Supplementary Information

Automated Genotype QC pipeline

We have built an additional pipeline that can be used to filter data as a precursor to running *Tractor* (https://github.com/eatkinson/Post-QC). To run this optional precursor step to clean post-QC genotype data before LAI, the user inputs paths to their binary plink format sample genotypes file and a reference panel on the command line. The pipeline will perform post-genotyping QC, intersecting, phasing, and LAI with RFmixv2 to consistently prepare the data for downstream analysis. The Post-QC steps that are implemented are: (1) Extract autosomes; (2) Find and remove duplicate loci; (3) Update SNP IDs to dbsnp 144[1]; (4) Orient data to 1000 genome reference[2]; (5) Remove ambiguous A/T and G/C loci. This involves three substeps at step 4: a. Find and remove indels; b. Find and remove loci not found in 1000 genomes or that have different alleles; c. Flip alleles that are on the wrong strand. This script outputs warnings if more than 2% of sites are incongruous between the input dataset and 1000G locations, as this can indicate genome build discrepancies. Post-QCed cohort data is then intersected and phased with a user-specified reference panel. The merged dataset is then filtered to include only SNPs present in both the cohort data and reference panel using a MAF filter of 0.5% and a genotype missingness cutoff of 90%. Shapeit2[3] is implemented for phasing. Implementation is described in more detail at https://github.com/eatkinson/Post-QC.

Recovery of long-range haplotypes disrupted by statistical phasing results

To replicate standard analytical procedures employed on cohort data, we statistically phased our truth dataset using SHAPEIT2[3] software with a balanced reference panel composed of EUR and AFR continental individuals from the 1000 Genomes Project[2]. We modeled the expected distributions of EUR tracts within AA individuals as a Poisson process with rate equal to the number of generations ago when the pulse of admixture occurred. The waiting time until a recombination event disrupts a tract is expected to follow this distribution, which allowed us to quantify the probability of obtaining the observed number of EUR tracts given the data treatment as compared to truth expectations (See *Methods*). Analyzing the less common ancestry 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, $p=5.0\times10^{-26}$. After correcting phase switch errors, the likelihood of the tract distribution improved ($p=2.7\times10^{-11}$, 96 tracts) to contain approximately

half the excess tracts without phase error correction. After implementing one additional round of LAI on the corrected genotype files, the number of excess tracts was even further reduced (p=0.009, 62 tracts). Thus, our procedure for correcting phase switch errors statistically significantly recovers long-range haplotypes and better approximates the true tract length distributions (**Figure** S2; **Table** S2).

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 decreases the significance of the difference between the observed versus true tract length distributions (**Figure** S1, S2, **Table** S2). We note that running an additional round of LAI after recovering haplotypes produced the most accurate tract length distributions (62 EUR tracts compared to the truth set of 42, versus 131 after phasing and 96 before the additional LAI iteration – see Table S2). 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.

Characterizing the landscape of Tractor power across additional genomic and disease contexts

To evaluate *Tractor* power gains, we ran similar sets of simulations as described in the main text, varying effect size differences across ancestries, MAF differences, admixture fractions, and disease prevalence (**Figures** 3, S3-S5). Additional discussion on these simulation results is presented here to provide further detail on those not discussed in the main text.

Varying absolute MAF: We next fixed all other parameters and modified the absolute MAF of the simulated risk allele, with the relative difference in MAF between ancestries remaining constant. We changed our MAF from 20% to 10% and 40% under both the models of an effect only in the AFR background and with matching effect sizes between EUR and AFR haplotype context. We note that different MAFs affect power as they would under a regular GWAS; i.e. power tends to increase with MAF (**Figure** S3).

MAF differences between groups: To see if having a difference in the MAF between the two ancestral groups affected GWAS power, we varied the MAF in the EUR background to be 10, 20, and 30% while keeping the AFR MAF set to 20%. Changing the MAF differences when the effect sizes are equal across

ancestries does not dramatically affect power (Figure S3). Given unequal effect sizes, however, when the effect is present only in the AFR, as the MAF decreases in EUR, the power gain also decreases (**Figure** S4). And vice versa, higher MAF in the non-effect group affords our method more power gains.

Varying admixture proportions: Admixture proportion affects power in our method because it determines the effective number of samples per ancestry group. Consider the scenario again where there is no effect in EUR but there is an effect in AFR. A populations with 80% AFR ancestry has a greater number of haplotypes in which there is an effect compared to a population with 50% AFR ancestry. As a consequence, the population with 50% AFR admixture has lower power because there are fewer AFR tracts in general, meaning 50% of the sample does not contribute signal as opposed to only 20% of the sample not contributing signal (**Figure** 3C).

A related observation of note is that under the assumption of AFR only allelic effects, the simulated power of *Tractor* tended to be quite close to power predicted at *N**average African ancestry proportion using Quanto[4]. This implies that the power of *Tractor* under these conditions is effectively the power of an analysis stratified to just AFR ancestry individuals; i.e. the effective *N* given the AFR ancestral tracts.

