Title

Ubiquitous genome-wide variation at short tandem repeats is causally linked to changes in gene expression, blood cell counts and serum biomarkers in human populations

Author

Margoliash, Jonathan Bernard

Publication Date
2024

UNIVERSITY OF CALIFORNIA SAN DIEGO

Ubiquitous genome-wide variation at short tandem repeats is causally linked to changes in gene expression, blood cell counts and serum biomarkers in human populations

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Computer Science

by

Jonathan Benard Margoliash

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Alon Goren, Co-Chair
Professor Tiffany Amariuta-Bartell
Professor Vineet Bafna
Professor Abraham Palmer

2024

The dissertation of Jonathan Bernard Margoliash is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ABF – approximate Bayes factor

bp – base pairs (as a unit of length)

eQTL – expression QTL

GTEx – Genotype-Tissue Expression Project

GWAS – genome-wide association study

Indel – insertion or deletion

kb – kilobases (thousands of bases, as a unit of length)

LD – linkage disequilibrium

LMM – linear mixed model

mb – megabases (millions of bases, as a unit of length)

mQTL/meQTL – methylation QTL

PC – (genetic) principal component

pQTL – protein QTL

PRS – polygenic risk score

QC – quality control

QTL – quantitative trait locus

SNP/SNV – single nucleotide polymorphism/variant

STR – short tandem repeat

TF – transcription factor

TR – tandem repeat

TWAS – transcriptome-wide association study

UKB – UK Biobank

UKB RAP – UK Biobank Research Analysis Platform

VNTR – variable number tandem repeat

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

margoliash_chapter_1_supplementary_tables.xlsx – these are also available from https://www.nature.com/articles/s41588-019-0521-9

margoliash_chapter_1_supplementary_data_1.xlsx – this is also available from https://www.nature.com/articles/s41588-019-0521-9

margoliash_chapter_1_supplementary_data_2.gz – this is also available from https://www.nature.com/articles/s41588-019-0521-9

margoliash_chapter_1_supplementary_data_3.xlsx – this is also available from https://www.nature.com/articles/s41588-019-0521-9

margoliash_chapter_3_supplementary_tables.zip – this includes Supplementary Tables 3.1, 3.2, 3.4, 3.5, 3.7, 3.9, 3.11, 3.12, 3.14, which are also available from https: //www.cell.com/cell-genomics/fulltext/S2666-979X(23)00302-6#supplementaryMaterial . Other Chapter 3 Supplementary Tables are reproduced directly in this thesis

Supplementary Datasets for Chapter 3 can be accessed through the links in the Key Resources Table in that Chapter

ACKNOWLEDGEMENTS

I would like to thank my PhD mentors Professors Melissa Gymrek and Alon Goren, with whom all of this work has jointly been done, and without whom none of this work would have been possible. They took me into their labs when I had no knowledge of genetics and little background in research and gave me an opportunity to prove myself. In the years since they have given me countless hours of their time, where I have asked for their insight on any and every topic that has come to mind. I hope I have made this journey worth their while.

I have had a wonderful network of family, friends and lab mates during my PhD studies at UCSD – I am grateful for you all. I specifically thank my mom for reading this thesis while I was drafting it; that was such fun, and the writing is clearer for it. Lastly, I would like to thank my loving partner, Nora, who has been a wonderful support and gracious roommate throughout the writing process. I will remember this experience so much more fondly because of her.

Formal acknowledgments for the papers reproduced herein:

Chapter 1 contains a full reprint of the publication Fotsing, S. F.; Margoliash, J.; Wang, C.; Saini, S.; Yanicky, R.; Shleizer-Burko, S.; Goren, A.; Gymrek, M. The Impact of Short Tandem Repeat Variation on Gene Expression. *Nat. Genet.* **2019**, *51* (11), 1652–1659. https://doi.org/10.1038/s41588-019-0521-9.

Chapter 2 contains a full reprint of the publication Mousavi, N.; Margoliash, J.; Pusarla, N.; Saini, S.; Yanicky, R.; Gymrek, M. TRTools: A Toolkit for Genome-Wide Analysis of Tandem Repeats. *Bioinformatics* **2021**, *37* (5), 731–733. https://doi.org/10.1093/bioinformatics/btaa736.

Chapter 3 contains a full reprint of the publication Margoliash, J.; Fuchs, S.; Li, Y.; Zhang, X.; Massarat, A.; Goren, A.; Gymrek, M. Polymorphic Short Tandem Repeats Make Widespread Contributions to Blood and Serum Traits. *Cell Genomics* **2023**, *3* (12), 100458. https://doi.org/10.1016/j.xgen.2023.100458.

VITA

2013	Bachelor of Science with Honors in Mathematics, University of Chicago
2013-2015	Software Developer Engineer I, Amazon
2016-2017	Software Developer, Cardiovascular Health Studies Coordinating Center, University of Washington
2019	Master of Science in Computer Science, University of California San Diego
2024	Doctor of Philosophy in Computer Science, University of California San Diego

PUBLICATIONS

Fotsing, S. F.; **Margoliash, J.**; Wang, C.; Saini, S.; Yanicky, R.; Shleizer-Burko, S.; Goren, A.; Gymrek, M. The Impact of Short Tandem Repeat Variation on Gene Expression. *Nat. Genet.* **2019**, *51* (11), 1652–1659. https://doi.org/10.1038/s41588-019-0521-9.

Mousavi, N.*; **Margoliash, J***.; Pusarla, N.; Saini, S.; Yanicky, R.; Gymrek, M. TRTools: A Toolkit for Genome-Wide Analysis of Tandem Repeats. *Bioinformatics* **2021**, *37* (5), 731–733. https://doi.org/10.1093/bioinformatics/btaa736.

*__Equal contributors__

Margoliash, J.; Fuchs, S.; Li, Y.; Zhang, X.; Massarat, A.; Goren, A.; Gymrek, M. Polymorphic Short Tandem Repeats Make Widespread Contributions to Blood and Serum Traits. *Cell Genomics* **2023**, *3* (12), 100458. https://doi.org/10.1016/j.xgen.2023.100458.

ABSTRACT OF THE DISSERTATION

Ubiquitous genome-wide variation at short tandem repeats is causally linked to changes in gene expression, blood cell counts and serum biomarkers in human populations

by

Jonathan Bernard Margoliash

Doctor of Philosophy in Computer Science

University of California San Diego, 2024

Professor Melisssa Gymrek, Chair

Professor Alon Goren, Co-Chair

Short tandem repeats (STRs) are ubiquitous throughout the human genome and routinely vary within human populations but have largely been excluded from genome-wide analyses of variant contributions to human phenotypes. In my thesis work, I and my collaborators demonstrate that current bioinformatic advances now allow for the inclusion of

STRs in such studies. We present evidence suggesting that STRs are likely causal for 5.2-7.6% of signals for human blood traits as well as making widespread impacts on gene expression across different tissues. We demonstrate how to carefully interpret and maximize the reliability of statistical fine-mapping to overcome high degrees of correlation between nearby variants, showing its central utility for the study of complex traits. So doing, we uncover many putatively causal STRs strongly affecting human phenotypes.

Introduction

In this introduction I begin by providing a very brief overview of genomics, a review for some readers, as motivation for the performance of genome-wide association studies (GWAS). I then discuss how GWAS are conducted, what they hope to achieve, and what their limitations are. I show how including missing variant types can help address some of those limitations and focus on the study of short tandem repeats. Statistical fine-mapping is a recent field of work that has proven necessary to GWAS of complex traits, so I follow with a discussion of its development and current state. I conclude the introduction with a brief preview of the work I and my collaborators have contributed over the course of my doctorate to the study of STRs in GWAS. Those works are reproduced, with new forwards, in Chapters 1 through 3. I conclude my thesis with a discussion of current and future trends in these areas.

The Human Genome

Chromosomes, Bases, RNA and Proteins

With few exceptions, each of the trillions of cells in the human body has 46 long, string-like DNA molecules called *chromosomes* in its nucleus. This collection of chromosomes possesses the striking property that it is almost identical between the different cells in the same individual, and so it can be thought of as a single entity: that person’s *genome*. Most commonly, a person's genome consists of 22 pairs of *autosomal chromosomes*, or *autosomes*, labeled 1 through 22, and two *sex chromosomes*. Chromosomes are categorized this way as any two autosomal chromosomes with the same number are very similar, though not identical, to one another, both within an individual and between people. Two chromosomes that are similar in this manner are called *homologs* (n.) and *homologous* (adj.). This further allows for a conceptual *human reference genome*, containing a single representative copy of each autosome and each sex chromosome.

At the molecular level each chromosome is an enormously long string of *nucleotide* molecules bound together in a line, each nucleotide either an adenosine, thymidine, cytidine or guanosine, abbreviated A, T, C or G and collectively called *bases*. The first sequence of bases making up a human genome was published in 2001, and since then the genetics community has assigned each chromosome in the reference genome a sequence that, after many updates, roughly corresponds to the most common sequence of bases among its homologs in the wider population. A specific end of each chromosome has been arbitrarily selected as its start, and bases are counted from that end forward. Any numeric location (either specifying an individual base, or a range of bases) on a chromosome is called a *locus* (pl. *loci*). One copy each of the 22 autosomes is about 3 billion bases in length, and the sex chromosomes contain an additional few hundred million bases (*mb*, *megabases*, also *kb* for *kilobases*). The similarity between homologous chromosomes can be quantified in terms of bases – 99.9% of bases in two homologous chromosomes are the same and in the same order[1].

The genome contains *genes* which play a crucial role in the inner workings of our bodies. Genes are sections of the genome which are *transcribed* into RNA molecules which leave the vicinity of the DNA from which they were transcribed and can act elsewhere in the cell. Genes are said to be *expressed* when they are transcribed, and the amount of RNA transcribed from a gene and circulating in the cell at any one point is called that gene's *expression*. The most famously studied type of genes are *protein-coding genes* whose RNAs are called *messenger RNAs* (mRNAs). Proteins are made from mRNA molecules through the process of *translation* and are essential to the many processes that constitute the lives of cells and organisms. The desire to understand how proteins function, how differences in genetic variation affect those functions, and the hope to be able to change those functions through medical interventions are the main motivations for genetics.

The Non-Coding Genome and The Genome's Structure

Yet the genome does not just contribute to cell functions through its *protein-coding regions*, that is, the 1% of the genome[2] residing within protein-coding genes which directly corresponds to protein structure. Protein-coding genes also contain *introns*, sections within genes that are transcribed into the mRNA but are *spliced out* (removed) from the mRNA before it is translated into proteins. And the content of these introns contributes to the regulation of splicing[3].

Even among the non-intronic sections of a gene, called *exons*, the beginning exonic sequence and the ending exonic sequence in each gene are also not translated into proteins. These are called the *5'* (pronounced "five prime") and *3' untranslated regions* (UTRs), respectively, named after the molecular properties of each of those ends of the RN, and assist in regulating translation[4], among other functions.

Further, it is estimated that nearly 70% of human genes are transcribed into RNAs which are not mRNAs[5] and thus do not code for proteins. These *non-coding RNAs* of these *non-coding genes*, while much less studied than mRNAs, do show wide varieties of function[6,7], and so also contribute an important but not well quantified portion of the functioning of our cells. In summary, the non-coding transcribed regions of the genome, the introns, UTRs and non-coding RNAs, play important functional roles.

