Polymorphic short tandem repeats make widespread contributions to blood and serum traits

Jonathan Margoliash[1], Shai Fuchs[2], Yang Li[1,3], Arya Massarat[4], Alon Goren[3,*], Melissa Gymrek[1,3,*]

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA

[2]Pediatric Endocrine and Diabetes Unit, Sheba Medical Center, Edmond and Lily Safra Children's Hospital, Tel-Hashomer, Israel.

[3]Department of Medicine, University of California San Diego, La Jolla, CA

[4]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA

*Correspondence should be addressed to agoren@ucsd.edu and mgymrek@ucsd.edu.

Abstract

Short tandem repeats (STRs), genomic regions each consisting of a sequence of 1-6 base pairs repeated in succession, represent one of the largest sources of human genetic variation. However, many STR effects are not captured well by standard genome-wide association studies (GWAS) or downstream analyses that are mostly based on single nucleotide polymorphisms (SNPs). To study the involvement of STRs in complex traits, we imputed genotypes for 445,735 autosomal STRs into SNP data from 408,153 White British UK Biobank participants and tested for association with 44 blood and serum biomarker phenotypes. We used two fine-mapping methods, SuSiE and FINEMAP, to identify 118 high-confidence STR-trait associations predicted as causal variants under all fine-mapping settings tested. Using these results, we estimate that STRs drive 5.2-9.7% of GWAS signals for these traits. Our high confidence STR-trait associations implicate STRs in some of the strongest hits for multiple phenotypes, including a trinucleotide STR in *APOB* associated with LDL cholesterol and a *CGG* repeat in the promoter of *CBL* associated with multiple platelet traits. Replication analyses in additional population groups and orthogonal expression data further support the role of a subset of the candidate STRs we identify. Together, our study suggests that polymorphic tandem repeats make widespread contributions to complex traits, provides a set of stringently selected candidate causal STRs, and demonstrates the need to routinely consider a more complete view of human genetic variation in GWAS.

Introduction

Genome-wide association studies (GWAS) have become an indispensable tool for identifying which genes and non-coding regions in the genome influence complex human traits. While GWAS routinely identify tens to hundreds of genomic regions associated with individual traits, biological interpretation of GWAS results remains challenging[1]. Further, variants identified by GWAS still only explain modest amounts of trait variability for most phenotypes[2].

A major challenge is that typical GWAS pipelines only consider a subset of common genetic variants. The majority of GWAS have been based on common single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels) either genotyped using microarrays or imputed from population reference databases based on whole genome sequencing (WGS) data. However, detailed follow-up of individual GWAS signals has often revealed complex variants that were absent from the original analysis, such as repeats[3–5] or structural variants[6–8], to be the causal drivers of those signals. Indeed, a recent study showed that polymorphic protein-coding variable number tandem repeats (VNTRs) are likely causal drivers of some of the strongest GWAS signals identified to date for multiple traits[3].

Short tandem repeats (STRs) are a type of complex variant that consist of repeat units between 1-6bp duplicated many times in succession. Hundreds of thousands of STRs occur in the human genome[9], each spanning from tens to thousands of base pairs. STRs undergo frequent mutations resulting in gain or loss of repeat units[10], with per-locus mutation rates several orders of magnitude higher than average rates for SNPs[11] or indels[12]. Large repeat expansions at STRs are known to result in Mendelian diseases such as Huntington's disease, muscular dystrophies, hereditary ataxias and intellectual disorders[10,13]. Further, recent evidence suggests that more modest but highly prevalent variation at multiallelic non-coding STRs can also be functionally relevant. We and others have found associations between STR length and both gene expression[4,14] and splicing[15,16]. The impact of STRs on gene expression is hypothesized to be mediated by a variety of mechanisms including modulation of nucleosome positioning[17], altered methylation[14], affecting transcription factor recruitment[4] and impacting the formation of non-canonical DNA and RNA secondary structures[18,19]. Together, these suggest that STRs potentially play an important role in shaping complex traits in humans.

Despite this potential, STRs are not well-captured by current GWAS. Because STRs are not directly genotyped by microarrays and are challenging to analyze from WGS, STRs have been

largely excluded from widely used reference haplotype panels[20,21] and downstream GWAS analyses. While some STRs are in high linkage disequilibrium (LD) with nearby SNPs, many highly multiallelic STRs can only be imperfectly tagged by individual common SNPs, which are typically bi-allelic. Thus, underlying effects driven by variation in repeat length, especially at highly polymorphic STRs, have likely not been fully captured.

Recent technological advances can now enable incorporation of STRs into GWAS. We and others have created a variety of bioinformatic tools to genotype STRs directly from WGS by statistically accounting for the noise inherent in STR sequencing[22–27]. We recently leveraged these tools to develop a reference haplotype panel consisting of both SNP and STR genotypes that allows for imputation of STRs from SNP genotype data[28] in samples for which WGS is unavailable. We found that all but the most highly polymorphic STRs are amenable to imputation in European cohorts, with an average per-locus imputation concordance of 97% with genotypes obtained from WGS.

Here, we leverage our SNP-STR reference haplotype panel to impute genome-wide STRs into SNP array data from 408,153 White British individuals obtained from the UK Biobank (UKB) for which deep phenotype information is available[29]. Whereas a recent publication studied the effects of protein-coding VNTRs (118 total VNTRs with repeat units of 7+ base pairs; total length of up to several kilobases) on complex traits[3], our study focuses on a distinctset of repeats (namely 445,735 STRs with repeat units of 1bp to 6bp) which are mostly non-coding. We test for association between imputed STR lengths and 19 blood cell count and 25 biomarker traits. These traits provide multiple advantages: they are broadly and reliably measured, continuous, highly polygenic and have variants with relatively large effect sizes, thus enabling well-powered association testing.

We performed fine-mapping on these associations and estimate that STRs account for 5-10% of signals identified by GWAS for the traits we studied. We observed that fine-mapping results are more sensitive to data-filtering thresholds and meta-parameter choices than commonly acknowledged and thus require careful interpretation. After restricting to the signals which were consistently fine-mapped across settings, we identified 95 unique STRs strongly predicted to be causal for at least one trait. We highlight multiple STRs in this set which we predict contribute to some of the strongest hits for LDL cholesterol, platelet count, and other traits. Overall, our study demonstrates the widespread role of polymorphic tandem repeats and the need to consider a broad range of variant types in GWAS and downstream analyses such as fine-mapping.

Results

Performing genome-wide STR association studies in 44 traits

We imputed genotypes for 445,735 autosomal STRs into phased SNP array data from 408,153 White British individuals from the UKB using Beagle[30] in combination with our published SNP+STR reference haplotype panel[28] (**Methods**; **Fig. 1a**; **Supplementary Fig. 1**; **URLs**). Compared to common SNPs which are typically bi-allelic, many of the STRs imputed from our panel are highly multiallelic (**Fig. 1b**). We tested STRs for association with 44 quantitative blood cell count and biomarker traits (**Supplementary Table 1**) for which phenotype information was available for between 304,658-335,585 genetically unrelated subsets of individuals. To facilitate this and other STR association studies, we developed associaTR (**URLs**), an open-source custom software pipeline capable of testing for association between STR length and phenotype.

For each STR-trait pair, we used associaTR to test for a linear association between STR dosage (the sum of the imputed allele length dosages of both chromosomes) and the measured trait value (**Fig. 1c-d**). For comparison, we used plink[31] to perform similar association tests using 70,698,786 SNP and short indel variants that were imputed into the same individuals. For all associations (STR and SNP and indel), we included as covariates SNP genotype principal components, sex, and age (**Methods**). Additional covariates were included on a per-trait basis (**Supplementary Table 1**). We compared the output of our SNP analysis pipeline to previous results reported by Pan-UKB[32] and found that our pipeline produced similar results (**Supplementary Fig. 2**).

We then compared signals identified by SNPs and indels to those identified by STRs. For each trait we defined peaks as 500kb intervals centered on the lead genome-wide significant variant (a SNP, indel or STR with p≤5e-8) in that interval (**Methods**). We identified an average of 389 peaks per trait, with blood cell count traits generally more polygenic than biomarkers (**Fig. 1e**). Of these peaks, 65.8% contained both a significant STR and a significant SNP or indel, 32.5% contained only significant SNPs or indels, and 1.7% contained only significant STRs. The majority of strong peaks were identified by both STRs and SNPs and/or indels. No new strong peaks were identified only by STRs (**Fig. 1f**), which is expected given that SNP and indel genotypes were used to impute the STRs. Overall, p-values of the lead SNP or indel and lead STR were similar for most peaks. Thus, we focused on fine-mapping to determine which variants might be causally driving the identified signals.

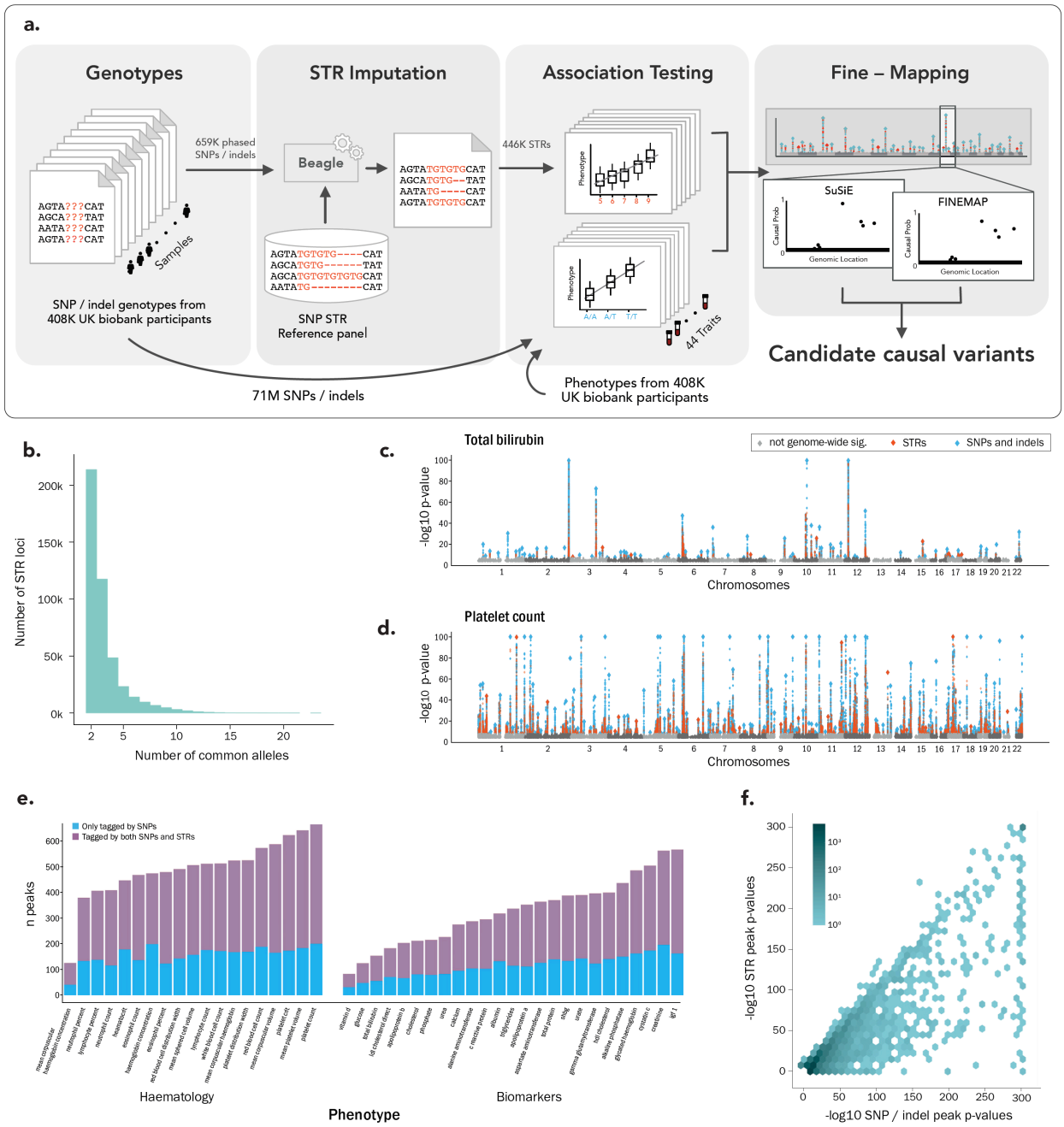


Figure 1: Genome-wide association tests identify STRs, SNPs and indels associated with blood and biomarker traits in the UKB. (a) Schematic overview of this study. STRs are imputed into phased hard-called SNPs. GWAS is performed on SNPs and STRs in parallel. Regions with significant signals are identified and then fine-mapped by two independent methods each under multiple scenarios, resulting in candidate causal STRs. **(b) Distribution of the number of common alleles at each imputed STR.** Common alleles are defined as alleles with estimated frequency ≥1% (**Methods**). For clarity we omitted from this figure the 237 imputed STRs with only a single imputed allele with frequency ≥1%. **(c-d) Representative association results.** Manhattan plots are shown for phenotypes (c) total bilirubin (an example moderately polygenic trait) and (d) platelet count (an example highly polygenic trait). Large diamonds represent the lead variants (pruned to include at most one lead variant per 10Mb for visualization). -log10 p-values are truncated at 100. Blue=SNPs and indels; orange=STRs. **(e) Summary of signals identified for each trait.** Bars show

the number of peaks per phenotype. Blue denotes peaks only containing genome-wide significant SNPs and indels, purple denotes peaks containing both significant SNPs and indels and STRs. The number of peaks only containing significant STRs is too small to be visible in this display. **(f) Comparison between lead SNP and indel and STR p-values at each peak.** If there are no STRs in a peak, the y coordinate is set to zero (same for SNPs and indels). p-values are capped at 1e-300, the maximum precision of our pipeline. The shade represents the number of peaks falling at each position on the graph.

Fine-mapping suggests 5-10% of significant signals are driven by STRs

We applied statistical fine-mapping to identify candidate causal variants that may be driving the GWAS signals detected above. We used two fine-mapping methods: SuSiE[33] and FINEMAP[34]. These methods differ in their modeling assumptions and thus provide partially orthogonal predictions. For each trait we divided its genome-wide significant variants (SNPs, indels and STRs) and nearby variants into regions of at least 500kb (**Methods**). This resulted in 14,494 fine-mapping trait-regions (**Supplementary Table 2**), with some trait-regions containing multiple nearby peaks. To compare outputs between fine-mappers in downstream analyses, we defined the causal probability (CP) of each variant to be a number between 0 and 1 that indicates the variant's chance of causality. For FINEMAP we defined a variant's CP to be the FINEMAP posterior inclusion probability (PIP) calculated for that variant. For SuSiE we defined a variant's CP to be the maximal SuSiE alpha value for that variant across pure credible sets in the region (**Supplementary Figs. 3-4**). We explain the rationale behind this choice in **Supplementary Note 1**.

We used two approaches to study the contribution of STRs vs. SNPs and indels to fine-mapped signals. First, we focused on the genome-wide significant variants (STR, SNP, or indel) with CP ≥ 0.8. (These accounted for a minority of the 21,030 total signals detected by SuSiE). SuSiE identified 4,490 such variants and FINEMAP identified 6,240. Of these, 7.4% (range 1.3-13.0% across traits; SuSiE) and 9.7% (range 1.2-14.9%; FINEMAP) are STRs (**Supplementary Table 3**). Among the subset of variants identified by both methods (4,028), 5.6% (range 0.9-12.8%) are STRs. Second, we considered the sum of CPs from all genome-wide significant variants, thereby taking into account the many signals which were not resolved to a single variant. STRs make up 5.2% (range 1.1-6.8%) of the total SuSiE CP sum and 8.3% (range 3.1%-10.2%) of the total FINEMAP CP sum. A potential limitation of this second metric is that variants with small CPs (CP ≤ 0.1) represent a large fraction (29.3% for SuSiE, 27.7% for FINEMAP) of these totals (**Supplementary Fig. 5**). Additionally, our results below suggest that a sizable subset of variant CPs are unstable or discordant between fine-mappers, particularly for STRs (**Supplementary Notes 2-3**), impacting the totals in both metrics. Nevertheless, these results above suggest that

between 5.2-9.7% of genome-wide significant signals can be explained by an STR, regardless of the fine-mapping method or metric used.

We next evaluated the robustness of our fine-mapping results. While SuSiE and FINEMAP tended to output similar results, they assigned highly discordant CPs to a subset of variants (**Supplementary Note 2**; **Supplementary Figs. 6-8**). Therefore, we performed additional analyses to identify a high-confidence set of causal STR candidates. We first conservatively restricted to the 177 candidate STRs with association p-values ≤ 1e-10 and with CP ≥ 0.8 in both FINEMAP and SuSiE. We then reran SuSiE and FINEMAP under a range of alternative settings, such as using best-guess STR genotypes instead of dosages and varying the prior distribution of effect sizes. These and other alternative settings are described in the **Methods**. The different settings we evaluated tended to produce concordant results, but again, for a subset of STRs, we observed highly inconsistent CPs (**Supplementary Figs. 9-12**). These discrepancies, which are discussed in detail in **Supplementary Note 3**, suggest that fine-mapping results can in some cases be highly sensitive to input filtering, model settings and imputation quality. Thus, we further restricted to those STR-trait associations which maintained CP ≥ 0.8 across a range of alternative fine-mapping conditions (**Methods**; **Fig. 2**; **Supplementary Table 4**). We refer to these below as confidently fine-mapped STR associations. Lastly, we added an STR in the *APOB* gene to this set as we noticed this variant only failed to meet the above criteria because it was simultaneously represented in both our STR reference panel and in the SNP and indel set generated by the UKB team (**Supplementary Note 4**). This left us with 118 confidently fine-mapped STR-phenotype associations corresponding to 95 distinct STRs.

Next, we evaluated our fine-mapping results by measuring their replication rates in populations besides White British individuals, with the expectation that causal associations will replicate at higher frequencies in other populations than non-causal associations due to having common biological functionality. The UKB includes genetically unrelated, self-identified groups of 7,562 Black, 7,397 South Asian, 1,525 Chinese, 11,978 Irish and 15,838 Other White participants (**Methods**). For each of those five groups we performed association testing for each STR against each trait (**Supplementary Table 5**). As expected, signals replicate at a higher rate in groups most closely related to our discovery cohort (i.e. Irish and Other White). Encouragingly, fine-mapped associations replicate at higher rates than non-fine-mapped associations in the Black, South Asian, and Chinese populations, even after stratifying by the discovery p-value (**Fig. 3**, **Supplementary Fig. 13**). To quantitatively measure this trend, for each population we fit a logistic regression model using whether signals replicated in that population as the outcome, the fine-

mapping status of those associations as the independent variable, and their -$\log_{10}$(p-value) in the discovery cohort as a covariate (**Supplementary Table 6)**. This analysis further supports the conclusion that fine-mapped associations replicate at higher rates. Additionally, the model consistently predicts that confidently fine-mapped STR associations replicate at higher rates than STRs fine-mapped by either fine-mapper alone, although only a subset of those predictions reached nominal significance, likely due to the small number of fine-mapped STR associations.

Next, we sought to characterize the set of confidently fine-mapped STRs. This set contains 62 poly-A repeats, 13 poly-AC, 5 poly-CCG, and 15 repeats with other units. Nine of these STRs overlap coding or untranslated regions (UTRs) (**Table 1**; **Supplementary Table 7**). Compared to all genome-wide significant STRs, confidently fine-mapped STRs were more likely to be exonic trinucleotide STRs (two-sided two-sample test of difference between proportions p=5e-05). No other annotation categories that we tested showed significant enrichment or depletion after multiple hypothesis correction (**Methods**; **Supplementary Fig. 14**). Lastly, we observed that 17 of these confidently fine-mapped STRs are significant quantitative trait loci (QTLs) for the expression of nearby genes in the Genotype-Tissue Expression (GTEx) dataset[35] (**Supplementary Tables 8-9**; **Methods**). We note that both of these analyses were underpowered, due to the low number of confidently fine-mapped STRs and the low sample sizes for the most relevant tissue types (e.g. kidney, liver).

STR coordinate (hg19 chr:pos)	Reference allele	Called repeat unit	Trait	Association P-value	Association Z-score	Gene (annotation)
1:204527033	$(TAA)_9$	AAT	platelet crit	5.76e-17	-8.37	*MDM4* (3'UTR)
2:21266752	$(CAG)_6$(CGCAGGCAG) $[CGC(CAG)_2]_2CGC$	CTG (Poly-Leucine)	apolipoprotein B	1.37e-279	-35.76	*APOB* (Coding)
2:106510441	$(AC)_6GTG(CA)_{10}C(TA)_7T$	AC	mean platelet volume	6.93e-29	-11.15	*NCK2* (3'UTR)
2:111878544	$(CGC)(CGCTGC)_2(CGC)_{13}C$	CCG	eosinophil count	4.96e-58	+16.06	*BCL2L11* (5'UTR)
			eosinophil percent	5.88e-75	+18.32	
2:204311891	$T_4CT_4CT_3CT_{18}$	T	IGF-1	3.97e-11	-6.61	*ABI2* (3'UTR*)
11:119077000	$(CGG)_{11}C$	CGG	mean sphered cell volume	6.88e-16	-8.07	*CBL* (5'UTR*)
			platelet count	3.77e-83	+19.32	
			platelet crit	6.07e-103	+21.55	
16:67229794	$(CAG)_{13}(CAA)(CAG)(TAA)(CAG)_3$	AGC (Poly-Serine)	mean sphered cell volume	3.07e-23	+9.93	*E2F4* (Coding)
			red blood cell count	1.08e-13	-7.43	
			mean corpuscular haemoglobin	2.83e-23	+9.94	
			mean corpuscular volume	9.27e-26	+10.49	
17:30469471	$(CCG)_{16}CC$	CCG	red blood cell distribution width	6.57e-13	+7.19	*RHOT1* (5'UTR)
17:33871548	T_{17}	A	mean platelet volume	4.30e-62	-16.63	*SLFN14* (3' UTR)
			platelet distribution width	1.18e-249	-33.78	

Table 1: Confidently fine-mapped STRs are identified in coding regions and untranslated regions (UTRs). Imputed alternate alleles and rsIDs are provided in **Supplementary Table 7.** Repeat units here are calculated as described in the **Methods**, except that they are required to be on the strand in the direction of transcription of the overlapping gene. Asterisks next to UTRs in the last column denote STRs which overlap UTRs of only noncanonical transcript(s) from Ensembl release 106.

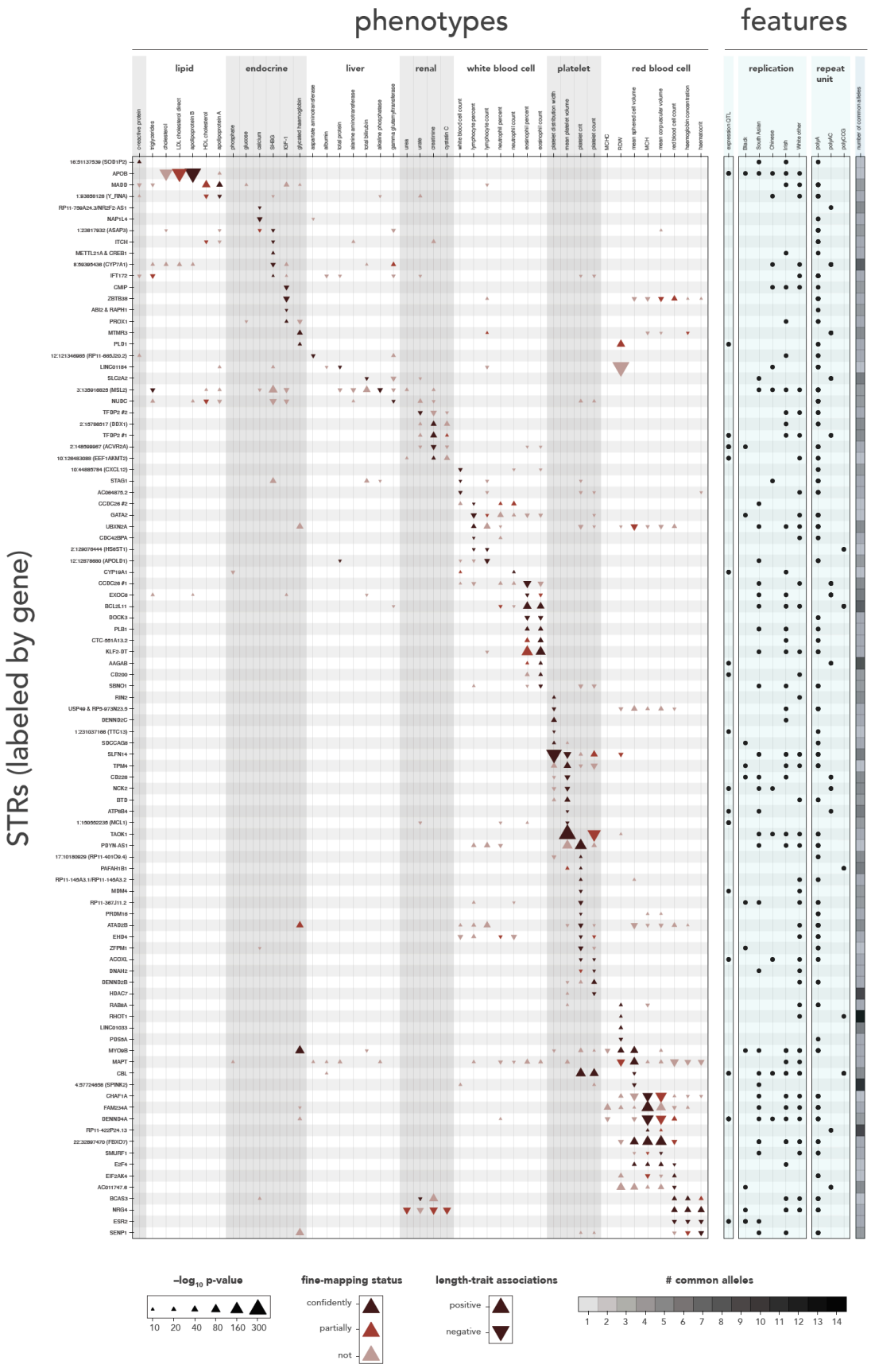


Figure 2: STRs are confidently fine-mapped to causally impact many traits. (a) Overview of confidently fine-mapped STRs. Only STRs with at least one confidently fine-mapped association are shown. Each triangle represents an STR-trait association with association p-value ≤ 1e-10. Black=confidently fine-mapped, red-brown=CP ≥ 0.8 in either initial FINEMAP or SuSiE run, light-tan=all other associations with p≤1e-10. Triangle direction (up or down) indicates the sign of the association between STR length and the trait. Triangle size scales with association p-value. Similar traits are grouped on the x-axis by white and light-grey bands. STRs are grouped on the y-axis according to the traits they were confidently fine-mapped to. STRs are labeled by the genes they reside in (protein coding genes preferred) or by chromosomal location and the nearest gene for intergenic STRs. *CCDC26* and *TFDP2* each contain two confidently fine-mapped STRs and appear twice. Light blue rows indicate (from left to right): whether each STR is associated with expression of a nearby gene (adjusted p<0.05; **Supplementary Table 8**), replicates with the same direction of effect in other populations (adjusted p<0.05; **Methods**), repeat unit, and the number of common alleles for each STR (as defined in **Fig. 1**; see scale beneath).