Varying the simulated disease prevalence: A comparison of two simulated diseases with prevalences of 10% and 20% shows that the relative differences in power between populations remains similar. Because this was a case/control sampling design with no misclassification of disease status, the prevalence of the disease did not notably impact power. In other words, prevalence affects power in our model as it would under a traditional GWAS framework. A theoretical situation in which admixture influenced misdiagnosis of a disease may affect power.

Validation of Tractor using additional phenotypes

Though we focus on TC in the main text to allow for expanded discussion of hits for this phenotype, we additionally validated *Tractor*'s performance in two other blood panel phenotypes in admixed UKB individuals, LDLC and HDLC. Both of these followed a similar trend as the main text example, and we will expand on results for LDLC here.

Tractor GWAS runs were again able to identify the major signals for LDLC in this cohort (Figure S11), including at canonical lipid genes *APOE, LDLR,* and *SORT1*. To investigate potential signal resolution improvement, we

investigated the hit for *PCSK9* on chr1. Proprotein convertase subtilisin-kexin 9 (*PCSK9*), produces an enzyme which regulates cholesterol through playing a key role in lipid metabolism. Specifically, it enhances the degradation of the intracellular LDL receptor, which mediates 70% of LDLC ingestion into cells, thereby leading to elevated blood LDLC levels[5,6]. The lead SNP identified by standard GWAS (rs11804420) lies in an intronic region of *PCSK9*. Conversely, the variant identified by *Tractor* AFR and meta-analysis runs (rs505151, NC_000001.10:g.[55529187G>A]) lies 8.74kb downstream of this variant in the exon 12 of *PCSK9* and results in an amino acid substitution from glutamate to glycine. We next attempted validation of rs505151 as being the putative causal variant through comprehensive fine-mapping of LDLC in two separate large cohorts: 345,235 white British individuals from UKBB and 135,808 Japanese individuals from Biobank Japan[7]. rs505151 was successfully fine-mapped to a 95% credible set in both populations.

There is substantial functional support for a putative causal role of the *Tractor* lead SNP in affecting blood LDLC levels. rs505151 is an established gain-of-function mutation in *PCSK9*, and has been shown to significantly raise LDLC levels in multiple human cohorts[8–11], leading to elevated risk for cardiovascular disease. This variant is at 3.83% frequency in the gnomAD[12] non-Finnish Europeans but at 26.43% frequency in the AFR. It therefore again appears that the substantially higher MAF in AFR as compared to EUR drove *Tractor*'s power to identify the putative causal SNP for LDLC. Of note, in both this example for LDLC and the TC example described in the main text, *Tractor* was able to identify the putative causal variant through initial GWAS procedures alone without the need to further fine-map (Figure S12).

References Cited

1. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–11 (2001).
2. Project, T. T. G. C. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135**, (2012).
3. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
4. WJ Gauderman & JM Morrison. Quanto.
5. Joseph, L. & Robinson, J. G. Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9) Inhibition and the

Future of Lipid Lowering Therapy. *Prog. Cardiovasc. Dis.* **58**, 19–31 (2015).

6. Weinreich, M. & Frishman, W. H. Antihyperlipidemic Therapies Targeting PCSK9. *Cardiol. Rev.* **22**, 140–146 (2014).
7. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
8. Peters, B. J. M. *et al.* Genetic variability within the cholesterol lowering pathway and the effectiveness of statins in reducing the risk of MI. *Atherosclerosis* **217**, 458–464 (2011).
9. Yin, R. X. *et al.* Interactions of several single nucleotide polymorphisms and high body mass index on serum lipid traits. *BioFactors* **39**, 315–325 (2013).
10. Qiu, C. *et al.* What is the impact of PCSK9 rs505151 and rs11591147 polymorphisms on serum lipids level and cardiovascular risk: A meta-analysis. *Lipids Health Dis.* **16**, 111 (2017).
11. Ben Djoudi Ouadda, A. *et al.* Ser-phosphorylation of PCSK9 (Proprotein Convertase Subtilisin-Kexin 9) by Fam20C (Family with Sequence Similarity 20, Member C) kinase enhances its ability to degrade the LDLR (Low-Density Lipoprotein Receptor). *Arterioscler. Thromb. Vasc. Biol.* **39**, 1996–2013 (2019).
12. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

Supplementary Tables

Supplementary Table 1: Local ancestry accuracy across demographic contexts. LAI accuracy, defined as the proportion of the time the ancestry of a variant was correctly called in a simulated truth dataset, was calculated for all sites across a range of demographic scenarios modifying the timing of the pulse of admixture or the overall fraction of ancestry contribution from the component ancestries. We also calculated ancestry-specific LAI accuracy under the realistic AA demographic scenario of 1 pulse of admixture 9 generations ago, with 84% contribution from Africa and 16% from Europe. Modified admixture fractions show the proportion of AFR and EUR ancestries, respectively.

