J. *Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations*. (2014). doi:10.1186/s13073-014-0091-5

34. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111-9 (2015).

35. Schaid, D. J., Chen, W. & Larson, N. B. *From genome-wide associations to candidate causal variants by statistical fine-mapping*. *Nature Reviews Genetics* **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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

651 38. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic
652 architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
653 39. Project, T. T. G. C. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135**,
654 (2012).
655 40. Skotte, L., Korneliussen, T. S. S., Moltke, I. & Albrechtsen, A. Ancestry specific association mapping in
656 admixed populations. *bioRxiv* 1–33 (2018). doi:10.1101/014001
657 41. Zhang, J. & Stram, D. O. The Role of Local Ancestry Adjustment in Association Studies Using Admixed
658 Populations. *Genet. Epidemiol.* **38**, 502–515 (2014).
659 42. Smith, E. N. *et al.* Genome-wide association study of bipolar disorder in European American and African
660 American individuals. *Mol. Psychiatry* **14**, 755–763 (2009).
661 43. Szulc, P., Bogdan, M., Frommlet, F. & Tang, H. Joint genotype- and ancestry-based genome-wide
662 association studies in admixed populations. *Genet. Epidemiol.* **41**, 555–566 (2017).
663 44. Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I. & London, S. J. Joint testing of genotype and
664 ancestry association in admixed families. *Genet. Epidemiol.* **34**, 783–791 (2010).
665 45. Coram, M. A. *et al.* Genome-wide Characterization of Shared and Distinct Genetic Components that
666 Influence Blood Lipid Levels in Ethnically Diverse Human Populations. *Am. J. Hum. Genet.* **92**, 904–916
667 (2013).
668 46. Aschard, H., Gusev, A., Brown, R. & Pasaniuc, B. Leveraging local ancestry to detect gene-gene
669 interactions in genome-wide data. *BMC Genet.* **16**, 1–9 (2015).
670 47. Zaitlen, N., Pas, B., Gur, T., Ziv, E. & Halperin, E. ARTICLE Leveraging Genetic Variability across
671 Populations for the Identification of Causal Variants. *Am. J. Hum. Genet.* **86**, 23–33
672 48. Pasaniuc, B. *et al.* Enhanced statistical tests for GWAS in admixed populations: assessment using
673 African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7**, e1001371 (2011).
674 49. Pasaniuc, B. *et al.* Analysis of Latino populations from GALA and MEC studies reveals genomic loci with
675 biased local ancestry estimation. *Bioinformatics* **29**, 1407–1415 (2013).
676 50. Chimusa, E. R. *et al.* Genome-wide association study of ancestry-specific TB risk in the South African

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coloured population. *Hum. Mol. Genet.* **23**, 796–809 (2014).

51. Shriner, D. Overview of admixture mapping. in *Current Protocols in Human Genetics* 1.23.1-1.23.8 (2013). doi:10.1002/cphg.44

52. Chen, M. *et al.* Admixture mapping analysis in the context of GWAS with GAW18 data. in *BMC Proceedings* **8**, S3 (BioMed Central Ltd., 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.* Strong Association of Socioeconomic Status and Genetic Ancestry in Latinos: Implications for Admixture Studies of Type 2 Diabetes. in *Racial Identities, Genetic Ancestry, and Health in South America: Argentina, Brazil, Colombia, and Uruguay* (eds. Gibbon, S., Santos, R. V. & Sans, M.) 137–153 (Palgrave Macmillan US, 2011). doi:10.1057/9781137001702_7