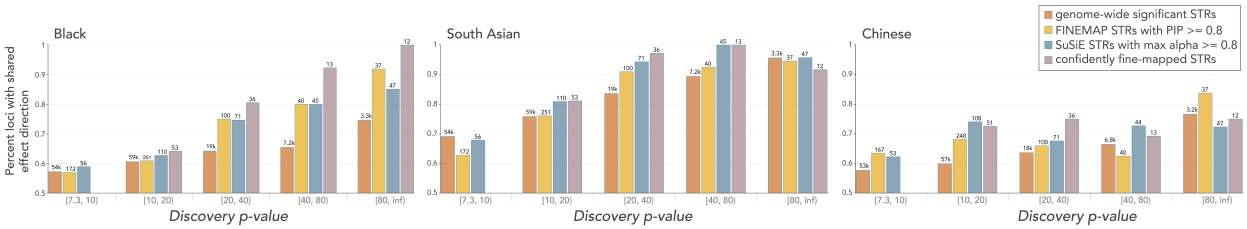


Figure 3: Concordance of White British STR effect directions in Black, South Asian and Chinese populations. The y-axis gives the fraction of STR associations measured in the discovery cohort that have the same direction of effect when measured in the replication population regardless of p-value. Brackets beneath the x-axis denote the binning of discovery -$\log_{10}$(p-values). Brown=genome-wide significant associations (discovery p≤5e-8), orange=FINEMAP fine-mapped STR associations (discovery p≤5e-8 and FINEMAP CP ≥ 0.8), teal=SuSiE fine-mapped STR associations (discovery p≤5e-8 and SuSiE CP ≥ 0.8) and purple=confidently fine-mapped STR associations. Annotations above each bar indicate the number of STR-trait associations considered. We required confidently fine-mapped STR associations to have p-value ≤ 1e-10, thus they do not appear in the left-most bin. The trends in these figures are somewhat sensitive to the choice of p-value bin boundaries so we additionally analyze this data using logistic models (**Supplementary Table 6**).

Fine-mapped STRs capture known associations

We identified multiple confidently fine-mapped STRs that were previously demonstrated to have functional roles, providing supporting evidence for the validity of our pipeline. For instance, our fine-mapping predicts a protein-coding CTG repeat (**Supplementary Table 7**) to be the causal variant for one of the strongest signals for LDL-cholesterol (LDL-C; two-sided association t-test p-value = 2e-235) and apolipoprotein B (p=1e-279), which forms the backbone of LDL-C lipoproteins[36]. This repeat is bi-allelic in the UKB cohort with an alternate allele corresponding to deletion of three residues (Leu-Ala-Leu) in the signal peptide coded in the first exon of the apolipoprotein B (*APOB*) gene[37]. This deletion occurs in an imperfect region of the CTG repeat, with sequence CTGGCGCTG. In agreement with previous studies[38,39], we found that the short allele is associated with high levels of both analytes. This STR also obtains association p-values ≤ 0.05 with apolipoprotein B and LDL-C in each of the five other populations we considered.

As another example, our confidently fine-mapped STR set implicates a multiallelic AC repeat (**Supplementary Table 7**) 6bp downstream of exon 4 of *SLC2A2* (also known as *GLUT2*, a gene that is most highly expressed in liver) as causally impacting bilirubin levels (p=8e-18). The potential link between GLUT2 and bilirubin is described in **Supplementary Note 4.** Previous studies in HeLa and 293T cells showed that inclusion of exon 4 of *SLC2A2* is repressed by the binding of mRNA processing factor hnRNP L to this AC repeat[40,41], implicating this STR in *SLC2A2* splicing. Notably, these studies did not investigate the impact of varying repeat copy number. We examined this STR in GTEx liver samples and did not find a significant linear association between repeat count and the splicing of exon 4, though we did find evidence for association with the splicing of exon 6 (**Supplementary Fig. 15**).

A trinucleotide repeat in CBL *regulates platelet traits*

Most of the confidently fine-mapped STR associations identified by our pipeline have, to our knowledge, not been previously reported. For example, this set includes positive associations between the length of a highly polymorphic CGG repeat in the promoter of the gene *CBL* (which encodes an E3 ubiquitin ligase) and both platelet count (p=4e-83) and platelet crit (p=6e-103; **Supplementary Table 7**; **Fig. 4a-b**; **Supplementary Fig. 16**). Compared to other types of STRs, CG-rich repeats in promoter and 5' UTR regions have been strongly implicated in transcriptomic[42,43] and epigenomic regulation[44]. This repeat is also confidently fine-mapped to an association with mean sphered cell volume (p=7e-16; **Supplementary Fig. 17**), but this is comparatively much weaker and we do not discuss it here. For both platelet crit and platelet count, the two fine-mappers identify two signals in this region, one of which they both localize to this STR. After conditioning on the lead variant from the other signal (rs2155380) the STR becomes the lead variant in the region by a wide margin (**Fig. 4c-d**). Conditioning on both the lead variant and the STR accounts for all the signal in the region (**Fig. 4e**). This supports the fine-mappers' prediction that there is a second signal in this region which is driven by the STR. The association between this STR's length and platelet crit replicated with $p \leq 0.05$ in all of the non-Black populations tested, and the association with platelet count replicated in three of those four populations. While these associations did not replicate in the Black population, this STR has shorter alleles in that population (**Fig. 4a**) and it appears that the relationship between allele length and platelet count may only be present at intermediate allele lengths (**Fig. 4b**). Population-specific distributions of allele lengths based on genotypes obtained directly from whole genome sequencing in the 1000 Genomes Project[21] (**Methods**) are highly similar to those obtained from

imputed data in the UKB, suggesting imputed genotypes at this locus are accurate across populations (**Supplementary Fig. 18**).

This STR contains a common imperfection (rs7108857, which changes the second CGG copy to TGG). That variant is in weak LD with the length of the STR (r^2 ranging between 0.023 (White British) and 0.175 (Chinese)) (**Fig. 4a**) and in strong LD with the lead variant of the other signal in this region (rs2155380). While rs7108857 is more strongly associated with the platelet traits than the STR's length (platelet count p=9e-86, platelet crit p=4e-98), given the fine-mappers' results that the STR length association is an independent signal, it is unsurprising that the STR-length association remains after stratifying on the presence of this imperfection (**Fig. 4f**). This suggests that imperfections and repeat lengths are different characteristics of repeats that may have distinct associations.

While the alleles present in our reference panel at this STR all have between 5 and 31 CCG repeat copies, much rarer large expansions of this repeat (>100 repeats) have been previously implicated in Jacobsen Syndrome[45,46], a disorder characterized in part by the deletion of *CBL* which has been observed together with platelet abnormalities[47]. Similarly, loss of function mutations of *CBL* have been associated with increased platelet count[48]. These observations directly implicate *CBL* as a negative regulator of platelet production. We found that increased CCG length was negatively associated with *CBL* expression in three tissues in the GTEx cohort[35] (each with p-value ≤ 0.05 after multiple hypothesis correction; **Supplementary Table 8**; **Fig. 4g**). Intriguingly, this association replicated (p=0.007) in Europeans and was only modestly significant in African (p=0.048) individuals in the Geuvadis cohort[49] (**Fig. 4h**), where we observed that African individuals have much higher overall *CBL* levels. This could explain why this STR's associations with platelet traits did not replicate in the Black population. Intriguingly, the association signals for both the platelet traits and expression show similar non-linear patterns, with linear effects for medium-sized repeats but with plateauing effects for the shortest and longest alleles. Overall, our results support the hypothesis that longer CCG repeat alleles contribute to increased platelet count in non-Black populations by decreasing CBL expression (**Fig. 4i**), matching the direction of the gene-trait correlation observed previously[48].

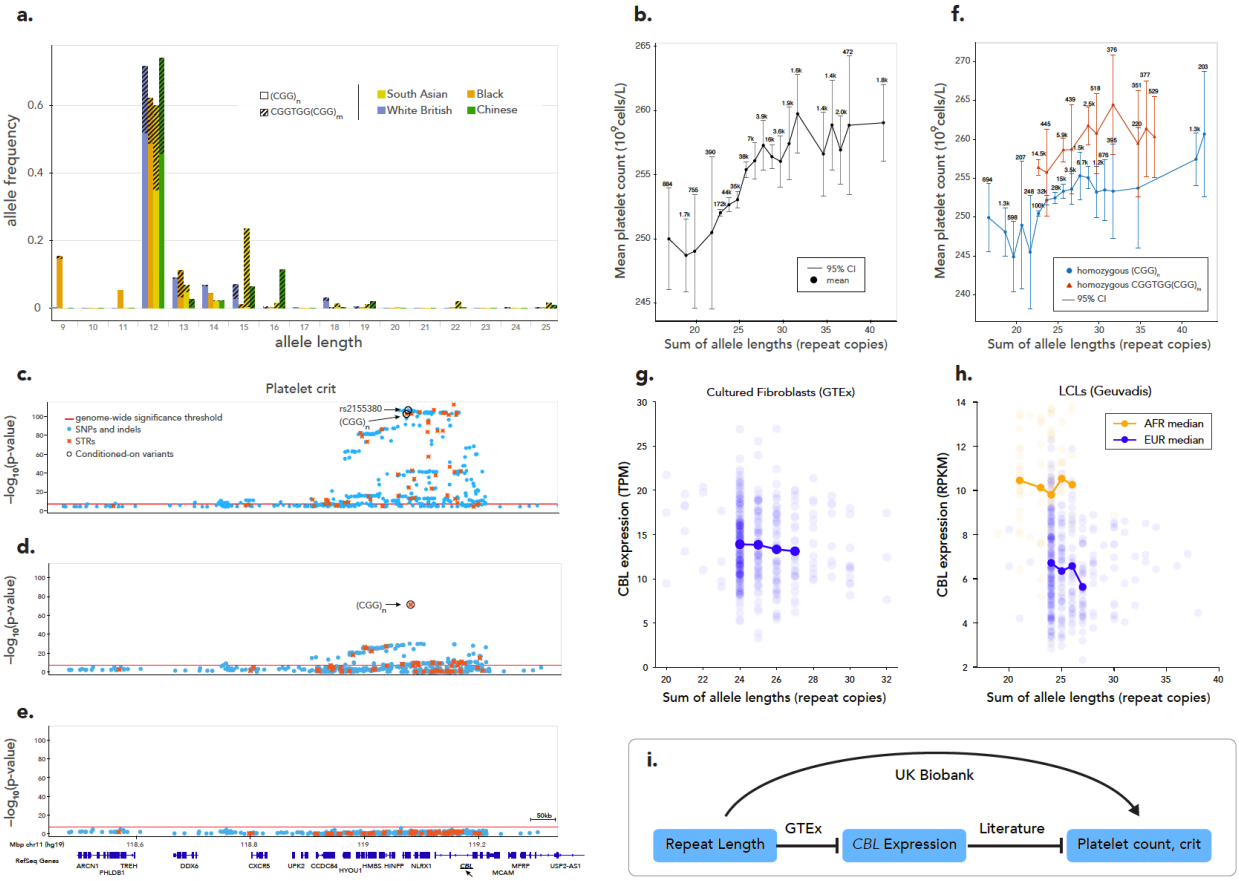


Figure 4: A highly polymorphic CGG repeat in the promoter of *CBL* influences platelet traits. (a) Distribution of STR alleles across populations. The x-axis gives STR length (number of repeat units) and y-axis gives the population frequency. The hatched portion of each bar corresponds to the alleles of that length that include a "TGG" imperfection at the second repeat (rs7108857). Colors denote different UKB populations. Extreme allele lengths 6, 8, 26, 27, 29, 30, 31, 32 each have frequency less than 1% in all populations and have been omitted. **(b) STR length vs. mean platelet count.** The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's probability of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each dosage. **(c-e) Association of variants at the *CBL* locus with platelet crit.** Association plots in the White British population are shown before conditioning **(c)**, after conditioning on rs2155380 **(d)**, and after conditioning on both rs2155380 and STR length **(e)**. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circles=the $(CGG)_n$ STR and rs2155380. **(f) STR length vs. mean platelet count conditioned on the TGG Imperfection rs7108857.** Blue=individuals homozygous for no imperfection (n=190,280); green=individual homozygous for the imperfection (n=26,824). Results are shown for platelet crit, similar results were obtained for platelet count (not shown). Individuals are categorized as being homozygous based on their most probable imputed genotype. 82% of individuals categorized as homozygous for the imperfect allele and 99% of those categorized as homozygous for the reference allele have an imputation probability of ≥ 95% for their genotypes. For each category, only length dosages with a frequency of 0.1% or greater in that category are displayed. **(g-h) STR length vs. *CBL* expression.** Associations are shown for Cultured Fibroblasts from GTEx (n=393) **(g)** and LCLs from Geuavdis (n=447) **(h).** Orange=African, blue=European. Solid lines give median expression values for each STR dosage with at least 5% frequency in each population group. **(i) Proposed pathway for effect of STR length on platelet traits.** The arrow denotes a positive association, the capped lines denote negative associations. Interactions are captioned by their information sources.

Additional examples of confidently fine-mapped STR-trait associations

We observe another 5' UTR CCG repeat in *BCL2L11* (also known as *BIM*) that is confidently fine-mapped to eosinophil percentage (p=6e-75) and eosinophil count (p=5e-58) (**Supplementary Table 7**). This repeat is the most strongly associated variant in the region for both traits, and conditioning on it accounts for the entire signal in this region (**Supplementary Fig. 19a**). Proteins in the BCL-2 family are known to act as anti- or pro-apoptotic regulators. *BIM* in particular is required in the tightly regulated lifespan of myeloid lineage cells, which include eosinophils. Loss of repression of *BCL2L11* was specifically shown to be associated with marked decline in eosinophil counts[50], and increased *BIM* expression in mice has been shown to increase eosinophil counts[51,52]. Similar to the CGG repeat in *CBL*, this STR is highly polymorphic and only shows a linear association with the trait across a subset of the range of possible allele lengths (**Supplementary Fig. 19b**).

While exonic repeats are potentially easier to interpret, a majority of our confidently fine-mapped STRs fall in intronic regions. We resolve one of the strongest signals for mean platelet volume (p<1e-300) to a multiallelic poly-A STR in an intron of the gene *TAOK1* (**Supplemental Table 7**; **Supplementary Fig. 20a**). This association replicated in all examined populations except the Black population. Furthermore, conditioning on the length of this STR demonstrates that it explains the majority of the signal in this region (**Supplementary Fig. 20b**). The same STR is also strongly associated with platelet count (p=2e-181), which reached a CP of 1 in 7 out of the 8 fine-mapping tests we ran and replicated in all examined populations except for the Black and Chinese populations.

TAOK1 is a protein kinase that plays a role in regulating microtubule dynamics[53] which is known to be critical to platelet generation[54]. The STR is in an intron of the canonical *TAOK1* transcript but lies immediately downstream of a non-protein coding transcript (ENST00000577583; a retained intron) and is approximately 2.4kb upstream of a differentially spliced exon. The STR also bears the hallmarks of a regulatory element: it is located in a DNase hypersensitivity cluster and overlaps a transcription factor binding site for ESR1 (**Methods**). This location is suggestive for the way that variation in the length of this STR could affect *TAOK1* gene regulation, potentially via impacting splicing or modulating enhancer activity. However, we could not test the impact of this STR on *TAOK1* regulation in GTEx as the STR was filtered due to low call rate (11%).

In another confidently fine-mapped example we identify a previously unreported association between a GTTT repeat in an intron of estrogen receptor beta (*ESR2*) (**Supplementary Table 7**; **Supplementary Note 4; Supplementary Fig. 21**) and haemoglobin concentration (p=1.15e-24), red blood cell count (p=3.21e-24) and haematocrit (p=1.00e-26), where additional repeat copies correspond to lower measurements of all three traits. Despite the relatively weak discovery p-value and differing allele distributions between the White British and Black populations (**Supplementary Fig. 21c**), all three associations replicate in the Black population with p-values ≤ 0.05. The associations do not consistently replicate in the other four examined populations, suggesting that the effect of this STR on red blood cell traits is potentially larger in the Black population. Consistent with these associations, *ESR2* has been implicated in the regulation of red blood cell production[55,56]. We found a significant negative association between STR length and *ESR2* expression in two tissues in GTEx (each with p-value ≤ 0.05 after multiple hypothesis correction; **Supplementary Table 8**). While evidence suggests a link between *ESR2* and red blood cell production, the expected direction of effect is unclear given the highly tissue-specific isoform usage and functions of this gene (**Supplementary Note**). Nevertheless, our results support a role of this STR in red blood cell production through regulation of ESR2.

We observed many additional associations of interest amongst the confidently fine-mapped STRs. For example, a highly polymorphic CCG repeat in the 5' UTR of *RHOT1* is associated with red blood cell distribution width (**Supplementary Table 7**). This repeat overlaps a CTCF binding site, is located within a nucleosome depleted region of a H3K27ac peak in LCLs, and shows a strong association with the expression of *RHOT1* in these cells (p=2e-44 in Europeans, p=0.035 in Africans; **Supplementary Fig. 22**). We also find multiple AC repeats in our set that are significantly associated with expression of nearby genes. This includes a polymorphic AC repeat located in the 3' UTR of *NCK2* which is associated with platelet distribution width and mean platelet volume (**Supplementary Table 7**). This repeat overlaps a PABPC1 binding site and has a significant negative association with *NCK2* expression in multiple GTEx tissues (**Supplementary Fig. 23**; **Supplementary Table 8**). Finally, many STRs in our fine-mapped set consist of poly-A repeats. While traditionally these have been among the most challenging regions of the genome to genotype[57], many such STRs, including poly-A repeats in *MYO9B*, *DENND4A*, and *NRG4*, show strong statistical evidence of causality and replicate across multiple population groups (**Fig. 2**). Taken together, these loci exemplify the large number of potentially causal variants that our list of confidently fine-mapped STRs provides to future studies.

Discussion

In this study, we imputed 445,735 STRs into the genomes of 408,153 participants in the UK Biobank and associated their lengths with 44 blood cell and biomarker traits. Using fine-mapping, we estimate that STRs account for 5-10% of causal variants for these traits. We stringently filtered the fine-mapping output to produce 118 confidently fine-mapped STR-trait associations with strong evidence for causality across 95 distinct STRs. These associations include some of the strongest signals for apolipoprotein B, mean platelet volume and mean corpuscular haemoglobin. These confidently fine-mapped STRs replicated in the Black, South Asian and Chinese UKB populations at higher rates than non-fine-mapped STRs ($p<0.02$ in each). A subset of these STRs were associated with expression of nearby genes, providing evidence for their impact on regulatory processes and explanations for their effects on the studied traits.

Broadly, our study highlights the importance of considering a more complete set of genetic variants in complex trait analysis. Many variant types are often highly multiallelic and only imperfectly tagged by individual common SNPs, including the STRs studied here but also VNTRs[3], copy number variants[6], HLA types[58], and some structural variants[59]. While these variants are often excluded from analysis pipelines due to the technical challenges they pose, they likely represent an important source of causal variants and heritability[60,61] that has yet to be captured. Further, we expect incorporation of this additional source of causal variants, which we observe often exhibit population-specific allele distributions, will improve downstream applications such as polygenic risk scores, particularly in constructing scores that are more applicable across diverse populations.

While our results uncover many novel candidate causal STR variants, we do not believe these findings to be exhaustive. Our fine-mapping procedure was exceptionally conservative and excluded hundreds of STR-trait associations strongly predicted to be causal in some but not all settings tested. Additionally, we focused only on a subset of autosomal STRs ascertained to be polymorphic and amenable to imputation in European individuals[28]. This excluded most long repeats such as those implicated in pathogenic expansion disorders and likely excluded STR alleles that are common only in non-European populations. Emerging whole genome sequencing datasets from the UKB and biobanks spanning diverse populations[62,63] are beginning to enable direct genotyping, rather than imputation, of STRs. This data is likely to dramatically improve the ability to capture additional STRs, particularly in underserved populations.

Methodological advances are also needed to support the study of STRs. Here we developed associaTR, an open-source reproducible pipeline that enables future studies to conduct STR association tests. However, we envision that integrating support for STR-length based tests and other complex variant associations into widely used GWAS toolkits is required to enable routine analysis of the full spectrum of human genetic variation. Further improvements to our association testing models are also likely to reveal new insights. Adoption by linear-mixed model methods would increase the power to detect length-based associations. Additionally, in this study we only modeled linear associations between STR lengths and trait values. However, visualization of many of the associations we identify, including those at *CBL* and *BCL2L11*, suggests that linear models account for only a subset of those associations and that many STR effects may be best detected through non-linear models. We also only tested for associations with repeat length. However, inspection of individual loci reveals that complex repeat structures are common (**Table 1**). Systematic evaluation of the potential for epistasis between repeat imperfections and STR lengths, as well as between the lengths of neighboring repeats, would potentially enable better understanding of the phenotypic impact of STRs.

Importantly, our results highlight current challenges in performing statistical fine-mapping. We found that fine-mapping results were in some cases highly sensitive to choices of tool settings and filtering thresholds, where in some settings a variant may be identified as highly likely to be causal but identified as having no causal impact in others. This suggests results of statistical fine-mapping should be interpreted with caution and evaluated for sensitivity to model choices, and that further work is needed to make the process of fine-mapping more robust.

Although fine-mapping inconsistencies were identified for SNPs and indels as well as STRs, they were most prevalent for STRs. While this may in part be due to issues with imputing STR genotypes, more research is needed to further evaluate the performance of current fine-mapping tools on regions containing STRs. Additionally, there is a need for fine-mapping tools that can model effects of multiallelic variants. Existing fine-mapping frameworks in theory can accurately model linear repeat-length associations, but we hypothesize that more detailed modeling of LD between SNPs and individual STR alleles may enable more accurate model fitting procedures. Similarly, during model fitting, existing tools often compare models which trade one causal variant for another variant in close LD, but greater accuracy may be obtained by comparing models which trade off a single, potentially causal, multiallelic variant for multiple simultaneously-causal biallelic variants.

Overall, our study provides a statistical framework for incorporating hundreds of thousands of tandem repeat variants into GWAS studies, identifies dozens of novel candidate variants for future mechanistic studies, and demonstrates that STRs likely make a widespread contribution to complex traits.

Methods

Selection of UK Biobank participants

We downloaded the fam file and sample file for version 2 of the phased SNP array data (referred to in the UKB documentation as the 'haplotype' dataset) using the ukbgene utility (ver Jan 28 2019 14:09:15 - using Glibc2.28(stable)) described in UKB Data Showcase Resource ID 664 (**URLs**). The IDs from the sample file already excluded 968 individuals previously identified as having excessive principal component-adjusted SNP array heterozygosity or excessive SNP array missingness after call-level filtering[29] indicating potential DNA contamination. We further removed withdrawn participants, indicated by non-positive IDs in the sample file as well as by IDs in email communications from the UKB access management team. After the additional filtering, data for 487,279 individuals remained.