Demographic model	*Accuracy*
Realistic, global	0.979
Realistic, AFR	0.978
Realistic, EUR	0.984
Pulse at 3 gens ago	0.984
Pulse at 20 gens ago	0.977
Admixture of 30/70	0.979
Admixture of 50/50	0.982

Supplementary Table 2. Tract distributions across data treatments and demographic models. Results shown are for the numbers of EUR tracts in simulated AA individuals. *P* values indicate the significance of obtaining the observed number of tracts given the truth expectations as modeled with a Poisson process calibrated for that demographic model. Results are shown for each step in the *Tractor* pipeline.

	Truth	***Rephased***		***Tracts Recovered***		***Tracts Recovered + LAI***	
Demographic model	*# Tracts*	*p*	*# Tracts*	*p*	*# Tracts*	*p*	*# Tracts*
Pulse at 3 gens ago	13.16	2.931E-67	117.74	8.188E-34	78.34	1.252E-09	41.40
Pulse at 5 gens ago	21.76	9.547E-49	121.36	3.137E-24	85.52	2.286E-06	48.52
Pulse at 9 gens ago	42.22	4.997E-26	130.94	2.752E-11	95.84	9.039E-03	62.62
Pulse at 15 gens ago	69.92	2.369E-15	149.80	2.543E-06	116.72	8.106E-02	84.46
Pulse at 20 gens ago	91.06	0.000E+00	169.16	4.026E-05	134.46	3.121E-02	100.04
Admixture of 30/70	65.10	0.000E+00	195.78	6.034E-16	141.34	2.291E-02	88.72
Admixture of 50/50	82.50	0.000E+00	231.46	0.000E+00	167.04	2.242E-02	104.46

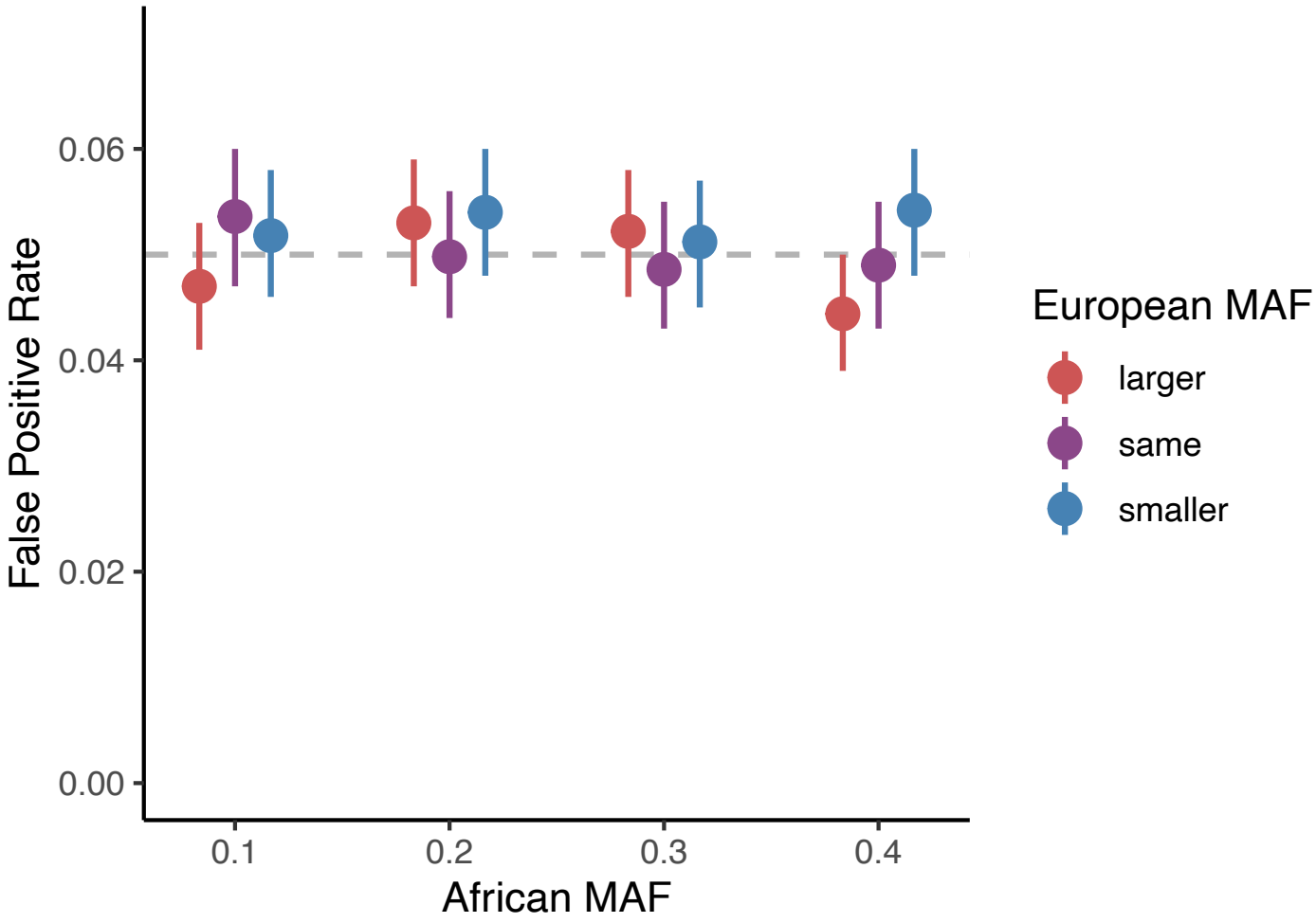
Supplementary Table 3. Genomic inflation factor values for empirical GWAS runs under different models.