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

703 62. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach
704 for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–88 (2013).
705 63. Geza, E. *et al.* A comprehensive survey of models for dissecting local ancestry deconvolution in human
706 genome. *Brief. Bioinform.* **20**, 1709–1724 (2019).
707 64. Schubert, R., Andaleon, A. & Wheeler, H. E. *Comparing local ancestry inference models in populations*
708 *of two- and three-way admixture*. *Research Square* (2018).
709 65. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for
710 whole human genomes. *PLOS Genet.* **14**, e1007308 (2018).
711 66. Andrés, A. M. *et al.* Understanding the accuracy of statistical haplotype inference with sequence data of
712 known phase. *Genet. Epidemiol.* **31**, 659–71 (2007).
713 67. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness.
714 *PLoS Genet.* **10**, e1004234 (2014).
715 68. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
716 69. Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324
717 individuals. *Nat. Commun.* **9**, 3391 (2018).
718 70. Musunuru, K. & Kathiresan, S. Genetics of Common, Complex Coronary Artery Disease. *Cell* **177**, 132–
719 145 (2019).
720 71. Rotimi, C. N. *et al.* The genomic landscape of African populations in health and disease. *Hum. Mol.*
721 *Genet.* **26**, R225–R236 (2017).
722 72. Superko, H. R., Momary, K. M. & Li, Y. Statins Personalized. *Medical Clinics of North America* **96**, 123–
723 139 (2012).
724 73. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of
725 gene expression. *PLoS Genet.* **8**, (2012).
726 74. Avery, C. L. *et al.* A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated
727 with metabolic syndrome phenotype domains. *PLoS Genet.* **7**, (2011).
728 75. Lettre, G. *et al.* Genome-Wide association study of coronary heart disease and its risk factors in 8,090

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

729 african americans: The nhlbi CARe project. *PLoS Genet.* **7**, (2011).

730 76. Talmud, P. J. *et al.* Gene-centric Association Signals for Lipids and Apolipoproteins Identified via the
731 HumanCVD BeadChip. *Am. J. Hum. Genet.* **85**, 628–642 (2009).

732 77. Sandhu, M. S. *et al.* LDL-cholesterol concentrations: a genome-wide association study. *Lancet* **371**,
733 483–491 (2008).

734 78. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects
735 variants that double the explained heritability. *PLoS Genet.* **7**, (2011).

736 79. Fox, C. S. *et al.* Genome-wide association to body mass index and waist circumference: the
737 Framingham Heart Study 100K project. *BMC Med. Genet.* **8**, S18 (2007).

738 80. Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the Framingham
739 Heart Study. *BMC Med. Genet.* **8**, S17 (2007).

740 81. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to
741 complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

742 82. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of
743 loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019).
744 doi:10.1101/531210

745 83. Lin, M. *et al.* Population specific reference panels are crucial for the genetic analyses of Native
746 Hawai'ians: an example of the CREBRF locus. *bioRxiv* 789073 (2019). doi:10.1101/789073

747 84. Ntzani, E. E., Liberopoulos, G., Manolio, T. A. & Ioannidis, J. P. A. Consistency of genome-wide
748 associations across major ancestral groups. *Hum. Genet.* **131**, 1057–1071 (2012).

749 85. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal
750 variants. *PLoS Genet.* **9**, e1003566 (2013).

751 86. Waters, K., Stram, D., … M. H.-Pl. & 2010, undefined. Consistent association of type 2 diabetes risk
752 variants found in europeans in diverse racial and ethnic groups. *ncbi.nlm.nih.govPaperpile*

753 87. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European
754 populations. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0512-x

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

88. Liu, J. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *nature.comPaperpile*

89. 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**, (2013).

90. Easton, D., Pooley, K., Dunning, A., Nature, P. P.- & 2007, undefined. Genome-wide association study identifies novel breast cancer susceptibility loci. *nature.comPaperpile*

91. 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).

92. Wegmann, D. *et al.* Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.* **43**, 847–853 (2011).

93. Zhang, R. The ANGPTL3-4-8 model, a molecular mechanism for triglyceride trafficking. *Open Biol.* **6**, 150272 (2016).

94. 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).

95. Zhang, R. Lipasin, a novel nutritionally-regulated liver-enriched factor that regulates serum triglyceride levels. *Biochem. Biophys. Res. Commun.* **424**, 786–792 (2012).

96. Siddiqa, A. *et al.* Visualizing the regulatory role of Angiopoietin-like protein 8 (ANGPTL8) in glucose and lipid metabolic pathways. *Genomics* **109**, 408–418 (2017).

97. Yamada, H. *et al.* Circulating betatrophin is elevated in patients with type 1 and type 2 diabetes. *Endocr. J.* **62**, 417–421 (2015).

98. Espes, D., Martinell, M. & Carlsson, P.-O. Increased circulating betatrophin concentrations in patients with type 2 diabetes. *Int. J. Endocrinol.* **2014**, 323407 (2014).