The rest of the genome – the genome between genes, called the *intergenic genome* – is also functional. To understand this, it is important to know a bit about *chromatin*, the genome's 3D physical and chemical structure. Our chromosomes exist as double helices – two strands of DNA running in opposite directions, with a uniform pairing of As to Ts and Cs to Gs, and vice versa. This is our chromosome's canonical *secondary structure* and the structure it holds when unperturbed in the nucleus, though it can enter non-helical *non-canonical secondary structures* in specific contexts. From each double stranded chromosome one strand has arbitrarily been selected as the chromosome's canonical *forward strand*, and frequently only the bases on that strand are referred to, as the bases on the *reverse complement* or *opposite* strand can be perfectly inferred from that information. (Though either strand can be transcribed into RNA, so bases on the reverse strand are often referred to in the context of genes transcribed on that strand). This also leads to the term *base pair* being used interchangeably with the term *base*.

The double stranded DNA helices are wrapped around molecules called *nucleosomes,* each nucleosome being made up of eight units called *histones* and being wrapped by 146 base pairs of the double helix[8]. The unwrapped portions of the chromosomes and the millions of wrapped nucleosomes in turn are organized into higher order structures, eventually comprising the complicated 3D structure that is called chromatin. This structure determines which genes can be bound by molecular transcription machinery, and the intergenic genome influences this structure. Thus the intergenic genome plays a crucial role in determining which genes are transcribed and at what rate, influencing the quantity and type of RNAs circulating in a cell, and thus influencing how many proteins and circulating, functional, non-coding RNAs are available. Lastly, it is important to realize that two loci on a chromosome can be many bases apart but be in close 3D contact in the chromatin structure, and so can interact. This means that portions of the intergenic genome far from genes can still be functional.

A reductive but useful way to refer to any information known about chromatin and DNA function is with *functional annotations* or *features*. An annotation is simply a descriptive piece of information, assigned to a stretch of a chromosome, that has either been experimentally validated or algorithmically predicted[9]. For instance, the protein-coding regions of the genome can be thought of as regions of the genome annotated as being protein-coding. Other annotations mark where introns and exons begin and end, and where *transcription start sites* – the loci where gene transcription begins – are located.

While all cells in a human body have nearly identical copies of that person's genome, the chromatin structure in cells can vary substantially between cell types and tissues, can change as the cell enters different cell states, and can change over an individual's lifetime. Similarly, which regions of mRNAs are spliced out or left in can differ between cell types and tissues, leading to different mRNA and protein *isoforms* in different contexts[3]. Thus unlike annotations of protein-coding regions, many annotations of chromatin structure and of splicing are given in reference to the specific tissues, cell types, cell states and developmental time points they were measured in. This variation in chromatin structure and splicing is part and parcel of different cells performing different functions in different contexts.

There are a multitude of genomic annotations which I will reference at various points throughout my thesis. The ENCODE encyclopedias, both the in-development[10] and current versions[11], are good references on annotations, as is the latest ENCODE publication[12]. It is not necessary to memorize the specifics of each of the following annotations for this thesis, but it is important to understand the types of information annotations convey. Annotations often implicate their chromosomal region in up- or down-regulating the expression of a specific gene or nearby genes. These implications come with varying levels of certainty – for instance, a region of the intergenic genome being marked as inaccessible can, but does not necessarily, impact nearby gene expression. Further, these annotations often overlap one another, providing a hierarchy of evidence – a region may be annotated as being an enhancer of gene expression because it is more specifically annotated as both being bound by a transcription factor and being in open chromatin. Furthermore, it is important to recognize that annotations are often probabilistic, for example, saying that some percentage of cells in a tissue are methylated at a specific C to G bond. Lastly, it should be noted that this is a rapidly progressing field of study;

experimental techniques for measuring new annotations, or increasing the specificity, ease, and breadth of measuring existing annotations, are constantly in development.

High-level annotations include *promoters*, regions around the transcription start sites of genes to which much of the molecular transcription machinery binds, *enhancers*, farther away (more *distal*) regions of the genome which interact with promoters to increase gene expression, and *insulators* and *silencers*, which are like enhancers but reduce gene expression. Enhancers, insulators, silencers and other such annotations can be collectively grouped under the term *regulatory elements*, and *candidate regulatory elements* are those whose function has not been thoroughly validated.

A broad lower-level annotation is *chromatin accessibility*, which describes how accessible regions of the genome are to interacting with proteins and RNAs, with the general heuristic that non-accessible regions are less functional or need to be opened to become functional. More specific low-level annotations include annotations that describe which sections of the chromosome are in long range contact with one another, and often focus on the regions of the chromosome that contact nearby promoters. There are annotations describing where each of the few thousand unique *transcription factors* bind (often abbreviated as TFs), those proteins that influence, and are often necessary for, the process of transcribing specific genes. *Nucleosome occupancy* can be annotated, that is, which regions of the genome are wrapped around nucleosomes. There are *histone modification* annotations, which describe the different chemical modifications that can be made to the histones in nucleosomes which then alter the wrapping of DNA around those nucleosomes, with specific histone modifications being associated with enhancing or silencing gene expression. And there are annotations which describe which bonds between C and G nucleotides (called *CpG bonds*) are *methylated*, and thus associated with repressing gene expression, and which are not. This is by no means an exhaustive list of genetic annotations, but it is sufficient for the purpose of this thesis.

Genetic Variation and Genetic Contributions to Human Traits

The above discussion focused on the functions of the standard genetic sequence in different parts of the human genome. The opposite side of that coin is focusing on the *variation* in the human genome, that is, the 0.1% difference in bases between people. The study of variation is of interest because genetic variation plays an essential role in determining a portion of the differences between people, including physical proportions such as height, the likelihoods of suffering from various diseases, and personality traits. Of course, studying genomic functions is complementary to studying the genetic variants causal for human traits as it explains how those variants act. But it is often the case that research which focuses on impactful variation first starts with measuring traits, then links them to variation in the human genome, and only then asks what the molecular function of that genetic variation is.

In this context, any trait that varies between people is called a *phenotype*. A genetic *variant* at a locus is a difference between the sequence of bases in different individuals at that locus. Each differing sequence of bases at that locus is one of that variant's *alleles*. The *reference allele* is the allele present in the theoretical reference genome, and the *alternate alleles* are the other alleles. An alternative way of categorizing alleles is referring to the *major allele*, which is the most common allele for that variant in the dataset or population being studied (often, but not always, corresponding to the reference allele), and the *minor allele(s)*, which are the other alleles at that variant. The *minor allele frequency* of a variant in a population is the percentage of all alleles of that variant in that population which are minor alleles. A variant is *biallelic* if only one alternate allele exists (at least, if only one exists in the dataset that is being studied) and is *multiallelic* otherwise. A variant is said to be *causal* for a phenotype if which allele is present at that locus causally affects the observed trait. Lastly, *to call* or *genotype* a variant in a person means to determine which allele(s) that person has at that variant. A *caller* or *genotyper* is an algorithm designed to perform that task.

There are many types of genetic variants. The most commonly studied small genetic variants are *SNPs* (single nucleotide polymorphisms, also called SNVs, single nucleotide variants, a term used interchangeably by most authors), each of which is a change of a single base pair to another base pair. These are also the most numerous type of genetic variant, with an estimate of 13.75 million common SNPs (that is, SNPs with minor allele frequency > 1%) in the world's population[13]. Further, as more humans that are sequenced, it seems increasingly likely that nearly every base in the genome has a SNP variant in some individual, amounting to billions of loci with rare variation[14].

Another small variant type commonly studied is small *insertions* (gains) or *deletions* (losses) of base pairs compared to the reference genome, together called *indels*, with an estimate of 4.4 million common indel variants in the world's population[13]. Only a small minority of these small variants are known to impact human phenotypes.

On the other side of the spectrum are large variations in DNA. The most dramatic types of large genetic variation are in individuals who have more or fewer chromosomes than the standard number (referred to as *aneuploidies*), or where a large chunk of a chromosome has been *translocated* (moved) to another chromosome. In between these two extremes are a wide variety of types of variation, see Table Introduction.1, including *tandem repeats* variants, which I focus on below. There are fewer large variants in the genome than small variants but on average they tend to have larger impacts on human traits[15].

Table Introduction.1: Classes of Large Genetic Variants

Tandem Repeats	Short or large sections of the genome repeated a few or many times in sequence
Mobile Element Insertions	Specific sequences of DNA that have copied and reinserted themselves into other parts of the genome
Copy Number Variants	Large sections of the genome that are duplicated in part or in whole at other locations in the genome
Structural Variants	A term encompassing all variation affecting at least 50 bases[16], including moderate to large deletions, insertions, or inversions of sections of chromosomes

For this thesis it is also important to have a broad understanding of the technologies most commonly used for calling variants. The cheapest and oldest technology important to this work is the *microarray*. Microarrays are tiny chips consisting of arrays of hundreds of thousands to millions of biomolecular *probes* or *markers*[17]. Each probe is designed to bind a known, unique, ~50 base region of DNA from the reference genome. Further, each probe is designed to only bind to the region if that region contains a specific allele of a SNP or small indel. Thus after exposing the probes on a microarray to DNA fragments from the cells of an individual, each probe that is bound conveys a specific allele that individual possesses. Importantly, microarrays are not a reading technology: they only indicate the presence or absence of alleles tested by the array.

While reading the first human genome cost billions of dollars, in the last decade and a half the technology of *short-read whole-genome sequencing* (WGS) has become ever more affordable, with costs dropping below $1000 per individual. Short-read WGS involves taking the genomes of thousands of cells from one individual and breaking them into short segments of DNA, roughly 150 bases in length, called reads. The sequence of bases is read off each of these reads, and each sequence is aligned back to the reference genome to determine where it came from. After alignment, differences between the individual's genome and the reference genome indicate alternate alleles. Alternate alleles longer than the size of a single short read can be difficult to detect, though probabilistic methodologies can infer the presence of specific classes of alternate alleles based on the distributions of reads seen. Short-read sequencing has dropped in cost so dramatically that while at the beginning of my thesis it was only available in small cohorts of a few thousand individuals, it is now beginning to be applied to the largest biobanks containing hundreds of thousands of individuals[14].

The last technology I will discuss is *long-read* WGS. This is similar in concept to short-read WGS except that the reads are one to multiple orders of magnitude longer than the standard 150 base short-reads. This allows long reads to capture large genetic variation that cannot be precisely estimated from short reads and has led to enormous breakthroughs in variant calling[5]. But as of now long-read WGS only exists in cohorts of a few thousand individuals due to its currently prohibitive cost[18,19], and thus is not sufficiently wide-spread to support the type of analyses I have focused on in my thesis work.

Lastly, it is important to know that phenotypes can be categorized by their *genetic architectures* as either *Mendelian traits* or *complex traits*. *Mendelian traits*, also called *monogenic traits*, are those whose presence or absence is caused by variation around a single gene. This category includes many rare and debilitating diseases such as Huntington's disease. On the other end of the spectrum are *complex traits*, also called *polygenic traits*, which are so-named because they are influenced by large numbers of regions in the genome. Unlike Mendelian traits, complex traits cannot be strongly predicted by any single genetic region, but they can be strongly predicted by the alleles at many regions taken together. The category of complex traits includes a wide range of phenotypes, including anthropomorphic features such as height, and neurological diseases such as schizophrenia and bipolar disorder.