The λ_{GC} for standard GWAS as compared to *Tractor* 2-way joint analysis and meta-analysis implementations are presented for three blood lipid phenotypes, High-density lipoprotein (HDLC) and Total Cholesterol (TC).

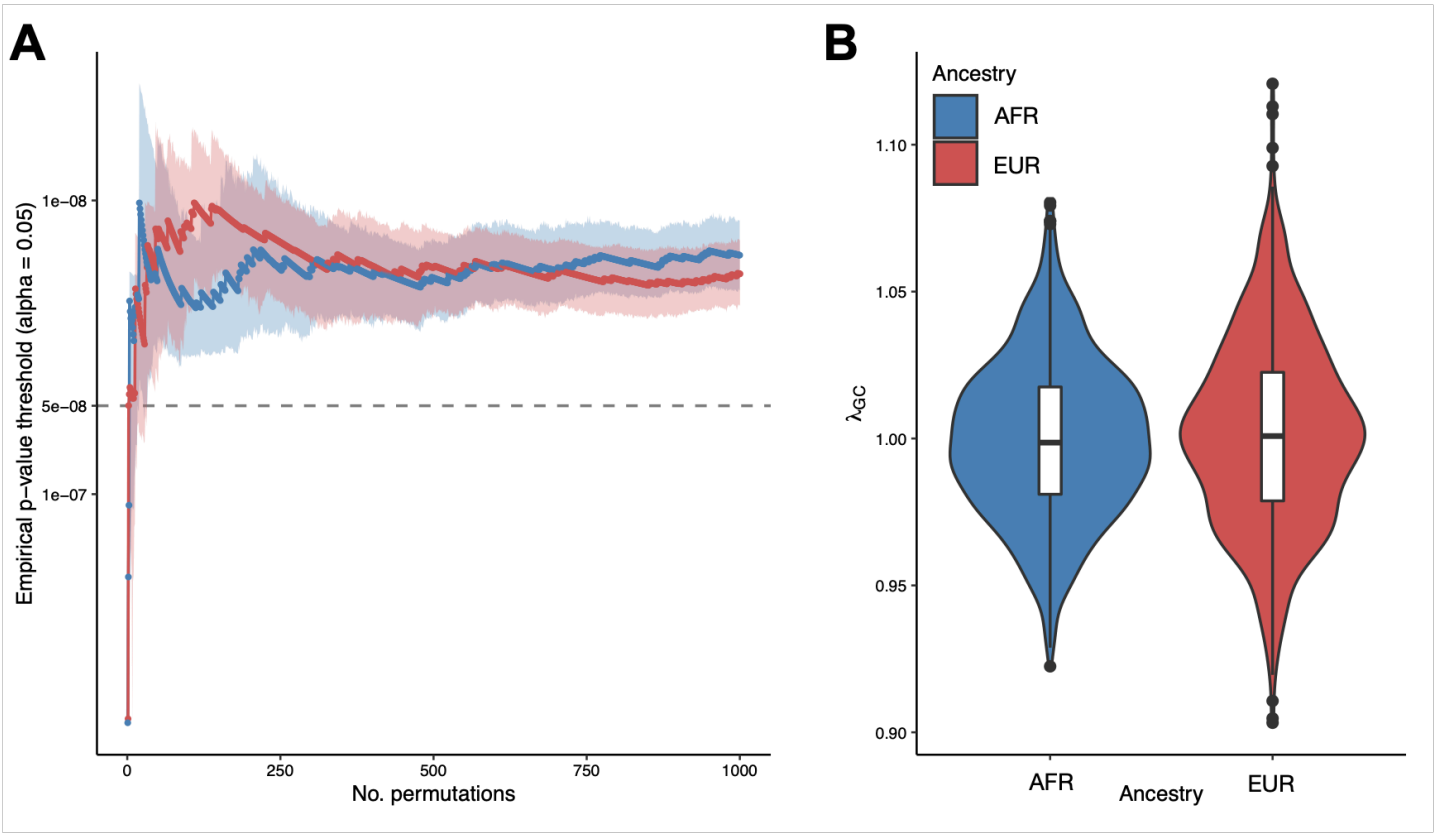
	HDLC	*TC*
Standard GWAS	1.029233	1.024476
Tractor AFR	1.024476	1.005142
Tractor EUR	1.007485	1.006547
Tractor meta	1.02971	1.009363

