Title

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

Journal

Cell Genomics, 3(12)

ISSN

2666-979X

Authors

Margoliash, Jonathan

Fuchs, Shai

Li, Yang

et al.

Publication Date

2023-12-01

DOI

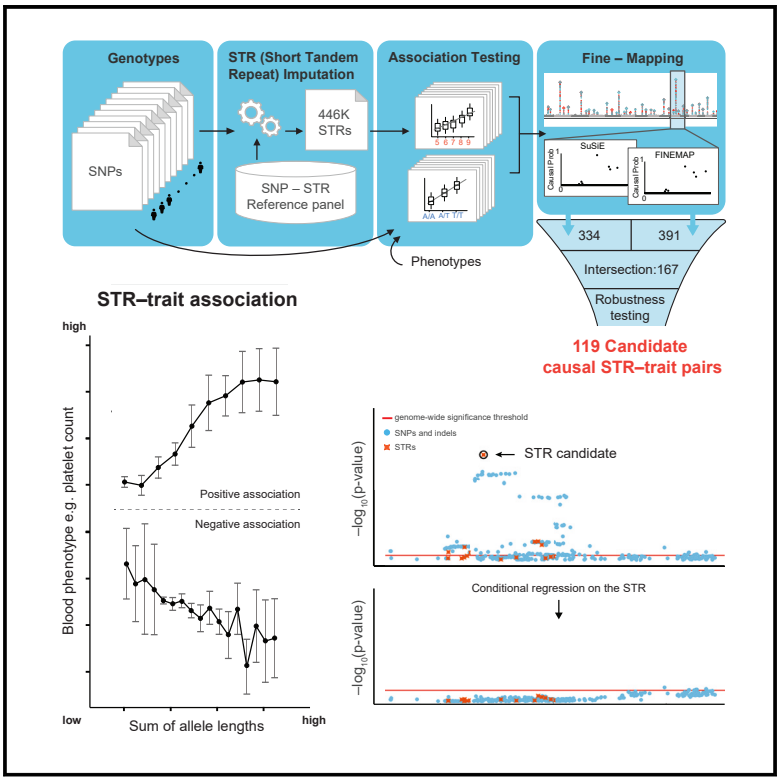
10.1016/j.xgen.2023.100458

Cell Genomics

Article

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

Graphical abstract



Highlights

- A novel framework enables incorporating short tandem repeat variants into GWASs
- Short tandem repeats comprise 5.2%–7.6% of candidate causal variants for blood traits
- Stringent fine-mapping identifies 119 candidate causal repeat-trait associations
- Incorporation of repeats into future GWASs is likely to reveal novel causal variants

Authors

Jonathan Margoliash, Shai Fuchs, Yang Li, Xuan Zhang, Arya Massarat, Alon Goren, Melissa Gymrek

Correspondence

agoren@ucsd.edu (A.G.), mgymrek@ucsd.edu (M.G.)

In brief

Margoliash et al. produce a framework for including short tandem repeat (STR) genetic variants in complex trait analysis. Using two fine-mapping methods, they estimate that STRs account for 5.2%–7.6% of causal variants identifiable for the studied traits and highlight 119 candidate causal STR-trait associations, resolving some of the strongest associations for multiple phenotypes. This study suggests that STRs play an important role in complex traits and demonstrates the need to include a more complete set of genetic variation in genome-wide association studies.

Margoliash et al., 2023, Cell Genomics *3*, 100458
December 13, 2023
https://doi.org/10.1016/j.xgen.2023.100458

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

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

Jonathan Margoliash,[1] Shai Fuchs,[2] Yang Li,[1,3] Xuan Zhang,[3] Arya Massarat,[4] Alon Goren,[3,*] and Melissa Gymrek[1,3,5,*]
[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA
[2]Pediatric Endocrine and Diabetes Unit, Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Ramat Gan, Israel
[3]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[4]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA 92093, USA
[5]Lead contact
*Correspondence: agoren@ucsd.edu (A.G.), mgymrek@ucsd.edu (M.G.)
https://doi.org/10.1016/j.xgen.2023.100458

SUMMARY

Short tandem repeats (STRs) are genomic regions consisting of repeated sequences of 1–6 bp in succession. Single-nucleotide polymorphism (SNP)-based genome-wide association studies (GWASs) do not fully capture STR effects. To study these effects, we imputed 445,720 STRs into genotype arrays from 408,153 White British UK Biobank participants and tested for association with 44 blood phenotypes. Using two fine-mapping methods, we identify 119 candidate causal STR-trait associations and estimate that STRs account for 5.2%–7.6% of causal variants identifiable from GWASs for these traits. These are among the strongest associations for multiple phenotypes, including a coding CTG repeat associated with apolipoprotein B levels, a promoter CGG repeat with platelet traits, and an intronic poly(A) repeat with mean platelet volume. Our study suggests that STRs make widespread contributions to complex traits, provides stringently selected candidate causal STRs, and demonstrates the need to consider a more complete view of genetic variation in GWASs.

INTRODUCTION

Genome-wide association studies (GWASs) are an indispensable tool for identifying which genes and non-coding regions in the genome influence complex human traits, yet biological investigation of those regions remains challenging.[1] A major limitation is that typical GWAS pipelines only consider single-nucleotide polymorphisms (SNPs) and short insertions or deletions (indels). However, detailed follow-up of individual GWAS signals has often revealed complex variants absent from the original analysis, such as repeats[2,3] or structural variants,[4,5] to be the causal drivers of those signals. Indeed, a recent study showed that polymorphic protein-coding variable number tandem repeats (VNTRs) likely drive some of the strongest GWAS signals for multiple traits.[2]

Short tandem repeats (STRs, also known as microsatellites) are a type of complex variant consisting of repeat units between 1 and 6 base pairs duplicated multiple times in succession. Over 1 million STRs occur in the human genome,[6] each spanning tens to thousands of base pairs. STRs frequently mutate, resulting in gains or losses of repeat units,[7] with average per-locus mutation rates orders of magnitude higher than rates for SNPs[8] or indels.[9] Large repeat expansions at STRs are known to result in Mendelian diseases such as Huntington's, muscular dystrophies, hereditary ataxias, and intellectual disorders.[10,11]

Recent evidence suggests that modest but ubiquitous variation at multi-allelic non-coding STRs is also relevant. We and others have associated STR lengths with both gene expression[3,12,13] and splicing.[14,15] The impact of non-coding STRs on gene expression is hypothesized to be mediated by a variety of mechanisms including modulating nucleosome positioning,[16] altering methylation,[12,17] affecting transcription factor recruitment,[3] and impacting non-canonical secondary DNA[18,19] and RNA[20,21] structure formation. This suggests that STRs potentially play an important role in shaping complex traits in humans.

Despite this, STRs are largely excluded from reference haplotype panels[22–24] and downstream GWAS analyses, as STRs are not directly genotyped by microarrays and are challenging to analyze from whole-genome sequencing (WGS) data. While some STRs are in high linkage disequilibrium (LD) with nearby SNPs, many are highly multi-allelic and imperfectly tagged by individual common SNPs, which are typically biallelic. Thus, effects driven by repeat-length variation have likely not been fully captured, especially for highly multi-allelic STRs.

Recent advances now enable incorporation of STRs into GWASs. We and others have created bioinformatics tools to genotype STRs directly from WGS by statistically accounting for the noise inherent in STR sequencing.[6,25–29] Using one of these tools, we developed a reference haplotype panel consisting of both SNP and STR genotypes that allows for imputing STRs into genotype array data[30] from samples lacking WGS data. In that study we found that all but the most highly polymorphic STRs are amenable to imputation in European cohorts, with an average per-locus concordance of 97% between imputed and WGS genotypes.

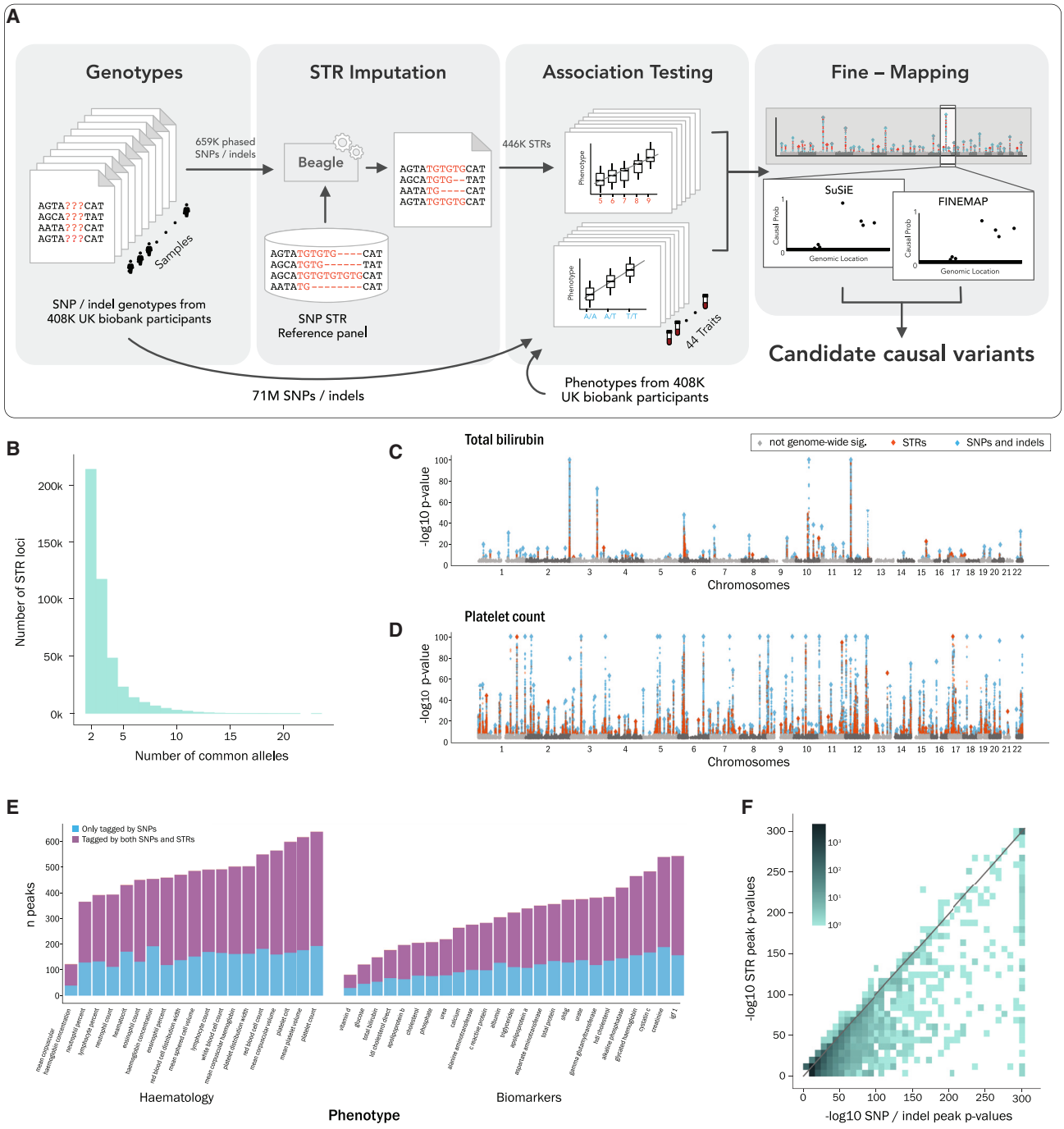


Figure 1. Genome-wide association tests identify STRs and SNPs associated with blood and biomarker traits in the UKB

(A) Schematic overview of this study. STRs are imputed into phased variants obtained from genotype arrays. GWASs are performed on SNPs and STRs in parallel. Regions with significant signals are identified and then fine-mapped using two methods each under multiple scenarios, resulting in candidate causal STRs.

(B) Distribution of the number of common alleles at imputed STRs. We define common alleles as alleles with estimated frequency ≥1% (STAR Methods). For clarity, we omitted the 237 imputed STRs with only a single common allele.

(C and D) Representative association results. Manhattan plots are shown for (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 variant per 10 Mb for visualization). $-\log_{10}$ p values are truncated at 100. Blue, SNPs; orange, STRs.