Most traits fall somewhere between these two ends of the spectrum, and many traits share properties of both sides of the spectrum. For example, 5-10% of breast cancer cases are caused by high-impact inherited mutations in genes such as *BRCA1* and *BRCA2*, and thus are Mendelian, while over a hundred different loci have small influences on the occurrence of breast cancer in the more than 90% of cases with no known high-impact inherited mutation[20]. Yet while the binary categorization of traits under the labels Mendelian and complex is not fully precise, it is still useful as those categories of traits are researched in different ways[21]. Once a Mendelian variant is found, much research can be devoted to understanding the cascading network of biological pathways it impacts. Whereas for complex traits, signals tend to be weaker, and much research is spent on simply trying to identify either a causal variant, a causal gene or causal mechanism at each signal. My thesis work has been devoted to the study of complex traits.

The Limits of Our Genetic Knowledge

As much as knowledge of the human genome drives further genetic findings, the limitations of that knowledge likewise guide genetic study. For the genome is vast and has large tracts that remain uncharted. In 2022, the existence of 93.2% of the roughly 20,000 predicted human protein-coding genes had been experimentally confirmed according to the Human Proteome Project[22]. Yet in 2023 only 82% of those genes coded for a protein with either an inferred molecular function, an inferred cellular location or an inferred biological process in the Gene Ontology knowledgebase, while only 68% had experimental validation for one of those three annotations[23]. Further, in 2023, only 56.2% of genes had a known molecular reaction categorized in the human Reactome Pathway Knowledgebase[24]. These statistics overstate our knowledge of the genome in that they ignore functional non-coding RNA molecules. But more critically, these estimates do not measure how much information is missing about the protein-coding genes themselves. There are many unknown isoforms of genes[7], unmeasured posttranslational modifications of proteins, and unnoticed reactions proteins are involved in.

Yet despite these limitations, understanding of the effect of genetic variation on protein functionality is much more complete than the understanding of variant effect on non-coding genomic functionality. The recent release of the AlphaFold algorithm marked a milestone in our ability to predict protein structures[25], and AlphaMissense is the corresponding attempt to predict the effect of every protein-coding SNP[26]. These are just the tip of a wide-range of protein variant effect predictors. The study of protein variant effect predictors is its own field of work, and as most of the variants I dealt with in my PhD work are not protein-coding, I refer the interested reader to a sampling of the many articles written on this topic[27–29].

While some variant effect predictors attempt to extend this functional insight to non-coding genetic variation, those efforts are not nearly as successful. These difficulties are in part due to the challenge of assaying the non-coding genome, but also because there are many cell types, cell states, and developmental time points that need to be studied to understand the non-coding genome's function, and because acquiring those biological samples is expensive. The GTEx (Genotype-Tissue Expression) project[30] is, to my knowledge, the largest general purpose repository of diverse human tissues with prior research authorization and already measured gene expression data. Yet GTEx currently only has samples from ~900 individuals for the most numerously sampled tissue types, and many relevant tissues only have samples from a hundred or fewer individuals. This means that many rare variants are simply not present in the GTEx cohort, and study of the variation that is present has limited statistical power. Further, GTEx does not currently have tissues sampled from different developmental time points, in response to different exposures, or in any of the many disease states of the body.

The current difficulty with annotating and understanding variation in the non-coding genome is well demonstrated by a recent perspective of the Encyclopedia of DNA Elements (ENCODE) project[31], a knowledgebase consortium dedicated to annotating the human genome. The perspective says that "very few examples of condition-specific activation or repression [have been found] … Similarly, information from human fetal tissue, reproductive organs and primary cell types is limited. In addition, although many open chromatin regions have been mapped, the transcription factors that bind to these sequences are largely unknown, and little attention has been devoted to the analysis of repetitive sequences." It says that even ENCODE's new phase will only focus on closing this knowledge gap "in a few reference cell lines", relying on prediction for the rest. In sum, they say that "although very large numbers of noncoding elements have been defined, the functional annotation of ENCODE-identified elements is still in its infancy".

All this is to say that, while knowledgebase consortia such as the Human Proteome Project, the Gene Ontology knowledgebase, the Reactome Pathway knowledgebase, and ENCODE have produced valuable and ever-growing amounts of information about the human genome, they are by no means complete. Thus research which prioritizes the study of genes and variants with known functional contributions will inevitably miss large swaths of functional genetic material whose import has yet to be discovered. This motivates alternative studies that are unbiased by the extent of our current knowledge. My thesis work has focused on one such method: genome-wide association studies.

Genome-Wide Association Studies

Genome-wide association studies (GWAS), like many analyses that study the relationship of genetic variation to human phenotypes, aim to further understanding of the biomolecular mechanisms of causal genomic variants and of the genes and cell types causally involved with human traits. GWAS are performed with the hope of better predicting phenotypes and of creating information that leads to better interventions for disease phenotypes. As knowledge of genomic function is limited, a motivating aspect of GWAS is the desire to study the involvement of variants from all genomic regions without restricting to only variants with known function or searching only within regions of known function. The GIANT consortium's recent GWAS of height, a landmark study because of the 5 million individuals it analyzed, is a good reference for consortium-standardized GWAS protocol, and it analyzed 1.4 million common variants from across the genome[32], or one variant roughly every two thousand bases. This type of blanketing examination of the genome makes GWAS one of the most effective methodologies for uncovering evidence of which genomic regions are causally involved with phenotypes, and the information generated by GWAS can then be used as a starting point for many follow-up analyses.

Currently, experimental assays that induce variation at spots in the genome and study the resulting effect are too slow and costly to be performed genome-wide. Thus researchers of genome-wide variation are pushed to statistically analyze existing variation found in nature. A GWAS is one such analysis. GWAS are performed in *biobanks* where individuals have donated their genetic and phenotypic information to science.

In a GWAS, for each alternate allele on each chromosome to be tested, each individual is assigned a *dosage*: the number 0, 1 or 2 which counts how many of the individual's two copies of that chromosome contain that alternate allele. Most GWAS only study biallelic variants as that is simpler and secondary alternate alleles tend to have very low minor allele frequencies, and thus refer to ‘variants’ being tested instead of ‘alternate alleles of variants’ being tested. As it is so ubiquitous, I will use that terminology going forward, just noting that multiple alternate alleles of a single variant can be individually tested.

In the prototypical GWAS of a quantitative trait, for each variant, researchers perform ordinary least squares regression of the trait values of each person against their dosage values for that variant, with additional covariates including sex, age, and genetic principal components. This gives an effect size for the variant’s dosage as well as the effect size’s standard error; the effect size can be interpreted as the average difference between the trait value of an individual with that of another individual who has one additional copy of the alternate allele for that variant among their chromosomes. When a binary trait is studied, such as presence or absence of a disease, logistic regressions are performed instead of linear regressions and this changes the interpretation of the regression coefficients, but GWAS analysis otherwise proceeds similarly. From the effect size and standard error researchers calculate a z-score and p-value and see if the p-value refutes, in-likelihood, the hypothesis that the effect size is exactly zero. See Figure Introduction.1 for an example association between a phenotype and a variant. These calculations are done individually for each alternate allele of each variant studied, leading to a table of *summary statistics* – effect size, its standard error, and the derived p-value – for each alternate allele of each variant in the genome. PLINK 2[33] is a standard tool for performing GWAS described in this manner.

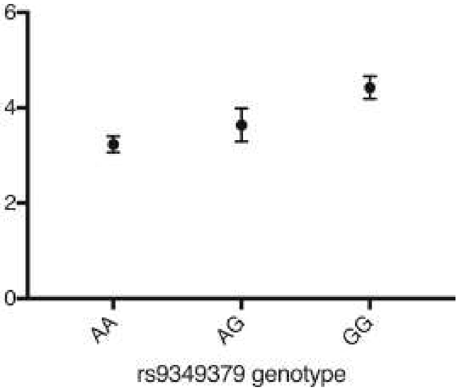


Figure Introduction.1: An association between a phenotype and the genotype of a biallelic variant. The biallelic variant is a SNP, identified by its ID rs9349379, with two alleles: A (the major allele) and G. The variant's possible diploid genotypes are plotted on the x-axis. The y-axis represents the value of the (in this case unnamed) phenotype. For each genotype, the mean and standard deviation of phenotypes from individuals with that genotype are plotted against the y-axis. In this example, individuals with G alleles have higher phenotype values; the p-value for the regression performed here was 0.00136. Adapted from Gupta et al.[34]

GWAS analyses of the last ten years have moved towards more sophisticated models than simple linear regression, but it is important to note that the overarching framework is the same. These days GWAS often run tests using linear mixed models (LMMs)[35,36], which incorporate random effects covariates based on a genetic measure of the relatedness of the individuals in the study, on top of the linear model used by ordinary least squares. This boosts statistical power by better accounting for the noise in the phenotype measurement due to genomic factors that correlate with genetic relatedness. REGENIE is another recent method which attempts to achieve the same goal as LMMs, but does so by linearly testing each variant's dosage against the residual of a prediction of the phenotype from other variation in the genome[37]. Special methods have also been developed to improve GWAS statistical reliability for binary phenotypes, using either the Firth[37,38] or saddle-point[35] corrections to logistic regression so that test statistics are not inflated in the presence of low counts of minor alleles in cases. GWAS also are routinely conducted via meta-analysis, where summary statistics from multiple underlying GWAS are reanalyzed jointly to heighten statistical power, e.g. the GIANT consortium height GWAS[32]. Another variation on GWAS, though not new, is to test variants for association separately within individuals of each sex, e.g.[39,40].

But regardless of the specific methodology, GWAS analyses share the common property that they test variants from all genomic regions and give each variant an estimated effect size and p-value that is unadjusted for local genetic structure (an issue discussed in more detail below). Under this paradigm, GWAS results have largely proven to replicate across datasets and between research groups[41,42]. Further, this has led to the creation of follow-up analyses which use GWAS summary statistics as input and are agnostic to the specific GWAS methodology those summary statistics were created with. (An exception to this rule are GWAS which test not only for linear associations between variants and phenotypes, but also non-linear dominant and recessive effects, e.g.[43]. The statistics output from such tests would need to be treated specially, and I do not discuss them further.)

The most direct use of GWAS summary statistics is to highlight the variants whose p-values are below a pre-fixed threshold designed to control for the false positives that come from performing so many independent tests. The community standard for this threshold is $p < 5 \times 10^{-8}$; variant associations more significant than this are often referred to as *genome-wide significant*. This threshold is stringent enough that GWAS must be performed in large cohorts to have sufficient power to detect most causal signals. Once GWAS researchers identify variants passing this threshold, they cluster them according to their genomic positions, and infer that the genomic regions containing those clusters are causally involved with the phenotype being studied. A region identified this way is called a *signal* or a *hit* and the variants in the region are said to *tag* that signal. GWAS signals will vary between tens of kilobases in length to a few megabases depending on the definition used, the strength of the signal and the local LD structure (discussed below). The *lead variant* for the signal is the variant with the lowest p-value among tested variants in that region.