Supplementary Figures

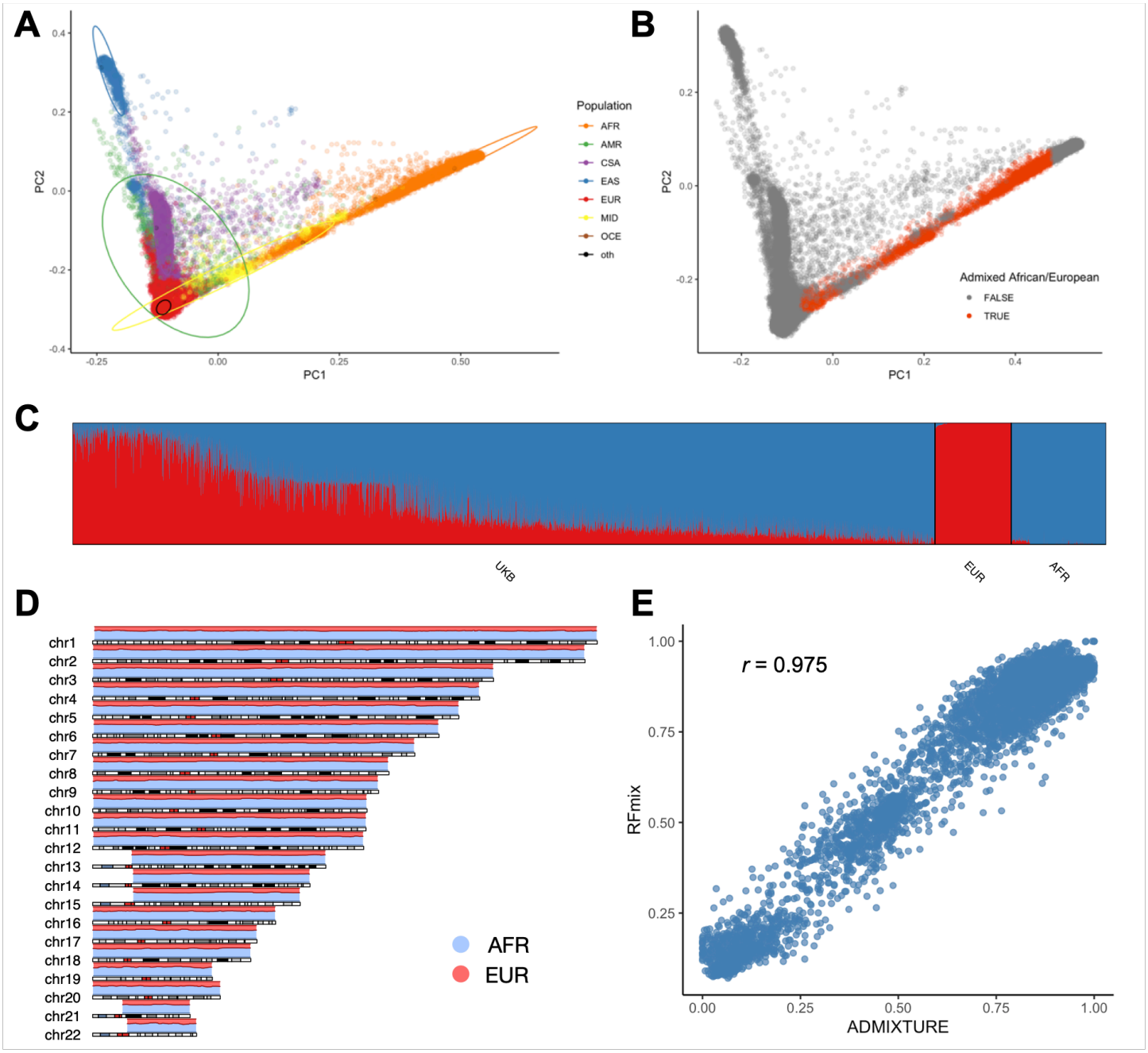
Supplementary Figure 1. **Type I error rate of *Tractor* across differing absolute and relative MAFs**. False positive rates (i.e. % of simulations for which p < 0.05) for *Tractor* and basic GWAS for 5000 replicates run under a realistic AA demographic model of 80/20 AFR/EUR admixture at a 10% disease prevalence with 12k cases and 30k controls. Error bars show the 95% confidence intervals; standard error was calculated as the expected SE of a binomial distribution, p*(1-p)/(n-1). The 5% frequency expectation is indicated with the dashed grey line.