99. Hu, H. *et al.* Increased circulating levels of betatrophin in newly diagnosed type 2 diabetic patients. *Diabetes Care* **37**, 2718–22 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

781 100. Fu, Z. *et al.* Elevated circulating lipasin/betatrophin in human type 2 diabetes and obesity. *Sci. Rep.* **4**,
782 5013 (2015).
783 101. Cannon, M. E. *et al.* Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at
784 the ANGPTL8 HDL-C GWAS Locus. (2017). doi:10.1534/g3.117.300088
785 102. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations
786 using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).
787 103. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids*
788 *Res.* **31**, 3812–4 (2003).
789 104. Atkinson, E. G. *et al.* No Evidence for Recent Selection at FOXP2 among Diverse Human Populations.
790 *Cell* **174**, 1424-1435.e15 (2018).
791 105. Deng, L., Ruiz-Linares, A., Xu, S. & Wang, S. Ancestry variation and footprints of natural selection along
792 the genome in Latin American populations. *Sci. Rep.* **6**, 1–7 (2016).
793 106. Jin, W. *et al.* Genome-wide detection of natural selection in African Americans pre- and post-admixture.
794 *Genome Res.* **22**, 519–527 (2012).
795 107. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common
796 disease. *Genome Biol.* **18**, 77 (2017).
797 108. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured
798 populations. **44**, 243–246 (2012).
799 109. The Hail team. Hail. (2018). Available at: https://github.com/hail-is/hail. (Accessed: 16th January 2019)
800 110. Kluyver, T. *et al. Jupyter Notebooks—a publishing format for reproducible computational workflows*.
801 *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016). doi:10.3233/978-
802 1-61499-649-1-87
803 111. International Hapmap Consortium, T. The International HapMap Project. *Nature* **426**, 789–796 (2003).
804 112. Tishkoff, S. a *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**,
805 1035–44 (2009).
806 113. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acad. Sci. U. S. A. **108**, 11983–8 (2011).

114. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. (2017). doi:10.1016/j.ajhg.2017.03.004

115. Chen, C. Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).

116. Williams, A. admix-simu: program to simulate admixture between multiple populations. (2016). doi:10.5281/ZENODO.45517

117. Skotte, L., Jørsboe, E., Korneliussen, T. S., Moltke, I. & Albrechtsen, A. Ancestry-specific association mapping in admixed populations. *Genet. Epidemiol.* **43**, 506–521 (2019).

118. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–2 (2002).

119. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–64 (2009).

120. Google Cloud Platform Blog. Google Compute Engine launches, expanding Google's cloud offerings. Available at: https://cloudplatform.googleblog.com/2012/06/google-compute-engine-launches.html. (Accessed: 16th January 2019)

121. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–75 (2007).

122. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

123. Bokeh Development Team. Bokeh: Python library for interactive visualization. (2019). Available at: https://bokeh.org/citation/. (Accessed: 31st March 2020)

124. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, (2015).

125. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

126. Benner, C., Havulinna, A., Salomaa, V., Ripatti, S. & Pirinen, M. Refining fine-mapping: effect sizes and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regional heritability. *bioRxiv* 318618 (2018). doi:10.1101/318618

127. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv* 501114 (2019). doi:10.1101/501114

128. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 1–11 (2019).

129. 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).

130. HARRELL, F. E. & DAVIS, C. E. A new distribution-free quantile estimator. *Biometrika* **69**, 635–640 (1982).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