As an example, the GIANT GWAS identified 7,209 non-overlapping genomic segments whose associations with human height pass this threshold[32]. These have a mean size of around 90 kilobases and cover about 21% of the genome, corroborating the notion that genome-wide scans are necessary for uncovering the diverse genetic architecture of highly polygenic traits. These researchers believe they have identified nearly all common variation associated with height in Europeans and use their results to explain ~45% of the variation in human height between European individuals, but doing so required studying 5.4 million individuals. Height is a prototypical complex trait, and many less-complex traits will require study of fewer individuals to achieve this saturation. However, the authors of the GIANT GWAS predict that traits even more complex than height, such as inflammatory bowel disease, schizophrenia and body mass index, will require tens of millions of individuals to reach this level of saturation.

There are many important consequences of having to impose the incredibly strict threshold of $p < 5 \times 10^{-8}$. For one, it forces researchers to perform GWAS in large biobanks. Collecting large biobanks of data is incredibly expensive, and so has only been funded by governments and pharmaceutical companies in the wealthier parts of the world. This has resulted in huge disparities in the size of biobanks containing people of European decent as compared to people of African, Hispanic and South Asian descent. (Biobank Japan[44], the Taiwan Biobank[45] and the China Kadoorie Biobank[46] each contain more than 100,000 genotyped individuals, placing East Asians somewhere in the middle of this spectrum). This has led to GWAS predominantly being performed on European populations and makes medical insights derived from GWAS results less applicable to individuals of other ancestry groups. This inequity in GWAS research is widely recognized in the field but is only slowly being overcome.

This strict p-value threshold is also especially problematic for the study of complex case-control phenotypes with low case rates and numerous but generally weak effect sizes (e.g. schizophrenia). A phenotype like schizophrenia has too few cases in general purpose biobanks to detect its signals, and thus needs a genetic biobank built of specifically schizophrenic individuals for GWAS of schizophrenia to have sufficient power. This limits the extent to which GWAS can provide insight into such traits.

Another consequence of this strict threshold is that rare causal variants are infrequently strongly-enough associated with a trait to pass this threshold, even if they highly impact that trait. As such, rare variants are often analyzed in aggregate to reduce the number of overall tests and loosen this threshold. These aggregate rare variant analyses often accompany GWAS, but are not the subject of my thesis; I refer the interested reader to an old but informative review article[47].

Yet once a GWAS association passes this threshold, researchers can believe that it tags a causal genetic signal. The key insight to this is that when a parent's genome contains two different alleles for a variant on their two homologous copies of a chromosome, a child of theirs will inherit one of those two at random, uninfluenced by choice or environmental factors. This fact roughly ensures that genetic variants are randomly distributed throughout the members of a population. Thus the genetic signals identified by a GWAS are causal for the trait being studied – not themselves caused by the trait, nor merely correlated with it – though the causal pathway from genetic variation to trait need not be direct.

This causal assumption does not hold when simultaneously studying multiple groups of individuals coming from distinct ancestries. In that case, the presence or absence of an allele is often correlated with the difference in ancestry, and thus ends up being correlated with all genetic differences between those ancestries, as well as all traits that differ between people of those ancestral groups, whether those differences are due to different genetic, environmental or social factors. However, GWAS researchers uniformly include covariates, called *genetic principal components* (PCs), that are indicative of ancestry and are thought to mostly account for these correlations, and thus it is still reasonable to assume that by and large GWAS signals identify genetic causation.

LD Tags Causal Signals and Obfuscates Causal Variants

Equally important to understanding GWAS results is that while a variant significantly associated with a trait more-or-less guarantees that there is a causal genetic influence on that trait in the genomic region surrounding the variant, the variant itself is not guaranteed to be, and is in fact unlikely to be, causal. This is due to the phenomenon of *LD* (linkage disequilibrium)[48]. To illustrate LD, consider the example of the two homologous copies of the chromosomes 8 and 14 in a parent, call them 8a, 8b, 14a and 14b. Consider one variant on each chromosome: call them v8 and v14, and suppose that the allele for v8 on chromosome 8a is not the same as the allele on 8b, i.e. that v8a ≠ v8b. Similarly assume v14a ≠ v14b. Because the chromosomes 8 and 14 are different molecules, which allele of v8 the parent passes on to a child of theirs is independent from which allele of v14 they pass on. This random assortment of v8a and v8b compared to v14a and v14b means that, even if the presence of v8a in people in one generation was correlated with the presence of v14a in those people, over successive generations, that correlation would disappear, i.e. those two variants would tend towards linkage *equilibrium*.

But now consider another variant v8' on chromosome 8, with alleles v8'a ≠ v8'b. Now v8a and v8'a are on the same molecule, as are v8b and v8'b. During the phenomenon of *recombination* during meiosis, each pair of homologous chromosomes are likely to exchange corresponding pieces of themselves. Thus if v8 and v8' are far enough apart from one another on chromosome 8, then similarly to variants on different chromosomes, v8 and v8' will be inherited relatively independently from one another, and the presence or absence of their alleles will become uncorrelated after successive generations. But if v8 and v8' are quite close to each other on the chromosome, then recombination will be unlikely to occur between v8 and v8', and a child will likely inherit either v8a and v8'a, or v8b and v8'b, but is unlikely to inherit v8a and v8'b or v8b and v8'a. This means that even over many generations, the presence of v8a in a person will be strongly predictive of the presence of v8’a in that person, and vice versa. For this reason, variants that are close together in the genome tend to remain in *linkage disequilibrium*.

LD is important to GWAS because if any variant in a region causally influences a trait, all the variants in partial LD with that variant will appear correlated with that trait, with the strength of the LDs partially determining the strength of those trait correlations. This leads to the phenomenon where any GWAS signal will be identified by many tagging variants, all with varying strengths. GWAS results are commonly visualized as *Manhattan plots* where each variant association is a point whose genomic position is on the x-axis (successive chromosomes arranged head-to-tail) and whose association’s $-log_{10}(p\ value)$ is on the y-axis, so that higher points are more strongly associated. The effect of having many tagging variants leads GWAS signals to appear as peaks on these plots, whose visual similarity to skyscrapers lends Manhattan plots its name. See Figure Introduction.2 as an example. Figure Introduction.3 shows a zoomed in view of an example Manhattan plot peak. Each point on a Manhattan plot corresponds to an association similar to the example in Figure Introduction.1.

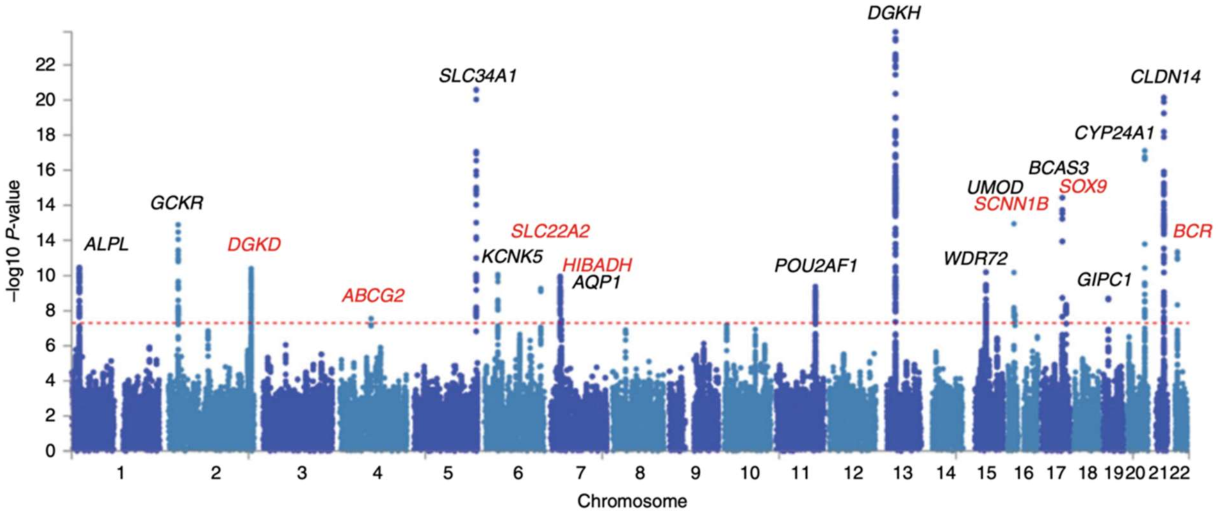


Figure Introduction.2: A Manhattan plot. The dotted red line denotes the genome-wide significance threshold. Each peak has been labeled with the name(s) of one or more genes at that locus which are plausibly involved with the trait under study, though a reader should assume there is not sufficient evidence to confidently identify the causal gene(s) at many of these loci. Adapted from Howles et al.[49]

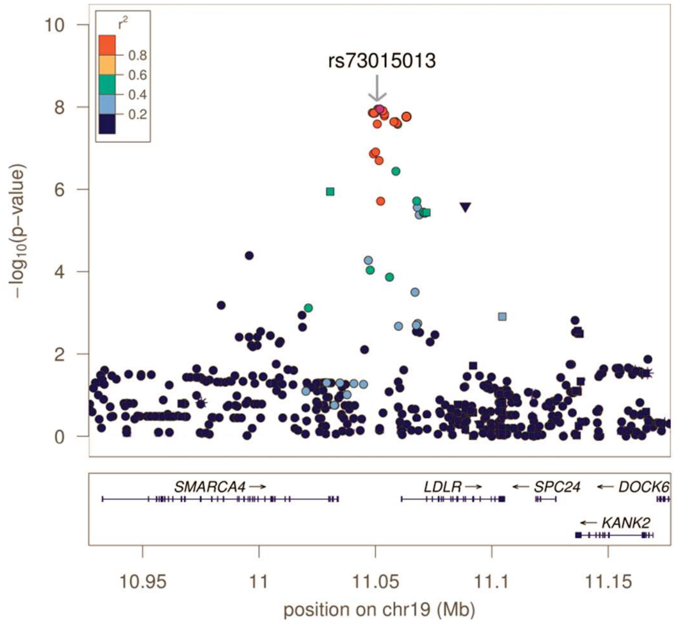


Figure Introduction.3: The association of variants in a genomic region with a phenotype. This is a zoomed in view of one of the peaks from a Manhattan plot. Each point is a variant, with its shape corresponding to an annotation of that variant (annotation names are omitted here for clarity). The lead variant is given by its ID, rs73015013. Each variant in the region is colored by its r^2 with the lead variant. The genes in the region and their directions of transcription are displayed underneath the variants – the exons of the genes are thick bars, with the introns being displayed as a thin line connecting them. Adapted from Sanna et al.[50]

LD between two variants is commonly measured as r^2 – the correlation between the dosages of the alternate alleles at those two variants across individuals in a population. The higher the r^2 value, the more both variants will appear correlated with a trait if either one is causal for that trait. As heuristics, $r^2 > 0.3$ is often considered high enough to say two variants are in at least partial LD, and $r^2 > 0.8$ is enough that it may be difficult to distinguish which variant from the pair is causal for an association and not merely correlated, though these cutoffs are arbitrary and just for the sake of intuition. LD should be handled analytically (using statistical fine-mapping, discussed below), not heuristically. LD can also be measured by pairwise statistics such as D′ (pronounced D-prime)[48], but that measure, while more informative of the relative historical origins of two variants, is less informative of GWAS signal tagging than the r^2 measure, and thus I do not discuss it further.