(E) Summary of signals identified per trait. Bars show the number of peaks per phenotype. Blue denotes peaks only containing genome-wide significant SNPs, and purple denotes peaks containing both significant SNPs and STRs. Peaks only containing significant STRs are too few to be visible in this display.

(legend continued on next page)

Here, we leveraged that reference panel to impute 445,720 genome-wide STRs into SNP array data from 408,153 White British individuals in the UK Biobank (UKB) for which deep phenotype information is available.[31] Whereas a recent publication studied the effects of 118 protein-coding VNTRs (with repeat units of 7+ base pairs) on complex traits,[2] our study focuses on genome-wide STRs (namely with repeat units of 1–6 bp), most of which are non-coding. We tested 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, and 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 5.2%–7.6% of signals identified by GWASs for these traits. We observed that some fine-mapping results are substantially influenced by the choice of fine-mapper or are sensitive to data-processing choices and fine-mapper instabilities, and thus require careful interpretation. After restricting to signals that consistently fine-mapped across multiple fine-mappers and settings, we identified 93 unique STRs strongly predicted to be causal for at least one trait. We highlight STRs from this set, which we predict drive some of the strongest hits for multiple traits, including apolipoprotein B and platelet traits. Overall, our study demonstrates the widespread role of polymorphic tandem repeats and highlights the need to consider a broad range of variant types in GWASs and fine-mapping.

RESULTS

Performing genome-wide STR association studies in 44 traits

We imputed genotypes for 445,720 autosomal STRs into phased genotype array data from 408,153 UKB White British individuals using Beagle[32] in combination with our published SNP-STR reference haplotype panel[30] (Figure 1A, STAR Methods, and Figure S1). This imputation yielded genotypes broadly similar to those of WGS (see below). Compared to common SNPs, which are typically biallelic, the imputed STRs are highly multi-allelic (Figure 1B). We tested STRs for association with 44 quantitative blood cell count and other biomarker traits (Table S1), which were available for between 304,658 and 335,585 genetically unrelated individuals. To facilitate this and other STR association studies, we developed associaTR (see key resources table), an open-source software package for identifying associations between STR lengths (measured by the number of repeat units) and phenotypes.

For each STR-trait pair, we used associaTR to test for linear association between STR dosage (the sum of the imputed allele length dosages of both chromosomes) and the trait measurement (Figures 1C and 1D). We used plink[33] to perform similar association tests for 70,698,786 SNP and short indel variants that were imputed into the same individuals[31] (hereafter referred to collectively as SNPs for brevity). For all associations, we included as covariates SNP-genotype principal components, genetic sex, and age (STAR Methods). Additional covariates were included on a per-trait basis (Table S1). We compared the output of our SNP analysis pipeline to results reported by Pan UKBB[34] and found that our pipeline produced similar results, with slightly weaker p values, likely due to not using a linear mixed model (Figure S2).

We compared signals identified by SNPs to those identified by STRs. For each trait we defined peaks as non-overlapping 250-kb intervals centered on the lead genome-wide significant variant (an SNP or STR with $p < 5e{-}8$) in that interval (STAR Methods). We identified 389 peaks per trait on average, with blood cell count traits generally more polygenic than other biomarkers (Figure 1E). Of these peaks, 65.9% contained both a significant STR and a significant SNP, 32.5% contained only significant SNPs, and 1.7% contained only significant STRs. The majority of strong peaks (containing any variant with $p < 1e{-}100$) were identified by both STRs and SNPs, in that they contain both an STR and an SNP with $p < 1e{-}80$. No new strong peaks were identified only by STRs (Figure 1F), which is unsurprising, since the STRs were imputed from SNP genotypes. Overall, p values of the lead SNP 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.

Fine-mapping suggests that 5.2%–7.6% of signals are driven by STRs

We applied statistical fine-mapping to identify causal variants that may be driving the GWAS signals detected above. We used two fine-mapping methods, SuSiE[35] and FINEMAP.[36] These methods differ in their modeling assumptions and thus provide partially orthogonal predictions. For each trait we divided its genome-wide significant variants (SNPs and STRs) and nearby variants into non-overlapping regions of at least 250 kb (STAR Methods). This resulted in 14,491 fine-mapping trait regions (Table S2), 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 for each fine-mapper to be the fine-mapper's prediction of that variant's chance of being causal. We defined a variant's FINEMAP CP to be the posterior inclusion probability FINEMAP calculated for that variant. We defined a variant's SuSiE CP to be the maximal SuSiE alpha value for that variant across pure credible sets (Figures S3 and S4). We further explain these choices in Note S1.

We used two approaches to study the contribution of STRs vs. SNPs to fine-mapped signals. First, we focused on the genome-wide significant variants (STRs or SNPs) with CP $\geq$ 0.8 (these accounted for a minority of the 21,045 pure signals detected by SuSiE and the 33,756 signals detected by FINEMAP). SuSiE identified 4,494 such variants and FINEMAP identified 5,170. Of these, 7.4% (range 1.3%–13.0% across traits; SuSiE) and 7.6% (range 1.4%–14.0%; FINEMAP) are STRs. Among the

(F) Comparison between lead SNP and STR p values at each peak. If there are no STRs in a peak, the y coordinate is set to zero (equivalently for SNPs). p values are capped at 1e−300, the maximum precision of our pipeline. Color shading represents the number of peaks falling at each position on the graph. The bottom-left tile (which only contains peaks whose lead SNP and STR variants fall in the least significant bin) has been removed so as not to not skew the color bar's scale. See also Table S1; Figures S1 and S2.

subset of variants identified by both methods (3,961), 5.4% (range 1.0%–11.1%) are STRs. Second, we considered the sum of CPs from all genome-wide significant variants in all trait regions, thereby accounting for the many signals not resolved to a single variant. STRs make up 5.2% (range 1.1%–6.8% across traits) of the total SuSiE CP sum and 7.4% (range 2.9%–9.0%) of the total FINEMAP CP sum. A potential limitation of this second metric is that variants with small CPs (CP $\leq$ 0.1) represent a large fraction (29.3%, SuSiE; 35.1%, FINEMAP) of these totals (Figure S5). Additionally, our results below suggest that a sizable subset of variant CPs are either discordant between fine-mappers or unstable, particularly for STRs (Notes S2 and S3), impacting both metrics. Nevertheless, these results suggest that 5.2%–7.6% of causal variants identifiable from GWASs can be attributed to an STR, regardless of the fine-mapping method or metric. This is comparable to the percentage of non-major alleles per person and is roughly half the percentage of per-person base-pair variation, accounted for by STR lengths as compared to SNPs in our study (Table S3). Table S4 reports the 511 genome-wide significant STR associations across 409 distinct STRs with either FINEMAP or SuSiE CP $\geq$ 0.8, and Table S5 shows a subset of those that pass stringent thresholding (see below).

To evaluate the reliability of our approach for determining the relative contributions of STRs vs. SNPs, we performed fine-mapping simulations assuming a simple additive model. We used two strategies for simulating phenotypes, in each case simulating only causal SNPs, and assessed to what extent STRs were incorrectly identified by SuSiE or FINEMAP as contributing to the underlying signals. For the first strategy, we randomly chose between one and three causal SNPs for a total of 1,644 simulations. For the second, we chose the causal variants to be those indicated by SuSiE as being potentially causal for a representative real trait (platelet count), thereby attempting to simulate properties of truly causal variants, for a total of 1,374 simulations. These procedures and rationales are described in STAR Methods, Table S6, and Figure S6.

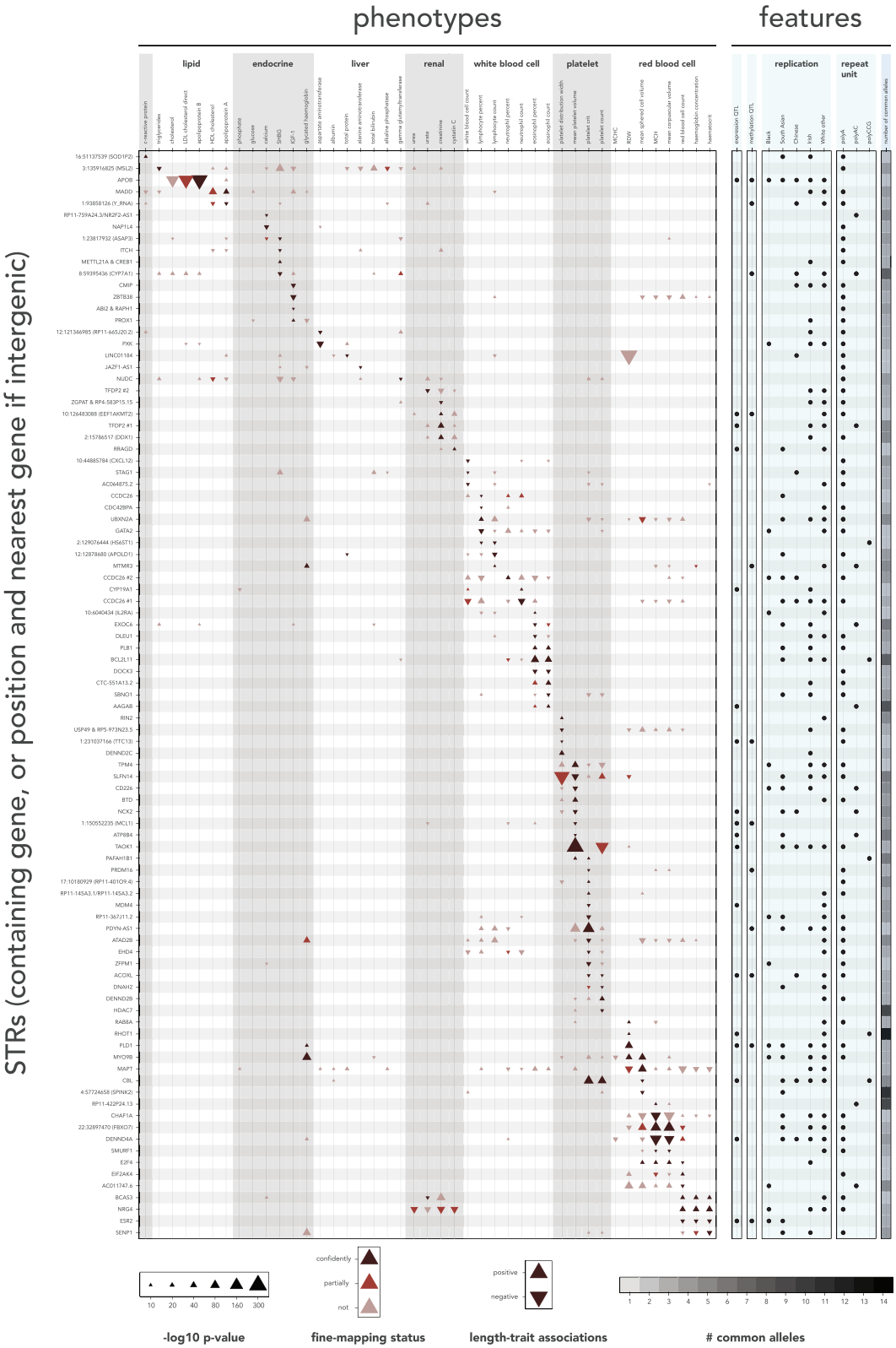
In simulations with randomly chosen causal SNPs, STRs comprised between 0% and 0.46% of genome-wide significant variants with CP $\geq$ 0.8. In contrast, using phenotypes simulated from the second strategy, 1.4%–3.2% were STRs (Table S7). Of the total CP assigned by SuSiE or FINEMAP to genome-wide significant variants, 0.50%–0.95% and 3.1%–3.2% were assigned to STRs in the first and second simulation strategies, respectively (Table S8). These numbers are uniformly lower than the 5.2%–7.6% contribution estimate above. This suggests that if the 44 traits studied here have genetic architectures similar to the simulated phenotypes, the results above are unlikely to be fully explained by systematic bias of fine-mapping in favor of STRs. However, we expect there are complexities of the genetic architecture of blood traits that we did not simulate, and we cannot rule out the possibility that they cause such bias. These results also suggest that some fine-mapped STRs likely are false positives. On the other hand, we observed that a large fraction (66%–81%) of simulated causal SNPs are not assigned CP $\geq$ 0.8 by fine-mapping and observed a similar lack of sensitivity in limited simulations including causal STRs (Table S7). We expect this low sensitivity is a greater source of uncertainty regarding the relative contribution of variant types than the false-positive rates.