We downloaded the sample quality control (QC) file (described in the sample QC section of UKB Data Showcase Resource ID 531 (**URLs**)) from the European Genome-Phenome Archive (accession EGAF00001844707) using pyEGA3[64]. We subsetted the non-withdrawn individuals above to the 408,870 (83.91%) participants identified as White-British by column `in.white.British.ancestry.subset` of the sample QC file. This field was computed by the UKB team to only include individuals whose self-reported ethnic background was White British and whose genetic principal components were not outliers compared to the other individuals in that group[29]. In concordance with previous analyses of this cohort[29] we additionally removed data for:

- 2 individuals with an excessive number of inferred relatives, removed due to plausible SNP array contamination (participants listed in sample QC file column `excluded.from.kinship.inference` that had not already been removed by the UKB team prior to phasing)

- 308 individuals whose self-reported sex did not match the genetically inferred sex, removed due to concern for sample mislabeling (participants where sample QC file columns `Submitted.Gender` and `Inferred.Gender` did not match)

- 407 additional individuals with putative sex chromosome aneuploidies removed as their genetic signals might differ significantly from the rest of the population (listed in sample QC file column `putative.sex.chromosome.aneuploidy`)

Following these additional filters the data for 408,153 individuals remained (99.82% of the White British individuals considered above).

SNP and indel dataset preprocessing

We obtained both phased hard-called and imputed SNP and short indel genotypes made available by the UKB.

Phased hard-called genotypes: We downloaded the bgen files containing the hard-called SNP haplotypes (release version 2) and the corresponding sample and fam files using the ukbgene utility (UKB Data Showcase Resource 664 (**URLs**)). These variants had been genotyped using microarrays and phased using SHAPEIT3[65] with the 1000 genomes phase 3 reference panel[21]. Variants genotyped on the microarray were excluded from phasing and downstream analyses if they failed QC on more than one microarray genotyping batch, had overall call-missingness rate greater than 5% or had minor allele frequency less than 0.01%. Of the resulting 658,720 variants, 99.5% were single nucleotide variants, 0.2% were short indels (average length 1.9bps, maximal length 26bps), and 0.2% were short deletions (average length 1.9bps, maximal length 29bps).

Imputed genotypes: We similarly downloaded imputed SNP data using the ukbgene utility (release version 3). Variants had been imputed with IMPUTE4[29] using the Haplotype Reference Consortium panel[20], with additional variants from the UK10K[66] and 1000 Genomes phase 3[21] reference panels. The resulting imputed variants contain 93,095,623 variants, consisting of 96.0% single nucleotide variants, 1.3% short insertions (average length 2.5bps, maximum length 661bps), 2.6% short deletions (average length 3.1bps, maximum length 129bps). This set does not include the 11 classic human leukocyte antigen alleles imputed separately.

We used bgen-reader[67] 4.0.8 to access the downloaded bgen files in python. We used plink2[31] v2.00a3LM (build AVX2 Intel 28 Oct 2020) to convert bgen files from both hard-called and imputed

SNPs to the plink2 format for downstream analyses. For hard-called genotypes, we used plink to set the first allele to match the hg19 reference genome. Imputed genotypes already matched the reference. Unless otherwise noted, our pipeline worked with imputed genotypes as non-reference allele dosages, i.e. $\Pr(\text{heterozygous}) + 2 * \Pr(\text{homozygous alternate})$ for each individual.

STR imputation

We previously published a reference panel containing phased haplotypes of SNP variants alongside 445,735 autosomal STR variants in 2,504 individuals from the 1000 Genomes Project[21,28] (**URLs**). This panel focuses on STRs ascertained to be highly polymorphic and well-imputed in European individuals. Notably, this excludes many STRs known to be implicated in repeat expansion diseases, STRs that are primarily polymorphic only in non-European populations, or STRs that are too mutable to be in strong linkage disequilibrium (LD) with nearby SNPs.

To select shared variants for imputation, we note that 641,582 (97.4%) of variants hard-called and phased in the UKB participants were present in our SNP-STR reference panel. As a quality control step, we filtered variants that had highly discordant minor allele frequencies between the 1000 Genomes European subpopulations (**URLs**) and White British individuals from the UKB. We first took a maximal unrelated set of the White British individuals (see **Phenotype Methods** below) and then visually inspected the alternate allele frequency of the overlapping variants (**Supplementary Fig. 1**) and chose to remove the 110 variants with an alternate allele frequency difference of more than 12%.

We used Beagle[30] v5.1 (build 25Nov19.28d) with the tool's provided human genetic maps (**URLs**) and non-default flag `ap=true` to impute STRs into the remaining 641,472 SNPs and indels from the SNP-STR panel into the hard-called SNP haplotypes. Though we performed the above comparison between reference panel Europeans and UKB White British individuals, we performed this STR imputation into all UKB participants using all the individuals in the reference panel. We chose Beagle because it can handle multiallelic loci. Due to computational constraints, we ran Beagle per chromosome on batches of 1000 participants at a time with roughly 18GB of memory. We merged the resulting VCFs across batches and extracted only the STR variants. Lastly, we added back the `INFO` fields present in the SNP-STR reference panel that Beagle removed during imputation.

Unless otherwise noted, our pipeline worked with these genotypes as length dosages for each individual, defined as the sum of length of each of the two alleles, weighted by imputation probability. Formally, $dosage = \sum_{a \in A} len(a) * [Pr(hap_1 == a) + Pr(hap_2 == a)]$, where A is the set of all possible STR alleles at the locus, $len(a)$ is the length of allele a, and $Pr(hap_i == a)$ is the probability that the allele on the ith haplotype is a, output by Beagle in the `AP1` and `AP2` `FORMAT` fields of the VCF file.

Estimated allele frequencies (**Fig. 1b**) were computed as follows: for each allele length L for each STR, we summed the imputed probability of the STR on that chromosome to have length L over both chromosomes of all unrelated participants. That sum is divided by the total number of chromosomes considered to obtain the estimated frequency of each allele.

Standardized k-mers and inferred repeat units

Each STR in the SNP-STR reference panel was previously annotated with a repeat period - the length of its repeat unit - but not the repeat unit itself. We inferred the repeat unit for each STR in the panel as follows: we considered the STR's reference allele and given period. We then took each k-mer in the reference allele where k is the repeat period, standardized those k-mers, and took their counts. We define the standardization of a k-mer to be the sequence produced by looking at all cyclic rotations of that k-mer and choosing the first one lexicographically. For example, the standardization of the k-mer CTG would be AGC. If the most common standardized k-mer was less than twice as frequent as another standardized k-mer, we did not call a repeat unit for that STR (11,962 STRs; 2.68%). This produced the strand-dependent repeat unit for that STR. To infer a strand-independent repeat unit for the STR we looked at all rotations of the strand-dependent repeat unit in both the forward and reverse directions, and chose whichever came first lexicographically. For example the repeat unit for the STR TGTGTGTG would be AC, while the strand-dependent repeat unit would be GT.

Phenotypes and covariates

IDs listed in this section refer to the UKB Data Showcase (**URLs**).

We analyzed a total of 44 blood traits measured in the UKB. 19 phenotypes were chosen from Category Blood Count (ID 100081) and 25 from Category Blood Biochemistry (ID 17518). We refer to them as blood cell count and biomarker phenotypes respectively. The blood cell counts were measured in fresh whole blood while all the biomarkers were measured in serum except for

glycated haemoglobin which was measured in packed red blood cells (details in Resource ID 5636). The phenotypes we analyzed are listed in **Supplementary Table 1**, along with the categorical covariates specific to each phenotype that were included during association testing.

We analyzed all the blood cell count phenotypes available except for the nucleated red blood cell, basophil, monocyte, and reticulocyte phenotypes. Nucleated red blood cell percentage was omitted from our study as any value between the bounds of 0% and 2% was recorded as exactly either 0% or 2% making the data inappropriate for study as a continuous trait. Nucleated red blood cell count was omitted similarly. Basophil and monocyte phenotypes were omitted as those cells deteriorate significantly during the up-to-24-hours between blood draw and measurement. This timing likely differed consistently for different clinics, and different clinics drew from distinct within-White British ancestry groups, which could lead to confounding with true genetic effects. See Resource ID 1453 for more information. Reticulocytes were excluded from our initial pipeline. This left us with 19 blood cell count phenotypes. For each blood cell count phenotype we included the machine ID (1 of 4 possible IDs) as a categorical covariate during the association tests to account for batch effects.

Biomarker measurements were subject to censoring of values below and above the measuring machine's reportable range (Resource IDs 1227, 2405). **Supplementary Table 1** includes the range limits and the number of data points censored in each direction. Five biomarkers (direct bilirubin, lipoprotein(a), oestradiol, rheumatoid factor, testosterone) were omitted from our study for having >40,000 censored measurements across the population (approximately 10% of all data), since those would require analysis with models that take censoring into account. The remaining biomarkers had less than 2,000 censored measurements. We excluded censored measurements for those biomarkers from downstream analyses as they consisted of a small number of data points. For each serum biomarker except LDL cholesterol and total bilirubin we included aliquot number (1-3) as a categorical covariate during association testing as an additional step to mediate the dilution issue (described in Resource ID 5636). LDL cholesterol and total bilirubin were run on a version of our analysis pipeline prior to accounting for the aliquot covariate. Glycated haemoglobin was not subject to the dilution issue, being measured in packed red blood cells and not serum, so no aliquot covariate was published in the UKB showcase or included in our analysis.

For each phenotype we took the subset of the 408,153 individuals above that had a measurement for that phenotype during the initial assessment visit or the first repeat assessment visit,

preferentially choosing the measurement at the initial assessment when measurements were taken at both visits. We include a binary categorical covariate in association testing to distinguish between phenotypes measured at the initial assessment and those measured at the repeat assessment. Each participant's age at their measurement's assessment was retrieved from Data Field ID 21003.

The initial and repeat assessment visits were the only times the biomarkers were measured. The blood cell count phenotypes were additionally measured for those participants who attended the first imaging visit. We did not use those measurements and for each phenotype excluded the <200 participants whose only measurement for that phenotype was taken during the first imaging visit as we could not properly account for the batch effect of a group that small (**Supplementary Table 1**).

No covariate values were missing. Before each association test we checked that each category of each categorical covariate was obtained by at least 0.1% of the tested participants. We excluded the participants with covariate values not matching this criterion, as those quantities would be too small to properly account for batch effects. In practice, this meant that for each biomarker we excluded the <100 participants that were measured using aliquot 4, and that for 8 biomarkers we additionally excluded the ≤125 participants that were measured using aliquot 3 (**Supplementary Table 1**).

For each phenotype we then selected a maximally-sized genetically unrelated subset of the remaining individuals using PRIMUS[68] v1.9.0. Precomputed measures of genetic relatedness between participants (described in UKB paper supplement section 3.7.1[29]) were downloaded using ukbgene (Resource ID 664). We ran PRIMUS with non-default options `--no_PR -t 0.04419417382` where the `t` cutoff is equal to 0.5^9, chosen so that two individuals are considered to be related if they are relatives of third degree or closer. This left between 304,658 and 335,585 unrelated participants per phenotype (**Supplementary Table 1**).

Sex and ancestry principal components (PCs) were included as covariates for all phenotypes. Participant sex was extracted from the hard-called SNP fam file (see above). The top 40 ancestry PCs were extracted from the corresponding columns of the sample QC file (see the Participants **Methods** section above).

We then rank inverse normalized phenotype values for association testing. The remaining unrelated individuals for each phenotype were ranked by phenotype value from least to greatest

(ties broken arbitrarily) and the phenotype value for association testing for each individual was taken to be $normal\ quantile\left(\frac{sample\ rank\ +0.5}{n\ samples}\right)$. We use rank inverse normalization as it is standard practice, though it does not have a strong theoretical foundation[69] and only moderate empirical support[70–73].

For each phenotype and its remaining unrelated individuals we standardized all covariates to have mean zero and variance one for numeric stability.

Association testing

We performed STR and SNP association testing separately. In both cases, we used simple linear models instead of linear mixed model (LMM) methods[74], as existing tools implementing LMM-based associations do not handle STR length-based tests, and our downstream analyses require STR and SNP associations to be computed using the same model to enable accurate comparisons. For STR association testing, the VCFs produced by Beagle were accessed in python by cyvcf2[75] 0.30.14 and a modified version of our TRTools library[76] v3.0.2. In line with plink's recommendation for SNP GWAS (**URLs**), 6 loci with non-major allele dosage < 20 were filtered. For each STR, we fit the linear model $y = g * \beta_g + C * \beta_C + \epsilon$ where y is the vector of rank-inverse-normalized phenotype values per individual, g is the vector of STR length dosage genotypes per individual, β_g is the effect size of this STR, C is the matrix of standardized covariates, β_C is the vector of covariate effect sizes, and ϵ is the vector of errors between the model predictions and the outcomes. Models were fit using the `regression.linear_model.OLS` function of the Python statsmodels library v0.13.2 (**URLs**). Per GWAS best-practices, we used imputation dosage genotypes instead of best-guess genotypes[77].

We used plink2[31] v2.00a3LM (build AVX2 Intel 28 Oct 2020) for association testing of imputed SNPs and indels. For each analysis, plink first converts the input datasets to its pgen file format. To avoid performing this operation for every invocation of plink, we first used plink to convert the SNP and indel bgen files to pgen files a single time. We invoked plink once per chromosome per phenotype. We used the plink flag `--mac 20` to filter loci with minor allele dosage less than 20 (**URLs**). Plink calculates minor allele counts across all individuals before subsetting to individuals with a supplied phenotype, so this uniformly filtered 22,396,837 (24.1%) of the input loci from each phenotype's association test leaving 70,698,786 SNPs and indels. Plink fit the same linear

model described above in the STR associations, except that g is the vector of dosages of the non-reference SNP or indel allele.

For conditional regressions, we fit the model $y = g * \beta_g + f * \beta_f + C * \beta_C + \epsilon$ where all the terms are as described above, except f is the vector of per-individual genotypes of the variant being conditioned on, and β_f is its effect size.

Comparison with Pan-UKB pipeline

We compared the results of our pipeline to results available on the Pan-UKB[32] website (see **URLs**) using bilirubin as an example trait. We matched variants between datasets on chromosome, position, reference and alternate alleles. We found our pipeline produced largely similar p-values to those reported for European participants in Pan-UKB (**Supplementary Fig. 2**).

Defining significant peaks

Given a peak width w (bps) and a p-value threshold t, we selected variants to center peaks on in the following manner:

1. Order all variants (of all types) from most to least significant. For variants which exceed our pipeline's precision (p<1e-300), order them by their chromosome and base pair from first to last. (These variants will appear at the beginning of the list of all variants).

2. For each variant: If the variant has p-value > t, break. If there is a variant in either direction less than $w/2$ bps away which has a lower p-value, continue. Otherwise, add this variant to the list of peak centers.

We define peaks to be the w (base pair) width regions centered on each selected variant. The statistics given in the **Results** are calculated using $w = 250kb$ and $t = 5e-8$. The identification of peaks in **Fig. 1c-d** was made with $w = 20mb$ and $t = 5e-8$ for visualization purposes.

Identifying indels which are STR alleles

Some STR variant alleles are represented both as alleles in our SNP-STR reference panel and as indel variants in the UKB imputed variants panel. We excluded the indel representations of those alleles from fine-mapping, as they represent identical variants and could confound the fine-mapping process. For each STR we constructed the following interval:

$$\begin{cases} (start - 3, end + 3), & period = 1 \\ (start - 2 * period, end + 2 * period), & period > 1 \end{cases}$$

where $period$ is the length of the repeat unit. and $start$ and end give the coordinates of the STR in base pairs. We call an indel an STR-indel if it only represents either a deletion of base pairs from the reference or an insertion of base pairs into the reference (not both), overlaps only a single STR based on the interval above, and represents an insertion or deletion of full copies of that STR's repeat unit. We conservatively did not mark any STR-indels for STRs whose repeat units were not called (see above) or for which the insertion or deletion was not a whole number of copies of any rotation of the repeat unit.

Fine-mapping

For each phenotype, we selected contiguous regions to fine-map in the following manner:

1. Choose a variant (SNP or indel or STR) with p-value < 5e-8 not in the major histocompatibility complex (MHC) region (chr6:25e6-33.5e6).
2. While there is a variant (SNP or indel or STR) with p-value < 5e-8 not in the MHC region and within 250kb of a previously chosen variant, include that variant in the region and repeat.
3. This fine-mapping region is (min variant bp – 125kb, max variant bp + 125kb).
4. If the resulting region has no STR variants with p ≤ 5e-4, exclude it from downstream analyses.
5. Start again from step 1 to create another region, starting with any variant with p-value < 5e-8 not already in a fine-mapping region.

This is similar to the peak selection algorithm above but is designed to produce slightly wider regions so that we could fine-map nearby peaks jointly. We excluded the MHC because it is known to be difficult to effectively fine-map. Steps 1-3 produced 14,494 trait-regions, of which 13,283 passed step 4 and were analyzed downstream. Due to computational challenges during fine-mapping (see below), we also excluded three regions (urate 4:8165642-11717761, total bilirubin 12:19976272-22524428 and alkaline phosphatase 1:19430673-24309348) from downstream analyses (see below), leaving 13,280 trait-regions.

We used two fine-mapping methods to analyze each region:

SuSiE[33]: For each fine-mapping trait-region, for each STR and SNP and indel variant in that region that was not filtered before association testing, was not an STR-indel variants (see above) and had p-value ≤ 5e-4 (chosen to reduce computational burden), we loaded the dosages for that variant from the set of participants used in association testing for that phenotype. For those regions we also loaded the rank-inverse-normalized phenotype values and covariates corresponding to that phenotype. We separately regressed the covariates out of the phenotype values and out of each variant's dosages and streamed the residual values to HDF5 arrays using h5py v3.6.0 (**URLs**). We used rhdf5 v2.38.0 (**URLs**) to load the h5 files into R. We used an R script to run SuSiE v0.11.42 on that data with non-default values `min_abs_corr=0` and `scaled_prior_variance=0.005`. `min_abs_corr=0` forced SuSiE to output all credible sets it found so that we could determine the appropriate minimum absolute correlation filter threshold in downstream analyses. We set `scaled_prior_variance` to 0.005 which is a more realistic guess of the per-variant percentage of signal explained than the default of 20%, although we determined that this parameter had no effect on the results (**Supplementary Note 3**). The SuSiE results for some regions did not converge within the default number of iterations (100) or produced the default maximum number of credible sets (10) and all those credible sets seemed plausible (minimum pair-wise absolute correlation > 0.2 or size < 50). We reran those regions with the additional parameters `L=30` (maximum number of credible sets) and `max_iter=500`. No regions failed to converge in under 500 iterations. We re-analyzed several loci that produced 30 plausible credible sets again with `L=50`. No regions produced 50 plausible credible sets. SuSiE failed to finish for two regions (urate 4:8165642-11717761, total bilirubin 12:19976272-22524428) in under 48 hours; we excluded those regions from downstream analyses. A prior version of our pipeline had applied a custom filter to some SuSiE fine-mapping runs that caused SNPs with total minor allele dosage less than 20 across the entire population to be excluded. For consistency, any regions run with that filter which produced STRs included in our confidently fine-mapped set were rerun without that filter. Results from the rerun are reported in **Supplementary Table 4**.

SuSiE calculates credible sets for independent signals and calculates an alpha value for each variant for each signal – the probability that that variant is the causal variant in that signal. We used each variant's highest alpha value from among credible sets with purity ≥ 0.8 as its casual probability (CP) in our downstream analyses (or zero if it was in no such credible sets). See **Supplementary Note 1**.

FINEMAP[34]: We selected the STR and SNP and indel variants in each fine-mapping region that were not filtered before association testing and had p-value < 0.05 (chosen to reduce computational burden). We excluded STR-indels (see above). We constructed a FINEMAP input file for each region containing the effect size of each variant and the effect size's standard error. All `MAF` values were set to `nan` and the `ref` and `alt` columns were set to `nan` for STRs as this information is not required. We then took the unrelated participants for the phenotype, loaded their dosage genotypes for those variants and saved them to an HDF5 array with h5py v3.6.0 (**URLs**). To construct the LD input file required by FINEMAP, we computed the Pearson correlation between dosages of each pair of variants. We then ran FINEMAP v1.4 with non-default options `--sss --n-configs-top 100 --n-causal-snps 20`. In regions which FINEMAP gave non-zero probability to their being 20 causal variants, we reran FINEMAP with the option `--n-causal-snps 40` and used the results from the rerun. FINEMAP did not suggest 40 causal variants in any region. FINEMAP caused a core dump when running on the region alkaline phosphatase 1:19430673-24309348 so we excluded that region from downstream analyses. (For convenience, for the regions containing no STRs, we directly ran FINEMAP with `--n-causal-snps 40`, unless those regions contained less than 40 variants in which case we ran FINEMAP with `--n-causal-snps` <#variants>).

We used the PIP FINEMAP output for each variant in each region as its CP in downstream analyses.

Alternative Fine-mapping Conditions

We reran SuSiE and FINEMAP using alternative settings on trait-regions that contained one or more STRs with p-value ≤ 1e-10 and CP ≥ 0.8 in both the original SuSiE and FINEMAP runs. Each new run differed from the original run in exactly one condition. We restricted our set of high-confidence fine-mapped STRs (**Supplementary Table 4**) to those that had p-value ≤ 1e-10 and CP ≥ 0.8 in the original runs and maintained CP ≥ 0.8 in a selected set of those alternate conditions.

For SuSiE, we evaluated using best-guess genotypes vs. genotype dosages as input. For FINEMAP, we tested varying the p-value threshold, choice of non-major allele frequency threshold, effect size prior, number of causal variants per region, and stopping threshold.

See **Supplementary Note 3** for a more detailed discussion of these various settings and their impact on fine-mapping results.

Replication in other populations

We separated the participants not in the White British group into population groups using the self-reported ethnicities summarized by UKB showcase data field 21000 (**URLs**). This field uses UKB showcase data coding 1001. We defined the following five populations based on those codings (counts give the maximal number of unrelated QC'ed participants, ignoring per-phenotype missingness):

- Black (African and Caribbean, n=7,562, codings 4, 4001, 4002, 4003)
- South Asian (Indian, Pakistani and Bangladeshi, n=7,397, codings 3001, 3002, 3003)
- Chinese (n=1,525, coding 5)
- Irish (n=11,978, coding 1002)
- Other White (White non-Irish non-British, n=15,838, coding 1003)

Self-reported ethnicities were collected from participants at three visits (initial assessment, repeat assessment, first imaging). The above groups also exclude participants who self-reported ethnicity at more than one visit and where their answers corresponded to more than one population (after ignoring 'prefer not to answer' `code=-3` responses). We did include any participants who were neither in the White British population nor any of the above populations. Unlike for the determination of White British participants, genetic principal components were not used as filters for these categories.

For the association tests in these populations we applied the same procedures for sample quality control, unrelatedness filtering, phenotype transformations, and preparing genotypes and covariates as in the White British group. The only changes in procedure were that (a) we removed categorical covariate values where there were fewer than 50 participants with that value, (in which case we also removed those participants from analysis, as that would be too few to properly control for batch effects), whereas for White British individuals we used a cutoff of 0.1% instead, (b) we also applied this cutoff to the visit of measurement categorical covariate, resulting in some association tests that excluded individuals whose first measurement of the phenotype occurred

outside the initial assessment visit and (c) we included the aliquot covariate for LDL cholesterol and total bilirubin, which had been excluded in our initial run in White British (see above). See **Supplementary Table 5** for details.

STRs were marked as replicating in another population (**Fig. 2a**) if any of the traits confidently fine-mapped to that STR share the same direction of effect as the White British association and reached association p-value ≤ 0.05 after multiple hypothesis correction (i.e. if there are three confidently fine-mapped traits, then an STR is marked as replicating in the Black population if any of them has association p-value ≤ 0.05/3 = 0.0167 in the Black population).

Logistic regression analysis of replication direction

We used logistic regression to quantitatively assess the impact of fine-mapping on replication rates while controlling for discovery p-value. For this analysis, to have sufficient sample sizes, we defined that an STR-trait association replicates in another population if it had the same direction of effect in that population as in the White British population, regardless of the replication p-value.

For each of the five replication populations, we compared four categories: all gwsig (genome-wide significant associations in the discovery population, i.e. p-value ≤ 5e-8), FINEMAP (discovery p-value ≤ 5e-8 and FINEMAP CP ≥ 0.8), SuSiE (discovery p-value ≤ 5e-8 and SuSiE CP ≥ 0.8) and confidently fine-mapped STR (STR associations in our confidently fine-mapped set).

For each comparison, we used the function `statsmodels.formula.api.logit` from statsmodels v0.13.2 (**URLs**) to fit the logistic regression model:

$$\texttt{replication\_status} \sim \texttt{STR\_in\_target\_category} + \log_{10}(\text{p-val}) + \log_{10}(\text{p-val})^2$$

where `replication_status` is a binary variable indicating whether or not the given STR-trait association replicated in the other population, `p-val` is the discovery p-value, and `STR_in_target_catgegory` is a binary variable indicating if the STR is in the target category.

For each replication population, we considered various models:

- All gwsig STRs with either FINEMAP, SuSiE, or confidently fine-mapped STRs as the target category.