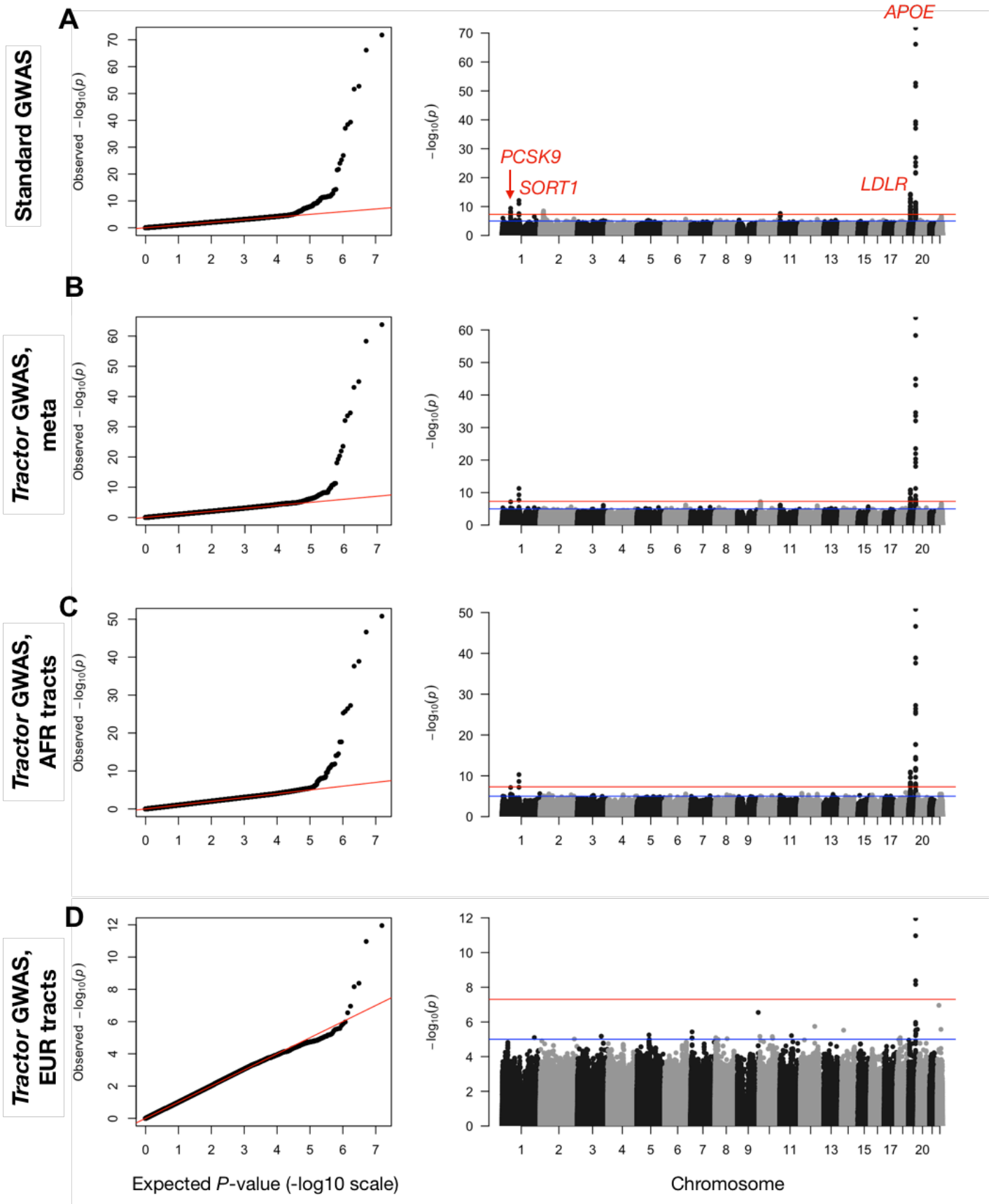
Supplementary Figure 2. **Empirical assessment of the ideal genome-wide *p*-value threshold for ancestral segments**. (A) Each line represents an estimated empirical *p*-value threshold as we increase the number of permutations, shading represents bootstrapping 95% CI of the estimates. (B) For each ancestry, the distribution of λ_{GC} values across 1,000 permutations is shown with a boxplot where the center is the median, the bounds of box represent the first quantile to third quantile, and whiskers are 1.5 * IQR.