We evaluated imputation quality at the 409 STRs in Table S4 by comparing imputed genotypes to genotypes obtained from recently released WGS data for 200,025 UKB individuals. At each locus we computed the Pearson r^2 between imputed length dosages and WGS length sums, in addition to other metrics (Table S4). Per-locus r^2 values are greater than 0.9 for 78.7% of these STRs and greater than 0.8 for 92.7%. Other imputation concordance measurements perform comparably (Figure S7 and STAR Methods). Overall, they suggest that fine-mapping with imputed data is unlikely to systematically differ from fine-mapping with hard-called genotypes for these loci. Results below are based on imputed genotypes unless otherwise stated.

Identifying and characterizing confidently fine-mapped STRs

We performed additional analyses to identify high-confidence causal STR candidates. First, we noticed that SuSiE and FINEMAP assigned highly discordant CPs to a subset of variants (Note S2 and Figures S8–S10). Thus, we conservatively narrowed our focus to the 167 candidate STR associations with association p values <1e−10 and with CP $\geq$ 0.8 in both FINEMAP and SuSiE. Second, to confirm that the fine-mappers' settings did not appreciably influence our results, we reran SuSiE and FINEMAP under a range of alternative settings (STAR Methods). These additional runs tended to produce concordant results, but again for some STRs produced highly inconsistent CPs (Figures S11–S14), which we mostly attribute to imputation uncertainty and FINEMAP instability (Note S3). Thus, we further restricted our focus to the 118 (70.7%) of the 167 STR-trait associations that also maintained CP $\geq$ 0.8 across these additional runs. We refer to the STR-trait associations meeting these criteria as confidently fine-mapped STR associations. Lastly, we added an association with an STR in the *APOB* gene to this set, as it only failed to meet these criteria because the STR was simultaneously represented in both our imputed STRs and in the SNP and indel set (Note S4). In total, we report 119 confidently fine-mapped STR-trait associations corresponding to 93 distinct STRs, which we display in Figure 2 and Table S5.

We evaluated these results by measuring their replication rates in populations besides White British individuals, with the expectation that causal associations replicate at higher frequencies in other populations than non-causal associations, due to shared biological functionality. The UKB includes self-identified groups of 8,043 Black, 7,952 South Asian, 1,568 Chinese, 12,957 Irish, and 16,051 Other White participants who passed quality control, noting that we have chosen to use the population labels that participants saw and self-ascribed to in the UKB intake survey (STAR Methods). About 40% of each population has WGS data, similar to the White British population. Using that WGS data we validated the imputed genotypes of those populations for the STRs in Table S4, finding that 78.2% and 93.9% of per-locus dosage r^2 values are greater than 0.8 and 0.6, respectively, in the South Asian population, 45.5% and 84.8% in the Black population, and 64.3% and 84.8% in the Chinese population (Figure S7). These metrics



(legend on next page)

are weaker than in the White British population; this is expected given our largely European reference haplotype panel. Nevertheless, our results suggest that imputed genotypes are sufficiently accurate across these groups for downstream analysis.

For each trait, for each fine-mapping region for that trait identified among White British individuals, we tested each STR in that region for association with that trait in each of the other populations (Table S9; individual loci in Tables S4 and S5). As expected, signals replicate at a higher rate in the groups most closely related to our discovery cohort (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 (Figures 3 and S15). To quantitatively measure this trend, for each population we fit a logistic regression model using whether signals replicated in that population as the outcome, those associations' fine-mapping statuses as the independent variable, and their $-\log_{10}$ p value in the discovery cohort as a covariate. Those regressions further support that fine-mapped associations replicate at higher rates (Table S10). Additionally, the models predict that confidently fine-mapped STR associations replicate at higher rates than STR associations 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 confidently fine-mapped STRs. This set contains 62 poly(A) repeats, 11 poly(AC) repeats, 5 poly(CCG) repeats, and 15 repeats with other units. Twelve of these overlap coding or untranslated regions (UTRs) (Tables 1 and S11; the two protein-coding repeats are described in Note S4 and Figure S16). Compared to genome-wide significant STRs, confidently fine-mapped STRs were more likely to be exonic trinucleotide STRs, in 5′ UTR regions or in non-protein-coding genes (two-sided two-sample test of difference between proportions: p = 2e−26, 1e−3, and 2e−4, respectively) (Figure S17 and STAR Methods). No other annotations showed significant signal after multiple hypothesis correction, likely due to the small number of confidently fine-mapped STRs. Lastly, we observed that 18 confidently fine-mapped STRs are significant *cis* expression quantitative trait loci (QTLs) and 12 are significant *cis* DNA methylation QTLs in the Genotype-Tissue Expression (GTEx) dataset[37] (Figure 2, Tables S12–S14, Figure S18, and STAR Methods). We note that the GTEx analyses were underpowered due to low sample sizes, particularly for relevant tissue types (e.g., kidney and liver).

Fine-mapped STRs capture known associations

We identified multiple fine-mapped STRs previously demonstrated to have functional roles, supporting the validity of our pipeline. For instance, our confidently fine-mapped set implicates a protein-coding CTG repeat (Table S11) to be causal for one of the strongest apolipoprotein B signals (two-sided association t test, p = 1e−279; in one of four apolipoprotein B peaks with minimal p value exceeding our numeric precision). Apolipoprotein B forms the backbone of low-density lipoprotein (LDL) cholesterol lipoproteins,[38] and this locus is also one of the strongest LDL signals (p = 6e−236; fifth most significant peak), with this STR marked as causal in eight of nine LDL fine-mapping runs. This repeat is biallelic in the UKB cohort, with a three-residue deletion (Leu-Ala-Leu) in the signal peptide in the first exon of the apolipoprotein B gene as the alternative allele.[39] It is an imperfect deletion in the CTG repeat, with sequence CTGGCGCTG. In agreement with previous findings,[40] we found the short allele to be associated with higher levels of both analytes. We discuss this locus further in Note S4.

As another example, our initial fine-mapping implicates a multi-allelic AC repeat (Table S11) 6 bp downstream of exon 4 of *SLC2A2* (also known as *GLUT2*, a gene most highly expressed in liver) as causally impacting bilirubin levels (p = 9e−18). However, this repeat was not confidently fine-mapped due to its FINEMAP CP of 0.61 not passing our 0.8 threshold, despite its SuSiE CP of 0.99. The potential link between *SLC2A2* and bilirubin is described in Note S5. Previous studies in HeLa and HEK293T cell lines showed that inclusion of exon 4 of *SLC2A2* is repressed by the binding of mRNA processing factor hnRNP L to this repeat,[41,42] 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 exon 4 splicing, although we did find evidence for association with exon 6 splicing (Figure S19).

A trinucleotide repeat in *CBL* regulates platelet traits

Most confidently fine-mapped STR associations identified here have, to our knowledge, not been previously reported. This includes positive associations between the length of a highly polymorphic CGG repeat in the promoter of the gene *CBL* and both platelet count (p = 4e−83) and platelet crit (p = 6e−103; 11th most significant platelet-crit peak; Figures 4A and 4B; Table S11; Figure S20). This finding fits the trend of CG-rich repeats in promoter and 5′ UTR regions being strongly implicated

Figure 2. STRs are confidently fine-mapped to causally impact many traits

Only STRs with a confidently fine-mapped association are shown. Triangles represent 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 values <1e−10. Triangle direction (up or down) indicates the sign of the association between STR length and the trait. Triangle size represents association p value. Similar traits are grouped on the x axis by white and light-gray bands. STRs are grouped on the y axis according to the traits to which they were confidently fine-mapped. STRs in genes are labeled by those genes (protein-coding genes preferred), intergenic STRs by chromosomal location and nearest gene. *CCDC26* and *TFDP2* each contain two confidently fine-mapped STRs and appear twice. Light-blue rows indicate (from left to right): which STRs are associated with the expression of a nearby gene (adjusted p < 0.05; Table S12), associated with the methylation of a nearby CpG site (Table S14), replicate with the same direction of effect in other populations (adjusted p < 0.05; STAR Methods), repeat unit, and the number of common alleles (defined in Figure 1; see scale beneath). Additionally, we mark the STRs in *TAOK1* and *RHOT1* as expression QTLs although they failed WGS call-rate filters in GTEx, as the *TAOK1* STR was associated with *TAOK1* expression when imputed into GTEx (STAR Methods) and the *RHOT1* STR was associated with *RHOT1* expression in the Geuvadis dataset (STAR Methods). The data summarized here are available in Tables S4, S5, S12, and S14.

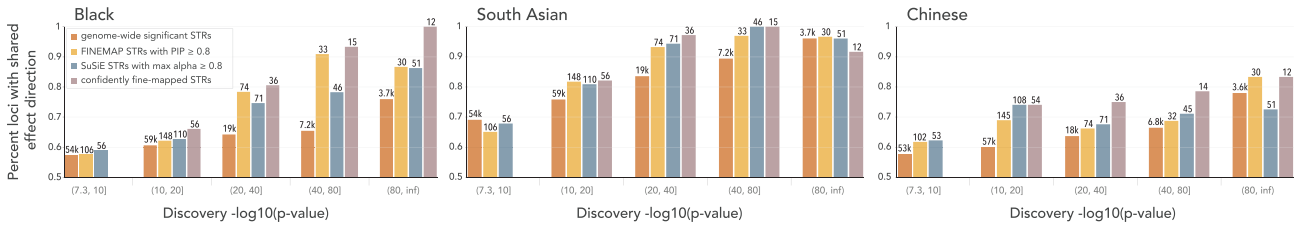


Figure 3. Concordance of White British STR effect directions in Black, South Asian, and Chinese populations

The y axis gives the fraction of STR associations measured in the White British discovery population that have the same effect direction when measured in the replication population (regardless of p value). Parentheses beneath the x axis denote the binning of discovery $-\log_{10}$ p values. Brown, genome-wide significant associations (discovery p < 5e−8); orange, FINEMAP STR associations (discovery p < 5e−8 and FINEMAP CP ≥ 0.8); teal, SuSiE STR associations (discovery p < 5e−8 and SuSiE CP ≥ 0.8); 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 values <1e−10; thus, they do not appear in the leftmost bin. This figure is somewhat sensitive to the choice of p-value bin boundaries, so we additionally analyze these data using logistic models (Table S10). See also Figure S15.

in transcriptomic regulation,[13] often via epigenomic regulation.[43,44] This repeat's association with mean sphered cell volume is also confidently fine-mapped (p = 7e−16; Figure S21), but that signal is weaker and we do not discuss it. For both the platelet crit and platelet count phenotypes, SuSiE and FINEMAP identify two genome-wide significant signals in this region, one of which they both localize to this STR. After conditioning on a lead variant from the other signal (rs2155380), this STR becomes the lead variant in the region by a wide margin (Figures 4C and 4D). Conditioning on both rs2155380 and this STR accounts for all the signal in the region (Figure 4E), supporting the fine-mappers' prediction that there are two signals in this region, one of which is driven by this STR.

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 in imputed genotypes between 0.023 [White British] and 0.175 [Chinese]) (Figure 4A) and in strong LD with the lead variant of the other signal (rs2155380, White British r^2 = 97.8%). 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) and is associated with *CBL* expression in the GTEx cohort (minimum p = 2.04e−18 in esophagus muscularis), 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 this imperfection (Figure 4F). This suggests that imperfections and repeat lengths are different characteristics of STRs and may have distinct associations.

The imputation of this STR displays relatively modest levels of concordance with WGS data (r^2 = 0.582 between imputation length dosages and WGS length sums; Table S5). Yet, reassuringly, hard-called genotypes from WGS show similar trends with both platelet traits (Figures 2B, S20A, S20C, and S20D). Further, this STR's allele length distributions in the UKB are highly concordant with those in the 1000 Genomes Project (Figures 4A and S22).