- All FINEMAP STRs with confidently fine-mapped STRs as the target category.
- All SuSiE STRs with confidently fine-mapped STRs as the target category.

For each model, we performed a one-sided t-test for the hypothesis that the coefficient for the covariate `STR_in_target_category` was greater than zero, i.e. testing that being in the target category increased the predicted chance of replicating in the chosen population (**Supplementary Table 6**).

Gene and transcription factor binding annotations

For all analyses not using GTEx data, gene annotations were based on GENCODE 38[78] (**URLs**). Transcription factor binding sites identified by ENCODE[79] overlapping several loci (*TAOK1*, *RHOT1* and *NCK2*) were identified through visual inspection of the "Txn Factor ChIP" track in the UCSC Genome Browser[80] and using the "Load from ENCODE" feature of the Integrative Genomics Viewer[81].

Enrichment testing

We tested the following categories for enrichment in STRs identified by our association testing pipeline:

- Genomic feature: We grouped records by feature type and restricted to features with support level 1 or 2 except for genes which don't have a support level. We used bedtools[82] to compute which features intersect each STR and the distance between each STR and the nearest feature of each feature type.
- Repeat unit: unit length and standardized repeat unit were defined as described above. Repeat units occurring in <1000 STRs were grouped by repeat length. Repeats whose unit could not be determined were considered as a separate category.
- Overlap with expression STRs (eSTR): we tested for overlap with either all eSTRs or fine-mapped eSTRs as defined in our previous study to identify STR-gene expression associations in the Genotype Tissue Expression (GTEx) cohort[42].

Enrichment p-values were computed using a Chi-squared test (without Yate's continuity correction) if all cells had counts ≥ 5. A two-sided Fisher's exact test was used otherwise. Chi-squared and Fisher's exact tests were implemented using the `chi2_contingency` and `fisher_exact` functions from the Python scipy.stats package v1.7.3 (**URLs**).

Expression association analysis in GTEx

We had previously analyzed associations[42] between STRs and gene expression in GTEx V7. Here we reanalyzed those associations using GTEx V8. We obtained 30x Illumina whole genome sequencing (WGS) data from 652 unrelated participants in the Genotype-Tissue Expression project (GTEx)[35] through dbGaP accession number phs000424.v8.p2. WGS data was accessed using fusera (**URLs**) through Amazon Web Services. We genotyped STRs using HipSTR[24] v0.5 with HipSTR's hg38 reference STR set (**URLs**). All individuals were genotyped jointly using default parameters. GTEx's whole genome sequencing procedure is not PCR-free, which likely contributed to low call rates at long poly-A and GC-rich STRs. The resulting VCFs were filtered using DumpSTR from TRTools[76], using the parameters `--filter-hrun --hipstr-min-call-Q 0.9 --hipstr-min-call-DP 10 --hipstr-max-call-DP 1000 --hipstr-max-call-flank-indel 0.15 --hipstr-max-call-stutter 0.15 --min-locus-callrate 0.8 --min-locus-hwep 0.00001` . We also removed STRs overlapping segmental duplication regions (UCSC Genome Browser[83] h38.genomicSuperDups table). Altogether, 728,090 STRs remained for downstream analysis.

For each tissue, we obtained gene-level and transcript-level transcripts-per-million (TPM) values, exon-exon junction read counts, and exon read counts for each participant from GTEx Analysis V8 publicly available from the GTEx project website (**URLs**). Gene annotations are based on GENCODE v26[78]. We focused on 41 tissues with expression data for at least 100 samples (**Supplementary Table 9**). We restricted our analysis to protein-coding genes, transcripts and exons that did not overlap segmental duplication regions.

To control for population structure, we obtained publicly available genotype data on 2,504 unrelated individuals from the 1000 Genomes project[21] genotyped with Omni 2.5 SNP genotyping arrays. We performed the following principal components analysis jointly on that data and the SNP genotypes based on WGS of the 652 individuals above. We removed all indels, multiallelic SNPs, and SNPs with minor allele frequency less than 5%. We then used plink v.1.90b3.44 to subset these remaining SNPs to a set of SNPs in approximate linkage equilibrium with the command `--indep 50 5 2`. We excluded any remaining SNPs with missingness rate 5% or greater. We lastly ran principal component analysis using smartpca70[84] v.13050 with default parameters.

We removed genes with TPM less than 1 in more than 90 percent of individuals. PEER factors[85] were calculated using PEER v1.0 from the TPM values which remained after filtering. For each gene, we tested for association with each STR within 100kb. For each test we performed a linear regression between the STR's dosage (sum of allele lengths) and gene expression (TPM). We included the loadings of the top five genotype principal components as computed above and the top N/10 PEER factors as covariates. The number of PEER factors was chosen to maximize the number of significant associations across a range of tissues. We did not include sex or age as covariates.

For each STR we computed Bonferroni-adjusted p-values to control for the number of gene × tissue tests performed. Associations that remained with adjusted $p \leq 0.05$ are shown in **Supplementary Table 8**.

We additionally used the GTEx cohort to test for an association between length of the bilirubin-associated dinucleotide repeat identified in *SLC2A2* with splicing efficiency in liver. We obtained exon-exon junction read counts and exon read counts from the GTEx website (**URLs**). We calculated the percent spliced in (PSI) value for each exon in the manner suggested by Schafer et al.[86]. We performed a linear regression to test between the STR's dosage and PSI of each exon within 10kb, using the top 5 ancestry principal components as covariates.

Expression analysis of the CBL *and* RHOT1 *STRs in Geuvadis*

We applied HipSTR[24] v0.6.2 to genotype STRs from HipSTR's hg38 reference STR set (**URLs**) in 2,504 individuals from the 1000 Genomes Project[87] for which high-coverage WGS data was available. Gene-level reads per kilobase per million reads (RPKM) values based on RNA-seq in lymphoblastoid cell lines for 462 1000 Genomes participants were downloaded from the Geuvadis website (**URLs**). Of these, 449 individuals were genotyped by HipSTR.

Similar to the GTEx analysis, we performed a linear regression between STR dosage (sum of allele lengths) and RPKM, adjusting for the top 5 genotype principal components (computed as above for the GTEx analysis, but only on populations included in Geuvadis and separately for Europeans and Africans) and N/10 (45) PEER factors as covariates. PEER analysis was applied using PEER v1.0 to the matrix of RPKM values after removing genes overlapping segmental duplications and those with RPKM less than 1 in more than 90% of LCL samples. We performed a separate regression analysis for African individuals (YRI) and European individuals (CEU, TSI, FIN, and GBR). After restricting to individuals with non-missing expression data and STR

genotypes and who were not filtered as PCA outliers by smartpca[84,88] included in EIGENSOFT v6.1.4, 447 LCL samples remained for analysis in each case (num. EUR=358, and AFR=89 for *CBL*, EUR=359 and AFR=88 for *RHOT1*).

URLs

- 1000 genomes individuals: https://www.internationalgenome.org/data-portal/sample using the "Download the list" tab
- associaTR: https://github.com/LiterallyUniqueLogin/ukbiobank_strs/
- Beagle Human genetic maps: https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/
- fusera: https://github.com/ncbi/fusera
- GENCODE 38 (hg19): http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh37_mapping/gencode.v38lift37.annotation.gff3.gz
- Geuvadis: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/files/analysis_results/?ref=E-GEUV-1
- GTEx v8:
  - https://www.gtexportal.org/home/datasets
  - https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz
  - https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_STARv2.5.3a_junctions.gct.gz
  - https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_exon_reads.parquet
- h5py https://github.com/h5py/h5py
- HDF5: https://www.hdfgroup.org/HDF5/
- HipSTR STR reference https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/hg38.hipstr_reference.bed.gz
- Pan-UKB:
  - Overview: https://pan.ukbb.broadinstitute.org/downloads
  - manifest: https://docs.google.com/spreadsheets/d/1AeeADtT0U1AukliiNyiVzVRdLYPkTbruQSk38DeutU8
  - bilirubin SNP summary statistics: https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/biomarkers-30840-both_sexes-irnt.tsv.bgz and https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files_tabix/biomarkers-30840-both_sexes-irnt.tsv.bgz.tbi

- Plink association testing best practices: https://www.cog-genomics.org/plink/2.0/assoc#glm
- rhdf5: https://www.bioconductor.org/packages/release/bioc/html/rhdf5.html
- Scipy.stats: https://docs.scipy.org/doc/scipy/reference/stats.html
- SNP-STR reference panel: https://gymreklab.com/2018/03/05/snpstr_imputation.html
- Statsmodels: https://www.statsmodels.org/stable/index.html
- UKB Data Showcase Search Page: https://biobank.ctsu.ox.ac.uk/crystal/search.cgi

Acknowledgments

Research reported in this publication was supported in part by NIH/NHGRI grants R01HG010885 (M.G. and A.G.) and 1RM1HG011558 (M.G.). This research has been conducted using the UK Biobank Resource under application number 46122. We thank R. Wachs for helping with illustrations. We also thank K. Frazer and M. D'Antonio for helpful discussions and comments.

Author contributions

J.M. led, designed and performed the analyses and wrote the manuscript. S.F. helped oversee physiological interpretation of individual signals. A.M. assisted with analysis of the *APOB* locus. Y.L. performed expression analyses of the GTEx data. A.G. and M.G. conceived the study, supervised analyses, and wrote the manuscript. All authors read and approved this manuscript.

Competing interests

The authors have no competing financial interests to declare.

Code availability

The associaTR tool and code for generating the figures in this paper can be found at: https://github.com/LiterallyUniqueLogin/ukbiobank_strs/

References cited

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

2. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

3. Mukamel, R. E. *et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes*. 2021.01.19.427332 https://www.biorxiv.org/content/10.1101/2021.01.19.427332v1 (2021) doi:10.1101/2021.01.19.427332.

4. Grünewald, T. G. P. *et al.* Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. *Nat. Genet.* **47**, 1073–1078 (2015).

5. Song, J. H. T., Lowe, C. B. & Kingsley, D. M. Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *Am. J. Hum. Genet.* **103**, 421–430 (2018).

6. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).

7. Boettger, L. M. *et al.* Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48**, 359–366 (2016).

8. Leffler, E. M. *et al.* Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**, eaam6393 (2017).

9. Willems, T. *et al.* The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).

10. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).

11. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).

12. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci.* **107**, 961–968 (2010).

13. Malik, I., Kelley, C. P., Wang, E. T. & Todd, P. K. Molecular mechanisms underlying nucleotide repeat expansion disorders. *Nat. Rev. Mol. Cell Biol.* **22**, 589–607 (2021).

14. Quilez, J. *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44**, 3750–3762 (2016).

15. Tw, H., Jd, G., Ce, Y. & Gr, C. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc. Natl. Acad. Sci. U. S. A.* **101**, (2004).

16. Hui, J., Stangl, K., Lane, W. S. & Bindereif, A. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat. Struct. Biol.* **10**, 33–37 (2003).

17. Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–1216 (2009).

18. Murat, P., Guilbaud, G. & Sale, J. E. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biol.* **21**, 209 (2020).

19. Rothenburg, S., Koch-Nolte, F., Rich, A. & Haag, F. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl. Acad. Sci.* **98**, 8985–8990 (2001).

20. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

21. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

22. Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 1–13 (2018).

23. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).

24. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).

25. Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

26. Tang, H. *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).

27. Tankard, R. M. *et al.* Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *Am. J. Hum. Genet.* **103**, 858–873 (2018).

28. Saini, S., Mitra, I., Mousavi, N., Fotsing, S. F. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).

29. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

30. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

31. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

32. Pan-UKB team. Pan-ancestry genetic analysis of the UK Biobank. https://pan.ukbb.broadinstitute.org/ (2020).

33. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

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

35. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

36. Berberich, A. J. & Hegele, R. A. A Modern Approach to Dyslipidemia. *Endocr. Rev.* bnab037 (2021) doi:10.1210/endrev/bnab037.

37. Boerwinkle, E. & Chan, L. A three codon insertion/deletion polymorphism in the signal peptide region of the human apolipoprotein B (APOB) gene directly typed by the polymerase chain reaction. *Nucleic Acids Res.* **17**, 4003 (1989).

38. Niu, C. *et al.* Associations of the APOB rs693 and rs17240441 polymorphisms with plasma APOB and lipid levels: a meta-analysis. *Lipids Health Dis.* **16**, 166 (2017).

39. Riches, F. M. *et al.* Apolipoprotein B signal peptide and apolipoprotein E genotypes as determinants of the hepatic secretion of VLDL apoB in obese men. *J. Lipid Res.* **39**, 1752–1758 (1998).

40. Hui, J. *et al.* Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.* **24**, 1988–1998 (2005).

41. Huang, Y. *et al.* Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Mol. Cell* **45**, 459–469 (2012).

42. Fotsing, S. F. *et al.* The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019).

43. Sutcliffe, J. S. *et al.* DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum. Mol. Genet.* **1**, 397–400 (1992).

44. Garg, P. *et al.* A Survey of Rare Epigenetic Variation in 23,116 Human Genomes Identifies Disease-Relevant Epivariations and CGG Expansions. *Am. J. Hum. Genet.* **107**, 654–669 (2020).

45. Jones, C. *et al.* Co-localisation of CCG repeats and chromosome deletion breakpoints in Jacobsen syndrome: evidence for a common mechanism of chromosome breakage. *Hum. Mol. Genet.* **9**, 1201–1208 (2000).

46. Jones, C. *et al.* Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene CBL2. *Nature* **376**, 145–149 (1995).

47. Laleye, A. *et al.* Giant platelets in a case of deletion 11q24-qter confirmed by fluorescence in situ hybridization. *Am. J. Med. Genet.* **110**, 170–175 (2002).

48. Plo, I. *et al.* Genetic Alterations of the Thrombopoietin/MPL/JAK2 Axis Impacting Megakaryopoiesis. *Front. Endocrinol.* **8**, 234 (2017).

49. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

50. Kotzin, J. J. *et al.* The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. *Nature* **537**, 239–243 (2016).

51. Bouillet, P. *et al.* Proapoptotic Bcl-2 Relative Bim Required for Certain Apoptotic Responses, Leukocyte Homeostasis, and to Preclude Autoimmunity. *Science* **286**, 1735–1738 (1999).

52. Ogilvy, S. *et al.* Constitutive Bcl-2 expression throughout the hematopoietic compartment affects multiple lineages and enhances progenitor cell survival. *Proc. Natl. Acad. Sci.* **96**, 14943–14948 (1999).

53. Draviam, V. M. *et al.* A functional genomic screen identifies a role for TAO1 kinase in spindle-checkpoint signalling. *Nat. Cell Biol.* **9**, 556–564 (2007).

54. Favier, R. & Raslova, H. Progress in understanding the diagnosis and molecular genetics of macrothrombocytopenias. *Br. J. Haematol.* **170**, 626–639 (2015).

55. Azad, P., Villafuerte, F. C., Bermudez, D., Patel, G. & Haddad, G. G. Protective role of estrogen against excessive erythrocytosis in Monge's disease. *Exp. Mol. Med.* **53**, 125–135 (2021).

56. Mukundan, H., Resta, T. C. & Kanagy, N. L. 17β-Estradiol decreases hypoxic induction of erythropoietin gene expression. *Am. J. Physiol.-Regul. Integr. Comp. Physiol.* **283**, R496–R504 (2002).

57. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).

58. D'Antonio, M. *et al.* Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. *eLife* **8**, e48476 (2019).

59. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).

60. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).

61. Press, M. O., Carlson, K. D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* **30**, 504–512 (2014).

62. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

63. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

64. Foix, Anna & Blachly, James. pyEGA3: EGA download client. (2021).

65. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).

66. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).

67. Horta, D. bgen-reader: Bgen file format reader.

68. Staples, J., Nickerson, D. A. & Below, J. E. Utilizing Graph Theory to Select the Largest Set of Unrelated Individuals for Genetic Analysis. *Genet. Epidemiol.* **37**, 136–141 (2013).

69. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behav. Genet.* **39**, 580–595 (2009).

70. Bishara, A. J. & Hittner, J. B. Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol. Methods* **17**, 399–417 (2012).

71. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. *PLOS ONE* **9**, e85150 (2014).

72. Bishara, A. J. & Hittner, J. B. Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality. *Educ. Psychol. Meas.* **75**, 785–804 (2015).

73. Auer, P. L., Reiner, A. P. & Leal, S. M. The effect of phenotypic outliers and non-normality on rare-variant association testing. *Eur. J. Hum. Genet.* **24**, 1188–1194 (2016).

74. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

75. Pedersen, B. cyvcf2: fast vcf parsing with cython + htslib.

76. Mousavi, N. *et al.* TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37**, 731–733 (2021).

77. Zheng, J., Li, Y., Abecasis, G. R. & Scheet, P. A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. *Genet. Epidemiol.* **35**, 102–110 (2011).

78. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

79. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

80. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

81. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

83. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).

84. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

85. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

86. Schafer, S. *et al.* Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). *Curr. Protoc. Hum. Genet.* **87**, 11.16.1-11.16.14 (2015).

87. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. 2021.02.06.430068 Preprint at https://doi.org/10.1101/2021.02.06.430068 (2021).

88. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet.* **2**, e190 (2006).

Supplementary Notes

Supplementary Note 1: Summary of fine-mapping models

We applied two different fine-mapping methods, SuSiE[1] v0.11.42 and FINEMAP[2] v1.4. FINEMAP assumes a priori that each variant has an equal chance of being causal and that each variant's chance of causality is independent from the causal status of the other variants. It then attempts to stochastically walk over all reasonably-probable choices of collections of causal variants and assigns each causal configuration a posterior probability based on the observed associations and that prior. It then calculates posterior inclusion probabilities (PIPs) by summing over the walked configurations. For downstream analyses we required a single measurement of causality for each variable from both fine-mappers, which we called those variables' causal probabilities (CPs). For FINEMAP, we took each variable's PIP to be its FINEMAP CP.

While FINEMAP models each region as a collection of causal variants, SuSiE models each region as a collection of causal signals (called effects in the SuSiE manuscript), enabling SuSiE to study variants' contributions to each signal separately. To fit this model, SuSiE alternates between updating its model of each signal, attempting with each update to improve how the collection of all signals fits the observed data. As SuSiE only allows for the possibility of one variant being considered causal in any signal, if two variants are both estimated to be causal, they are forced during model fitting into different signals from one another. SuSiE calculates a value, alpha, for each variant in each signal – the probability that variant causes that signal – and then calculates a single PIP for each variant which gives the probability that the variant is causal in at least one signal. For reasons we explain below, unlike for FINEMAP, we chose an alpha value (or zero) as the SuSiE CP for each variant, rather than a PIP.

SuSiE reports a purity value for each signal, and we used that value to discard signals which were not well fine-mapped. SuSiE constructs 90%-credible sets for each signal so that the estimated probability of the credible set containing a variant causal for that signal is at least 90% (other values, such as 95%-credible sets, could be constructed similarly). SuSiE defines the purity of the credible set for each signal to be the minimum absolute correlation between any pair of variants in the set. The SuSiE manuscript suggests discarding signals with purity less than 0.5, but also states that the threshold is arbitrary. Looking at the distribution of credible set purities across all of our trait-regions (**Supplementary Fig. 3**) we decided to discard credible sets with purity less than 0.8, reasoning that the upper mode of the distribution is well above that threshold

and that a signal containing two variants with correlation less than 0.8 has not been acceptably resolved.

SuSiE's PIPs are calculated across all credible sets regardless of purity, while we wished to conservatively only consider variants which had passed this added layer of scrutiny. Additionally, we saw that the PIP metric is sensitive to values of `L` (the number of signals fit per locus) – for one extreme example, in a locus with 57 variants, SuSiE run with `L=50` assigned each variant a PIP ≥ 0.5, which is likely unrealistic. So instead of using SuSiE's PIPs, we took each variant's highest alpha score from among credible sets with purity at least 0.8 as its SuSiE CP (or zero if it was in no such credible sets). This choice was uniformly conservative; CPs defined this way must be less than SuSiE's PIPs. We also found it to be less sensitive to `L` – we examine this more thoroughly in **Supplementary Note 3** below, but in the above example we note that there was only one credible set containing less than 50 variants and it was not pure, so each variant in that region has a SuSiE CP of 0. We compared our SuSiE CP metric to SuSiE's PIP metric in **Supplementary Fig. 4** and saw that these two measures only strongly differed for variants whose contribution to any single pure signal was small. As our downstream analyses focused on variants with large alpha values in pure credible sets, this means that our use of alpha values instead of PIPs was not strongly impactful in analyzing those variants. The impact is that we conservatively restricted which variants we examined. Lastly, we note that for high purity thresholds such as the one we use, our metric should be very similar to calling SuSiE's `susie_get_pip` function with the flag `prune_by_cs=TRUE`, a method not examined in the SuSiE manuscript and one we did not encounter until after performing this work.

Supplementary Note 2: Comparing results across fine-mapping methods

To assess the reliability of our fine-mapping results, we measured how often the two fine-mapping methods agreed with one another, and how sensitive they were to model settings. First, we used SuSiE's credible sets as a proxy for the truly independent signals in our data. We observed that while SuSiE and FINEMAP were in agreement for most of the signals, their results were strongly discordant for a sizable number of signals (**Supplementary Fig. 6**). In particular, for 10.4% of 90%-credible sets returned by SuSiE (which by definition are assigned at least a 90% chance of containing a causal variant), the sum of FINEMAP's assigned CPs for all variants in each of those sets was less than 0.1, indicating that FINEMAP concluded those sets had a < 10% chance of containing a causal variant.

Second, we looked at the variant level and saw that for most variants, the CPs from FINEMAP and SuSiE were similar (**Supplementary Fig. 7**), with FINEMAP assigning slightly higher CPs overall (possibly due to our use of SuSiE alpha values per variant instead of the overall PIPs). However, we again saw that SuSiE and FINEMAP markedly disagree at a subset of loci. For instance, among all SNPs and indels which at least one fine-mapping method assigned a CP ≥ 0.95 and the other method was decisive about their causality (assigning either CP ≥ 0.95 or CP ≤ 0.05), 20% of those were assigned a CP ≥ 0.95 by one method and a CP ≤ 0.05 by the other. For STRs, the fine-mapping methods disagreed at more than half of the loci (58%) that were assigned CP ≥ 0.95 by one method and decisively scored by the other, suggesting the CPs for STRs are even less reliable. This highlights the need for additional quality control before stating that variants assigned a high posterior probability by a single fine-mapper are likely to be causal. Without any prior on which fine-mapper to believe when the two disagreed, we focused only on the 177 trait-STR associations for which association p-values were well below the genome-wide significance threshold (p-value≤1e-10) and both fine-mappers assigned high CPs (CPs≥0.8) (**Supplementary Fig. 8a;** the 177 associations can be extracted from **Supplementary Table 3**).

Supplementary Note 3: Assessing robustness of fine-mapping results

We further assessed how robust our fine-mapping results were to differences in the fine-mapping conditions, data filtering thresholds and algorithm metaparameters used. For SuSiE, we modified the inputs (1) `scaled_prior_variance`, (2) `tol`, (3) `residual_variance`, and (4) `L`, and also (5) changed the input genotypes from dosage genotypes to best-guess genotypes and (6) changed the prior to favor SNPs and indels over STRs as causal variants. For FINEMAP, we modified the inputs (1) `--prior-std` and (2) `--prob-conv-sss-tol` and also (3) filtered input variants with total non-major allele dosage less than 100, (4) filtered variants with p-value ≤ 5e-4, (5) set the prior on the number of causal variants per region to 4, and (6) changed the prior to favor SNPs and indels over STRs as causal variants.

We were encouraged that a few of the SuSiE settings had minimal impact on the results. Specifically, we tested the following changes on a subset of mean platelet volume fine-mapping regions:

- `scaled_prior_variance` – This is the initial value for the estimation of the prior variance of the causal effect sizes relative to the variance of the phenotype. We changed this from the default of 0.2 to 5e-4 which resulted in no change to observed CPs.

- `tol` – This determines what amount of change in the objective function between optimization rounds is small enough to cause SuSiE to terminate. We reduced this from the default of 1e-3 to 1e-4 and saw only miniscule changes in the results (**Supplementary Fig. 9a**).
- `residual_variance` – This is the initial value for the estimation of the residual variance of the phenotype after controlling for all effects at the locus. By default, the `residual_variance` is initialized to the full variance of the phenotype, which in our study was slightly less than 1 (rank-inverse normalization set it to 1, and then regressing covariates out of the phenotype before running SuSiE reduced it slightly). We ran SuSiE with alternate `residual_variance` values of 0.95 and 0.8 and saw very small changes in the results (**Supplementary Fig. 9b,c**)
- `L` – This is the number of signals SuSiE fits in a region, or equivalently, the upper bound on the number of causal variants SuSiE attempts to find (**Supplementary Fig. 9d**). In our original fine-mapping runs, we ran SuSiE with `L=10`, and only increased `L` if needed. In the comparison below, we ran SuSiE with `L=50`. The SuSiE manuscript[1] states that inflated `L` values should not adversely impact model fitting because extraneous signals contribute small probabilities dispersed over many variants, thus not strongly changing any single variant's prediction, and also the learned effect sizes of these extra signals are shrunk towards zero. We see in our comparison that this only induces a large change in CP for a small fraction of variants. Of those, almost all of them are variants with non-zero CP values under the `L=10` case and zero CP in the `L=50` case. Thus, if they have any effect, this indicates that in most cases inflated values of `L` should lead to more conservative fine-mapping results.