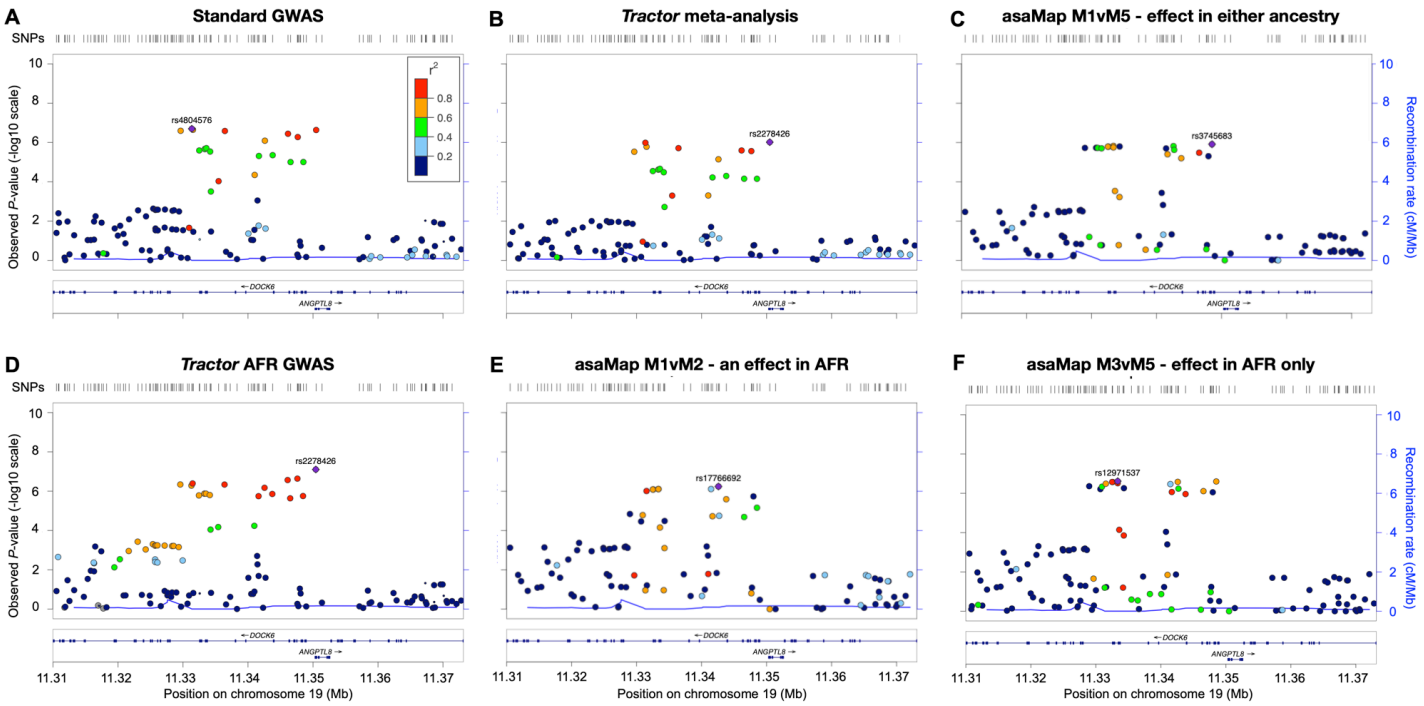
Supplementary Figure 3: Selection of admixed African-European UKBB individuals. (**A**) Ancestry assignment by a random forest model trained on 1000G and HGDP reference data. (**B**) Individuals selected for inclusion here are highlighted in red. (**C**) Admixture fractions at *k*=2 across UKBB individuals included in analyses. The 1000 Genomes EUR and AFR superpopulation individuals were used as a reference for ADMIXTURE. (**D**) Local ancestry distribution across the genome. Some edges of chromosomes were trimmed due to uncertainty in calls. (**E**) Correlation in AFR admixture fraction estimates obtained from ADMIXTURE as compared to RFmix. In panels C-E, blue is AFR ancestry, red is EUR ancestry.