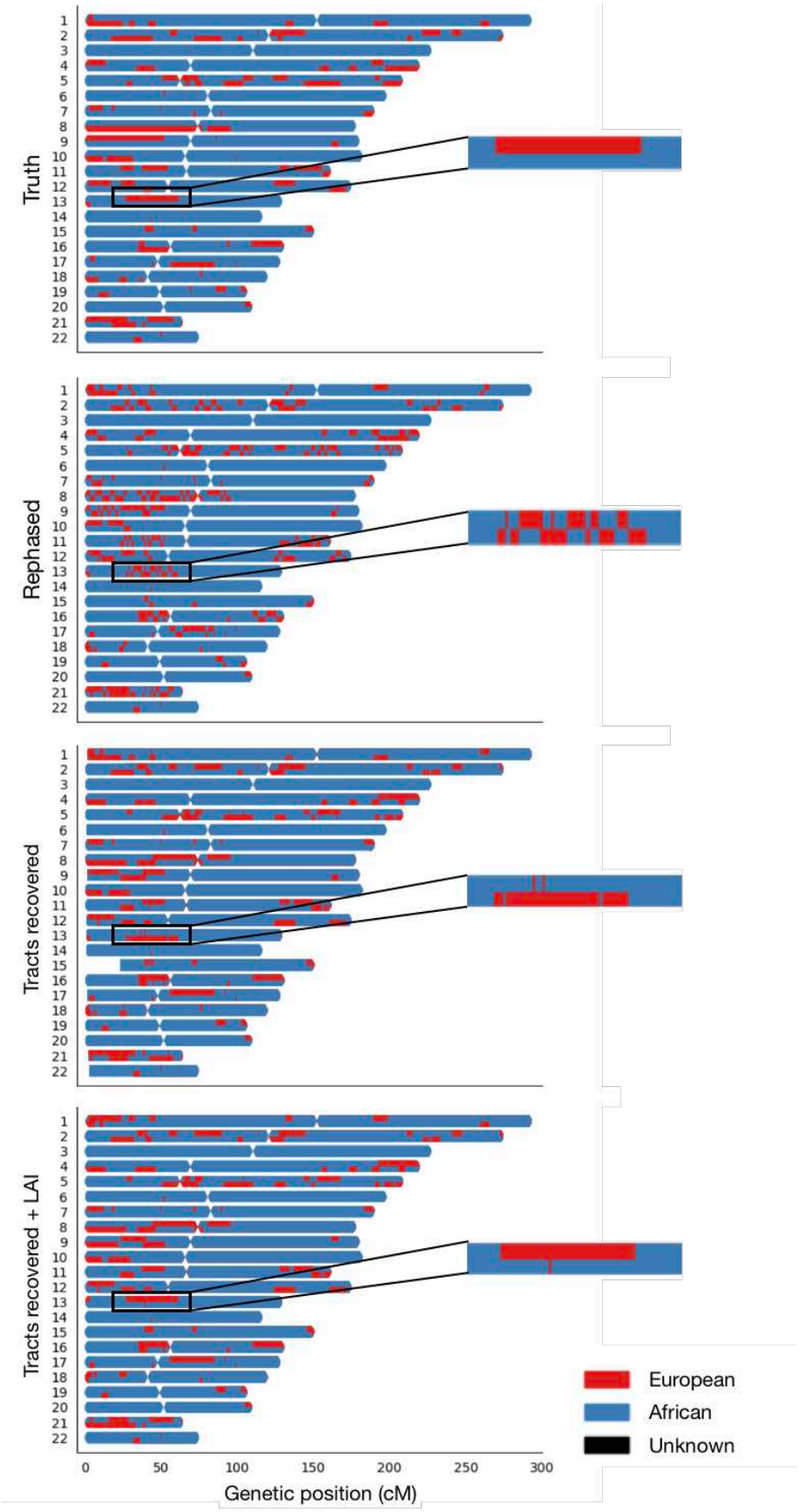