CBL codes for a protein in the RING finger subfamily of E3 ubiquitin ligases—a class of proteins, each with specific target molecule(s), that ubiquitinate their targets, priming them for downstream degradation. CBL targets the thrombopoietin receptor MPL,[45] thereby downregulating thrombopoietin signaling.[46] As thrombopoietin is the primary positive regulator of platelet production,[47] this implicates *CBL* as a negative regulator of platelet production. As further evidence, controlled experiments in mice demonstrate that loss of *CBL* function in megakaryocytes, the bone marrow platelet progenitor cells, results in increased platelet counts.[48] Further, we observed that increased CGG length is negatively associated with *CBL* expression in three GTEx cohort tissues[37] (p values <0.05 after multiple hypothesis correction; Figure 4G and Table S12) and in European individuals in the Geuvadis cohort[49] (p = 0.007; Figure 4H). Combined, all these data lead to an overall hypothesis that longer CGG repeat alleles contribute to increased platelet count by decreasing *CBL* expression (Figure 4I).

Additional confidently fine-mapped STR-trait associations

We observe a 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) (Table S11). 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 (Figure S23). *BCL2L11* is a pro-apoptotic regulatory protein and is required in the tightly regulated lifespan of myeloid lineage cells,[50] which include eosinophils. One mouse-model study showed that loss of repression of *BCL2L11* lowered eosinophil counts,[51] and another showed that *BCL2L11* knockout increased granulocyte counts, a class of cells including eosinophils.[52] This implicates *BCL2L11* in the regulation of eosinophil count, supporting the connection we observe between eosinophil count and this STR's length.

While exonic repeats are easier to interpret, most of our confidently fine-mapped STRs fall in intronic regions. We resolve one of the strongest signals for mean platelet volume (p < 1e−300; one of 12 peaks with p values exceeding our numeric precision) to a multi-allelic poly(A) STR in an intron of the gene *TAOK1* (Table S11 and Figure S24A). Conditioning on this STR's length demonstrates that it explains most of the signal in this region (Figure S24B). The same STR also shows a strong association with platelet count, with p = 2e−181 and a SuSiE CP of 1.

TAOK1 is a protein kinase that plays a role in regulating microtubule dynamics,[53] and microtubule function is known to be

Table 1. Confidently fine-mapped STRs are identified in coding regions and UTRs

STR coordinate (hg19 chr:pos)	Reference allele	Repeat unit	Trait	Association p value	Association *Z* score	Gene (annotation)	Transcription direction
1:204527033	$(TAA)_9$	AAT	platelet crit	5.76e−17	−8.37	*MDM4* (3′ UTR)	+
2:21266752	$(CAG)_6$(CGCAGGCAG) $[CGC(CAG)_2]_2$CGC	CTG (polyleucine)	apolipoprotein B	1.37e−279	−35.76	*APOB* (coding)	−
2:106510441	$(AC)_6$GTG$(CA)_{10}$C$(TA)_7$T	AC	mean platelet volume	6.93e−29	−11.15	*NCK2* (3′ UTR)	+
2:111878544	(CGC)$(CGCTGC)_2(CGC)_{13}$C	CCG	eosinophil count; eosinophil percent	4.96e−58; 5.88e−75	+16.06; +18.32	*BCL2L11* (5′ UTR)	+
2:204311891	$T_4CT_4CT_3CT_{18}$	T	IGF-1	3.97e−11	−6.61	*ABI2* (3′ UTR*) ENST00000295851.10 (1)	+
6:90121977	$(TC)_7$	TC	cystatin C	1.24e−16	+8.28	*RRAGD* (3′ UTR)	−
11:119077000	$(CGG)_{11}$C	CGG	mean sphered cell volume; platelet count; platelet crit	6.88e−16; 3.77e−83; 6.07e−103	−8.07; +19.32; +21.55	*CBL* (5′ UTR*) ENST00000634586.1 (5)	+
15:40312923	T_{16}	T	red blood cell count	2.27e−20	+9.25	*EIF2AK4* (not protein coding*) ENST00000558743.1 (2)	+
16:67229794	$(CAG)_{13}$(CAA)(CAG)(TAA)$(CAG)_3$	AGC (polyserine)	mean sphered cell volume; red blood cell count; mean corpuscular haemoglobin; mean corpuscular volume	3.07e−23; 1.08e−13; 2.83e−23; 9.27e−26	+9.93; −7.43; +9.94; +10.49	*E2F4* (coding)	+
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)	−
20:32971954	A_{20}	A	shbg	6.37e−15	−7.80	Y RNA ENST00000364628.1 (3)	−

Imputed alternative alleles and rsIDs are provided in Table S11. Here repeat units are calculated as in STAR Methods, except that they are required to be in the direction of transcription of the containing gene. For STRs fine-mapped to multiple traits, we list those traits and their corresponding p values and *Z* scores separated by semicolons. We denote with asterisks the STRs that only appear in non-canonical transcripts for their genes from Ensembl release 106. Additionally, two STRs in this list only appear in transcripts or genes that are not protein coding. For all those STRs, we provide Ensembl transcript numbers followed by parentheses containing the Ensembl transcript support level, a number from 1 to 5, with larger numbers indicating lower levels of evidence. The protein-coding repeats in *APOB* and *E2F4* are further analyzed in Note S4. See also Table S5.

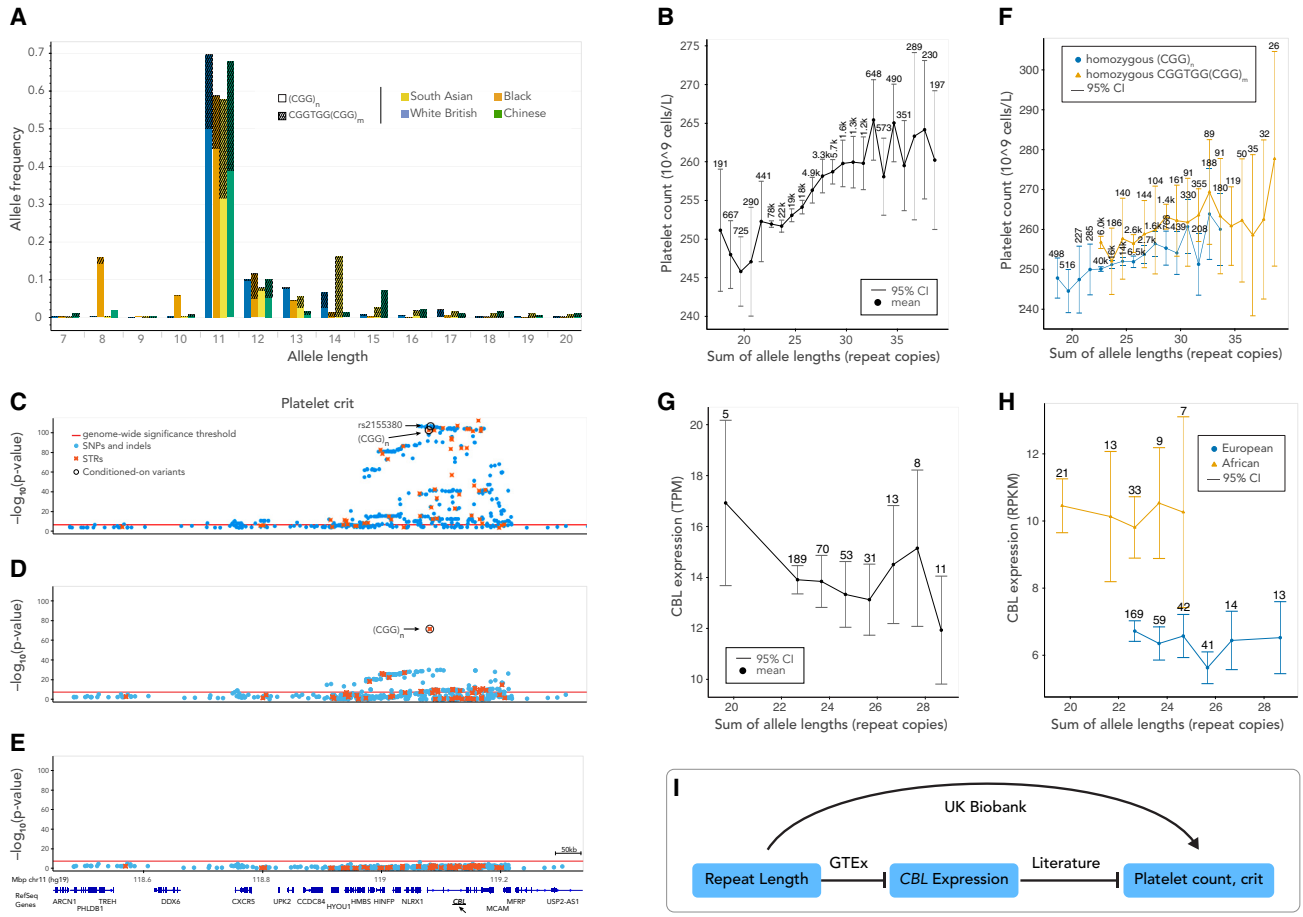


Figure 4. A polymorphic CGG repeat in the promoter of *CBL* influences platelet traits

(A) Distribution of STR alleles across populations. The x axis gives STR length (in number of full repeat unit lengths, using WGS data), and the y axis gives the population frequency. The hatched portion of each bar corresponds to those alleles that include a TGG imperfection at the second repeat (rs7108857). We label the ultra-rare alleles with T imperfections at other locations as "perfect" for these analyses. Colors denote different populations. Allele lengths 3–6, 21–33, 36, and 37 each have frequency less than 1% in all populations and are omitted.

(B) STR length vs. platelet count. STR-length sums were calculated from WGS data on (potentially related) White British participants that passed quality control. Error bars correspond to 95% confidence intervals. Only allele length sums with a frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum.

(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; orange, STRs. Red line, genome-wide significance threshold; black circles, the $(CGG)_n$ STR and rs2155380.

(F) STR length vs. platelet count conditioned on the TGG imperfection rs7108857. STR-length sums were calculated as in (B). Blue, individuals homozygous for no imperfection (n = 86,974); orange, individuals homozygous for the imperfection (n = 11,778). For each category, only length sums with a frequency of 0.2% or greater in that category are displayed.

(G and H) STR length vs. *CBL* expression. Associations are shown for cultured fibroblasts from GTEx (n = 393) (G) and LCLs from Geuavdis (n = 447) (H). Black, all available GTEx data regardless of population; yellow, African; blue, European. Only allele length sums with at least five corresponding participants are displayed.

(I) Proposed pathway for the effect of STR length on platelet traits. The arrow denotes a positive association, and the capped lines denote negative associations. Interactions are captioned by their information sources.

See also Figures S20–S22.

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 of *TAOK1* (ENST00000577583, which contains a retained intron) and is approximately 2.4 kb 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 an ERS1 transcription factor binding site (STAR Methods). Although this STR was filtered from our initial GTEx callset due to low call rate (11%), we imputed it into SNP data from that cohort (STAR Methods). While we did not identify significant associations between repeat length and splicing of any nearby exons, STR lengths showed significant negative correlation with *TAOK1* expression in five tissues (strongest p value 8e−6 in thyroid; Figures S24C and S24D). The repeat also showed significant associations with the expressions of nearby genes *ANKRD13B* and *TP53I13*,

although their potential role in platelet regulation is less clear (Table S12).

Another confidently fine-mapped example identifies an association between a GTTT repeat in an intron of estrogen receptor beta (*ESR2*) and haemoglobin concentration (p = 1e−24), red blood cell count (p = 3e−24), and haematocrit (p = 1e−26), where additional repeat copies correspond to lower measurements of all three traits (Note S5, Figure S25, and Table S11). Despite the weak discovery p value and differing allele distribution with the White British population (Figure S25C), these associations replicate in the Black population with p values <0.05. Further, consistent with these associations, *ESR2* ligand 17β-estradiol has been implicated in the regulation of red blood cell production.[55,56] We found significant negative associations between STR length and *ESR2* expression in two GTEx tissues (p values <0.05 after multiple hypothesis correction; Table S12). The expected direction of the effect of *ESR2* on red blood cell production is unclear given the highly tissue-specific isoform usage and functions of this gene (Note S5). Nevertheless, our results support a role for this STR in red blood cell production through regulation of *ESR2*.