LD in a region can be measured as a matrix of r^2 values corresponding to the pair-wise r^2 between each pair of variants. Clusters of variants which are all in high mutual r^2 are called *LD blocks*. As seen in the discussion of figure 7 in the HapMap paper[48], not all nearby variants in a region will have similar LD values – LD values will dramatically depend on the historical order mutations occurred in. Thus even in regions of high LD, only some variants will segregate together in LD blocks. In addition, the farther the minor allele frequencies of two variants are from one another, the less it is possible for them to be in high LD, and thus the less it is possible for them to be in the same LD block. Nonetheless, LD blocks often contain tens of tightly correlated variants spread over tens of kilobases or more. And due to tagging, GWAS signals are biased towards appearing in significantly larger than average LD blocks[51]. For example, most of the signals highlighted in our Chapter 3 GWAS span hundreds of kilobases which, for reference, is much larger than the median (~26kb) and mean (~67kb) lengths of genes in the human genome[52].

It is because of LD that GWAS with a few million variants, such as the GIANT height GWAS[32], work at all. For there are ~18 million common SNPs and indels[13], and even ignoring uncommon causal variants and causal variants of other variant types, a GWAS which tests only a few million variants is only testing a small percentage of all common variants. To account for this, such a GWAS tests variants that are spread evenly and densely enough throughout the genome. This way, any causal variant which is even relatively common in the population will be in strong enough LD with one of the tested variants that the tested variant will be strongly associated with the studied trait. This *tagging* allows GWAS to reliably detect causal genetic regions without necessarily testing any causal variants.

While some GWAS rely on tagging to make sure they identify common causal signals, others try to test many variants to increase their likelihood of being able to identify the underlying causal variants. In the last few years biobanks have begun to generate whole genome sequencing (WGS) data for their cohorts, allowing recent GWAS in those biobanks to test hundreds of millions of WGS-called variants, e.g.[14]. However, up until recently the standard has been for large biobanks to have called a smaller number (on the order of hundreds of thousands to a few million) of common variants in their cohorts using microarrays, and this is still the best data available for many biobanks. In order to test many more variants in such a biobank, GWAS will use *imputation*, pairing the microarray data with WGS-called variants from other (usually smaller) publicly available datasets called imputation panels. For example, before it had generated WGS data for each of its participants, the UK Biobank imputed over 90 million variants into their cohort from less than a million microarray variants[53]. Historically, imputation panels would include the HapMap[54], 1000 Genomes[13] and Haplotype Reference Consortium panels[55]; now perhaps the largest such panel is the TOPMed imputation panel[56].

Imputation is possible for the same reasons that underlie LD. If it is known from an imputation panel that the alleles of two variants are correlated, then one of those alleles being measured in an individual by the microarray dataset (called a *hard-called* allele) can be used to help infer the presence of the other allele (called an *imputed* allele). Thus information about a small number of variants can be turned into information about very many variants. Imputation tools are more sophisticated than this description implies; in particular, they jointly infer the presence of sets of nearby alleles that appear together in the imputation panel, called *haplotypes*, instead of instead of relying on pairwise correlations, but the underlying principle is similar. IMPUTE5[57] is a tool often used for imputation, but our lab has primarily relied on Beagle[58], which is similarly accurate, sufficiently fast and also allows for the imputation of multiallelic variants which IMPUTE5 has not supported.

Imputation tools are statistical models and so, at each imputed variant, for each individual, they estimate the probability of that person having zero, one or two alternate alleles at that variant. This allows researchers to remove calls or variants whose which were uncertainly imputed, which is more common when imputing rare variants from common variation. These probabilities are also important during the testing step of a GWAS. Instead of assigning each person the dosage that they are most likely to have at any variant (a *best-guess* call), it has been shown that GWAS perform better when they test for associations with the *expected* number of alleles (an average in the range 0 to 2) for each person at each imputed variant[59]

Searching for Causality and Therapeutic Implications

As discussed, random allele assortment during reproduction means that the regions identified by GWAS are causal for the trait being studied. But while GWAS are good at identifying causal regions, LD means that the effects measured for each individual variant are merely correlative and cannot be assumed to be causative.

Despite this difficulty, GWAS routinely attempt to *fine-map* their signals, that is, to take the causal regions they uncover and try to discover which variants, genes, tissues or cell types, and molecular mechanisms the causal signals at those regions act through. This has been true ever since the first widely recognized GWAS paper[60], published in 2007, which spent much of its effort trying to identify the genes influencing disease within the causal disease signals it discovered. This desire for fine-mapping is to be expected: geneticists aim to better understand the human body and wish to discover better therapeutics for diseases. Yet if understanding health and disease is a primary goal of GWAS, it is infrequently discussed by GWAS publications. GWAS publications routinely develop hypotheses for and sometimes succeed in experimentally confirming the biological mechanisms by which GWAS signals act, but they only infrequently discuss the downstream health implications of those mechanisms.

I have found this disconnect to be disconcerting at times. However, an article reviewing the first 10 years of GWAS discoveries[21] can help reframe this apparent contradiction. This review highlights the contributions GWAS have made to drug discovery in two domains: type 2 diabetes and autoimmune disorders. But as prelude to discussing these drug successes the review first highlights the enormity of studies in the GWAS and broader genetics communities dedicated to understanding those diseases. The large number of studies is testament to the inherent difficulty of causally fine-mapping genetic signals, testament to the difficulty of connecting causal signals to mechanistic understanding of underlying biological processes. Hence it is unsurprising that there are relatively few GWAS studies which also successfully identify the mechanisms behind the signals they discover, much less attempt or succeed at applying such knowledge to improve health outcomes.

Even so, the review article of the first 10 years of GWAS highlights strong connections between GWAS results and drug development, if only in a few cases. These case studies can be paired with evidence that shows broad correlation, though not causation, between genetics research and drug development. In 2015, 8% of drugs on the market had some level of support from human genetic evidence at the gene level as compared to 2% of drugs in preclinical stages[61], and in 2021, over two-thirds of drugs approved by the US FDA's Center for Drug Evaluation and Research had some level of support from human genetic evidence at the gene or protein levels[62].

Thus there is some evidence that GWAS are connected to therapeutic successes. In the context of drug development, that moderate level of evidence is motivation enough. Drug development is hugely expensive: drugs often fail in late stage clinical trials due to lack of efficacy[61] despite functional evidence in preclinical stages[63]. Further, a reasonable fraction of drug discoveries are still serendipitous[64] despite attempts to systematize the drug discovery process. So GWAS research does not need to routinely lead to drug discovery or repurposing to be worthwhile. Rather, any additional evidence from GWAS or other genetic analyses that can lead to even slightly more frequent drug discovery and prioritization successes is of help.

So GWAS efforts are valuable despite not being able to directly suggest therapeutic applications of their work. Instead, GWAS focus on taking the crucial first step of fine-mapping as many causal variants, genes, tissues or cell types, and molecular mechanisms for as many of their signals as possible. That guides my organization of the rest of this section of the Introduction, where I describe how fine-mapping is performed despite confounding LD and how these types of fine-mapping evidence interplay with one another. I begin by discussing the fine-mapping of causal variants, next discuss complex variants missing from causal variant analyses, and then conclude by overviewing methods for identifying causal genes. For readers who are interested in the field of study which frames GWAS results more directly as a means to drug repurposing, I refer them instead to this review article[65].

Causal Variants

There are two general ways GWAS attempt to identify causal variants. The first is through strength of signal alone. Heuristically, if one variant’s association with a phenotype is much stronger than the other variants in the region, (perhaps after conditioning on variants already presumed to be causal), i.e. if the strength of the effect of that variant comparatively overcomes the LD present in the region, then the variant is a good candidate for causality.

However, studies have shown that lead variants may often not be causal[66,67], which is especially plausible in regions where multiple causal variants are present[68,69]. Thus instead of heuristically gauging signal strength, researchers often employ statistical fine-mapping, discussed in its own section of the Introduction below, which can be used to quantitatively disentangle LD patterns and assess the level of evidence that the lead variant is causal, as well as being able to identify multiple causal variants in a region. A variant identified this way, either by strength of p-value or by statistical fine-mapping evidence, can motivate follow up studies focused on resolving the mechanism of that variant. Such is the case in a study which followed up on a specific locus identified by a vascular disease GWAS[34].

However, it is often the case that there are many strongly trait-associated variants in a region which are in tight LD and have similar association strengths, and that neither ranks based on p-values nor statistical fine-mapping can sufficiently resolve the causal variant from that cluster. In such a case, if one of the variants overlaps an annotation which gives the variant a known or plausible mechanism of action (e.g. the variant is protein coding), then that strengthens the hypothesis of that variant's causality compared to the rest. In addition to pinpointing the causal variant, if the overlapped annotation is present only in relation to one gene in the region but not others, or in some cell type(s) but not others, this can help resolve both the genes and cell type(s) the signal is likely to act through. This is demonstrated by a GWAS in type 1 diabetes GWAS[43] as well as a GWAS in a type 2 diabetes[70].

It is important to mention that annotations must be considered within the context of LD confounding and not instead of it. For example, one GWAS of type 2 diabetes coding variants[71] demonstrated through statistical fine-mapping that at least a third of the coding variants they found to be strongly associated with that disease were likely not themselves causal.

It is also important to realize that multiple estimates suggest that >90% of GWAS signals lie outside of coding regions[72,73]. These estimates refer to all associated variants tested by a GWAS rather than the unknown percentage of causal variants, and thus cannot be precise.

Nonetheless, they suggest that a large majority of GWAS signals are non-coding. As discussed above our understanding and annotation of non-coding variation is very limited, so matching GWAS-prioritized variants to known annotations should only be expected to succeed for a limited subset of signals. For example, overlapping 52 type 1 diabetes GWAS signals with chromatin accessibility signals measured in the most relevant accessible cell type only provided evidence for 5 of the 52 signals[43]. Yet though this provided insight into less than 10% of the regions studied, a strong candidate for causality was identified using this chromatin accessibility and follow-up 3D chromatin contact mapping data.

Given all this, the most common scenario is that the top associations at a GWAS signal cannot be distinguished due to LD confounding, and there are no annotations sufficiently convincing as to pinpoint the causal variants from among them. For this reason, many GWAS focus only on a few of the signals they identify, leaving the rest unresolved[43,70]. Even so, GWAS generates hypothesis: GWAS provides a list of the most associated variants (or variants most prioritized by statistical fine-mapping) and directs future research towards identifying which of those variants might be causal. In the long run, this hopefully will lead to the identification of more causal variant mechanisms, and the translation of that learning to understanding and annotating the rest of the genome.

Still, it would be preferable if causal variant fine-mapping was more successful. My thesis work on complex variants is one of many different avenues for increasing that likelihood.

Complex Variants and Missing Variants

In our lab, we use the umbrella term *complex (genetic) variant* to refer to any genetic variant that is smaller than whole chromosomal loss or duplication and is not a SNP or short indel, encompassing many distinct variant classes such as tandem repeats, copy number variants, mobile element insertions and other structural variants. We use the term complex variation because these larger DNA variants are difficult to naively call from short read sequencing and have all generally been omitted from GWAS to-date. Note that, despite using the same word, complex variants and complex traits are different ideas.