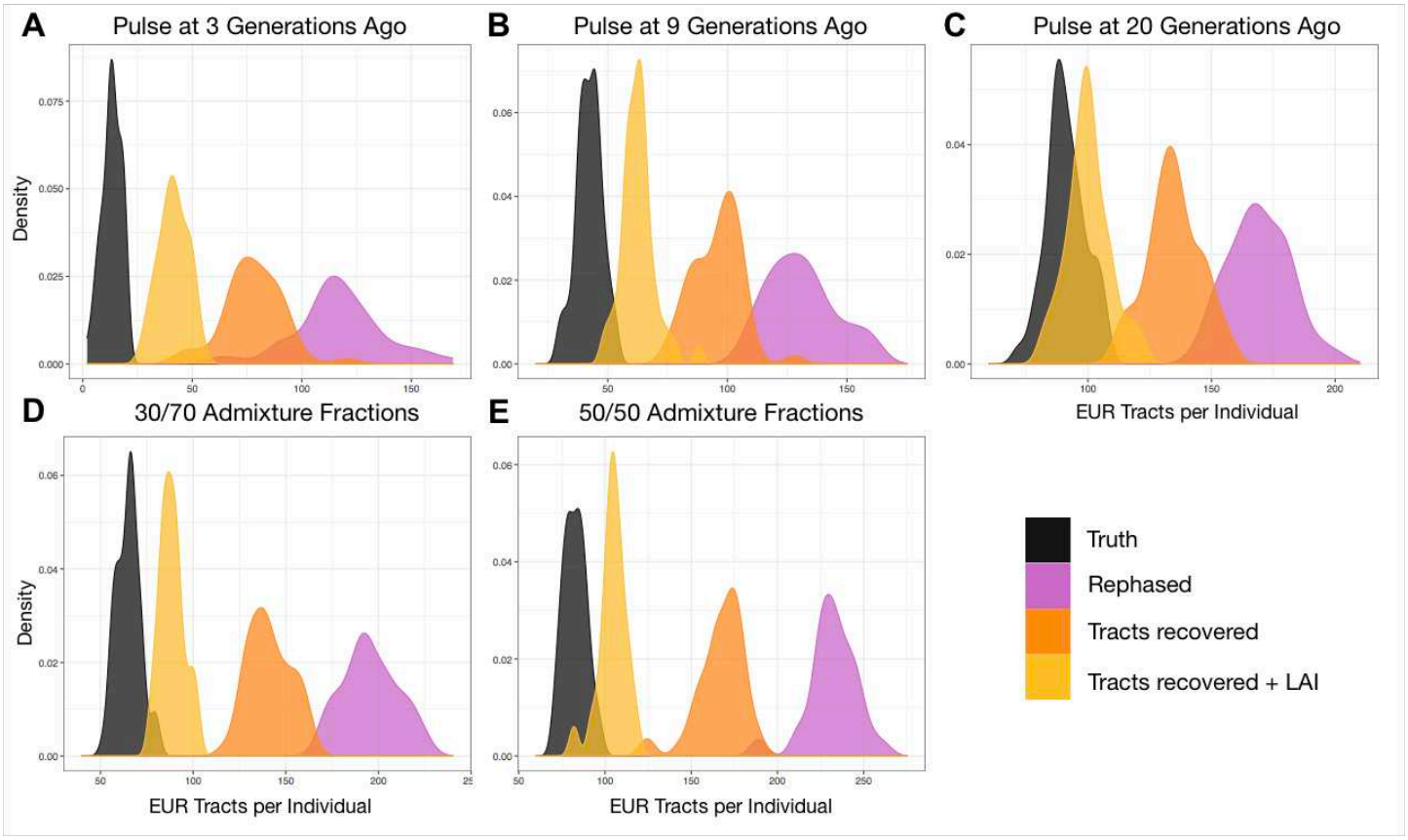