However, many of the fine-mapping conditions (individually documented below) did impact the end results. We ran fine-mapping under each of those conditions on the trait-regions of the 177 STR-trait associations above. We present supplementary figures showing how the CPs of variants changed under those conditions (**Supplementary Figs. 10, 11a-e**). Because we would expect true signals to be robust to these choices, we restricted our set of confidently fine-mapped STRs to the 118 that had CP ≥ 0.8 under each of those conditions (**Supplementary Table 4**). While the set of trait-regions used for running these tests was chosen to identify candidate causal STRs, **Supplementary Figs 10-11e** identify similar trends for SNPs and indels in those regions. Thus, we hypothesize that these comparisons are relevant for fine-mapping of all variant types.

SuSiE with best-guess genotypes vs dosage genotypes

We ran SuSiE with the best-guess genotypes from our imputation pipeline instead of the dosage genotypes from that pipeline (**Supplementary Fig. 10**). Discrepancies in best-guess vs. dosages

reflect imputation uncertainty. As we did not have ground truth STR genotypes, we could not resolve these discrepancies and thus discarded loci where this choice strongly impacted the results.

Note: We ran SuSiE on each best-guess trait region with the parameters `L=50` and `max_iter=500`. While these are different values than those for the baseline runs, the results should still be comparable:

- Both the dosage and best-guess SuSiE runs converged in fewer iterations than their respective `max_iter` values, and larger values of `max_iter` than the convergence number should not affect the results.
- We discuss above why overestimating L should not importantly impact the results.

Additionally, as noted in the **Methods**, baseline SuSiE runs for some trait-regions were run with the dosage < 20 SNP filter while others were not. To control for this, we ran the SuSiE best-guess comparison with the same set of variants as the original runs in each trait-region.

For the FINEMAP comparisons below, as in our baseline runs, we ran each trait-region with `--n-causal-snps 20`, and then reran it with `--n-causal-snps 40` if that run's results included a possibility of at least 20 causal variants.

FINEMAP with alternative p-value thresholds

By default, we chose to filter as few variants as possible from our fine-mapping runs while still controlling for computational costs, which meant filtering variants with p>5e-2 from our FINEMAP runs and variants with p>5e-4 from our SuSiE runs, as FINEMAP was less computationally intensive. To check if this difference impacted the fine-mappers' results we ran FINEMAP having filtered all variants with p>5e-4 and compared it to our default FINEMAP runs (**Supplementary Fig. 11a**). Unexpectedly, this change strongly impacted the CPs of some variants. This CP difference occurred despite the large difference between the p-values of the impacted variants and the p-values of the omitted variants.

FINEMAP with alternative choice of non-major allele frequency threshold

To test whether FINEMAP results were strongly influenced by rare variants, we excluded all variants with total non-major allele dosage < 100 (population frequency less than approximately 0.015%) on top of the filter excluding variants with p-value > 0.05 (**Supplementary Fig. 11b**).

(Note that variants with total non-major allele dosage < 20 were excluded from association testing and thus from all fine-mapping runs).

FINEMAP with alternative choice of effect size prior

FINEMAP's default `--prior-std` value is 0.05 which gives causal variants a default effect size of 0.25% of phenotypic variance. We modified this to `--prior-std 0.0224` to reflect published expected effect sizes for GWAS variants of about 0.05%[3] (**Supplementary Fig. 11c**).

FINEMAP with alternative prior on the number of causal variants per region

We ran FINEMAP with the prior of four causal variants per trait-region instead of one (**Supplementary Fig. 11d**). We did this by adding a column prob to the FINEMAP input Z file which contained the value 4/n for each variant, where n was the number of variants in the trait-region, and by running FINEMAP with the `--prior-snps` flag.

FINEMAP with alternative `--prob-conv-sss-tol` *stopping threshold*

We ran FINEMAP with the flag `--prob-conv-sss-tol` 0.0001 (reduced from the default of 0.001) (**Supplementary Fig. 11e**). This reduced what amount of change in the objective function over the last 100 rounds of optimization would be considered small enough to cause FINEMAP to terminate.

In summary, the dosages vs best-guess genotypes choice when running SuSiE, and the FINEMAP p-value threshold setting were strongly impactful. The FINEMAP threshold on non-major allele frequency, effect size prior, and prior on number of causal variants per region were moderately impactful settings. And the FINEMAP stopping threshold setting was minorly impactful. Overall, about a third of results that passed both fine-mappers failed to replicate in one of the alternate fine-mapping conditions above, again highlighting the need for careful inspection of fine-mapping settings prior to result interpretation. Encouragingly, for many of these comparisons we see that the default SuSiE and FINEMAP runs were more likely to agree that variants were causal (both CPs ≥ 0.95) for those variants that the alternate fine-mapping condition also agreed were causal. This suggests that concordance between different fine-mapping algorithms may be able to provide security against the instability in the results of any single algorithm. While we focused on fine-mapping results for STRs, which generally showed lower

concordance across methods than SNPs, our results suggest similar robustness checks should be performed when fine-mapping SNPs and other variant types.

There were several fine-mapping conditions we tested that strongly impacted the resulting CPs but that we did not use as filters when selecting our causal STR candidates since they represent unrealistic parameter choices. We report their values in **Supplementary Table 3**. Those conditions were:

- We ran FINEMAP with the flag `--prior-std 0.005`, corresponding to an expected effect size 0.0025% (**Supplementary Fig. 11f**). We concluded that this was much lower than the effect sizes we were hoping to detect.
- Both SuSiE and FINEMAP have the default assumption that each variant is as likely to be causal as any other variant (regardless of allele frequency). We instead conservatively ran SuSiE and FINEMAP with the prior assumption that SNPs and indels were 4x more likely to be causal than STRs. For this, we set the prior probability of causality for each SNP or indel to 4/(4*n_SNPs_indels + n_STRs) and for each STR to 1/(4*n_SNPs_indels + n_STRs). For SuSiE we did this by setting the `prior_weights` input to an array containing those probabilities. For FINEMAP we did this by adding a column `prob` to the FINEMAP input Z file which contained those probabilities, and by running FINEMAP with the `--prior-snps` flag. As expected, this resulted in overall decreased STR CPs (**Supplementary Fig. 12**). While we did not filter our candidate STRs based on this setting, we were encouraged to see that a majority of the strongest hits replicated despite this conservative setting.

Finally, we note there are other parameters which were not tested here but that could be tested for robustness. This includes whether FINEMAP results are sensitive to overestimating `--n-causal-snps` and testing if fine-mapping results are sensitive to the size of the trait-regions being fine-mapped.

Supplementary Note 4: Additional details for specific fine-mapped STRs

Coding trinucleotide repeat in APOB: This repeat did not initially appear in our list of confidently fine-mapped STRs due to our process for filtering indels. Our pipeline filtered "STR-indels" (**Methods**), which we defined as indels in the UKB dataset corresponding to differences in STR length. We did not filter an indel imputed by the UKB team that corresponded exactly to the short allele of the STR imputed from our reference panel, since the indel consists of an imperfect repeat sequence (GCCAGCAGC for a CAG repeat). The presence of this indel alongside the STR during fine-mapping caused SuSiE's (but not FINEMAP's) results to, in some cases, show low confidence as to which of the two variants were causal. Specifically, FINEMAP assigned a CP of 1 to the STR for both traits apolipoprotein B and LDL cholesterol under each FINEMAP run used for filtering down to the confidently fine-mapped set. For the original run for the apolipoprotein B trait and for the best-guess run for both traits, SuSiE created a credible set containing both the indel and the STR and assigned each a CP of less than 0.8, causing the association not to pass our filters for confidently fine-mapped STRs. However, if we sum the SuSiE CPs of both variants in those runs we get a CP of over 0.97 in each case, making the apolipoprotein B association pass our confidently fine-mapped thresholds. Thus, we added this association to our confidently fine-mapped set. We note that the original SuSiE run for the LDL trait assigned low CPs to both the STR and the indel. While that was the only fine-mapping of the eight runs used for filtering that did not assign the pair of variants a combined CP ≥ 0.8 for LDL, it precludes us from adding the LDL association to the confidently fine-mapped set. For both the apolipoprotein B and LDL cholesterol associations, we updated the CPs in **Supplementary Tables 3** and **4** to reflect the combined CPs for both variants.

While we manually resolved this issue for the *APOB* STR, similar issues are likely to have caused other STRs in our set not to fine-map appropriately. We expect the choice of which indel representations to filter and which to treat as distinct variants will be critical for proper analysis of many STR loci in the future.

Dinucleotide repeat in SLC2A2 *(*GLUT2*)*: We identified a dinucleotide repeat immediately upstream of exon 4 of *SLC2A2* as a confidently fine-mapped STR for bilirubin. While *SLC2A2* has not previously been causally linked to bilirubin levels, *SLC2A2* mediates glucose transport to hepatocytes, where glucose is stored in the form of glycogen. Glycogen degradation produces intermediates that are substrates in the process that regulates bilirubin conjugation and excretion[4] and thus could potentially impact bilirubin levels in the blood. This effect of *SLC2A2* on bilirubin

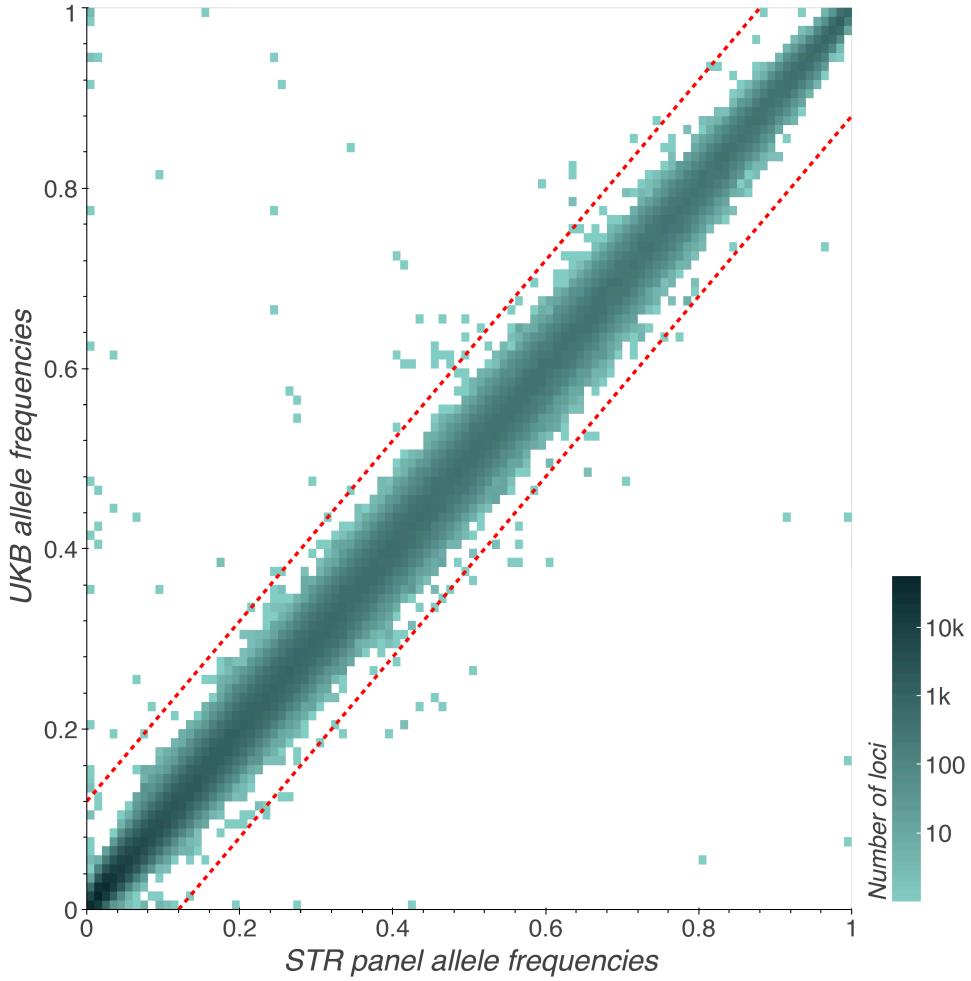
levels may be partially corroborated by a large cohort study on babies with congenital hyperinsulinemic hypoglycemia, a condition that inhibits glycogen breakdown, which reported elevated bilirubin in that population[5].

Tetranucleotide repeat in ESR2: We identified a GTTT repeat in an intron of *ESR2* whose length is negatively associated with haemoglobin concentration, red blood cell count, and haematocrit. *ESR2* is known to regulate red blood cell production. Studies conducted in populations chronically exposed to high altitude hypoxia, a driver of erythrocytosis (excess red blood cell production), demonstrated inhibition of erythrocytosis through activation of estrogen beta signaling in ex vivo models[6]. These observations are corroborated by a study of rat models under hypoxia, where beta-estrogen treatment reduced circulating levels of erythropoietin, a kidney-derived factor that stimulates red blood cell production[7].

We additionally identified a negative association between length of this STR and *ESR2* expression. However, the expected direction of this association is unclear. Multiple *ESR2* isoforms exist, either as a result of alternative splicing of the last coding exons (exon 8 and exon 9, respectively), deletion of one or more coding exons, or alternative usage of untranslated exons in the 5′ region[8]. One of the five isoforms found in humans has an undetectable affinity to estrogen. Rather, ESR2 antagonizes estrogen-alpha receptor signaling[9]**.** Thus, the tissue-specific effect of *ESR2* expression on estrogen-receptor signaling depends on the dominant isoform in that tissue.

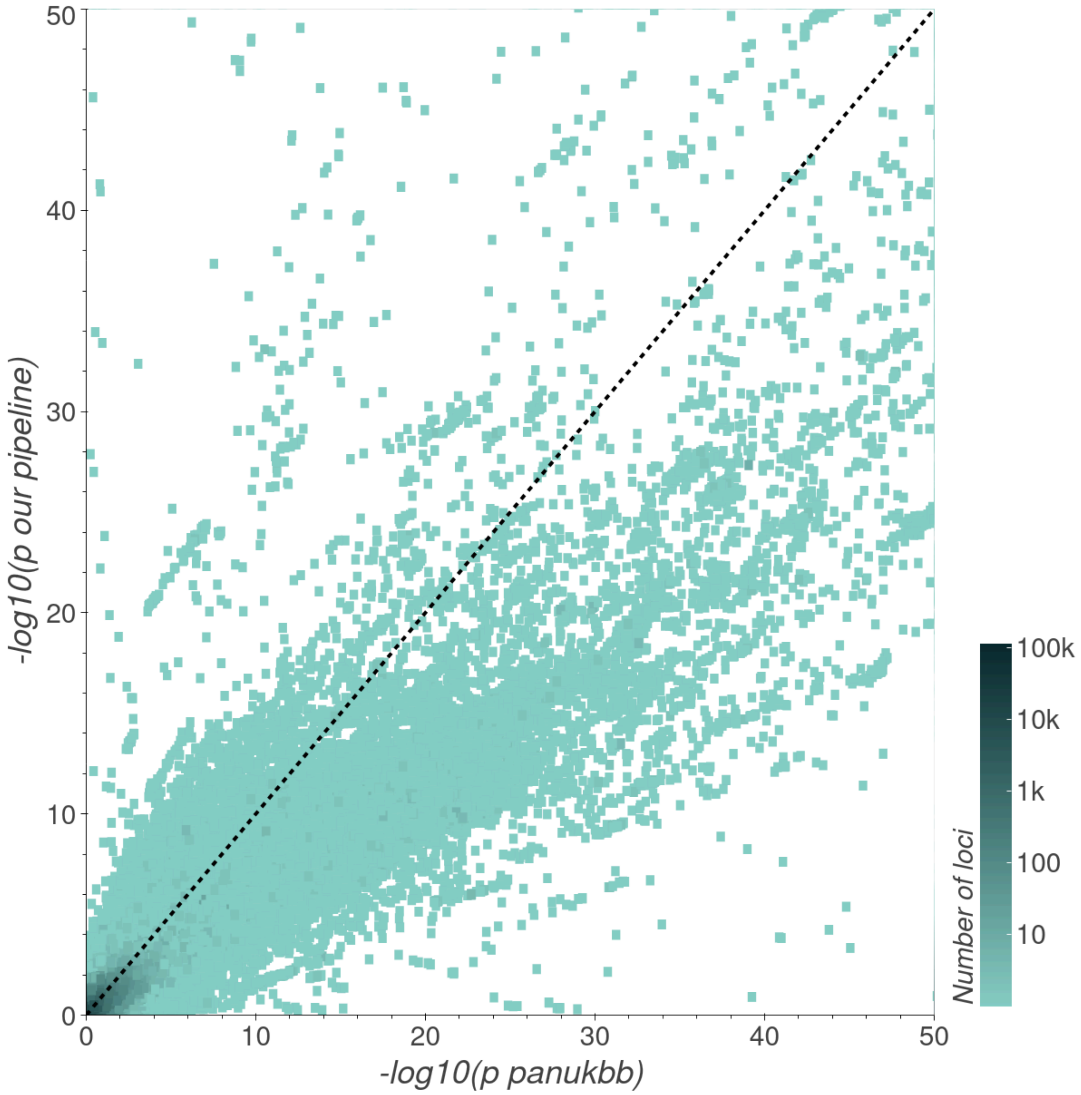
Supplementary Figures

Supplementary Figure 1: Comparison of SNP alternate allele frequencies between our SNP-STR reference panel and UKB phased hard-called variants



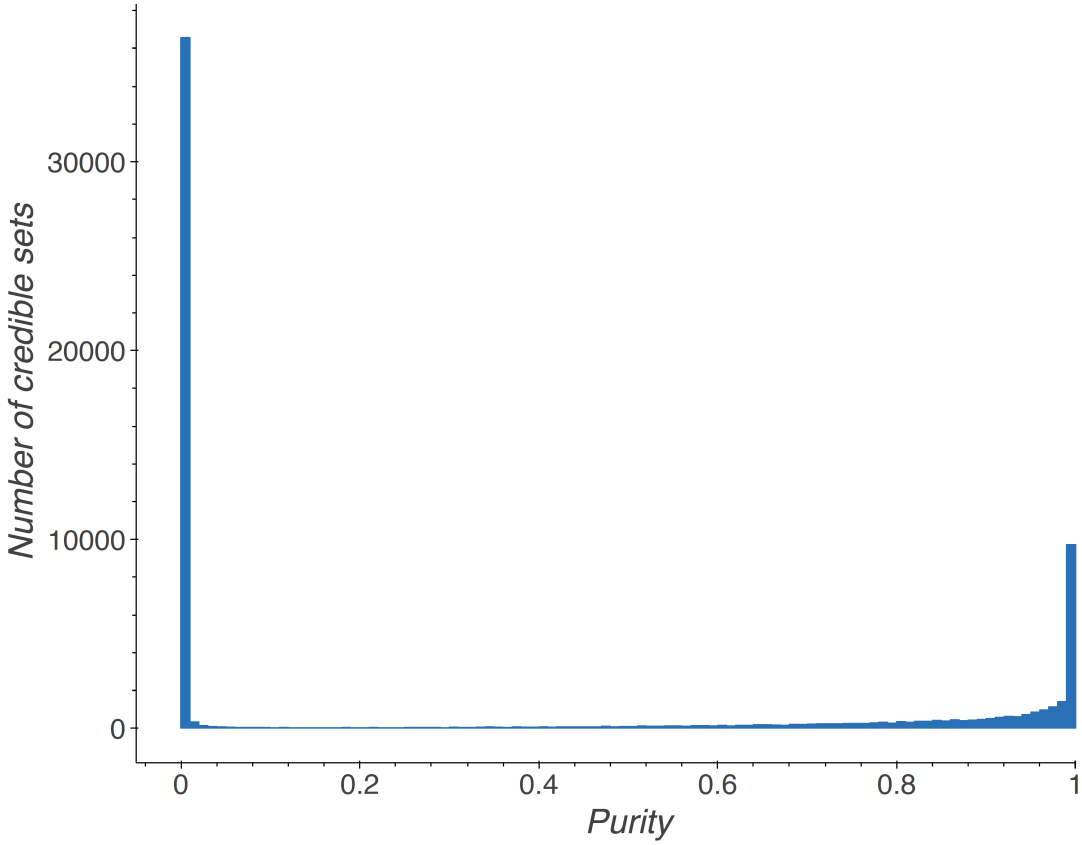
The x-axis indicates the alternate allele frequency of variants calculated from the European individuals in our SNP-STR reference panel[10] (**Main Text URLs**). The y-axis indicates their alternate allele frequency in unrelated participants in the White British population in the UKB. We filtered variants with more than a 12% difference in alternate allele frequency (indicated by the red diagonal line). The color gradient represents the number of variants ($\log_{10}$ scale) whose p-values fall in each region. White regions contain no variants.

Supplementary Figure 2: Comparison of association p-values between our pipeline and summary statistics published by Pan-UKB.



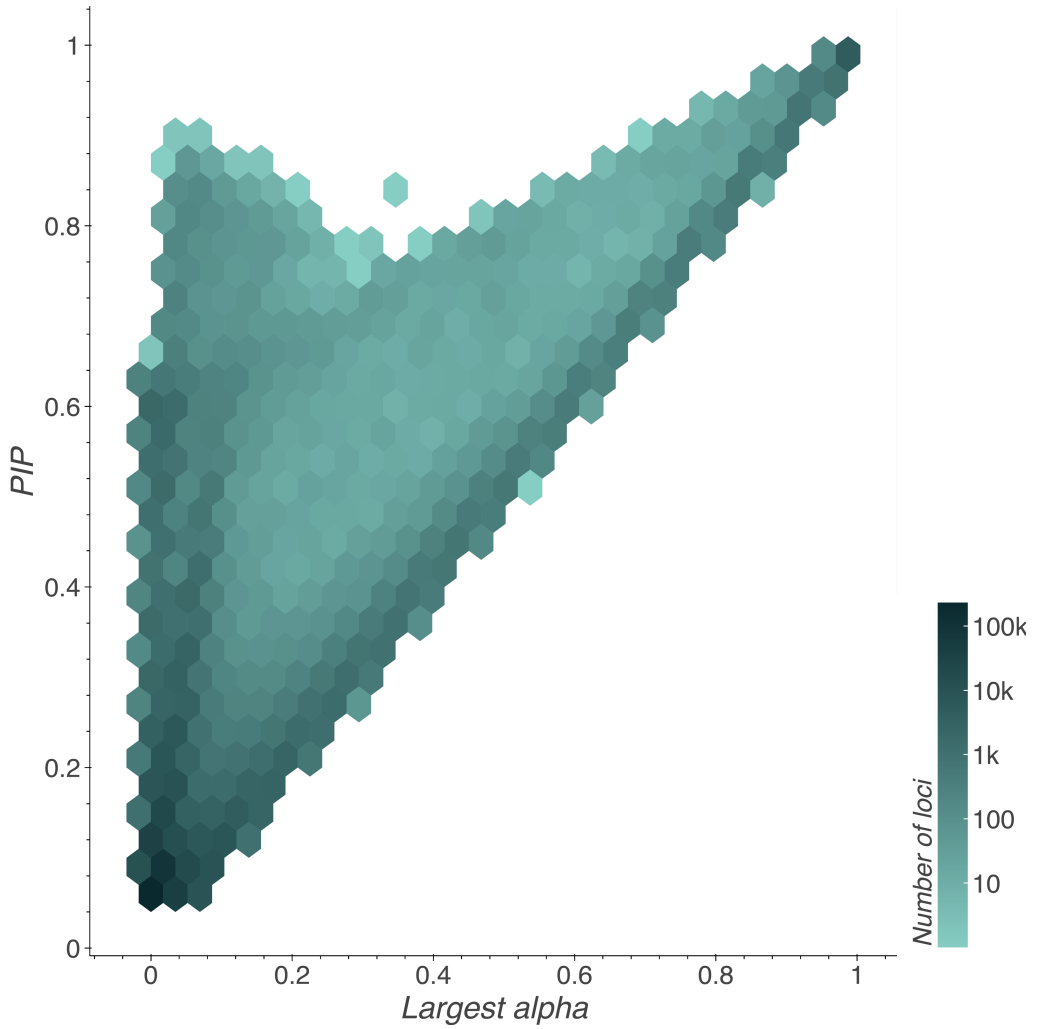
Heatmap of -$\log_{10}$ p-values obtained from the Pan-UKB[11] study of UKB data (x-axis) vs. from our study (y-axis) for total bilirubin associations. The color gradient represents the number of variants ($\log_{10}$ scale) whose p-values fall in each region. White regions contain no variants. P-values less than 1e-50 are truncated. Our pipeline's p-values are highly correlated with PanUKB's but are overall more conservative, which may be attributable to differences in models used (linear mixed model for Pan-UKB vs. linear model used here, see **Methods**).