The historical exclusion of complex variation from GWAS is partially due to their exclusion from the reference panels used for imputation. The Haplotype Reference Consortium's reference panel[55], widely used with ~1500 citations as of writing, only includes SNPs. This is also true of the older HapMap reference variant set[54], which despite being retired before the beginning of my thesis work[74], is still used in variety of current analyses as a database of common variation[32,75]. The TOPMed panel[56] may be the largest imputation panel at this time which does not have restricted controls, and while it includes indels in addition to SNPs, it does not yet include complex variants[76] (though the inclusion of structural variation in TOPMed is described in a current preprint[77]). To my knowledge, the only commonly used reference panel which includes complex variants at the moment is the 1000 Genomes panel[13] which has included structural variation in addition to indels since at least 2015[16]. Any GWAS using variants imputed from a panel that excludes complex variation will by necessity not study complex variation. But even beyond that, I suspect that smaller research teams look to the variant calling methodologies of these consortia for guidance. Thus the lack of inclusion of complex variation in these consortia likely contributes to standardizing the lack of study of complex variation across the research community.

Prior to my thesis work, there was already evidence that SNP- and indel-based GWAS would identify signals whose causal variants could not be identified, and which would later be resolved to a causal complex variant common in the population which was not studied by the initial GWAS but was tagged by its variants[78–80]. These findings became part of the motivation for my thesis work, where we took a class of common complex variants, in our case *short tandem repeats* (STRs), and included them in GWAS studies. In doing so, we hoped to identify causal complex variants using GWAS, and hoped to improve the overall rate at which GWAS signals could be resolved to causal variants. (I discuss motivation specific to STRs in the section of the Introduction on STRs below. Our results working with STRs are discussed in Chapters 1 and 3.) Other examples of causal complex variants tagged but overlooked by SNP- and indel-based GWAS have been discovered by the research community while I have pursued my PhD studies, e.g.[81,82].

Further, while it is expected that common SNPs strongly tag most missing common biallelic variants, our lab[83,84] and others[81,82] have demonstrated that when a complex variant such as an STR is present throughout the population at a range of common lengths, the presence or absence of any individual common SNP is unlikely to be strongly correlated with the complex variant's length due to the SNP's biallelic nature. This means that LD should be less confounding for such a causal complex variant, and suggested that if we included STRs in GWAS, we would be able to causally identify some of them through statistical fine-mapping.

The other common occurrence which leads to identifying causal variants is when they overlap genomic annotations which strongly implicate causality, such as being in a protein-coding region. This has allowed for the causal identification of coding complex variation[81,82]. A small subset of STRs are coding, are in 5' UTR regions, or are directly adjacent to splice sites. All of those are more likely to be causal than the average STR, are easier to identify when they are causal, and can help implicate causal genes.

However, my work focused on the study of STRs genome-wide. And the above categories only make up a very small fraction of the STRs in the genome. While I will discuss STRs in more depth below, for now it is sufficient to say that there are many mechanisms of action hypothesized for non-coding STRs, but similar to most variant types in the non-coding genome, these mechanisms are not well annotated or well validated. Thus in general most causal STRs cannot be identified through annotations. And even when likely-causal non-coding STRs would be identified by fine-mapping, without causal annotation, it would be difficult to link them to causal genes or understand their mechanisms of action. Still, even without being able to annotate putatively causal STRs, we hoped in the long run that identifying them would lead to follow up studies which could decipher the mechanisms by which they act.

There is also benefit even when statistical fine-mapping only indicates that a complex variant is one of many variants in an LD block that may be causal for GWAS signal. For that improves hypothesis generation at that locus, indicating that follow-up research should study that complex variant along with all the other variants in the region when seeking to identify the causal variant.

It is for all these reasons – identifying the occasional causal, well-annotated complex variant, the more frequent uncovering of evidence that an unannotated non-coding complex variant is likely causal, and the routine occurrence of identifying possibly causal complex variants in LD with other possibly causal SNPs – that I have focused my work on complex variant GWAS.

Causal Genes and QTL Studies

While my work has focused on identifying causal variants, GWAS researchers are more regularly interested in identifying causal genes, as the proteins they code for are the fundamental units which act on pathways in the body outside the nucleus, and as those proteins can be nominated as therapeutic targets[43,65]. The difficulty is that GWAS fundamentally do not test genes for associations with traits. Further, there are often many candidate genes near a GWAS signal, and distant genes cannot be ruled out as GWAS signals may act on genes over 100 kb[73] or 500 kb[34] away. To give a sense of the magnitude of this challenge, a review from 2017[21] suggested that the entire research community had identified causal genes for only one third of the hundred type 2 diabetes GWAS signals identified, and considered this a major success story.

One method to identify causal genes is to look for nearby genes which already have some evidence of relevance to the trait being studied. This is the strategy taken by the foundational GWAS study from 2007[60]. A similar recommendation is given by the type 2 diabetes study mentioned above, which suggests focusing on genes which can already be "plausibly linked to the … phenotype"[71]. Yet the issue with this approach is that our knowledge of the genome is limited, and it is often unknown which pathways a given gene is involved in.

A simple alternative is the oft discussed heuristic of nominating the gene closest to a signal as likely enough to be causal for that signal, without having to understand the mechanism of that causality. Unfortunately, how often this is correct is a matter of large disagreement – different gene prioritization studies suggest this is accurate anywhere from one fifth[85] to one third[86,87] to two thirds of the time[88] or more[73,89]. These estimates likely differ due to being biased by limitations of their data, their methodologies, and which types of genetic mechanisms their methodologies do or do not consider. Yet more fundamentally, the estimates will also vary according to the extent to which weaker GWAS signals are or are not included in the analyses.

Instead of relying on the closest gene to be causal, researchers have tried to find approaches that are more data driven. Perhaps the most common method is to test which variants influence gene expression, directly addressing the limitation that GWAS do not test gene expressions for association. This is called an expression QTL (*eQTL*) study[30]. QTL is an acronym for quantitative trait locus, which literally refers to any locus in the genome that is implicated in any quantitative trait being studied. However, in the context of studies in humans, a 'QTL study' most commonly refers to a study of molecular phenotypes of chromatin and transcription, including but not limited to the expression levels of genes (eQTL studies), the extent to which specific CpG bonds are or are not methylated[90] (called *mQTL* or *meQTL* studies), or the distribution of isoforms of mRNAs[91].

QTL studies of molecular traits are similar to GWAS of *organism-level* traits (e.g. traits such as height or heart failure that are properties of full organ systems or the whole body) but there are a few important differences between the two. First is that QTL studies inherently test many distinct outcomes simultaneously. For example, a QTL study of expression will individually test the expression of each gene in the genome (~20,000 traits), and QTL studies of methylation will individually test the methylation levels of each potentially methylated region in the genome (over 750,000 traits[90]). Secondly, the number of individuals accessible to QTL researchers is often much lower. For example, GTEx[30] is one of the prime sources of tissues for these analyses, and for most tissues, GTEx only has tissue samples from a few hundred individuals. (As an aside, the fact that human datasets of gene expression are not usually the same as the datasets GWAS are preformed in is a main reason why human genetics researchers do not test directly for associations between gene expression and traits).

The small number of individuals in QTL studies has downstream ramifications. QTL studies often lack power to overcome the multiple hypothesis burden of true genome-wide testing[92], and so only test each trait for association with nearby (called *cis*) genetic variants instead of all genetic variation in the genome (which would include *trans*, i.e. distant, variation). Here, the definition of nearby is up to the researcher, often in the 250kb-1mb range. Thus, like GWAS, QTL studies test variants from across the genome for association, but instead of testing each variant for association with a single trait, many test variants for associations only with the molecular traits that are anchored nearby. The lack of power in molecular QTL studies also necessitates slightly different methodologies for controlling for false positive rates. Often researchers control for false discovery rate at a threshold such as $p < 0.05$,[90] instead of controlling the family-wise error rate with the threshold $p < 5 \times 10^{-8}$ as used in organism-level GWAS.

As QTL studies often measure traits directly related to chromatin and transcription, they are interpreted differently than GWAS, providing direct evidence of the genetic mechanisms influenced by genetic variation, but not measuring the downstream influences of those mechanisms on organism-level traits. For instance, one of the main pieces of my thesis work was performing an eQTL study of STRs, a main goal of which was to try to elucidate the mechanisms of causal STRs (see Chapter 1). It is important to note that QTL evidence is confounded by LD, similarly to GWAS studies, so pinpointing causal variants for QTL signals can be difficult.

Another distinction with organism-level GWAS is that the molecular traits studied by QTL studies often vary between tissues and cell types – for instance gene expression, genetic methylation and mRNA splicing all can show distinct patterns in different cells. Thus QTL results are highly circumscribed to the cells and conditions the results were generated in, and consortia are constantly trying to expand the number of cell types that QTL studies have generated information for[31]. While the cell type specific nature of QTL studies limits their generalizability, it can also be a boon. For if a QTL has been studied in multiple cell types and is only present in one cell type and not others, that can help pinpoint the cell types relevant to that signal[39].

At this point, a reader may be confused as to what the difference is between a QTL study of chromatin and a chromatin annotation. To clarify: chromatin annotations, such as methylation levels and chromatin accessibility, are often measured in cells from one individual. Researchers then tentatively assume the likely presence of those chromatin annotations in other individuals. Further, researchers can hypothesize that variants in those regions may interfere with or modulate those annotations. But these assumptions and hypotheses are rarely verifiable from chromatin annotation information alone. On the other hand, QTL studies, such as methylation QTL studies or chromatin accessibility QTL studies, measure the differences in such annotations between individuals with different variants, directly testing those assumptions. In contrast, QTL studies are not designed to identify chromatin marks which are uniform throughout the individuals under study and not perturbed by the studied genetic variation.

A last note on QTL studies before moving back to the discussion of causal genes: despite being called QTL studies, *pQTL* studies which measure the quantities of different proteins, often follow the design and interpretation patterns of GWAS more closely than of molecular QTL studies. For pQTL studies are often performed on proteins circulating in the blood[93,94] and blood is an easy tissue to access, so pQTL studies can have similar sample sizes and power to organism-level GWAS. This allows pQTL studies to adopt stringent family-wise error rate thresholds similar to GWAS instead of laxer FDR thresholds used by other QTL studies. This also allows pQTL studies to look for *trans* signals in addition to *cis* signals.

As alluded to above, one motivation for QTL studies is to help fine-map which genes are causal for GWAS signals. Most naively, QTLs can be used similarly to annotations, where a researcher who identifies a GWAS signal that overlaps an eQTL signal can infer that the GWAS signal is caused by the eQTL signal, even if they cannot identify which variants are causal for either. A follow-up to a type 2 diabetes GWAS signal is a good example of this[39]. That study also demonstrates how this allowed them to also identify the causal cell type, as the eQTL signal was only present in that single cell type. The counter point is also true – if they had only studied eQTLs in non-relevant cell types, they would not have seen an eQTL signal here at all, and not been able to connect the GWAS signal to a gene. In fact, eQTL studies can even lead to misleading results when performed in less mechanistically relevant tissues[95].

Frequently, the simple approach of overlapping GWAS and eQTL signals suffers from the drawback of being unable to distinguish between genes causal for a GWAS signal, and non-causal genes whose expression is correlated with the same variants, but only due to LD. C*olocalization*[96] is the term for statistically distinguishing between these two possibilities. As it is in effect a type of *multi-trait statistical fine-mapping*, I discuss colocalization in more detail later in that section of the Introduction dealing with that topic. Colocalization is often used in fine-mapping GWAS signals, e.g.[70] and in our work in Chapter 1. However, due to worries that eQTL datasets are not sufficiently well powered to properly detect causal signals, when we incorporated eQTL data with our GWAS signals in Chapter 3 we only overlapped them and did not perform colocalization. I discuss this more in the Chapter 3 Forward.