We observed many additional interesting associations among the confidently fine-mapped STRs. For example, we find multiple confidently fine-mapped AC repeats that also significantly associate with expression of nearby genes. This includes a polymorphic AC repeat located in the 3′ UTR of *NCK2* that is associated with mean platelet volume (p = 7e−29; Figure S26 and Table S11). This repeat overlaps a binding site for the transcription factor PABPC1 and has a significant negative association with *NCK2* expression in multiple GTEx tissues (strongest p = 5e−7; Table S12). Separately, we find a highly polymorphic CCG repeat in the 5′ UTR of *RHOT1* that is associated with red blood cell distribution width (p = 7e−13; Table S11). WGS data show that our imputation of this locus is poor. Nonetheless, the effect of this STR is biologically plausible—it overlaps a CTCF binding site, is located within a nucleosome-depleted region of a H3K27ac peak in lymphoblastoid cell lines (LCLs), and shows strong association with the expression of *RHOT1* in LCLs in the Geuvadis dataset (p = 2e−44 in Europeans, p = 0.035 in Africans; Figure S27). Finally, many STRs in our fine-mapped set consist of poly(A) repeats. While traditionally these have been particularly challenging to genotype,[57] many such STRs, including poly(A) repeats in *MYO9B*, *DENND4A*, and *NRG4*, show strong statistical evidence of causality (Figure S2). Taken together, these loci exemplify the large number of confidently fine-mapped STRs our analysis provides for future study.

DISCUSSION

In this study, we imputed 445,720 STRs into the genomes of 408,153 UKB participants and associated their lengths with 44 blood cell and other biomarker traits. Using fine-mapping, we estimate that STRs account for 5.2%–7.6% of causal variants for these traits that can be identified by GWASs. We stringently filtered the fine-mapping output to produce 119 confidently fine-mapped STR-trait associations with strong evidence for causality, including some of the strongest signals for apolipoprotein B and platelet traits. These confidently fine-mapped STRs replicated in the Black, South Asian, and Chinese UKB populations at higher rates than non-fine-mapped STRs (each p < 0.02). A subset of these STRs is associated with the expression of nearby genes, explaining their effects via their plausible impact on regulatory processes.

Broadly, we highlight the importance of including more types of genetic variants in complex trait analysis. It has been proposed that STRs may represent an important source of the "missing heritability" in SNP-based GWASs.[58,59] Indeed, STRs, as well as VNTRs,[2] copy-number variants,[4] human leukocyte antigen types,[60] and some structural variants,[61] are often highly multiallelic and only imperfectly tagged by individual SNPs, suggesting that analyses omitting these variants may overlook important sources of causal variants and heritability. Further, we expect that incorporation of additional sources of causal variants, which often exhibit population-specific allele distributions, will improve applications such as polygenic risk scores, particularly in constructing scores that are transferable across populations.

Limitations of the study

While our results uncover many candidate causal STR variants, these findings are not 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. Further, whereas we performed association tests with a fixed-effects model, using a linear mixed model would increase power to detect additional associations. Additionally, computing constraints limited our analysis to a small number of traits. We hope that follow-up studies will extend this analysis to a wide variety of medically actionable traits.

Another limitation is that our study is based on imputed genotypes. Our SNP-STR reference panel only included 27.5% of the 1.6 million STRs in the HipSTR reference panel (see key resources table), due to the exclusion of STRs with low heterozygosity, non-autosomal STRs, most long repeats such as those implicated in pathogenic expansion disorders, and many STR alleles common only in non-European populations.[30] Further, imputed genotypes are inherently noisy, especially in non-European populations. Despite these limitations, analysis of WGS data released for 200,025 UKB participants[62] during the course of this study validated associations seen in imputed data. Subsequently, calls at 2.5 million STRs were released for 150,000 participants with WGS data.[62] Future studies performing STR-based GWASs solely using WGS datasets such as this will avoid these limitations.

Current challenges in statistical fine-mapping

Importantly, our study highlights that fine-mapping results are in some cases highly sensitive to the choice of fine-mapping tool and to a lesser extent to data-processing choices and fine-mapper instabilities, where one fine-mapping run would identify a variant as highly likely to be causal but a second would identify it as having no chance of causal impact. Further, our simulations suggest that fine-mappers have low sensitivity rates even when sample sizes are large and all model assumptions are met. This suggests that statistical fine-mapping results should be interpreted cautiously and evaluated for sensitivity to model choices

and that further work is needed to make statistical fine-mapping more robust.

Although fine-mapping inconsistencies existed for SNPs and STRs, they were more prevalent for STRs. While this may in part be due to STR imputation noise, more research is needed to evaluate the performance of fine-mapping tools on regions containing STRs. Additionally, there is need for fine-mapping tools that can model the effects of multi-allelic variants. In theory existing frameworks can handle 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. Similarly, existing tools often iteratively fit models by trading one causal variant for another variant in close LD, but greater accuracy may be obtained by trading a single, potentially causal, multi-allelic variant for multiple simultaneously causal biallelic variants.

Future directions

Methodological advances are needed to support the study of STRs. Here we developed associaTR, an open-source pipeline enabling studies to conduct STR-length-based association tests. However, integrating support for complex variants, including STR-length-based testing, into widely used GWAS toolkits would enable more routine analysis of the full spectrum of human genetic variation. Improvements to our models are also likely to reveal new insights. Here we only modeled linear associations between STR lengths and traits. Visualizations of the associations we identify suggest that linear models only partially approximate those signals and that they may be best described by non-linear models, such as quadratic or sigmoid relationships between repeat copy numbers and traits. However, fitting non-linear models requires modeling the effects of the two alleles at each locus separately while simultaneously controlling for overfitting and is a topic of ongoing work. We also only tested for associations with STR lengths. However, inspection of individual loci reveals that complex repeat structures are common (Tables 1 and S11). Systematic evaluation of potential epistasis between repeat imperfections and STR lengths, and between the lengths of neighboring repeats, would potentially improve our understanding of STR impacts.

Overall, our study provides a framework for incorporating hundreds of thousands of tandem repeat variants into GWASs, either via imputation or using WGS genotypes such as the newly released[62] UKB callset. Our study identifies dozens of candidate variants for future mechanistic studies and demonstrates that STRs likely make widespread contributions to complex traits.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Selection of UK Biobank participants
  - SNP and indel dataset preprocessing
  - STR imputation
  - Inferring repeat units
  - Phenotypes and covariates
  - Association testing
  - Comparison with Pan-UKB pipeline
  - Defining significant peaks
  - Identifying indels which are STR alleles
  - Fine-mapping
  - Alternative fine-mapping conditions
  - Fine-mapping simulations
  - WGS validation of imputed fine-mapped STRs
  - Replication in other populations
  - Logistic regression of replication direction
  - Gene, transcription factor binding annotation
  - Enrichment testing
  - Expression association analysis in GTEx
  - Methylation association analysis in GTEx
  - Targeted STR expression analysis in Geuvadis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100458.

ACKNOWLEDGMENTS

This research 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 her illustrations. We thank K. Frazer and M. D'Antonio for helpful discussions and comments. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this article were obtained from the GTEx portal and dbGaP accession numbers phs000424.v7.p2 and phs000424.v8.p2.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 21, 2022
Revised: September 9, 2023
Accepted: November 7, 2023
Published: December 13, 2023

REFERENCES

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

2. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Zheng, Y., McCarroll, S.A., and Loh, P.-R. (2021). Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science *373*, 1499–1505.
3. Grünewald, T.G.P., Bernard, V., Gilardi-Hebenstreit, P., Raynal, V., Surdez, D., Aynaud, M.-M., Mirabeau, O., Cidre-Aranaz, F., Tirode, F., Zaidi, S., et al. (2015). Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat. Genet. *47*, 1073–1078.
4. Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183.
5. Boettger, L.M., Salem, R.M., Handsaker, R.E., Peloso, G.M., Kathiresan, S., Hirschhorn, J.N., and McCarroll, S.A. (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat. Genet. *48*, 359–366.
6. Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods *14*, 590–592.
7. Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.
8. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., and Stefansson, K. (2012). A direct characterization of human mutation based on microsatellites. Nat. Genet. *44*, 1161–1165.
9. Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. USA *107*, 961–968.
10. Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.
11. Malik, I., Kelley, C.P., Wang, E.T., and Todd, P.K. (2021). Molecular mechanisms underlying nucleotide repeat expansion disorders. Nat. Rev. Mol. Cell Biol. *22*, 589–607.
12. Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R.S., Mittelman, D., and Sharp, A.J. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. *44*, 3750–3762.
13. Fotsing, S.F., Margoliash, J., Wang, C., Saini, S., Yanicky, R., Shleizer-Burko, S., Goren, A., and Gymrek, M. (2019). The impact of short tandem repeat variation on gene expression. Nat. Genet. *51*, 1652–1659.
14. Hefferon, T.W., Groman, J.D., Yurk, C.E., and Cutting, G.R. (2004). A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proc. Natl. Acad. Sci. USA *101*, 3504–3509.
15. Hui, J., Stangl, K., Lane, W.S., and Bindereif, A. (2003). HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. Nat. Struct. Biol. *10*, 33–37.
16. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K.J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. Science *324*, 1213–1216.
17. Martin-Trujillo, A., Garg, P., Patel, N., Jadhav, B., and Sharp, A.J. (2023). Genome-wide evaluation of the effect of short tandem repeat variation on local DNA methylation. Genome Res. *33*, 184–196.
18. Murat, P., Guilbaud, G., and Sale, J.E. (2020). DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. Genome Biol. *21*, 209.
19. Rothenburg, S., Koch-Nolte, F., Rich, A., and Haag, F. (2001). A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proc. Natl. Acad. Sci. USA *98*, 8985–8990.
20. Freudenreich, C.H. (2018). R-loops: Targets for Nuclease Cleavage and Repeat Instability. Curr. Genet. *64*, 789–794.
21. Niehrs, C., and Luke, B. (2020). Regulatory R-loops as facilitators of gene expression and genome stability. Nat. Rev. Mol. Cell Biol. *21*, 167–178.
22. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
23. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
24. Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.-F., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. *6*, 8111.
25. Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., Davis, M., Lamont, P., Clayton, J.S., Laing, N.G., et al. (2018). STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. *19*, 121.
26. Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. *47*, e90.
27. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., et al. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics *35*, 4754–4756.
28. Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., et al. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am. J. Hum. Genet. *101*, 700–715.
29. Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J., and Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am. J. Hum. Genet. *103*, 858–873.
30. Saini, S., Mitra, I., Mousavi, N., Fotsing, S.F., and Gymrek, M. (2018). A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat. Commun. *9*, 4397.
31. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
32. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348.
33. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.
34. Pan-UKB team (2020). Pan-ancestry Genetic Analysis of the UK Biobank. https://pan.ukbb.broadinstitute.org/.
35. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300.
36. Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501.
37. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
38. Berberich, A.J., and Hegele, R.A. (2022). A Modern Approach to Dyslipidemia. Endocr. Rev. *43*, 611–653. bnab037.
39. Boerwinkle, E., and Chan, L. (1989). 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.

40. Niu, C., Luo, Z., Yu, L., Yang, Y., Chen, Y., Luo, X., Lai, F., and Song, Y. (2017). Associations of the APOB rs693 and rs17240441 polymorphisms with plasma APOB and lipid levels: a meta-analysis. Lipids Health Dis. *16*, 166.

41. Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumüller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. *24*, 1988–1998.

42. Huang, Y., Li, W., Yao, X., Lin, Q.-J., Yin, J.-W., Liang, Y., Heiner, M., Tian, B., Hui, J., and Wang, G. (2012). Mediator complex regulates alternative mRNA processing via the MED23 subunit. Mol. Cell *45*, 459–469.

43. Sutcliffe, J.S., Nelson, D.L., Zhang, F., Pieretti, M., Caskey, C.T., Saxe, D., and Warren, S.T. (1992). DNA methylation represses FMR-1 transcription in fragile X syndrome. Hum. Mol. Genet. *1*, 397–400.

44. Garg, P., Jadhav, B., Rodriguez, O.L., Patel, N., Martin-Trujillo, A., Jain, M., Metsu, S., Olsen, H., Paten, B., Ritz, B., et al. (2020). 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.

45. Saur, S.J., Sangkhae, V., Geddis, A.E., Kaushansky, K., and Hitchcock, I.S. (2010). Ubiquitination and degradation of the thrombopoietin receptor c-Mpl. Blood *115*, 1254–1263.

46. Plo, I., Bellanné-Chantelot, C., Mosca, M., Mazzi, S., Marty, C., and Vainchenker, W. (2017). Genetic Alterations of the Thrombopoietin/MPL/JAK2 Axis Impacting Megakaryopoiesis. Front. Endocrinol. *8*, 234.