Figure 1. Painted karyograms of a simulated AA individual individual showing EUR (red) and AFR (blue) ancestral tracts across data treatments. The top panel shows the truth results for an example individual in our simulated AA cohort. A painted karyogram after statistical phasing is shown in the second row – note the disruption of long haplotypes. The third panel illustrates our recovery of tracts broken by switch errors in phasing. The bottom panel shows the smoothing and further improvement of tracts acquired through an additional round of LAI. The same section of chr13 showing an example tract at higher resolution is pictured on the right to highlight tract recovery.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 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 the realistic 9 generations since admixture. These can be compared to the inferred demographic model in AA with ~80% AFR ancestry 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.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. GWAS power gains across sample sizes, ancestral MAF differences, admixture proportions, and effect size differences. In all scenarios shown, dashed lines correspond to the power from the *Tractor* model incorporating local ancestry, 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 AA individuals: 80% AFR ancestry, an effect present only in the AFR genetic background, 12k cases and 30k controls, and 20% MAF. (**A**) There are similar gains in GWAS power when using the *Tractor* LAI-aware model across samples sizes of 4,000 (grey) and 12,000 (black) cases with 2x controls. (**B**) When there is a MAF difference between ancestries, the gains in power are even more pronounced. Gains vary across the allele frequency spectrum: black=MAF 10% AFR, 30% EUR; grey=MAF 20% AFR, 40% EUR. (**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 EUR background.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