Supplementary Figure 3: Distribution of SuSiE 90%-credible set purities



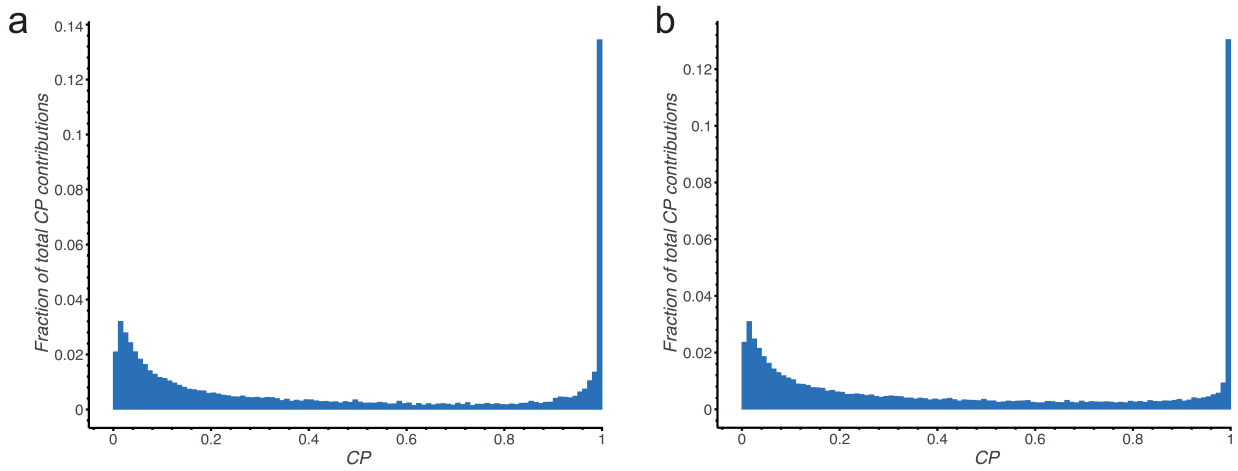
Distribution of SuSiE 90%-credible set purities across all trait-regions. (The rightmost bin is inclusive, containing SuSiE credible sets with purity up to and including 1, i.e. those that consist of a single variant.) Purity is defined as the minimum absolute correlation between any pair of variants in the set. For subsequent analyses, we discarded credible sets with purity < 0.8.

Supplementary Figure 4: PIP vs alpha values assigned by SuSiE



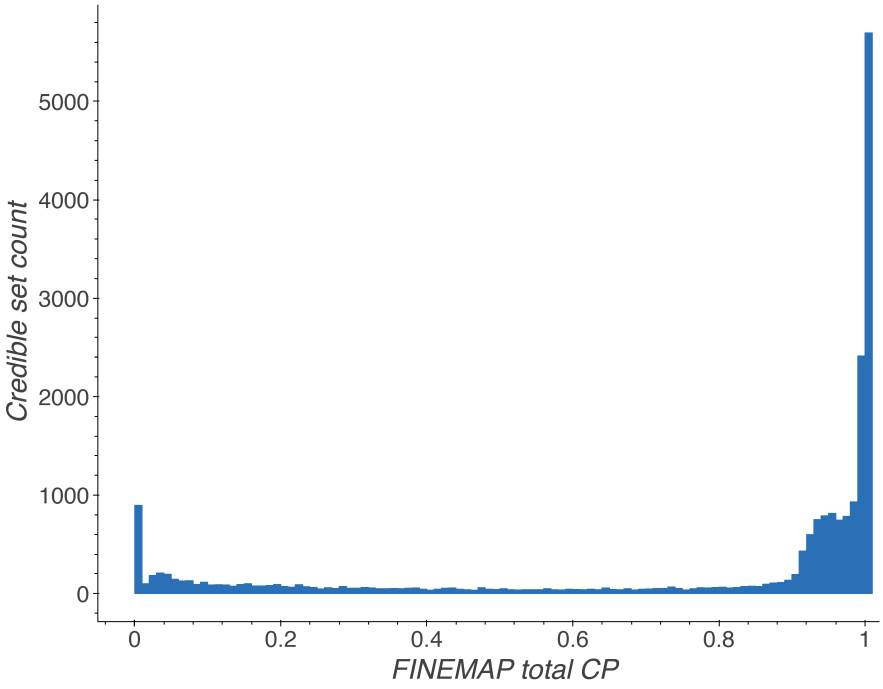
Largest alpha value (x-axis) vs. PIP (y-axis) for all variants obtaining PIP ≥ 0.05 across all trait regions. Color ($\log_{10}$ scale) indicates the number of data points falling in each bin (hexagon).

Supplementary Figure 5: Contribution of variants to signals genome-wide by variant CP



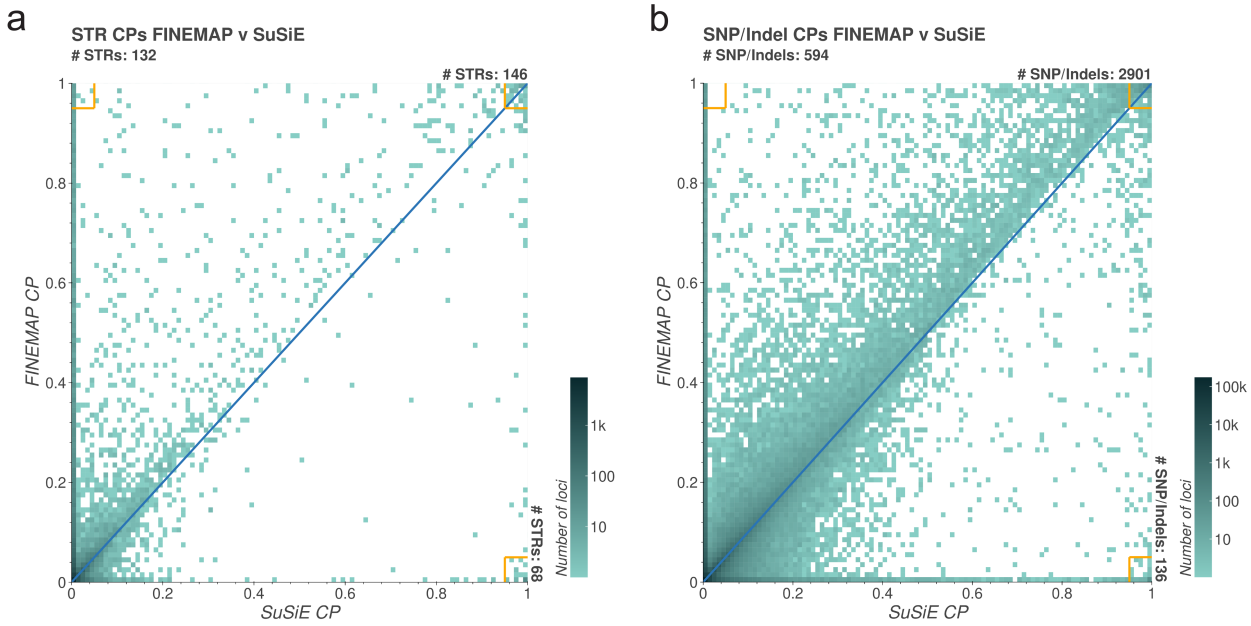
Summed contribution of genome-wide significant variants across all regions binned by variant CP as a fraction of the total CP of all genome-wide significant variants across all regions for SuSiE **(a)** and FINEMAP **(b)**. (The rightmost bin for each graph is inclusive, containing variants with CPs up to and including 1.) The total contribution of all variants across all regions with CP < 0.1 was 29.3% for SuSiE and 27.7% for FINEMAP.

Supplementary Figure 6: Total CPs assigned to SuSiE credible sets by FINEMAP



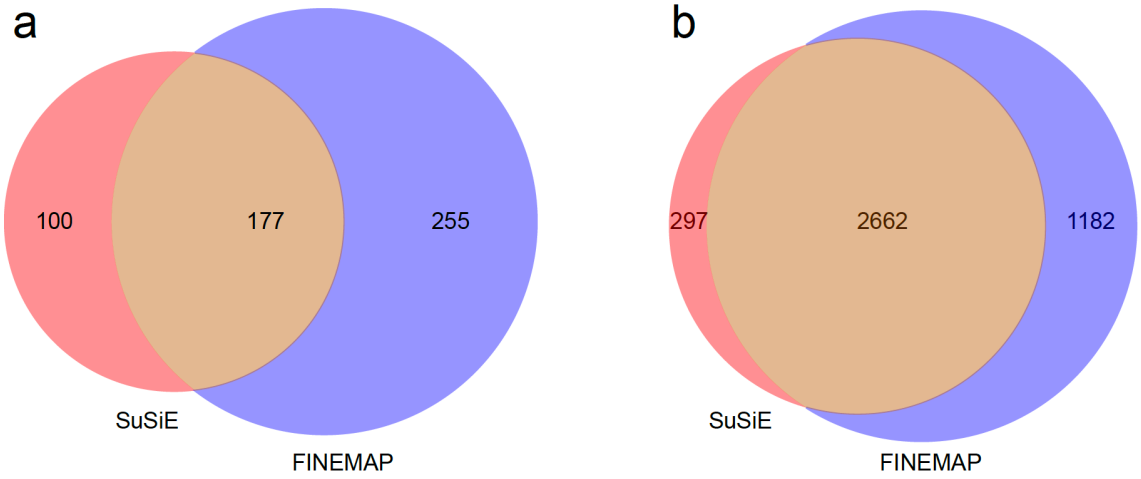
SuSiE 90%-credible sets across all trait-regions (with purity ≥ 0.8) were each binned by the total CP FINEMAP assigned to all variants in that set. Sets in the rightmost bin have FINEMAP total CP between 1 and 1.01 (i.e. FINEMAP predicts them to contain on average between 1 and 1.01 causal variants). FINEMAP assigned 3 SuSiE credible sets to have total CP greater than 1.01 (none of which attained total CP greater than 1.17); those 3 are omitted from the figure. By definition, SuSiE has estimated each 90%-credible set to have between a 90% and 100% chance of containing a single causal variant.

Supplementary Figure 7: Discordance between SuSiE and FINEMAP CPs



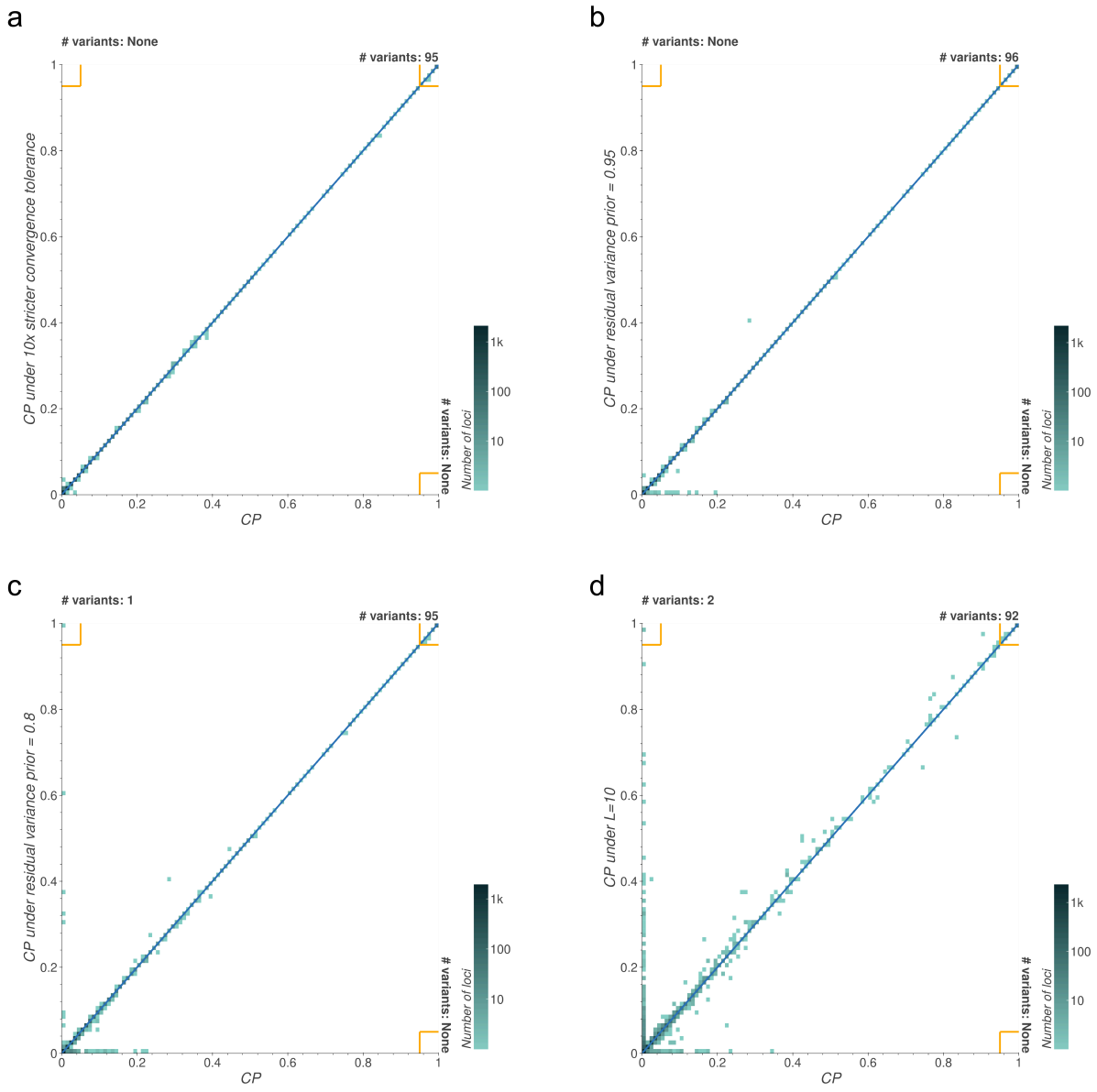
Comparison of CPs across all trait-regions between SuSiE (x-axis) and FINEMAP (y-axis) for genome-wide significant STRs **(a)** and SNPs and indels **(b)**. The blue line denotes equal CP. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes.

Supplementary Figure 8: Discordance between fine-mappers



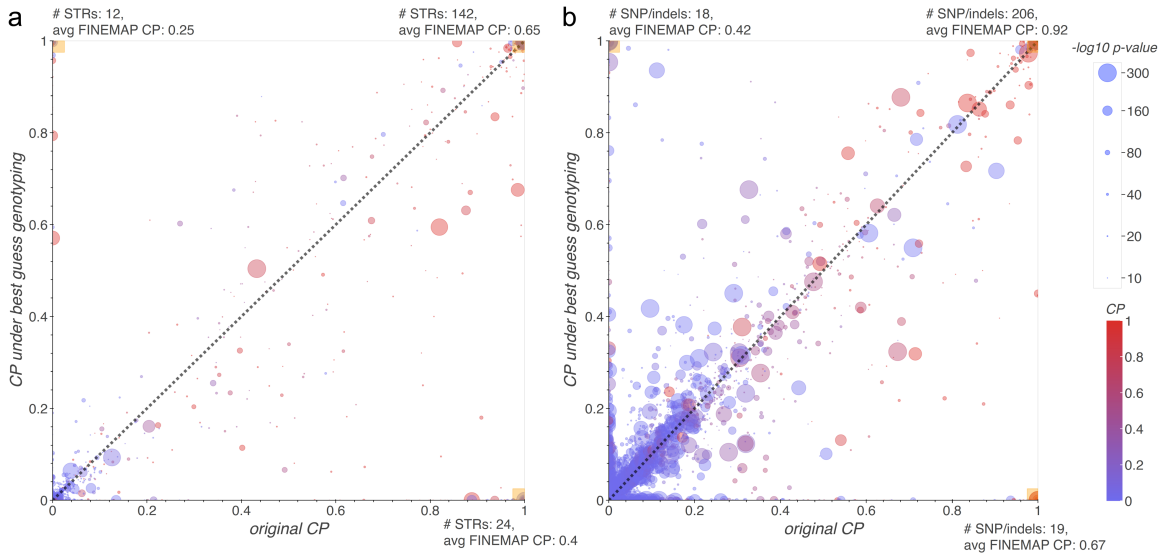
The number of STRs **(a)** and SNPs **(b)** with p-value ≤ 1e-10 assigned a CP ≥ 0.8 by only SuSiE (red), only FINEMAP (purple), or both (brown).

Supplementary Figure 9: SuSiE settings not used for filtering



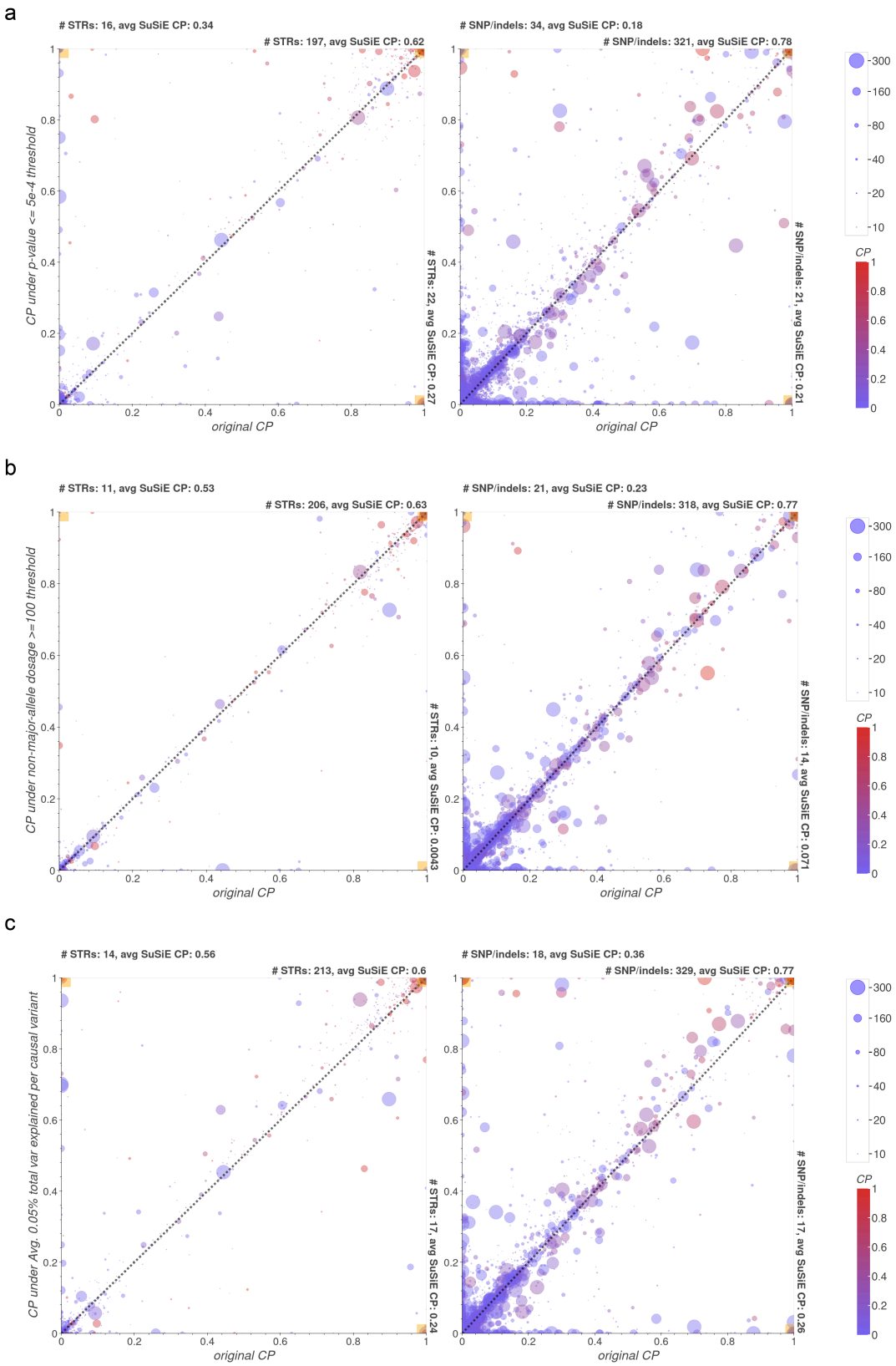
Concordance between SuSiE CPs for all genome-wide significant variants across most small-to-medium sized mean platelet volume fine-mapping regions under default settings on the x-axis (`tol=1e-3`, `residual_variance` slightly less than 1, and `L=50`) vs. a single alternate setting on the y-axis **(a)** `tol=1e-4`, **(b)** `residual_variance=0.95`, **(c)** `residual_variance=0.8` and **(d)** `L=10`. Blue lines denote equal CP. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes.

Supplementary Figure 10: Effect of best-guess genotypes on SuSiE results

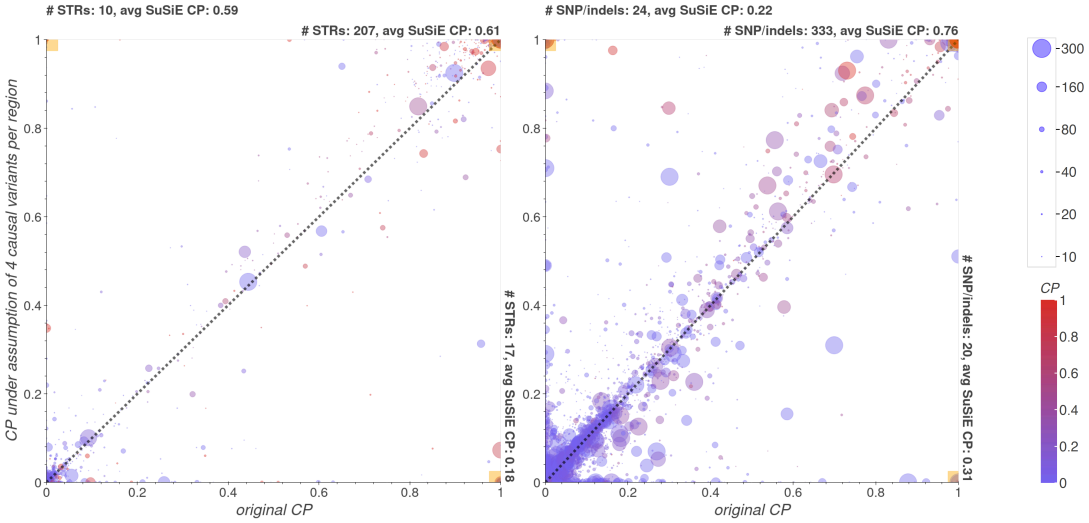


Discordance between SuSiE CPs for variants with p-value ≤ 1e-10 when run with dosage genotypes (x-axis) vs best-guess genotypes (y-axis) among STRs **(a)** and SNPs and indels **(b)**. These data points are taken from running SuSiE on the trait-regions containing the 177 STR-trait associations with p-value≤1e-10 and with both SuSiE and FINEMAP CPs≥0.8. Black lines denote equal CP. Larger circle sizes denote larger variant $-\log_{10}$ association p-values. Circle color denotes the CP of that variant from our default FINEMAP run. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes and the average FINEMAP CP value of those variants.