47. Kaushansky, K., Lok, S., Holly, R.D., Broudy, V.C., Lin, N., Bailey, M.C., Forstrom, J.W., Buddle, M.M., Oort, P.J., Hagen, F.S., et al. (1994). Promotion of megakaryocyte progenitor expansion and differentiation by the c-Mpl ligand thrombopoietin. Nature *369*, 568–571.

48. Märklin, M., Tandler, C., Kopp, H.-G., Hoehn, K.L., Quintanilla-Martinez, L., Borst, O., Müller, M.R., and Saur, S.J. (2020). C-Cbl regulates c-MPL receptor trafficking and its internalization. J. Cell Mol. Med. *24*, 12491–12503.

49. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

50. Shinjyo, T., Kuribara, R., Inukai, T., Hosoi, H., Kinoshita, T., Miyajima, A., Houghton, P.J., Look, A.T., Ozawa, K., and Inaba, T. (2001). Downregulation of Bim, a Proapoptotic Relative of Bcl-2, Is a Pivotal Step in Cytokine-Initiated Survival Signaling in Murine Hematopoietic Progenitors. Mol. Cell Biol. *21*, 854–864.

51. Kotzin, J.J., Spencer, S.P., McCright, S.J., Kumar, D.B.U., Collet, M.A., Mowel, W.K., Elliott, E.N., Uyar, A., Makiya, M.A., Dunagin, M.C., et al. (2016). The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. Nature *537*, 239–243.

52. Bouillet, P., Metcalf, D., Huang, D.C., Tarlinton, D.M., Kay, T.W., Köntgen, F., Adams, J.M., and Strasser, A. (1999). Proapoptotic Bcl-2 Relative Bim Required for Certain Apoptotic Responses, Leukocyte Homeostasis, and to Preclude Autoimmunity. Science *286*, 1735–1738.

53. Draviam, V.M., Stegmeier, F., Nalepa, G., Sowa, M.E., Chen, J., Liang, A., Hannon, G.J., Sorger, P.K., Harper, J.W., and Elledge, S.J. (2007). A functional genomic screen identifies a role for TAO1 kinase in spindle-checkpoint signalling. Nat. Cell Biol. *9*, 556–564.

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

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

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

57. Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nat. Biotechnol. *37*, 555–560.

58. Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability. Trends Genet. *26*, 59–65.

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

60. D'Antonio, M., Reyna, J., Jakubosky, D., Donovan, M.K., Bonder, M.-J., Matsui, H., Stegle, O., Nariai, N., D'Antonio-Chronowska, A., and Frazer, K.A. (2019). Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. Elife *8*, e48476.

61. Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.

62. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732–740.

63. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Preprint at bioRxiv.

64. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

65. Oliva, M., Demanelis, K., Lu, Y., Chernoff, M., Jasmine, F., Ahsan, H., Kibriya, M.G., Chen, L.S., and Pierce, B.L. (2023). DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. Nat. Genet. *55*, 112–122.

66. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598.

67. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*, e1001779.

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

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

70. Horta, D.. bgen-reader: Bgen File Format Reader.. https://bgen-reader.readthedocs.io/en/latest/index.html.

71. Pedersen, B.. cyvcf2: Fast Vcf Parsing with Cython + Htslib. http://brentp.github.io/cyvcf2/.

72. Collette, A.. Collaborators HDF5 for Python. https://www.h5py.org/.

73. The HDF Group (1997). Hierarchical Data Format. version 5. https://www.hdfgroup.org/HDF5/.

74. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

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

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

77. Fischer, B., Smith, M., and Pau, G. (2023). rhdf5: R Interface to HDF5. R Package Version 2.38.0. https://github.com/grimbough/rhdf5.

78. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

79. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proc. 9th Python Sci. Conf., 92–96.

80. Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., and Gymrek, M. (2021). TRTools: a toolkit for genome-wide analysis of tandem repeats. Bioinformatics *37*, 731–733.

81. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

82. Foix, A., and Blachly, J. (2021). pyEGA3: EGA Download Client.

83. O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.-F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nat. Genet. *48*, 817–820.

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

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

86. Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. PLoS One *9*, e85150.

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

88. Association Analysis - PLINK 2.0 https://www.cog-genomics.org/plink/2.0/assoc.

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

90. ENCODE Project Consortium; Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

91. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

92. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493–D496.

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

94. Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., and Hubner, N. (2015). Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). Curr. Protoc. Hum. Genet. *87*, 11.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
STR association testing results, by population and phenotype	This paper	https://gymreklab.com/science/2023/09/08/Margoliash-et-al-paper.html or as a dataset frozen at the time of publication on Dryad at the DOI https://doi.org/10.5061/dryad.z612jm6jk
Fine-mapping results by phenotype	This paper	https://gymreklab.com/science/2023/09/08/Margoliash-et-al-paper.html or as a dataset frozen at the time of publication on Dryad at the DOI https://doi.org/10.5061/dryad.z612jm6jk
1000 Genomes individuals	Auton et al.[23]	https://www.internationalgenome.org/data-portal/sample using the "Download the list" tab
1000 Genomes WGS data	Byrska-Bishop et al.[63]	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
Beagle-provided human genetic maps	Browning et al.[32]	https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/
GENCODE 38 (hg19)	Frankish et al.[64]	http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh37_mapping/gencode.v38lift37.annotation.gff3.gz
Geuvadis	Lappalainen et al.[49]	https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/files/analysis_results/?ref=E-GEUV-1
GTEx data portal	The GTEx Consortium[37]	https://www.gtexportal.org/home/datasets
GTEx expression data, exon read counts	The GTEx Consortium[37]	https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_exon_reads.parquet
GTEx expression data, junction read counts	The GTEx Consortium[37]	https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_STARv2.5.3a_junctions.gct.gz
GTEx expression data, TPM	The GTEx Consortium[37]	https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz
GTEx expression STRs, previously released	Fotsing et al.[13]	https://www.nature.com/articles/s41588-019-0521-9#Sec23
GTEx methylation data	Oliva et al.[65]	NCBI GEO database accession number GSE213478
GTEx methylation overview	Oliva et al.[65]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213478
GTEx methylation CpG locations	Oliva et al.[65]	https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE213478&format=file
GTEX WGS data	The GTEx Consortium[37]	dbGaP accession number phs000424.v8.p2
HipSTR STR reference	Willems et al.[6]	https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/
LiftOver chain file	Hinrichs et al.[66]	https://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz
LiftOver chain file	Hinrichs et al.[66]	ftp://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz
Methylation-STR dataset for validation	Martin-Trujillo et al.[17]	https://genome.cshlp.org/content/33/2/184.short
Pan-UKB manifest	Pan-UKB team[34]	https://docs.google.com/spreadsheets/d/1AeeADtT0U1AukliiNyiVzVRdLYPkTbruQSk38DeutU8
Pan-UKB overview	Pan-UKB team[34]	https://pan.ukbb.broadinstitute.org/downloads
Pan-UKB summary statistics for bilirubin	Pan-UKB team[34]	https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/biomarkers-30840-both_sexes-irnt.tsv.bgz
Pan-UKB summary statistics index for bilirubin	Pan-UKB team[34]	https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files_tabix/biomarkers-30840-both_sexes-irnt.tsv.bgz.tbi
SNP-STR reference panel	Saini et al.[30]	https://gymreklab.com/2018/03/05/snpstr_imputation.html
UKB data showcase search page	Sudlow et al.[67]	https://biobank.ctsu.ox.ac.uk/crystal/search.cgi

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
UKB genotypes, microarray and phased, release version 2	Bycroft et al.[31]	https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukbgene_instruct.html
UKB genotypes, imputed, release version 3	Bycroft et al.[31]	https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukbgene_instruct.html
UKB genotypes, whole genome sequencing	Halldorsson et al.[62]	https://ukbiobank.dnanexus.com/ under Bulk/Whole genome sequences/Whole genome CRAM files
UKB sample quality control file	Bycroft et al.[31]	European Genome-Phenome Archive accession EGAF00001844707
Software and algorithms		
AssociaTR, published as part of the TRTools[68] package	This paper	https://trtools.readthedocs.io/ and frozen at the time of publication on Zenodo at the DOI https://zenodo.org/records/10056105
Code for performing most of the analyses and generating most of the figures in this paper	This paper	https://github.com/LiterallyUniqueLogin/ukbiobank_strs/ and frozen at the time of publication on Zenodo at the DOI https://doi.org/10.5281/zenodo.8436632
Beagle v5.1 (build 25Nov19.28days)	Browning et al.[32]	https://faculty.washington.edu/browning/beagle/b5_1.html
Beagle v5.2 (beagle.28Jun21.220.jar)	Browning et al.[32]	https://faculty.washington.edu/browning/beagle/b5_2.html
bedtools	Quinlan et al.[69]	https://bedtools.readthedocs.io/en/latest/index.html
bgen-reader 4.0.8	Horta[70]	https://bgen-reader.readthedocs.io/en/latest/index.html
cyvcf2 0.30.14	Pedersen[71]	http://brentp.github.io/cyvcf2/
FINEMAP	Benner et al.[36]	http://christianbenner.com/
fusera	The MITRE Corporation	https://github.com/ncbi/fusera
h5py v3.6.0	Collette et al.[72]	https://github.com/h5py/h5py
HDF5	The HDF Group[73]	https://www.hdfgroup.org/HDF5/
HipSTR	Willems et al.[6]	https://github.com/gymrek-lab/HipSTR
Integrative Genomics Viewer	Robinson et al.[74]	https://igv.org/
LiftOver	Hinrichs et al.[66]	https://genome.ucsc.edu/cgi-bin/hgLiftOver accessed on 2023/03/09
PEER v1.0	Stegle et al.[75]	https://github.com/PMBio/peer/wiki/
plink v.1.90b3.44	Chang et al.[33]	https://www.cog-genomics.org/plink2/
plink2 v2.00a3LM (build AVX2 Intel 28 Oct 2020)	Chang et al.[33]	https://www.cog-genomics.org/plink/2.0/
PRIMUS v1.9.0	Staples et al.[76]	https://primus.gs.washington.edu/primusweb/
rhdf5 v2.38.0	Fischer et al.[77]	https://www.bioconductor.org/packages/release/bioc/html/rhdf5.html
Scipy.stats v1.7.3	Virtanen et al.[74]	https://docs.scipy.org/doc/scipy/reference/stats.html
smartpca included in EIGENSOFT v6.1.4	Price et al.[78]	https https://github.com/DReichLab/EIG
Statsmodels v0.13.2	Seabold et al.[79]	https://www.statsmodels.org/stable/index.html
SuSiE v0.11.42	Wang et al.[35]	https://stephenslab.github.io/susieR/index.html
TRTools v4.2.1 (including CompareSTR, DumpSTR and MergeSTR)	Mousavi et al.[80]	https://trtools.readthedocs.io/en/latest/
UCSC genome browser	Kent et al.[81]	https://genome.ucsc.edu/index.html
ukbgene utility (ver Jan 28 2019 14:09:15 - using Glibc2.28(stable))	UK Biobank	https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukbgene_instruct.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to Melissa Gymrek (mgymrek@ucsd.edu).

Materials availability

This study did not generate new reagents.

Data and code availability

- Original code has been deposited publicly on GitHub at https://github.com/LiterallyUniqueLogin/ukbiobank_strs and is available as a repository frozen at the time of publication on Zenodo at the DOI: https://doi.org/10.5281/zenodo.8436632
- STR association summary statistics and raw fine-mapping data have been deposited at https://gymreklab.com/science/2023/09/08/Margoliash-et-al-paper.html and are available as a dataset frozen at the time of publication on Dryad at the DOI: https://doi.org/10.5061/dryad.z612jm6jk.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

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[67] Resource ID 664 (see key resources table). 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[31] indicating potential DNA contamination. We further removed withdrawn participants, indicated by non-positive IDs in the sample file as well as by IDs in e-mail 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 (see key resources table)) from the European Genome-Phenome Archive (accession EGAF00001844707) using pyEGA3.[82] 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.[31] In concordance with previous analyses of this cohort[31] we additionally removed data for:

(1) 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)
(2) 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)
(3) 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. These variants were provided in reference genome hg19 coordinates, and all analyses in this study, unless otherwise specified, were performed with hg19 coordinates.