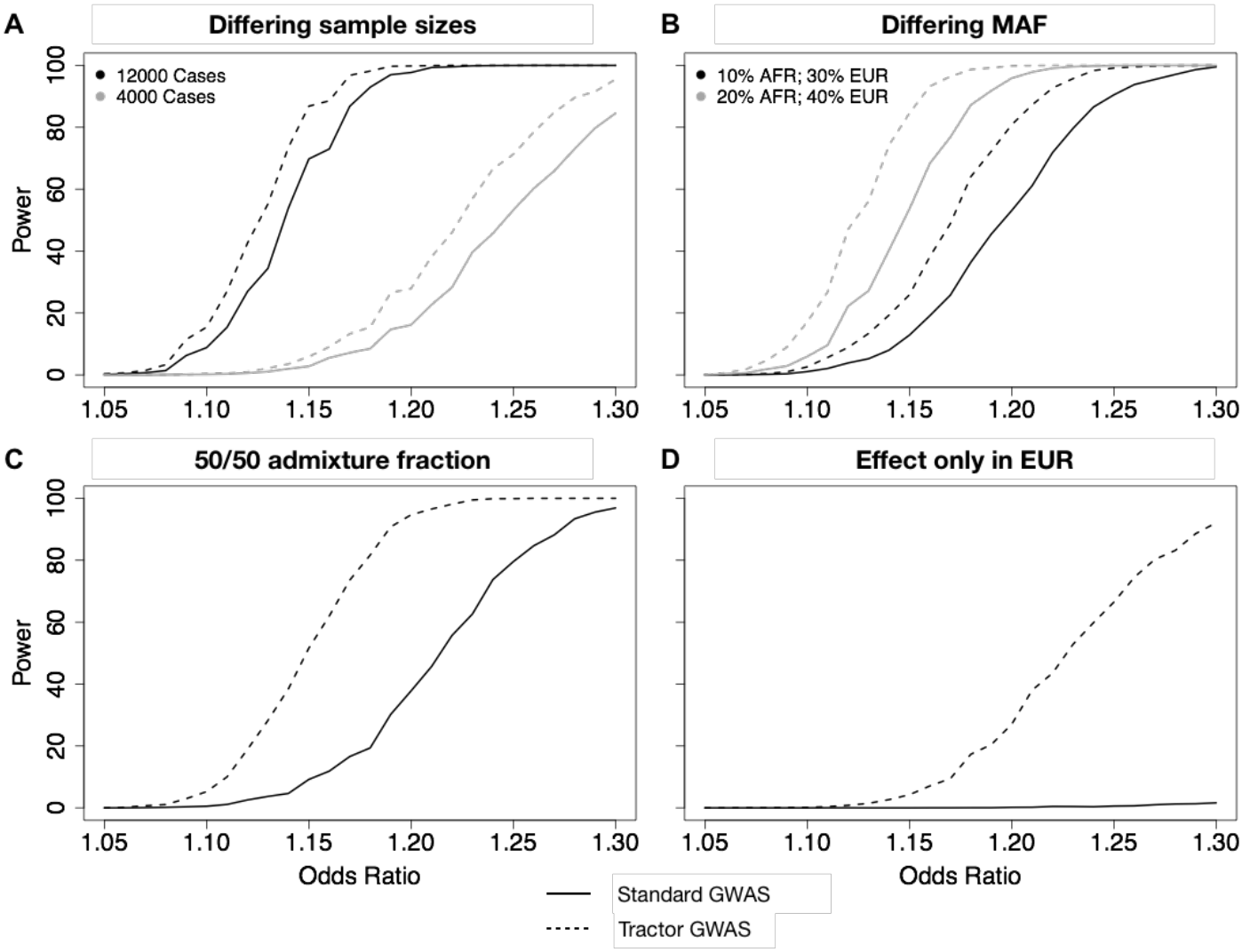
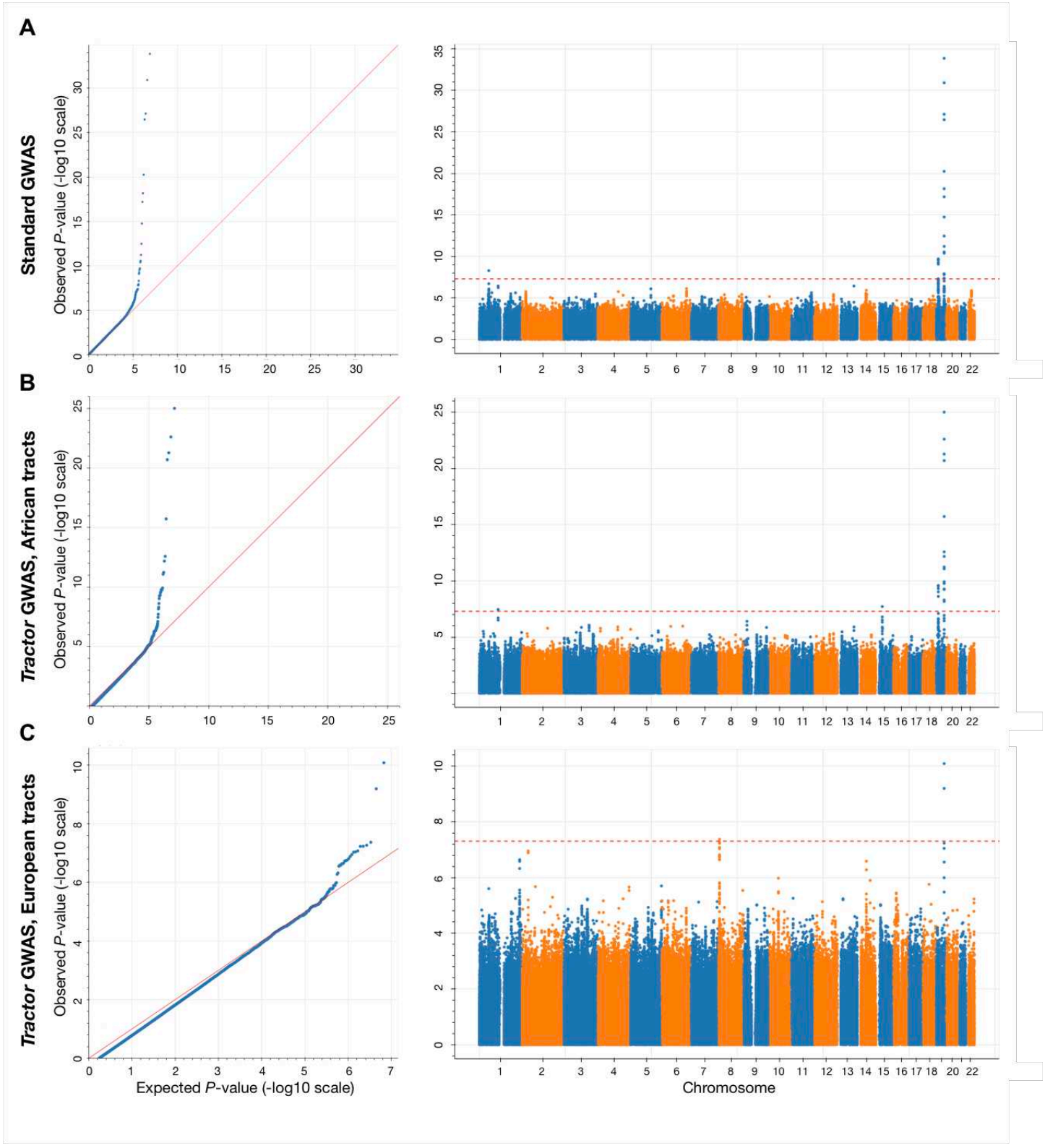