While overlapping and colocalization attempt to directly compare eQTL and GWAS signal patterns, transcriptome-wide association studies (TWAS) are a different class of methods which attempt to use variant GWAS and eQTL associations as a proxy for identifying gene associations. Specifically, these approaches use summary statistics from GWAS and eQTL datasets to impute gene expressions into the GWAS cohort, and then directly test the gene expressions for associations with the GWAS trait. TWAS can then, in theory, directly show which genes are involved with a trait. For more information, I refer the reader to a recent TWAS review[97]. I only highlight here that, like other eQTL base studies, TWAS can be confounded by data from tissues and cell types irrelevant to the trait under study[95]. Further, TWAS results are also susceptible to confounding due to LD[65,95], and so themselves need to be fine-mapped. Of a few recent attempts at fine-mapping TWAS signals[98,99], it is exciting to see that one such method, called cTWAS, achieves low power but very high precision in identifying causal genes by identifying the underlying variants which are causal for gene expression[100]. This is possibly another application of identifying causal non-coding variation such as causal STRs.

Lastly, there are methods for inferring causal genes from GWAS summary statistics aside from those that utilize eQTL data. For example, some methods cross GWAS summary statistics with knowledgebases of gene function and molecular pathways[101]. Some methods use all three types of data – GWAS summary statistics, eQTLs and knowledge bases[102].

Despite these efforts, causal gene prediction still remains a challenging problem – a recent gene prioritization effort using a variety of methods found that the different methodologies had relatively little overlap in the genes they prioritized[103]. Further, comparative studies of gene prioritization methods often suggest that the closest-gene heuristic performs as well as or nearly as well as methodologies which incorporate QTL and/or knowledgebase datasets[95,102,103]. And even making such comparisons is challenging due to the difficulty and biases in ascertaining curated sets of known-causal genes.

In summary, GWAS reliably generates knowledge of causal regions, while identification of causal variants and genes remains much more challenging. Many methodologies and data sources have been developed for those purposes. These methodologies can generate great insight when they provide strong evidence of which variants and genes are causal, but only do so for a relatively small proportion of GWAS signals. We envision that including STRs will increase that proportion slightly but significantly. More frequently, GWAS signals generate hypotheses for possible causal variants and genes and leave questions of causality for further, often experimental, research. We aim to include STRs in those hypotheses.

Polygenic Risk Scores

One important application of GWAS is generating *polygenic risk scores* (PRS), also called *genetic risk scores*[104], or *polygenic scores*[32]. While my thesis work does not directly involve PRS, due to their importance to population genetics, and because we hypothesize that the identification of causal STRs will improve PRS, I briefly discuss PRS here.

A PRS is a method that predicts either a phenotype (or future phenotype) of a person from the knowledge of which alleles that person has at a collection of variants, along with other covariates. PRS are called risk scores because the main interest in PRS is predicting which individuals will get specific diseases, which can allow for preventative treatment. However, PRS can be built for any phenotype, including those where the term risk is a misnomer. PRS, like many predictive models throughout the fields of statistics and machine learning, can achieve accurate predictions without identifying what subset of input features (in this case, genetic variants), are causal for the predicted phenomena. Nonetheless, I will explain why causal STR identification may improve PRS results.

The simplest PRS method, called *pruning and thresholding*, takes GWAS results, removes all variants below a tuned threshold (thresholding), selects a single variant from each associated LD block (pruning), and then uses the sum of the variants' GWAS effect sizes times individuals' dosages at those variants to predict those individuals' phenotypes[105]. There are many more sophisticated PRS methods; an interested reader could look to these reviews[106,107]. Further, more sophisticated PRS methods tend to perform better than pruning and thresholding, though different PRS methods perform best for different traits and in different contexts[105,107]. Nonetheless, many of these methods are similar to pruning and thresholding in that they build models from GWAS summary statistics, though they differ in how they select which variants to include their models (up to including all tested variants) and how they up- or down-weight GWAS effect sizes based on model priors or tuned parameters. While clinical usage of PRS is currently highly limited, there is much discussion of the future utility of PRS in the clinic[106,108–110]. And there are currently ongoing clinical trials for using PRS to predict breast cancer[111,112] and colorectal cancer[113] that could help bridge this gap, among other efforts.

I discuss PRS here not only because they are an important use case for GWAS, but also because they are one part of our motivation for attempting to identify causal complex variants. This is perhaps unintuitive, as PRS are predictive methods, and they should perform equally well whichever variants they include from any given highly correlated LD block, regardless of whether the variants they include are causal or merely correlated. While that is true, we[83,84] and others[81,82] have demonstrated that, a large subset of multiallelic STRs are not fully tagged by individual SNPs. Thus incorporating multiallelic STRs may yield marginal improvements to PRS for traits which those STRs are causal for. I am also encouraged that some existing PRS models only include sparse collections of variants[104,114–118]. I hypothesize that those methods would be particularly improved by swapping out individual variants for the causal STRs they tag; though it remains to be seen which traits, if any, those methods prove most successful for when STRs are incorporated.

That said, our lab hypothesizes that the benefits of causal variant identification may be most apparent for PRS *transferability*, which I will now define. Recall that there are huge disparities in the size of biobanks containing people of European decent as compared to people of African, Hispanic and South Asian descent, with East Asian biobanks falling somewhere in the middle. Due to these disparities, PRS used for people of non-European, non-East Asian descent are often first trained primarily on data from European individuals. Using a PRS in a population other than the population it was trained on is called *transferring* it to the target population. Creating PRS which transfer well will remain an important need until biobank sizes are more equitable across populations.

Currently, PRS do not transfer well[119,120], in that they show much lower accuracies in populations they are not trained on. This is widely hypothesized to be in part due to different LD patterns between training and target populations[119,121]. Said another way, PRS effect sizes for variants are reflections of the correlations between those variants and the causal variants they tag, and the expectation is that those correlation patterns may change for when moving between populations, thus rendering many of the PRS effect sizes inaccurate in the new (target) populations. I note that differing LD patterns are not thought to be the only cause of PRS accuracy loss between populations[121]. Nevertheless, the GIANT height GWAS demonstrates via simulation that differences in LD, as well as in minor allele frequencies, may be causal for an accuracy drop from 40% to 15% in their height PRS of Europeans vs Africans[32]. These simulation are complicated and their specifics are hard to verify due to the lack of knowledge of the truly causal variants for GWAS traits[122]. Still, the overall point is convincing, and the upshot is that if PRS could put more weight on causal variants as opposed to tagging variants, then their transferring inaccuracies would hypothetically be mitigated. That is further motivation for our attempts to identify causal STRs.

Short Tandem Repeats

My research has focused on the inclusion in GWAS of a type of complex variant called *short tandem repeats* (STRs). Above I have introduced GWAS, their role and utility in population genetics analyses, and the benefits of including heretofore missing variants in GWAS analysis. Here I introduce STRs.

Tandem repeats (TRs) are sections of the genome where the same sequence of bases is repeated many times in a row, tail to head. For example, the sequence …TTACAAACGACGACGACGTGAAC… contains four copies of an ACG repeat which can be highlighted using bolding and capitalization: …ttacaaACG**ACG**ACG**ACG**tgaac… . A tandem repeat is often discussed in terms of its *repeat unit* or *motif*, the length of that unit, the number of copies of that unit, and the total length of the repeat. For the example above, the repeat unit may be denoted by ACG, CGA or GAC, or if a researcher was considering the repeat on the reverse complement strand …CGTCGTCGTCGT… , then either CGT, GTC or TCG. Regardless of how it is named, the length of the repeat unit in this example is 3 bases, there are 4 copies of it, and the total length of the repeat is 12 bases.

Another important facet of tandem repeats is their purity. A tandem repeat is called *impure* if it contains one or more interruptions of the repeated sequence, say …ACGACGAC**A**ACG…, where the third G from the left has been replaced by an A. My work has focused on laying the groundwork for the study of repeats based on their lengths, and I do not focus much on the impurities within them. Nonetheless, it is important to recognize that impurities, in at least some instances, fundamentally change the biomolecular properties of repeats[80,123,124], and in the thesis Discussion I consider scanning for associations between repeat impurities and phenotypes. For the rest of this thesis, it is just important to know there is no precise agreed upon cutoff which delineates which sequences or regions of the reference genome are repeats with many impurities and which are non-repetitive sequences of bases. As such, numeric claims about classes of repeats fluctuate significantly from study to study depending on the extent of the set of variants being labeled repeats when making the claims.

My work has focused on length variation in *short tandem repeats* (STRs), also called *microsatellites*, *simple sequence repeats*[125] and *simple tandem repeats*. Our lab ascribes to the common definition that STRs are those tandem repeats whose repeat unit has a length of 6 or fewer bases. Note that this definition of STRs is irrespective of the total length of the repeat, which can range from tens of bases in common cases to hundreds or thousands of bases in extreme cases. Also note that other research groups often use similar but not entirely identical definitions for what constitutes an STR[126,127].

From well before the whole human genome was read, STRs were used as markers in forensics[128], genetic linkage analyses[129] and other applications as their high mutation rates cause them to frequently exist at different lengths in different individuals. A large body of research has also focused on *repeat expansions*, when an individual inherits an STR that is mutated to be hundreds or thousands of repeat units long, well beyond what is standard in the population. Repeat expansions in specific STRs are causal for over 50 severe Mendelian disorders, most of which primarily affect the central nervous system, such as Huntington's disease and ALS[127,130,131]. In contrast with these approaches, my research has focused on the causal properties of STRs instead of using them as markers, and has focused on common STR variation genome-wide instead of focusing on a few known pathogenic STR expansions.

Part of the motivation for working on STRs genome-wide is their numerousness. Somewhere on the order of 2.5%[5] to 6.77%[124] of bases in the genome lie in STRs, occurring at 1.6[132], 2.5[133] or 4.6[134] million distinct loci, depending on the definition used. These different loci can be characterized by their repeat unit: STRs with a repeat unit that is just a single nucleotide are called *homopolymers*, and are called *poly-As* when that nucleotide is an A. poly-As are important as they are the most individually numerous type of repeat in the genome, with 41.3%, 47% and 50.7% of the TRs in the Ensemble-TR v2 reference panel[84], human species table in the MicroSatellite DataBase[134] and HipSTR reference[132] being poly-A repeats, respectively. STRs with 2-6 base repeat units are called *dinucleotide*, *trinucleotide*, *tetranucleotide*, *pentanucleotide* and *hexanucleotide* repeats, respectively.

In addition to their quantity, individual STRs have very high per-generation mutation rates, commonly expanding or contracting by one or more repeat units. One estimate puts the average STR mutation rate at 5.6×10^{-5} mutations per locus per generation[135], much higher than the average genome-wide rate which is roughly 5×10^{-9} to 3×10^{-8} mutations per base pair per generation[136]. In particular, this is driven by the large number of STRs with shorter repeat units coupled with the fact that STRs with shorter repeat units have higher mutation rates than STRs with longer repeat units[135]. This leads to estimations that there are close to as many new STR mutations in each individual born as new SNP mutations (54[135] vs 73[137]) despite STR mutations only occurring at STR loci while SNP mutations can occur anywhere in the genome. This makes length variation in STRs a large fraction of genetic variation (see Chapter 3 Supplementary Table 3).