Phased hard-called genotypes: We downloaded the bgen files containing the hard-called SNP and indel haplotypes (release version 2) and the corresponding sample and fam files using the ukbgene utility (UKB Data Showcase Resource 664 (see key resources table)). These variants had been genotyped using microarrays and phased using SHAPEIT3[83] with the 1000 genomes phase 3 reference panel.[23] 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 1.9bp, maximal length 26bp), and 0.2% were short deletions (average length 1.9bp, maximal length 29bp).

Imputed genotypes: We similarly downloaded imputed SNP data using the ukbgene utility (release version 3). Variants had been imputed with IMPUTE4[31] using the Haplotype Reference Consortium panel,[22] with additional variants from the UK10K[24] and 1000 Genomes phase 3[23] reference panels. The resulting imputed variants contain 93,095,623 variants, consisting of 96.0% single nucleotide variants, 1.3% short insertions (average length 2.5bp, maximum length 661bp), 2.6% short deletions (average length 3.1bp, maximum length 129bp). This set does not include the 11 classic human leukocyte antigen alleles imputed separately.

We used bgen-reader[70] 4.0.8 to access the downloaded bgen files in python. We used plink2[33] 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 445,720 autosomal STR variants in 2,504 individuals from the 1000 Genomes Project[23,30] (see key resources table). 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.

The IDs listed in the 'str' column of Table S2 at that URL describe which variants in the reference panel are STRs and which are other types of variants. That produces a list of 445,715 unique variant IDs and 5 IDs which are each assigned to four separate variants in the reference panel VCFs. For the IDs with multiple assignments, we selected the variant that appeared first in the VCF and discarded the others, leaving 445,720 unique STR variants each with unique IDs.

While our analyses with these STRs were performed using hg19 coordinates unless otherwise stated, we also provide hg38 reference coordinates for these STRs in the supplemental tables. We obtained those coordinates using LiftOver[66] which resulted in identical coordinates as in HipSTR's[6] hg38 STR reference panel (see key resources table). All STRs successfully lifted over to hg38 coordinates.

To select shared variants for imputation, we note that 641,582 (97.4%) of SNP and indel variants that were 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 (see key resources table) 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 (Figure S1) and chose to remove the 110 variants with an alternate allele frequency difference of more than 12%.

We used Beagle[32] v5.1 (build 25Nov19.28days) with the tool's provided human genetic maps (see key resources table) 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 multi-allelic 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 i th haplotype is a, output by Beagle in the AP1 and AP2 FORMAT fields of the VCF file.

Estimated allele frequencies (Figure 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 chromosomal copies considered (equaling twice the number of unrelated participants) to obtain the estimated frequency of each allele.

Inferring 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 of 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 CAG would be AGC. If the most common standardized k-mer was less than twice as frequent as the second most common standardized k-mer, we did not call a repeat unit for that STR (11,962 STRs; 2.68%). Otherwise, the most common standardized k-mer was labeled as the forward-strand (based on the reference genome) repeat unit for that STR. To infer the strand-independent repeat unit for the STR we looked at all rotations of the forward-strand repeat unit in both the forward and reverse-complement directions and chose whichever comes first lexicographically. For example the repeat unit for the STR TGTGTGTG would be AC, while the forward-strand repeat unit would be GT. In the large majority of cases the repeat unit identified by this approach is the unit which is duplicated or deleted in alternate alleles, but this method of identifying repeat units does not consider alternate alleles and so does not make that guarantee.

Phenotypes and covariates

IDs listed in this section refer to the UKB Data Showcase[67] (see key resources table).

We analyzed a total of 44 blood traits measured in the UKB. 19 phenotypes were chosen from Category Blood Count (Data Field ID 100081) and 25 from Category Blood Biochemistry (Data Field 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 Table S1, 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). Table S1 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 we included aliquot number (0–3) as a categorical covariate during association testing as an additional step to mediate the dilution issue (described in Resource ID 5636). 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 for participants having measurements 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 (Table S1).

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 phenotype we excluded the <100 participants that were measured using aliquot 4, and that for 8 of the biomarker phenotypes we additionally excluded the $\leq$125 participants that were measured using aliquot 3 (Table S1).

For each phenotype we then selected a maximally-sized genetically unrelated subset of the remaining individuals using PRIMUS[76] v1.9.0. When multiple such maximal subsets existed (for instance, wherever a single individual needed to be chosen from a family of two), one subset was chosen arbitrarily, thus introducing some lack of reproducibility. Precomputed measures of genetic relatedness between participants (described in UKB paper supplement section 3.7.1[31]) 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 (Table S1).

Genetic sex and ancestry principal components (PCs) were included as covariates for all phenotypes. Participant sex was extracted from the fam file (described in the Participants Methods section 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 $\textit{normal quantile}\left(\frac{\textit{sample rank}+0.5}{\textit{n samples}}\right)$. We use rank-inverse normalization as it is standard practice, though it does not have a strong theoretical foundation[84] and only moderate empirical support.[68,85–87]

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. We developed associaTR to streamline performing association tests between STR length and quantitative traits. While our approach relies on a standard linear model, linear mixed models based on STR length dosages would likely result in increased power and will be considered in future studies. As our downstream analyses required STR and SNP associations to be comparable, we also used a standard linear model for SNP association testing.

For STR association testing, the imputed VCFs produced by Beagle were accessed in python with cyvcf2[71] 0.30.14 and v4.2.1 of our TRTools library.[80] In line with plink's recommendation for SNP GWAS,[88] 6 loci with non-major allele dosage <20 were filtered. For each STR, we fit the linear model $\vec{y} = \vec{g} * \beta_g + C * \vec{\beta_C} + \vec{\epsilon}$ where $\vec{y}$ is the vector of rank-inverse-normalized phenotype values per individual, $\vec{g}$ 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, $\vec{\beta_C}$ is the vector of covariate effect sizes, and $\vec{\epsilon}$ 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.[79] Per GWAS best-practices, we used imputation dosage genotypes instead of best-guess genotypes.[89]

We used plink2[33] 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. 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 $\vec{g}$ is the vector of dosages of the non-reference SNP or indel allele.

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

p values calculated from association testing are two-sided.

Comparison with Pan-UKB pipeline

We compared the results of our pipeline to results available on the Pan UKBB[34] Website (see key resources table) using bilirubin as an example trait. We matched variants between datasets on chromosome, position, reference and alternate alleles, excluding variants not present in both pipelines. We found our pipeline produced largely similar but somewhat less significant p values than those reported for European participants in Pan UKBB (Figure S2).

Defining significant peaks

Given a peak width w (bp), 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 $\geq 5e{-}8$, break. If there is a variant in either direction less than w bp 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 $w = 250kb$. The identification of peaks in Figures 1C and 1D was made with $w = 10mb$ for visualization purposes. Note that peaks centered on variants within $w/2$ bp of the end of a chromosome will necessarily be smaller than w bp in width.

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) 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. Note that peaks within 125kb of the end of a chromosome will necessarily be smaller than the minimum 125kb width in that direction.

This produced 14,494 trait-regions. Due to computational challenges during fine-mapping (see below), we excluded three regions (urate 4:8165642-11717761, total bilirubin 12:19976272-22524428 and alkaline phosphatase 1:19430673-24309348) from downstream analyses (see below), leaving 14,491 trait-regions.

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

SuSiE[35]: 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 $\leq$ 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[73] using h5py v3.6.0.[72] We used rhdf5 v2.38.0[77] 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 we considered 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 (Note S3). 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 pairwise absolute correlation $\geq$0.2 or size $\leq$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 Table S4.

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 $\geq$0.8 as its causal probability (CP) in our downstream analyses (or zero if it was in no such credible sets). See Note S1.

FINEMAP[36]: 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[73] with h5py v3.6.0.[72] 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-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 FINEMAP's posterior inclusion probability (PIP) 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 $\geq$ 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 (Table S5) to those that had p value < 1e−10 and CP $\geq$ 0.8 in the original runs and maintained CP $\geq$ 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. Additionally, we reran FINEMAP with no changed settings to examine potential FINEMAP instability.

See Note S3 for a more detailed discussion of these various settings and their impact on fine-mapping results.

Fine-mapping simulations

We simulated phenotypes under additive genetic models and fine-mapped those phenotypes separately at individual regions. Our simulations used real genotypes from White British UKB participants and focused on regions originally identified by our GWAS to maintain realistic LD patterns observed at regions with true signals. We used regions associated with platelet count as it was the phenotype with the maximal number of fine-mapping regions (n = 548).

Strategies for choosing causal variants and effect sizes

We applied three different strategies for choosing causal variants from these regions and choosing their effect sizes. For each strategy, we simulated phenotypes from those variants, ran SuSiE and FINEMAP on all SNPs and STRs in the region against the simulated phenotypes and determined whether the fine-mappers correctly identified the variants simulated to be causal.

For the first strategy we chose causal SNPs and indels at random, weighting by minor allele frequency (MAF). For this strategy, we did not simulate causal STRs. To begin, we took all SNPs/indels in all platelet count regions that had either FINEMAP CP $\geq$ 0.5 or SuSiE CP $\geq$ 0.5 and binned them by MAF (bin boundaries = [0.01%, 0.1%, 10%, 50%]), excluding all variants with MAF <0.01%. We assigned each bin a relative weight by the proportion of causal variants in that bin vs. in all bins as compared to the proportion of all variants in that bin vs. all bins, noting that these weights were relatively consistent across bins (within a factor of 2, Table S6). Using those bin weights, for each fine-mapping region, we then drew causal SNPs/indels at random from all SNPs/indels in the region, with each variant's chance of being drawn weighted by the bin that its MAF corresponds to. For each bin, we also collected all observed effect sizes of all variants falling in that bin, noting that as expected the effect sizes for common variants were smaller than those for rarer variants (Figure S6). For each variant chosen to be causal, we drew an effect size from the corresponding MAF bin. This strategy is designed so that the distributions of MAFs and effect sizes of causal variants in our simulations are similar to those observed for fine-mapped variants for the real phenotype. We repeated this strategy nine times for each simulation region, three times each choosing sets of one, two and three causal variants.

While the first strategy allows for a wide range of simulations by drawing causal variants at random, it may not capture systematic differences between the LD patterns of causal variants and the LD patterns of non-causal variants in causal regions. To address this, for the second strategy we chose variants fine-mapped by SuSiE for platelet count for simulating as causal as these may more closely capture LD patterns of truly causal variants. Specifically, we ran SuSiE on all the SNPs and indels in the fine-mapping region with $p < 0.0005$ against real platelet count data. Note that by only running SuSiE against the SNP and indel variants in the region, we forced SuSiE to give us the most plausibly causal set of SNPs/indels in the region under the condition that no STRs are causal. We discarded non-pure credible sets (those with variants in less than 0.8 r^2) as we expect them to be less reliable in identifying truly causal variants. In the 458/548 regions where there were any pure credible sets remaining, we took the top variant from each of the remaining credible sets to use as causal for simulations, using their effect sizes measured against the real platelet count trait as their effect sizes for simulation. For each region, we used its causal variant set to simulate three phenotypes (which are distinct due to different noise terms).

While this second strategy may capture more realistic causal LD patterns compared to choosing causal variants at random, it has the drawback that it relies on the accuracy of fine-mapping to choose the causal variants, which is what we are trying to assess. Strategy one relies on fine-mappers as well, but to a much lesser extent, using them only to identify causal variant MAF and effect size distributions. A second caveat to strategy two is that by restricting to pure credible sets, we likely omit real signals which SuSiE could not resolve well.

For our third strategy, we paralleled our second strategy, except instead of fine-mapping platelet count against only SNPs and indels, we fine-mapped it against all the variants in the region (including STRs), thus allowing it to select STRs as causal for simulation. We continued with simulations as in the second strategy for the 52/548 regions where SuSiE identified a causal STR. This third strategy is the only strategy we performed which simulated causal STRs. As the number of simulations performed with this third strategy was limited, we only use it to contribute briefly to our discussion in the main text.