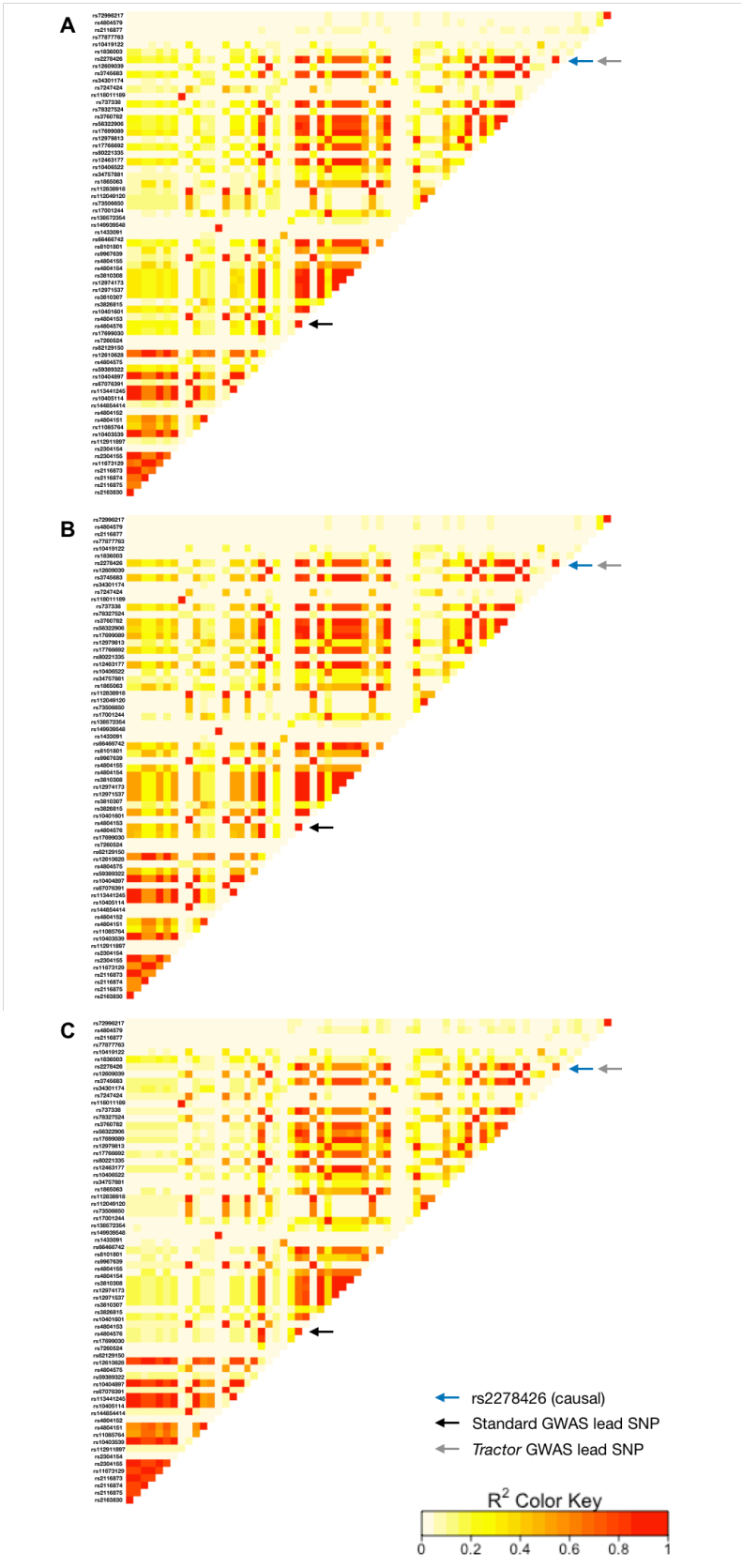
Supplementary Figure 4. *Tractor* GWAS replicates established hits for Low-Density Lipoprotein Cholesterol in admixed African-European individuals. QQ and Manhattan plots using (**A**) the standard GWAS model, (**B**)*Tractor* deconvolved segment meta-analysis, and *Tractor* joint-analysis results for the (**C**) AFR and (**D**) EUR backgrounds. Peaks are annotated in **A**, with the *PCSK9* peak examined for signal localization highlighted with an arrow. Traditional suggestive (1e-05) and stringent (5e-08) genome-wide significance thresholds are shown as the blue and red lines, respectively.


A
Standard GWAS
B
Tractor GWAS, meta
C
Tractor GWAS, AFR tracts
D
Tractor GWAS, EUR tracts
Observed −log10(p)
−log10(p)
APOE
PCSK9
SORT1
LDLR
Expected P-value (-log10 scale)
Chromosome

Supplementary Figure 5. Localization in *DOCK6* comparing *Tractor* performance to asaMap. The lead SNP for each model is indicated with a purple diamond. The putative causal SNP, rs2278426, was only successfully identified by *Tractor* analyses. 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. The recombination rate is shown as a blue line generated from the EUR superpopulation of the 1000 Genomes Project in A-C or the AFR superpopulation for panels D-F.



Supplementary Figure 6. Within-sample LD structure in the *DOCK6* region of interest. Pairwise LD calculated from **(A)** complete dataset of admixed UKBB individuals**, (B)** AFR tracts, and **(C)** EUR tracts. In all panels, the putative causal SNP, the lead SNP identified by standard GWAS, and the *Tractor* GWAS lead SNP are indicated with a blue, black, and gray arrow, respectively.


A
B
C
rs2278426 (causal)
Standard GWAS lead SNP
Tractor GWAS lead SNP
R2 Color Key
0
0.2
0.4
0.6
0.8
1

Supplementary Figure 7. *Tractor* better localizes a top hit for LDLC in the canonical gene *PCSK9*. Comparing runs on UKBB admixed individuals with (**A**) a standard GWAS model, (**B**) a meta-analysis of GWAS runs on deconvolved EUR and AFR tracts, (**C**) AFR-specific GWAS with *Tractor*, and (**D**) EUR-specific GWAS with Tractor. *Tractor* pinpoints a protein-coding lead SNP ~8.7kb downstream of the intronic variant identified by standard GWAS. 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 **C**, the recombination rate line was generated from the AFR superpopulation of the 1000 Genomes Project, for other panels the EUR superpopulation rate is shown.