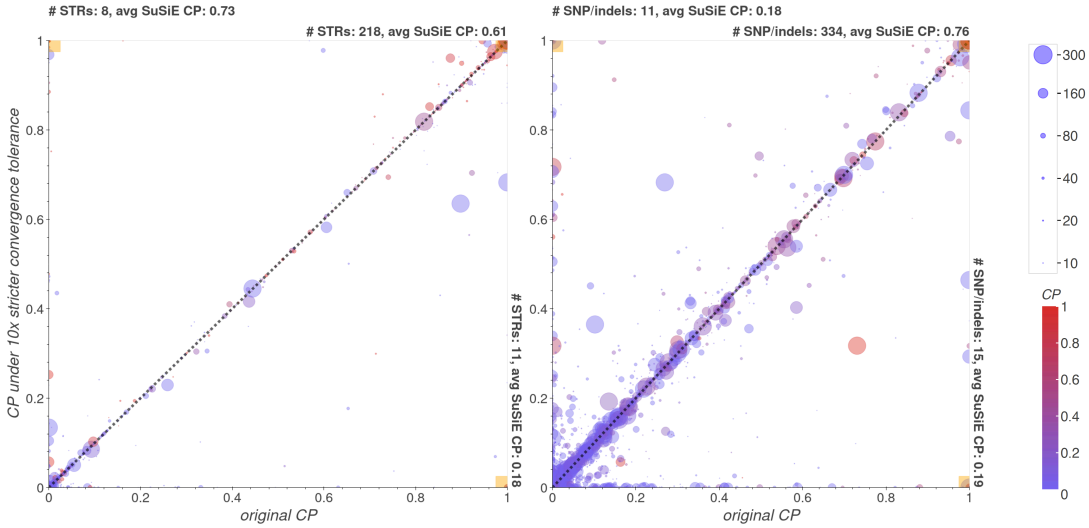
Supplementary Figure 11: Effect of alternate settings on FINEMAP results



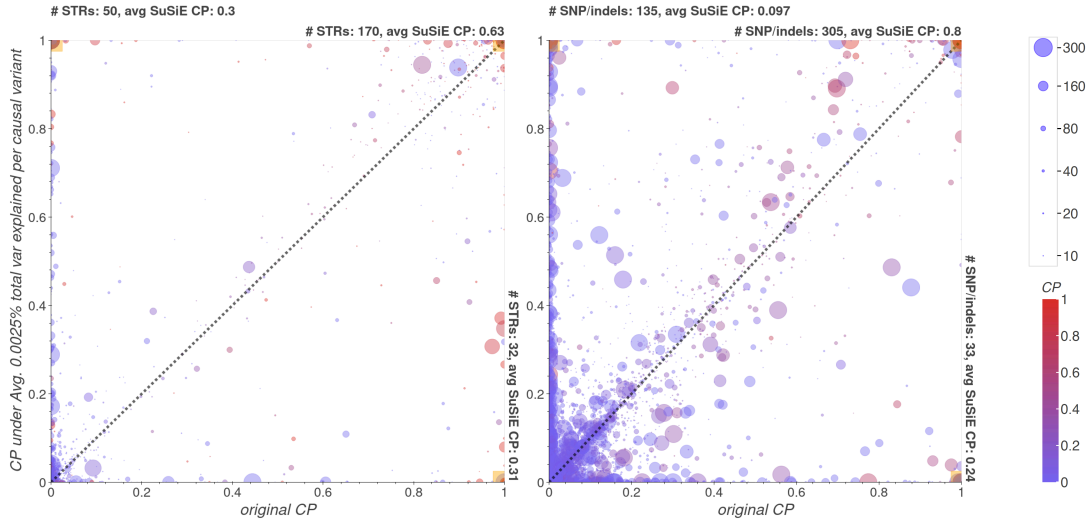
d



e

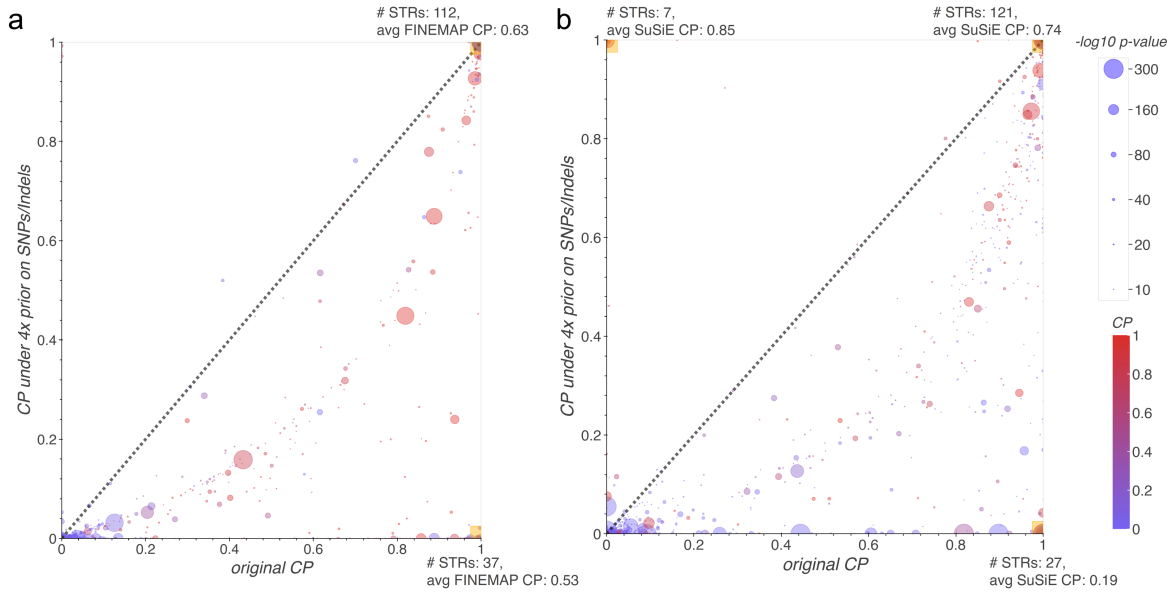


f



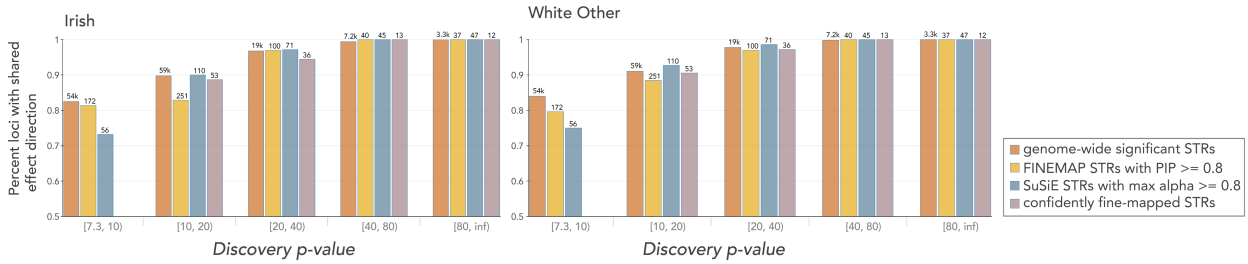
Discordance between FINEMAP CPs for variants with p-value ≤ 1e-10 when run under default settings on the x-axis (with a p-value > 5e-2 filter, `--prior-std 0.05`, prior of one causal variant per trait-region, `--prob-conv-sss-tol 0.001`) (x-axis) vs a single alternate setting on the y-axis **(a)** p-value > 5e-4 filter **(b)** additionally filtering those variants with total non-major-allele dosage < 100 **(c)** `--prior-std 0.0224` **(d)** prior of four causal variants per trait region **(e)** `--prob-conv-sss-tol 0.0001` and **(f)** `--prior-std=0.005`. These data points are taken from running FINEMAP on the trait-regions containing the 177 STR-trait associations with p-value≤1e-10 and with both SuSiE and FINEMAP CPs≥0.8. Discordance among STRs is plotted on the left, and among SNPs and indels is plotted on the right. Black lines denote equal CP. Larger circle sizes denote larger variant $-\log_{10}$ association p-values. Circle color denotes the CP of that variant from our default SuSiE run. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes and the average SuSiE CP value of those variants.

Supplementary Figure 12: Effect of conservative prior favoring SNPs and indels on estimated causality of STR variants



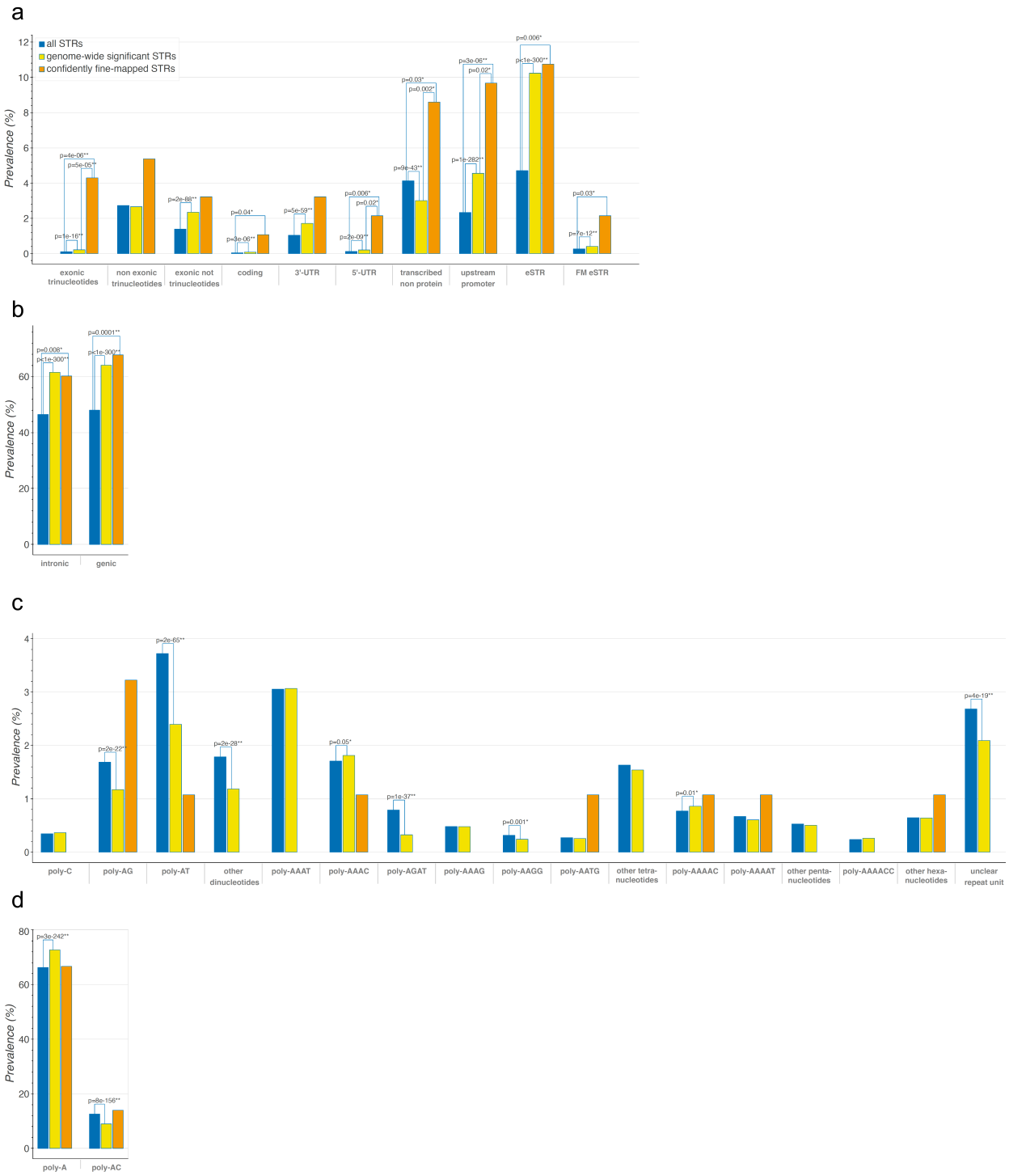
Discordance between CPs for STRs with p-value ≤ 1e-10 when run under default settings (x-axis) vs with a 4x prior of causality for SNPs and indels as compared to STRs (y-axis) in **(a)** SuSiE and **(b)** FINEMAP. These data points are taken from running SuSiE and FINEMAP on the trait-regions containing the 177 STR-trait associations with p-value≤1e-10 and with both SuSiE and FINEMAP CPs≥0.8. Black lines denote equal CP. Larger circle sizes denote larger variant -$\log_{10}$ association p-values. Circle color denotes the CP of that variant from the other fine-mapper's default run. Yellow boxes in the extreme corners are summarized by the number of variants residing in those boxes and the average SuSiE CP value of those variants.

Supplementary Figure 13: Replication of White British STR associations in other populations



The y-axis gives the fraction of STR associations measured in the discovery cohort that have the same direction of effect when measured in the replication population regardless of p-value (left=Irish, right=White Other, see **Fig. 3** for the non-White populations). Brackets beneath the x-axis denote the binning of discovery $-\log_{10}$ p-values. Brown=genome-wide significant associations (discovery p≤5e-8), orange=FINEMAP fine-mapped STR associations (discovery p≤5e-8 and FINEMAP CP ≥ 0.8), teal=SuSiE fine-mapped STR associations (discovery p≤5e-8 and SuSiE CP ≥ 0.8) and purple=confidently fine-mapped STR associations. Annotations above each bar indicate the number of STR-trait associations considered. We required confidently fine-mapped STR associations to have p-value ≤ 1e-10, thus they do not appear in the left-most bin. The trends in these figures are somewhat sensitive to the choice of p-value bin boundaries so we additionally analyze this data using logistic regression models (**Supplementary Table 6**).

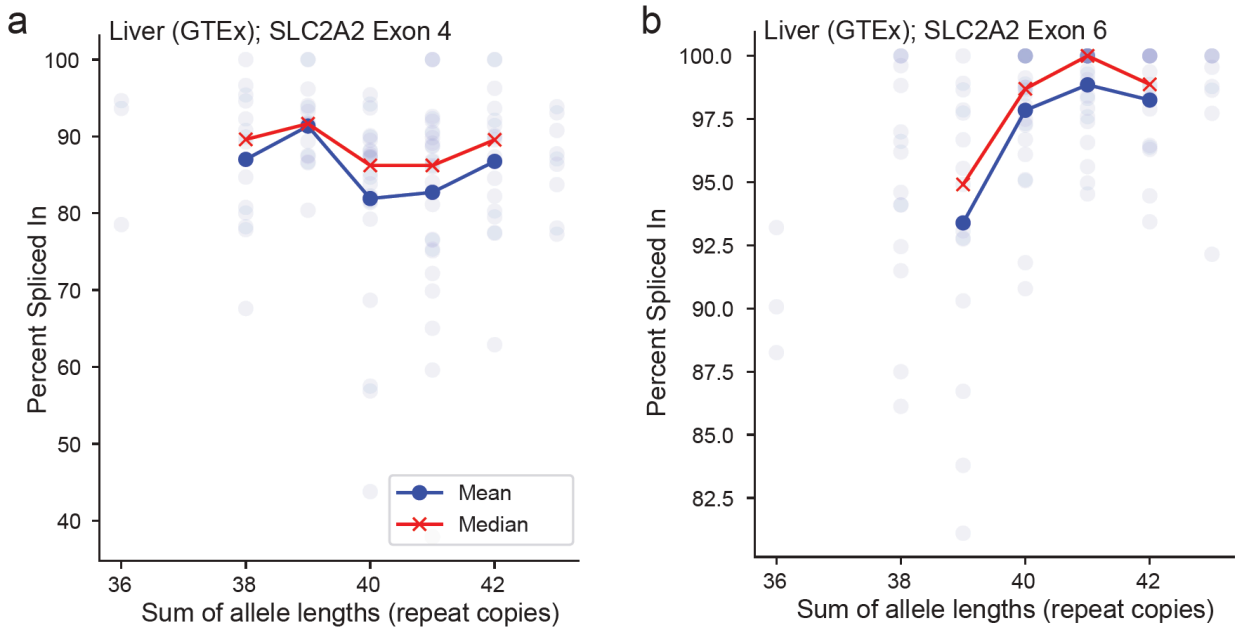
Supplementary Figure 14: Prevalence of STR features



Genomic annotation **(a-b)** and repeat unit **(c-d)** prevalences are shown for different categories of STRs. (Blue=all STRs in our imputation panel, yellow=genome-wide significant STRs for at least one trait, orange=confidently fine-mapped STRs). In **(a)**, "upstream promoter" is defined as the

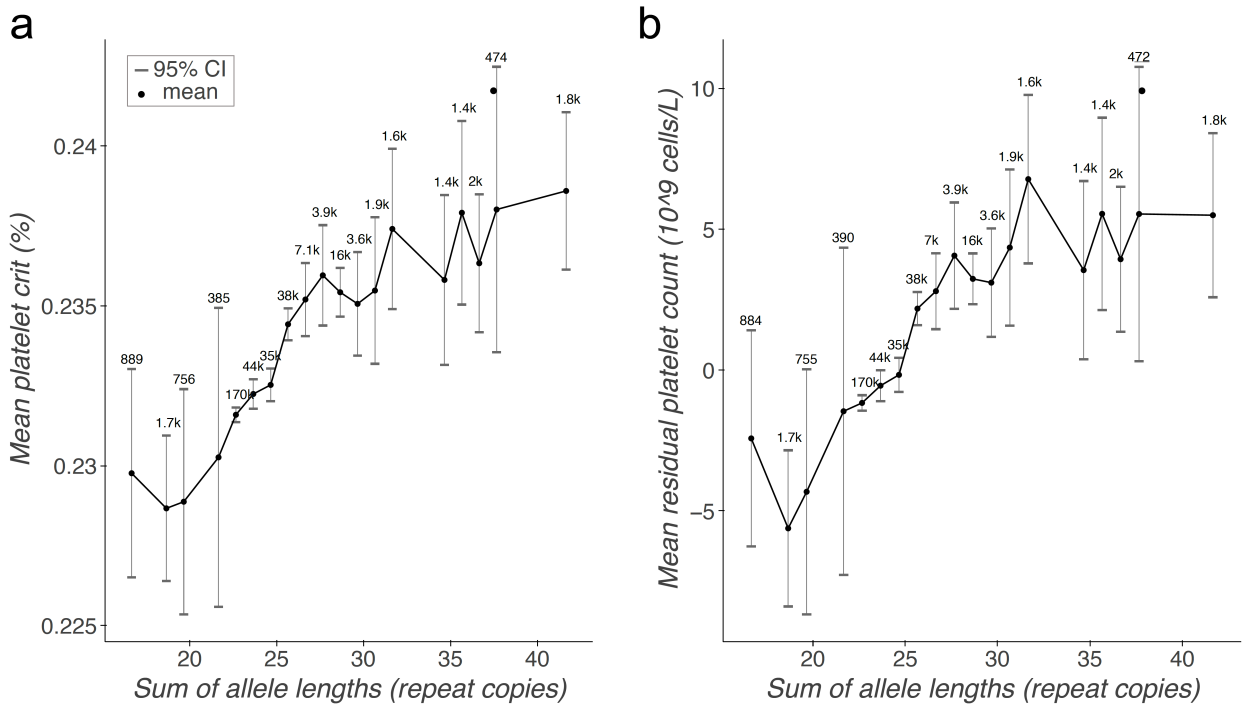
region 3kb upstream of a transcription start site. **(c-d)** contains all repeat units represented by at least one thousand STRs in our reference panel, except for trinucleotide STR repeat units, as enrichments for those could not be distinguished from the enrichment for exonic trinucleotide STRs as a whole. See the **Methods** for more details. P-values from two-sided tests of difference between proportions are only displayed when $p \leq 0.05$. Note that strong p-values for differences between the all STRs and genome-wide significant STRs categories could often be due to restricting to phenotypically-important genomic regions and not necessarily due to enrichment for causal variants.

Supplementary Figure 15: Splicing analysis of an STR in *SLC2A2*



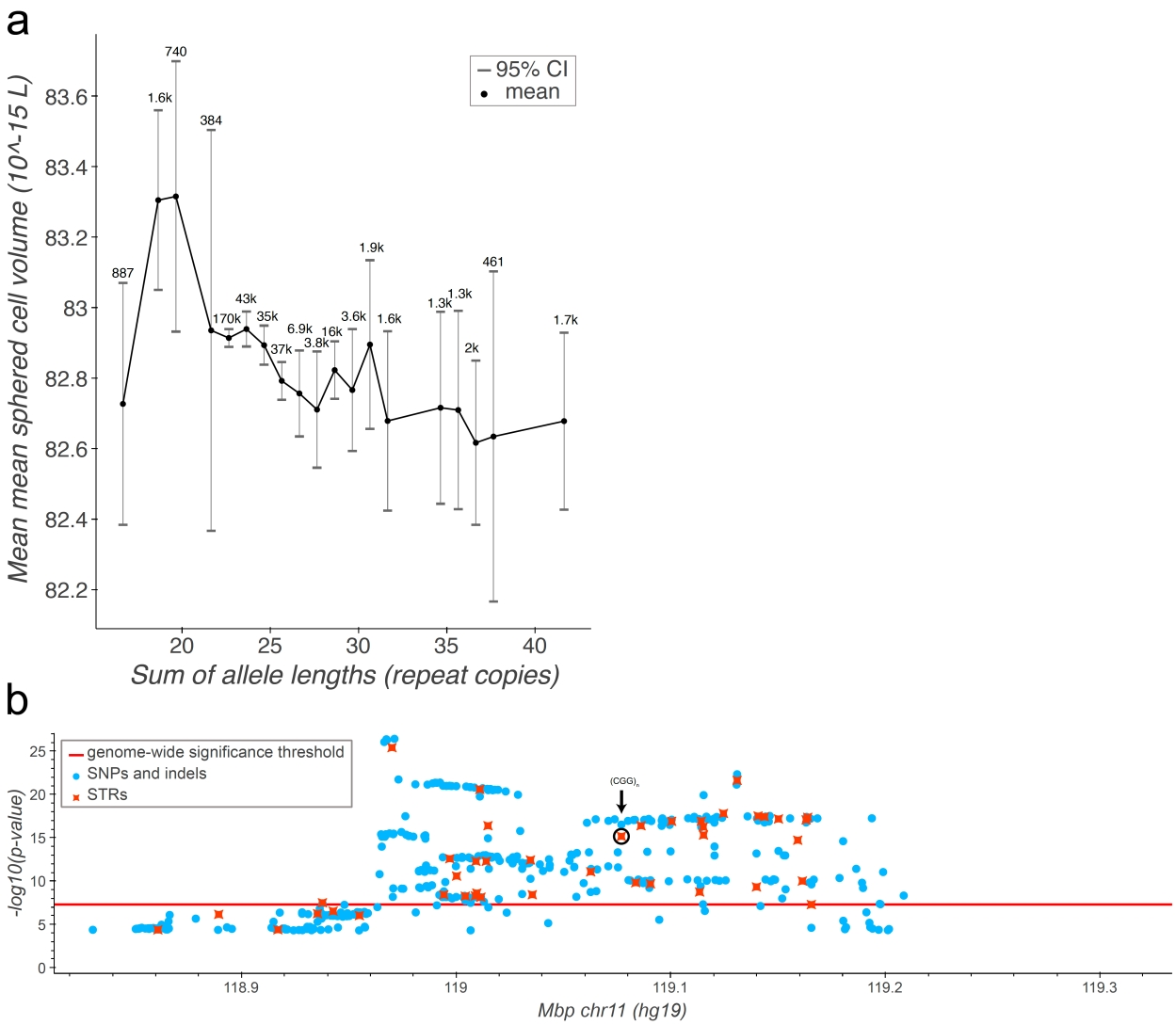
Scatter plots showing the association between the STR chr3:17100913 (GT repeat) and the splicing (percent spliced in, or PSI) of exon 4 (**a**; linear association p=0.77) and exon 6 (**b**; linear association p=8.7e-07) of *SLC2A2* in Liver samples from GTEx[12]. Blue dots represent single samples. Dots are transparent such that darker dots indicate multiple samples overlayed on the same point. For each plot, the x-axis represents the sum of repeat copies of STR in each individual and the y-axis represents percent spliced in (PSI) for the indicated exon. The red line shows the median and the blue line shows the mean PSI for each x axis value.

Supplementary Figure 16: Associations of an STR in *CBL* with platelet crit and residual platelet volume



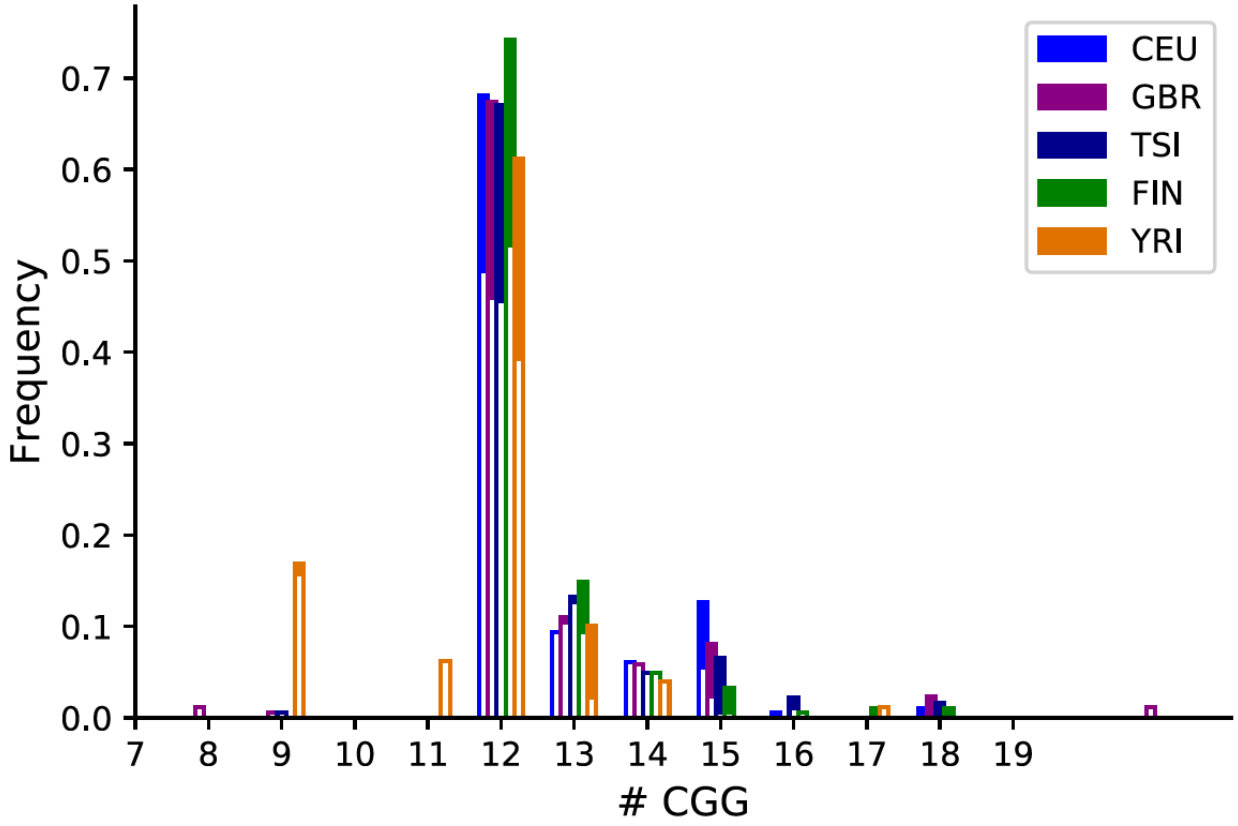
STR length vs mean platelet crit **(a)** and mean residual platelet count **(b)**. The trends are nearly identical to those in **Fig. 4b** for unadjusted platelet count. For **(b)** we calculated residuals by linearly regressing out the same covariates that were used in association p-value calculations (**Methods**), including sex, age, population principal components and categorical covariates for batch effects. We then calculated the weighted means for each dosage taking the residual values as fixed inputs. Note that in our association pipeline, p-values are calculated from regressions on rank inverse normalized phenotype values, while for this figure we do not use rank inverse normalization.