Simulating phenotypes

Let V represent the set of causal variants for a region. For each variant $v \in V$ let $\overrightarrow{g_v}$ represent a vector of participant genotype dosages and β_v denote the variant's chosen effect size. Assuming additive and independent contributions of each variant, we simulated a vector of phenotypes ($\overrightarrow{y}$) as $\overrightarrow{y} = \sum_{v \in V} \overrightarrow{g_v} * \beta_v + \overrightarrow{\epsilon}$, where $\overrightarrow{\epsilon} \sim N(0, \mathrm{diag}(1 - \sum_{v \in V} \beta_v^2 Var[\overrightarrow{g_v}]))$ so that similarly to the real, normalized, phenotypes used for our GWASs, the resulting phenotypes have mean 0 and variance 1.

Evaluating fine-mapping on simulated phenotypes

For each simulated phenotype and region we performed association testing of the variants in that region using the same methods as in the main analysis, excepting that we included no covariates and that the phenotypes were not subjected to rank-inverse normalization. We then ran FINEMAP and SuSiE against the variants in the region as described above (in particular, FINEMAP runs were restricted to variants with $p < 0.05$, SuSiE runs to variants with $p < 0.0005$), with the difference that the fine-mapping region was not recalculated from the simulated phenotype GWAS statistics but instead exactly matched to the causal region determined from the platelet count GWAS. Once fine-mappers were run, we calculated STR contribution statistics as for fine-mapping runs on the UKB blood traits (Tables S7 and S8).

Simulation caveats

Many choices in the design of these simulations affect the interpretation of their results. Notably, these simulated phenotypes make standard assumptions of additive genetic architectures, including no non-linear effects, no epistasis between variants, and that the

environmental contribution to each phenotype is both independent of an individual's genotypes and normally distributed. These simulations also assume that there are no confounding covariates. Additionally, these simulations choose the effect sizes of causal variants from effect sizes calculated in our platelet count GWAS. As effect sizes calculated in the GWAS were measured in mono-variant regressions against platelet count, they will be mis-estimated according to the corresponding variant's LD to all causal variants in the region in which it resides.

Further, we note that not recalculating the fine-mapping regions may artificially inflate the rate at which strategy one identifies causal variants, as when causal variants in strategy one were randomly chosen to fall near the edges of the region, there would be fewer variants in LD with those variants and fine-mapping them would be easier. This may contribute to the observation in Table S7 that both fine-mappers select STRs in simulation strategy two much more than in simulation strategy one. We also speculate that STRs truly causal for platelet count would contribute to that observation: if those STRs are well tagged by SNPs, strategy two's run of SuSiE would likely select those tagging SNPs for causal simulation. Then fine-mapping of those simulated phenotypes would have a relatively high chance of confusing those SNPs with the STRs they tag.

Lastly, we observe that FINEMAP mostly identifies variants with low p values, while a p value cutoff is necessary for accurate SuSiE results. Once a p value cutoff is applied, we see that the fine-mappers' results are almost entirely consistent with one another, in large distinction from how they perform when applied to real datasets, suggesting that there are some features of the architectures of blood traits are not captured by these simulations.

WGS validation of imputed fine-mapped STRs

We worked with WGS CRAM files for 200,025 UKB participants[62] on the UKB Research Analysis Platform cloud solution provided by DNA Nexus. This data was aligned to reference genome hg38. HipSTR was unable to load the index files for the CRAM files of 10 participants, possibly due to file corruption. Removing those participants left us with 200,015 participants. We inadvertently truncated the participant list, leaving 200,000 participants. From that participant list we called genotypes of the 409 STRs in Table S4 using HipSTR[6] in batches of 500 participants, using the flag --min-reads 10 and allowing HipSTR to estimate stutter-error models from the data. We merged batches using MergeSTR.[80] We performed call level filtering using DumpSTR[80] with the flags --hipstr-min-call-Q 0.9 --hipstr-min-call-DP 10 --hipstr-max-call-DP 10000 --hipstr-min-supp-reads 2 --hipstr-max-call-stutter 0.15 --hipstr-max-call-flank-indel 0.1. After calling all 200,000 individuals we summarized their genotypes separately per population, noting that 166,638 individuals were in our set of QC'ed (potentially related) White British UKB participants, accounting for 40.8% of the QC'ed White British participants.

We did not apply any locus-level filters, such as Hardy-Weinberg equilibrium, to our WGS results. We report per-locus WGS call rates for QCed (potentially related) individuals in each population. We used LiftOver[66] to lift the hg38 WGS calls to the hg19 reference genome (see key resources table). To compare the WGS calls to the imputed STR calls, we used CompareSTR from TRTools[80] branch compareSTR_upgrade using the flags --ignore-phasing --balanced-accuracy --vcf2-beagle-probabilities. We report multiple metrics at each locus, specifically concordance, the mean absolute summed-length difference, r^2 and dosage r^2.

For the following definitions, let X be the set of all samples, A be the set of all possible STR length alleles at a locus, let $S = \{a_1 + a_2 | a_1, a_2 \in A\}$ be the set of all summed-lengths possible at a locus (including the case of homozygous individuals when $a_1 = a_2$), for $x \in X$ let $s_{x,WGS}$ be the summed-length call for sample x from WGS data, and for $x \in X, s \in S$ let $Pr_{x,imp}(s)$ be the probability that sample x has a summed imputation length of s as output by the Beagle AP1 and AP2 FORMAT fields in the imputed VCF file.

We report (summed-length) per-locus concordances as $E_{x \in X}[Pr_{x,imp}(s_{x,WGS})]$. This metric has the advantage of being intuitive but is biased upwards for loci with a single very common allele and so should be interpreted cautiously for such loci. We also report mean absolute summed length differences as $E_{x \in X}[\sum_{s \in S} Pr_{x,imp}(s) \cdot |s_{x,WGS} - s|]$. This metric has similar caveats as the concordance metric. However, for highly multi-allelic loci where concordance is low, this metric can help quantify how close (or not) imputed calls are to the actual genotypes. We calculated r^2 as the square of the weighted Pearson correlation between $s_{x,WGS}$ and s for each sample $x \in X$ and all possible summed-lengths $s \in S$ (so that there are $|X| \cdot |S|$ total values being correlated), weighting by the imputation probabilities $Pr_{x,imp}(s)$. This correlation measure is more comparable across loci with different numbers of alleles than concordance. It has the downside of being less intuitive and of being more sensitive to the WGS-vs-imputation concordance of rare long and short alleles than the WGS-vs-imputation concordance of common average-length alleles. We report dosage r^2 as the square of the Pearson correlation between $s_{x,WGS}$ and the dosage $\sum_{s \in S} s \cdot Pr_{x,imp}(s)$ for each sample $x \in X$. Dosage r^2 is strictly greater than or equal to the weighted r^2 measure. While the weighted r^2 measure more directly measures the concordance of individual imputation probabilities with the WGS calls, the dosage r^2 measure better estimates how analyses like GWASs, which condense imputed probabilities into dosages, will perform.

Lastly, at each locus we report the frequency of each summed-length according to WGS calls, and for all samples with each WGS summed-length we report the probability that imputation concurs with that length: $E_{X|s_{x,WGS}=s}[Pr_{x,imp}(s_{x,WGS})]$.

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 (see key resources table). 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):

(1) Black (African and Caribbean, n = 7,562, codings 4, 4001, 4002, 4003)
(2) South Asian (Indian, Pakistani and Bangladeshi, n = 7,397, codings 3001, 3002, 3003)
(3) Chinese (n = 1,525, coding 5)
(4) Irish (n = 11,978, coding 1002)
(5) 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 not 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 and (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. See Table S9 for details.

STRs were marked as replicating in another population (Figure 2) 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).

We validated imputation STR lengths using WGS data in these populations as was done in the White British population, and report these results in Tables S4 and S5. The number of samples in our QC'ed set that had WGS data were 2,990 Black, 3,373 South Asian, 619 Chinese, 5,174 Irish and 6,428 Other White samples, all roughly 40% of their respective populations.

Logistic regression 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 $\geq$ 0.8), SuSiE (discovery p value < 5e−8 and SuSiE CP $\geq$ 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[79] to fit the logistic regression model:

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

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_category 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 (Table S10).

Gene, transcription factor binding annotation

For all analyses not using GTEx data, gene annotations were based on GENCODE 38[64] (see key resources table). Transcription factor binding sites and DNaseI hypersensitivity regions were identified by ENCODE[90] overlapping several loci (*TAOK1*, *RHOT1* and *NCK2*) through visual inspection of the "Txn Factor ChIP" and "DNase Clusters" tracks in the UCSC Genome Browser[81] and using the "Load from ENCODE" feature of the Integrative Genomics Viewer.[74]

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[69] 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.[13]

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.[91]

Expression association analysis in GTEx

We had previously analyzed associations[13] 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)[37] through dbGaP accession number phs000424.v8.p2. WGS data was accessed using fusera through Amazon Web Services. We genotyped STRs using HipSTR[6] v0.5 with HipSTR's hg38 reference STR set (see key resources table). 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,[80] 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[92] h38.genomicSuperDups table). Altogether, 728,090 STRs remained for downstream analysis.

The *TAOK1* STR locus was filtered from this genotyping for having an 11% call rate, so we imputed the genotypes at that locus into the GTEx cohort. GTEx V7 SNP files were downloaded from GTEx data portal (see key resources table). SNPs on chromosome 17 were extracted and filtered to remove using vcftools with the parameters --maf 0.01 --mac 3 --we 0.00001 --max-missing 0.8 --minQ 30. We used Beagle v5.2 (beagle.28Jun21.220.jar) with the tool's provided human genetic maps to impute STRs into the GTEx SNPs using the same reference panel used for imputation in the UKB cohort above.[30] From this imputation we took the best-guess genotypes of the *TAOK1* STR. We lifted the coordinates of the *TAOK1* STR from hg19 to hg38 using LiftOver.[66]

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 (see key resources table). Gene annotations are based on GENCODE v26.[64] We focused on 41 tissues with expression data for at least 100 samples (Table S13). 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[23] 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, multi-allelic 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 smartpca[78,93] included in EIGENSOFT v6.1.4 with default parameters.

We removed genes with TPM less than 1 in more than 90 percent of individuals. PEER factors[75] 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 genetic sex or age as covariates.

For each STR we computed Bonferroni-adjusted p values to control for the number of gene × tissue tests performed for that STR. Associations that remained with adjusted $p < 0.05$ are shown in Table S12.

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 (see key resources table). We calculated the percent spliced in value for each exon in the manner suggested by Schafer et al.[94] We performed a linear regression to test between the STR's dosage and the percent spliced in of each exon within 10kb, using the top 5 ancestry principal components as covariates.

Methylation association analysis in GTEx

This analysis used the same STR data and genotype principal components as the GTEx expression association analysis above.

We downloaded genome-wide DNA methylation (DNAm) profiling results from the NCBI GEO database under accession number GSE213478. This contained DNA methylation levels from the whole blood of 47 individuals who had been genotyped, including 754,054 autosomal CpG loci which passed quality control checks in that dataset (see key resources table).[65] We lifted those loci from hg19 to hg38. We performed per-locus inverse-normalization of the DNAm data prior to downstream analysis. We calculated 5 PEER factors from the normalized DNAm data across quality-controlled loci from all chromosomes (including sex chromosomes) using PEER v1.0,[75] choosing 5 factors to match the number of PEER factors used by the methylation study which generated this data.[65]

We tested for associations between the methylation of each autosomal CpG locus and the length of each STR located within 100kb of that locus. For each such pair, we performed a linear regression between the STR's dosage (sum of allele lengths across both chromosomes) and the inverse-normalized DNAm levels of that CpG locus, including the top five genotype principal components and the 5 PEER factors as covariates. We compared the effect sizes of these associations with those from another paper studying STR-methylation correlations in two separate cohorts in whole blood[17] and found that they were broadly consistent ($r = 0.73$, $p < 10^{-200}$, Figure S18C).

For each STR we computed Bonferroni-adjusted p values to control for the number of CpG tests performed for that STR. Associations that remained with adjusted $p < 0.05$ are shown in Table S14.

Targeted STR expression analysis in Geuvadis

We applied HipSTR[6] v0.6.2 to genotype STRs from HipSTR's hg38 reference STR set (see key resources table) in 2,504 individuals from the 1000 Genomes Project[63] for which high-coverage WGS data was available (see key resources table). 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 (see key resources table). 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, except that this was only performed for two STR-gene pairs (STRs identified by fine-mapping near the genes *CBL* and *RHOT1*). We adjusted 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[78,93] 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*).

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical tests are named as they are used and are described in the method details.