Figure 4. *Tractor* GWAS replicates established hits for Total Cholesterol in admixed African-European individuals and identifies new ancestry-specific loci. QQ and Manhattan plots for Total Cholesterol using the standard GWAS model (**A**) compared to *Tractor* joint-analysis results for the AFR (**B**) and EUR (**C**) backgrounds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.17.100727; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

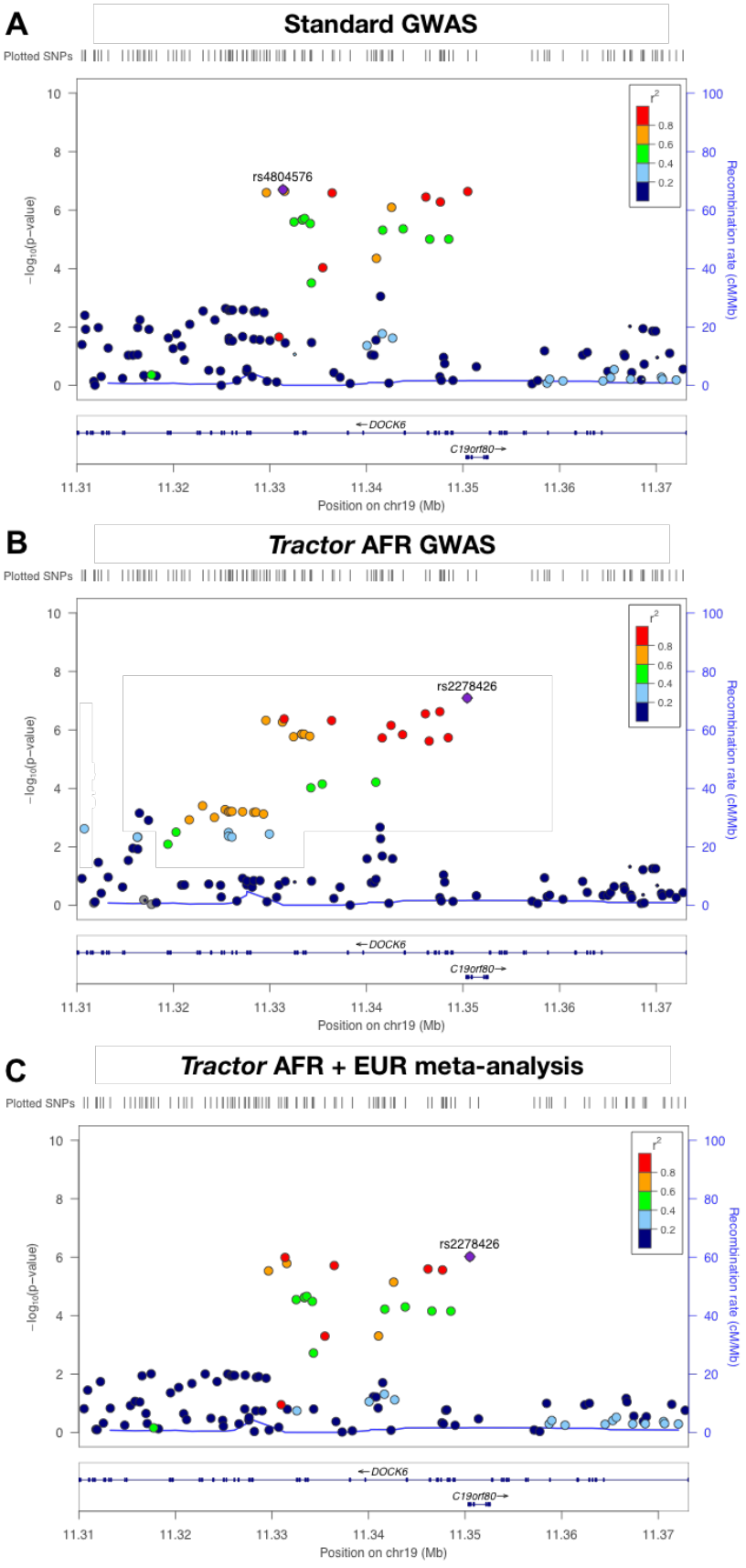


Figure 5. *Tractor* better localizes a top hit for Total Cholesterol. Previous hits for TC had pinpointed *DOCK6* as the gene of interest. Comparing runs on UKBB admixed individuals with a standard GWAS model (**A**), AFR-specific GWAS with *Tractor* (**B**), and a meta-analysis of GWAS runs on deconvolved EUR and AFR tracts (**C**), both ancestry-specific runs pinpoint a lead SNP ~20kb downstream in an intron of *DOCK6* spanning a better candidate gene *ANGPTL8* (also known as C19orf80) as the lead SNP. No significant signal was seen in the EUR segments. In all plots, point size is proportional to the number of samples included for that test, and color indicates r^2 to the named lead SNP. For **B**, the recombination rate line was generated from the AFR superpopulation of the 1000 Genomes Project, for other panels the EUR superpopulation rate is shown.