Supplementary Figure 17: Associations of an STR in *CBL* with mean sphered cell volume



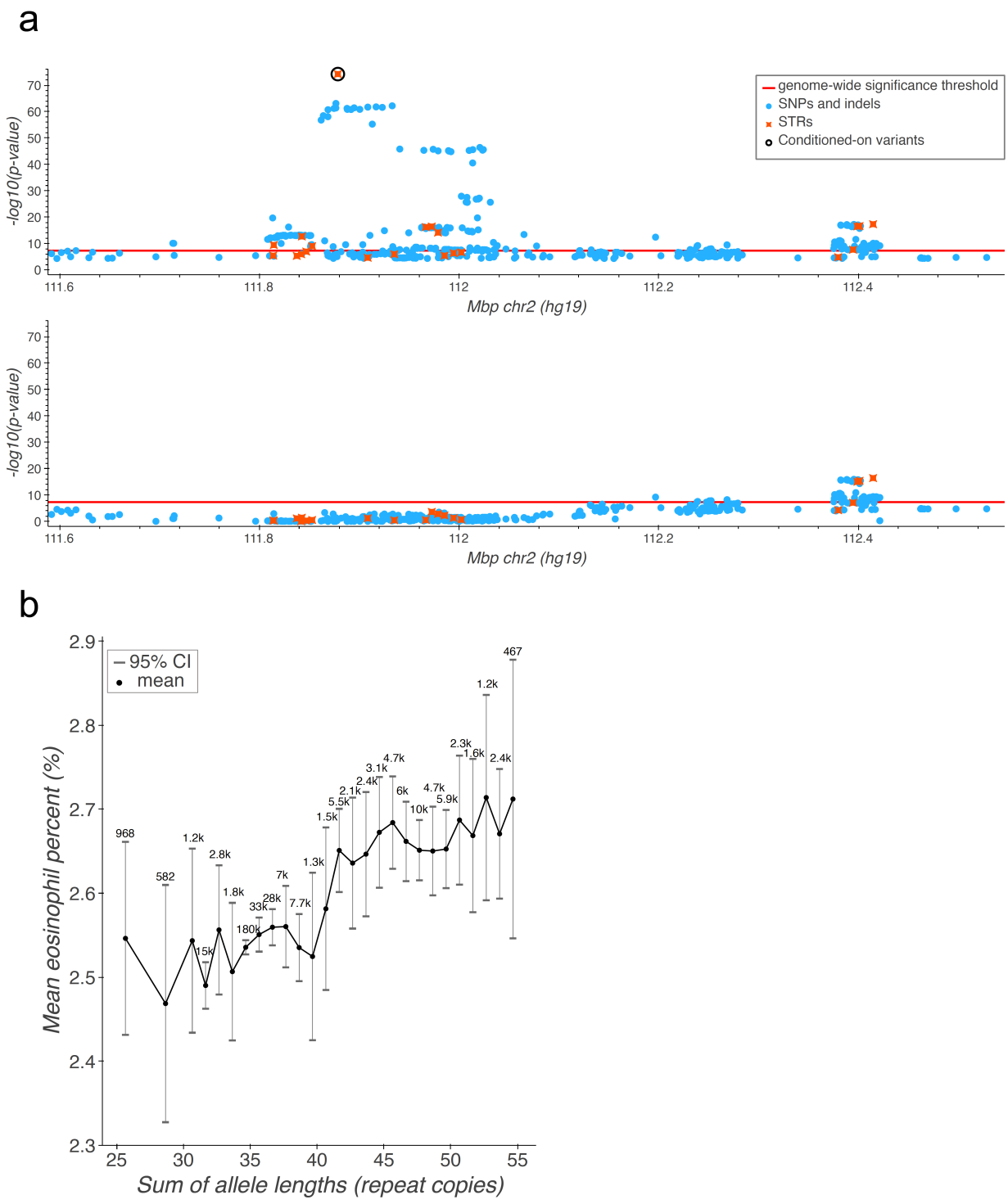
(a) Association between dosage of the CGG repeat at chr11:119077000 (hg19) and mean sphered cell volume. The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's likelihood of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each dosage. **(b)** Association of variants at the *CBL* locus and mean sphered cell volume. Light blue=SNP and indels; orange=STRs. Red line=significance threshold, black circle=the $(CGG)_n$ STR.

Supplementary Figure 18: Distribution of alleles of an STR in *CBL* across 1000 Genomes populations



The x-axis gives STR length (number of repeat units) and y-axis gives the population frequency. The solid portion of each bar corresponds to the alleles of that length that include a “TGG” imperfection at the second repeat (rs7108857). Colors denote 1000 Genomes populations that were included in the Geuvadis cohort[13].

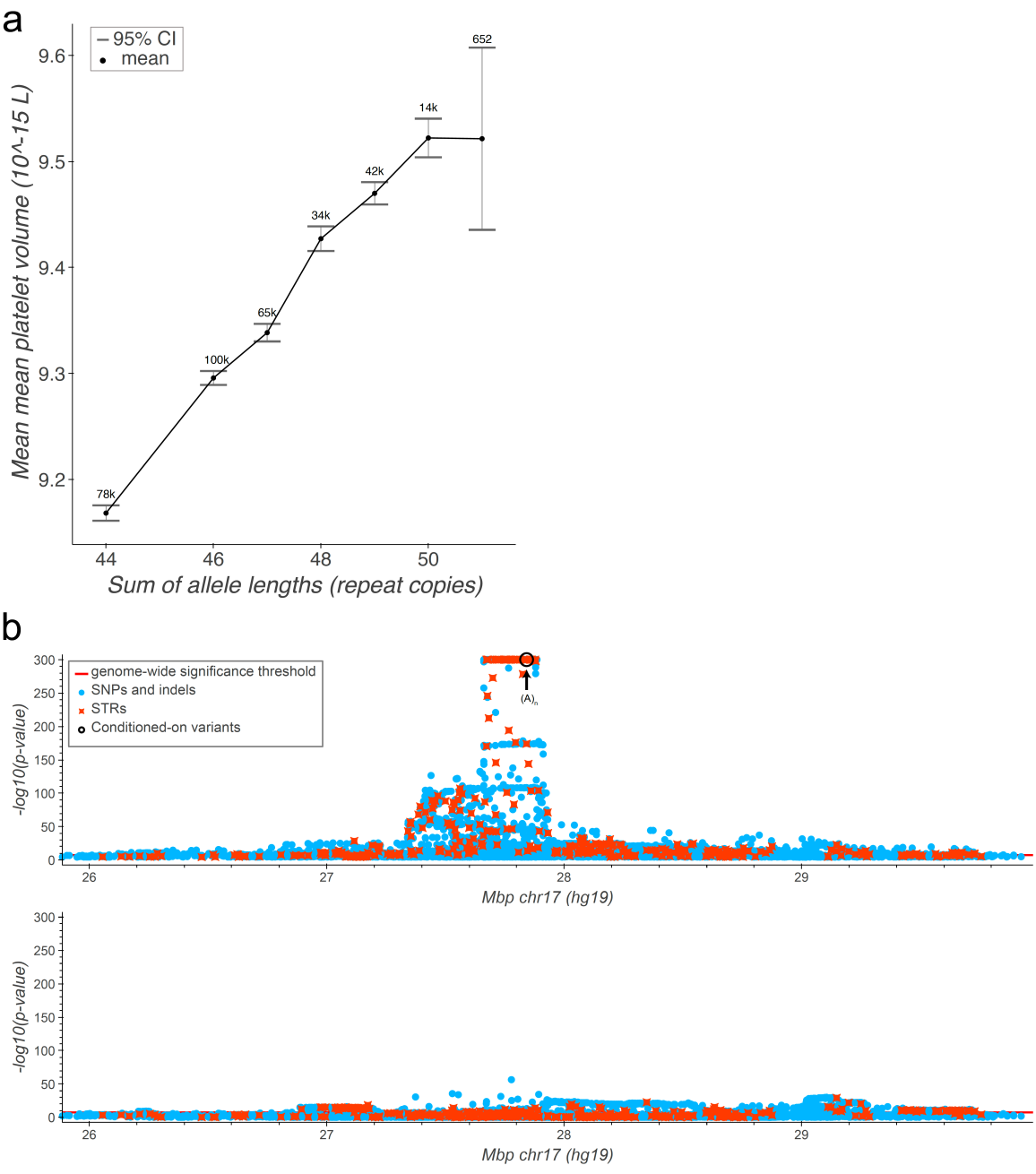
Supplementary Figure 19: Association of an STR in *BCL2L11*



(a) Association of variants at the *BCL2L11* locus and eosinophil percent, before (top) and after (bottom) conditioning on the CCG repeat at chr2:111878544 (hg19). Light blue=SNPs and indels;

orange=STRs. Red line=significance threshold, black circle=the $(CCG)_n$ STR. **(b)** Association between dosage of the CCG repeat and eosinophil percentage. The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's likelihood of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each dosage.

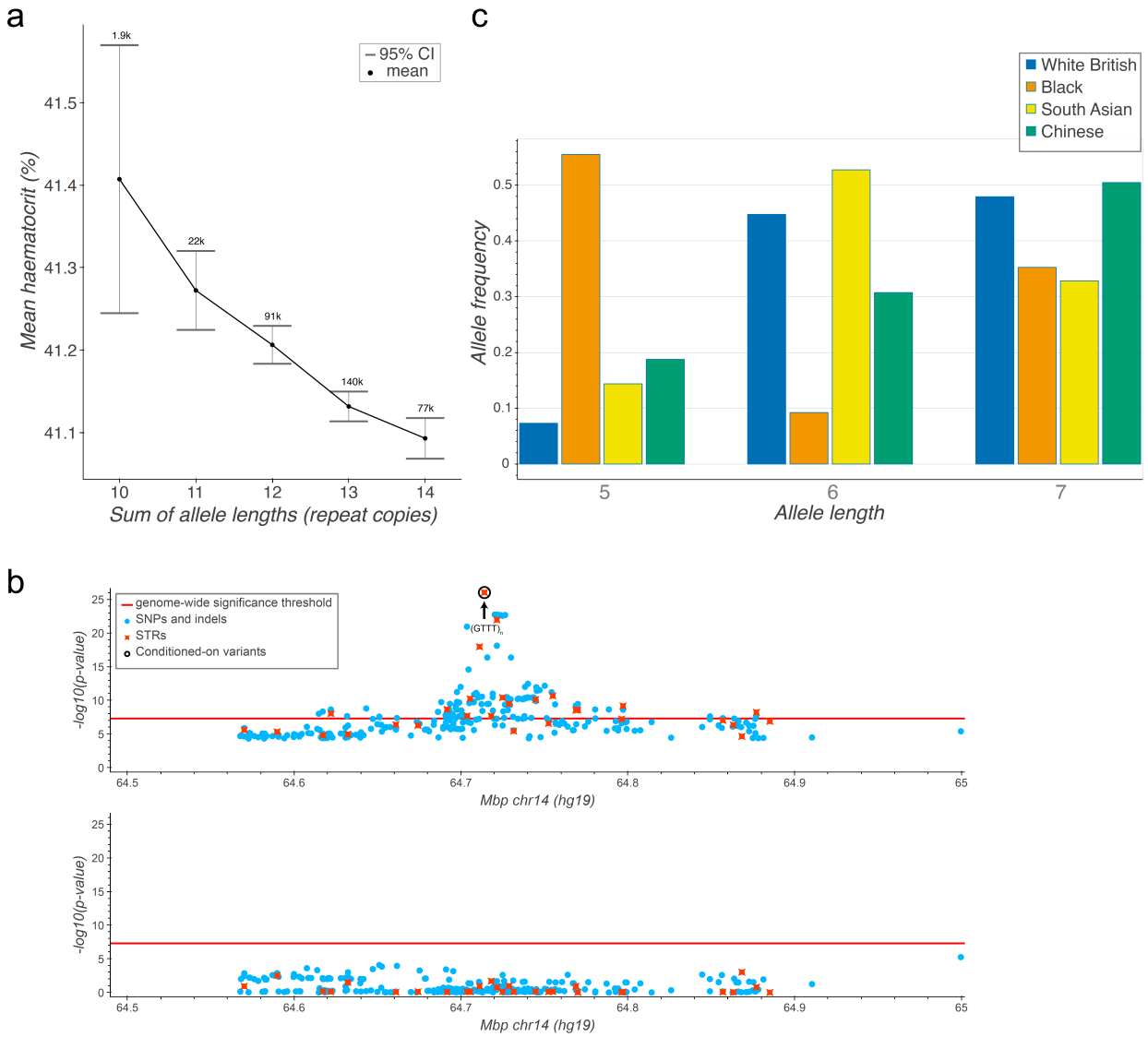
Supplementary Figure 20: Association of an STR in *TAOK1*



(a) Association between dosage of the A repeat at chr17:27842016 (hg19) and mean platelet volume. The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's likelihood of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide

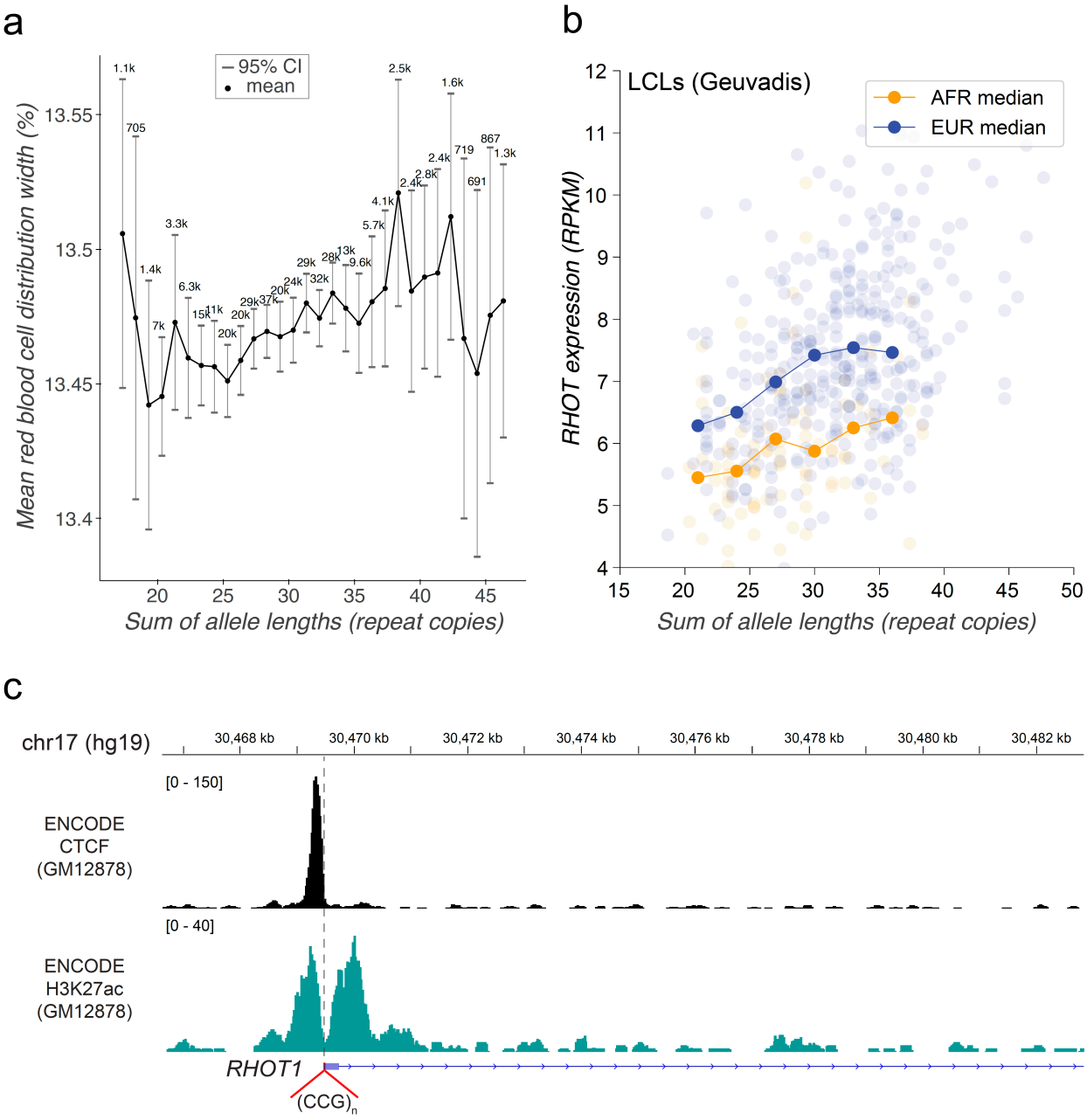
counts are displayed for each dosage. **(b)** Association of variants at the *TAOK1* locus and mean platelet volume before (top) and after (bottom) conditioning on the STR. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(A)_n$ STR.

Supplementary Figure 21: Association and allele distribution of an STR in *ESR2*



(a) Association between dosage of the GTTT repeat at chr14:64714051 (hg19) and haematocrit. The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's likelihood of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each dosage. **(b)** Association of variants at the *ESR2* locus and haematocrit. Conditioning on the repeat fully accounts for the signal seen in this region. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(GTTT)_n$ STR. **(c)** Distribution of STR length alleles in different populations (blue=White British, orange=Black, yellow=South Asian; green=Chinese).

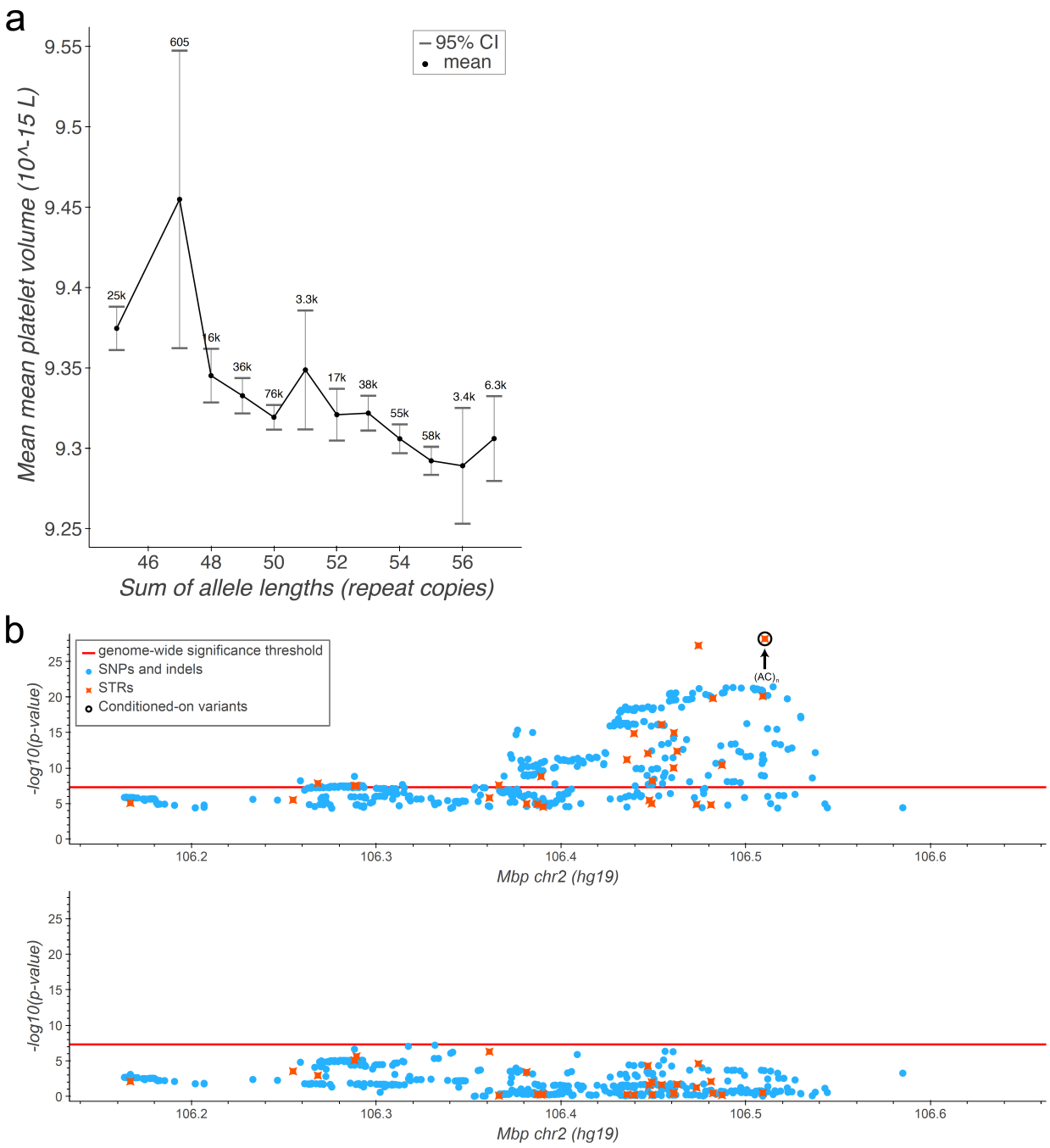
Supplementary Figure 22: Associations of an STR in *RHOT1*



(a) Association between dosage of the CCG repeat at chr17: 30469471 (hg19) and red blood cell distribution width. The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's likelihood of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each dosage. **(b)** Association between dosage of the repeat and *RHOT1* gene expression in the Geuvadis cohort[13] (LCLs; n=447). Solid lines give median

expression values for each STR dosage bin with at least 5% frequency in each group. Dosages were binned into groups spanning 3 repeat copies each since individually each genotype was relatively rare at this locus. **(c)** Positioning of the CCG repeat relative to the H3K27ac signal (note the localization within the nucleosome depleted region) and a CTCF binding site at the 5' UTR of *RHOT1*. The visualization was generated using the Integrative Genomics Viewer[14] loading the ENCODE[15] data for GM12878 LCLs. The image does not display the gene NR_136413 that also overlaps the STR as it is not expressed in LCLs.

Supplementary Figure 23: Association of an STR in *NCK2*



(a) Association between dosage of the AC repeat at chr2:106510441 (hg19) and mean platelet volume. The mean trait value for each STR dosage (sum of allele lengths) was calculated across White British participants, with each participant's contribution weighted by that participant's likelihood of having that dosage. 95% confidence intervals were calculated similarly. Only dosages with a population frequency of 0.1% or greater are displayed. Rounded population-wide

counts are displayed for each dosage. **(b)** Association of variants at the *NCK2* locus and mean platelet volume. Conditioning on the repeat fully accounts for the signal seen in this region. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(AC)_n$ STR.

References

1. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
2. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
3. Flint, J. GWAS. *Curr. Biol.* **23**, R265–R266 (2013).
4. Bánhegyi, G., Garzó, T., Antoni, F. & Mandl, J. Glycogenolysis - and not gluconeogenesis - is the source of UDP-glucuronic acid for glucuronidation. *Biochim. Biophys. Acta BBA - Gen. Subj.* **967**, 429–435 (1988).
5. Edwards, M., Falzone, N. & Harrington, J. Conjugated hyperbilirubinemia among infants with hyperinsulinemic hypoglycemia. *Eur. J. Pediatr.* **180**, 1653–1657 (2021).
6. Azad, P., Villafuerte, F. C., Bermudez, D., Patel, G. & Haddad, G. G. Protective role of estrogen against excessive erythrocytosis in Monge's disease. *Exp. Mol. Med.* **53**, 125–135 (2021).
7. Mukundan, H., Resta, T. C. & Kanagy, N. L. 17β-Estradiol decreases hypoxic induction of erythropoietin gene expression. *Am. J. Physiol.-Regul. Integr. Comp. Physiol.* **283**, R496–R504 (2002).
8. Lewandowski, S., Kalita, K. & Kaczmarek, L. Estrogen receptor β. *FEBS Lett.* **524**, 1–5 (2002).
9. Zhao, C., Dahlman-Wright, K. & Gustafsson, J.-Å. Estrogen Receptor β: An Overview and Update. *Nucl. Recept. Signal.* **6**, nrs.06003 (2008).
10. Saini, S., Mitra, I., Mousavi, N., Fotsing, S. F. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).
11. Pan-UKB team. Pan-ancestry genetic analysis of the UK Biobank. https://pan.ukbb.broadinstitute.org/ (2020).

12. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

13. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

14. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

15. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).