As mentioned above, length variation in STRs is one category of complex variation commonly excluded from GWAS. Partially this is because some imputation reference panels do not include any indels[54,55], of which STR variation is a subset. However, this is also due to STRs being difficult to naively call from short read sequencing data. Two facets of STRs in particular contribute to that difficulty. Firstly, if the process that generates reads for sequencing includes a step called *PCR* which was ubiquitous in older workflows, the process will often generate mutated reads with additional or fewer copies of the repeat. This phenomenon is known as *stutter error*, and is thought to be due to the same underlying biomolecular processes that cause repeats to mutate in the genome[138]. Stutter error leads to noisy short read sequencing data, which can often lead to STR loci being dropped from datasets due to low call quality. This is especially problematic for homopolymers: one estimate suggests that 17% of reads containing homopolymers experience stutter error when processed using PCR[139]. Secondly, the most common alleles for some STRs are nearly as long or longer than the length of the short reads used by WGS, leading to scenarios where no single read spans the repeat and thus the repeat cannot be called by standard genotypers which rely on information from spanning reads[140].

Further, even for GWAS which impute calls from references such as TOPMed[56] and 1000 Genomes[13] which contain indel calls, it remains unclear if the indel callers they used, which were likely not specialized to calling STRs, are sufficiently accurate and sensitive when at STR loci. The 1000 Genomes call set authors directly acknowledge this in their 2022 publication, saying that they "have not specifically included simple tandem repeats" in their ≥50bp structural variant call set as "accurate genome-wide discovery [of such repeats] remains a considerable challenge"[13]. The publication of the HipSTR STR genotyper in 2017 showed HipSTR to be more sensitive and accurate than the standard indel callers at the time[132], and this comparison only took place for repeats with total length less than 100 bases and repeat units of length two or more, excluding the STRs most prone to read (and thus call) errors[139]. Despite these pieces of evidence, there is need to reassess the capacity of today's general purpose indel callers to call STR loci.

Being part of the Gymrek lab, I have been well positioned to circumvent the challenges of STR calling, as one of our lab's specialties is tools for calling STRs from short-reads[132,140,141]. We are not the only lab to work on this task[142]. But much research by other labs has focused on creating tools that can detect repeat expansions[143–146], which is an important use case, but one which does not automatically lend itself to calling common alleles at STRs genome-wide. Our lab's specialization in STR callers also explains my focus on STRs as opposed to tandem repeats as a whole, for tandem repeats whose repeat unit length is 7 or more, referred to as *variable number tandem repeats* (VNTRs) or *minisatellites* (though again, definitions differ slightly between authors), often require different callers[147,148].

Motivating us to study common length variation in STRs was the ample evidence of the involvement of common differences in STR lengths in a wide range of genetic molecular mechanisms. Before the bulk of my thesis work, STRs lengths had already been shown to modulate splicing by inducing hairpins in RNAs[149] and by recruiting splicing activation factors[150], to modulate the affinities of the binding of transcription factors[80,151,152], and to tune genetic expression through the modulation of nucleosome positioning[153]. And STRs debatably were shown to repress genes through increasing CpG methylation[154]. Changing lengths in different STR repeat unit classes had also been shown to modulate a wide-range various DNA secondary structures, including Z-DNA[155], G-quadruplexes, hairpins and i-motifs[156], and the DNA-RNA hybrid structure R-loops[157]. These structures had been shown to promote the formation of mutations[157], interfere with transcription[155] and stall DNA replication during cell division[156]. Throughout the duration of my PhD further evidence of the mechanistic involvement of routine variation in STR lengths was produced, with a new study demonstrating wide-spread STR involvement with methylation[158], and another providing detailed evidence of STRs affecting the binding affinities of large classes of transcription factors[123].

Despite the breadth and strength of this evidence, it is important to note that only a few studies have attempted to link these STR mechanisms to GWAS signals directly in their native chromosomal context, e.g.[80], while most others have done so either in transfected plasmids[149,150,153,155,157] or via purpose-designed assays[123,151,156]. One study provided conclusive evidence linking the functionality of STR-mediated methylation in its native chromosomal context to human disease through the expression of a nearby gene using a CRISPR-based model[159]. But I note this study was investigating a repeat expansion disorder and not a GWAS signal driven by STR lengths common in the general population. All this is only to say that, like most types of non-coding variation, there is no obvious blueprint for identifying the molecular mechanisms of non-coding STRs, nor is there strong evidence for how widespread the impact of each of these mechanisms is expected to be.

Still, this evidence motivated our hypothesis that genome-wide analysis of STR associations would identify STRs causal for human traits. This motivated both the paper linking

STRs to gene expression in Chapter 1, as well as the blood traits paper in Chapter 3 that makes up the bulk of my thesis work.

In those papers we made the choice to, for each STR variant, test for a linear association between the phenotype under study and the sum of the lengths of the two STR alleles at the two copies of that variant. Summing over the two homologous chromosomes is analogous to the standard test for phenotype associations with the alternate alleles of biallelic variants, where the number of alternate allele copies present at each locus on both chromosomes is counted. However, the length-based testing stands in contrast to the standard GWAS approach for multiallelic variants. In PLINK 2 alternate alleles of multiallelic variants are tested separately[160]. This fails to pool information from across alleles and thus has reduced power to detect trends across multiple alleles, especially when three or more alleles are common. Many GWAS tools go one step further and require multiallelic variants to be split into multiple biallelic variants[161,162]. This confounds the presence of the reference allele with the presence of alternate alleles aside from the one being tested. Our choice of linear length-based testing avoids these losses of power, and our increased power to detect effects at STRs was another motivation for our work. I do note that linear length-based testing is not perfectly positioned to detect all, potentially non-linear, length-based trends, and I delve into alternative testing methods further in the Discussion.

Thus we were excited to perform GWAS with STRs due to our access to high-quality STR genotypes, our length-based testing model, and mounting previous evidence of the causal effects of common length variation in STRs. Still, as discussed in the previous sections, GWAS often struggles to identify causal variants. And this is amplified in the case of non-coding variants, especially STRs, which have so much mechanistic heterogeneity that it is routinely unclear what mechanisms any given associated STR may operate by. This challenge of distinguishing associated from causal STRs had already hampered our analyses in the past[163].

For that reason, and due to the need to quantitatively measure the probability of STR causality, our studies relied on *statistical fine-mapping* to select for causal STRs.

Statistical Fine-Mapping

As detailed above, *fine-mapping* of GWAS signals is the attempt to isolate the causal variant(s) and gene(s) in those signals from the many non-causal variants and genes they are in LD with. In this section I focus on *statistical fine-mapping* and specific statistical fine-mapping methods, called *statistical fine-mappers.* Each of these methods is built on a statistical model of the genetic associations in a genomic region. They use the associations between variants in a region and a trait, as well as LD patterns between the variants, to fit those models, and from the fitting they probabilistically decipher which of those variants are likely causal for the trait and which are merely in LD with other causal variants. While I note that there are tools for identifying causal genes which borrow from the field of statistical fine-mapping[100], and while causal variant identification can sometimes lead to causal gene identification, statistical fine-mappers have focused primarily on the discovery of causal variants, and that will be the focus of this section. I also note that some PRS models borrow from statistical fine-mapping[114], but again my focus here is on methods whose aim is to pinpoint causal variation.

Statistical fine-mapping has important strengths in comparison to other fine-mapping techniques. Wet-lab experiments that test mechanistic hypotheses are the gold standard of identifying and validating causal variants, genes and mechanisms. Yet despite constant technological advances, wet-lab experimentation both has limited throughput and often is costly and time-consuming[164,165]. In contrast to wet-lab based approaches, statistical fine-mapping is fast and cheap – with run times from seconds to hours depending on locus size, instead of weeks to months or years. There are also fine-mapping approaches which mine patterns of variant and gene function from existing knowledgebases and are roughly as fast and cheap as statistical fine-mapping. However, they rely on existing knowledge, while statistical fine-mapping is largely unbiased by existing hypotheses or information.

In turn statistical fine-mapping methods have their own drawbacks. The largest of these is that statistical fine-mapping results are not self-validating – the process of statistical fine-mapping does not innately develop understanding of what is going on at a locus, it simply produces a result and asks the researcher to trust that result. Like all fine-mapping techniques, statistical fine-mapping methods cannot distinguish the causal variant(s) at some loci. For statistical fine-mapping methods, this is when the signal being detected is weak enough that there is not enough statistical power to pull apart variants in high LD with one another. In practice, this means that while GWAS researchers already look for ever-larger datasets to increase their power to detect weak signals, at some loci, statistical fine-mapping applications will need even more data than that to successfully resolve those signals to their causal components. But when statistical fine-mapping cannot fully deconvolute LD blocks, it attempts to output lists of potentially causal variants as short as possible so that few follow-up wet-lab experiments are needed to test them.

Statistical fine-mapping is a field whose basic premises have changed in the last fifteen years and which has continued developing during the course of my PhD studies, so I will briefly trace its history before explaining the model that underpins current statistical fine-mappers. Statistical fine-mapping grew out of analyses of GWAS results in the late 2000s and early 2010s. The foundational 2007 Wellcome GWAS[60] is a prime example of a study which recognized the problem of LD confounding but did not perform statistical fine-mapping. That study took its GWAS associated variants that passed a specific p-value threshold and attempted to heuristically identify which of them were likely causal by which of them tagged known biology, lacking any straightforward way of making quantitative statements about causality probabilities. In subsequent years, an easy method for computing Bayes factors from GWAS summary statistics was derived[166]. This lead to the development of a method which could calculate posterior probabilities of causality for each variant in a GWAS signal region from just GWAS summary statistics[167], later called the approximate Bayes factor (ABF) method[168]. These posterior probabilities of causality allowed for the quantitative interpretation of GWAS summary statistics as predictors of causality.

Yet ABF makes the simplifying assumption that only one causal variant exists within each GWAS signal, and even in the early 2010s, many GWAS were using conditional forward stepwise regressions[169–171] to identify regions which likely contained multiple causal variants. Conditional forward stepwise regression iteratively regresses out the effects of all variants already marked as causal in a region (or in the genome), then designates as causal the variant with the strongest remaining association in the region (or in each region) and repeats until the new conditional summary statistics no longer pass a preset threshold.

However, the assumptions of this forward stepwise approach fail to hold up. In particular, it always designates the lead variant in a region as causal, despite estimates suggesting that this is often not the case[66,67] and despite knowing that the lead variant in a region with more than one causal variant may not be causal, and instead strongly associated due to being in partial LD with the multiple causal variants[68,69]. (I note that conditional regression can be useful for determining if a preselected set of variants explains all the GWAS signal in a region, but it is not reliable for determining if those variants are causal. We use conditional regression for this purpose in Chapter 3.) The COJO stepwise regression method[172] was developed to circumvent some of the drawbacks of forward stepwise methods by allowing for a potential backtracking step. But even this more sophisticated method has been shown to have worse precision and recall than more modern fine-mapping methods[168,173].

CAVIAR[165] from 2014 is the first fine-mapper, to my knowledge, that addressed both these problems by simultaneously assessing the chance of causality of multiple variants. CAVIAR's statistical model set the stage for many future statistical fine-mapping methods, so I describe it here. CAVIAR assumes that the measured trait value for each individual (represented as the vector y) is a noisy linear combination of individuals' measured genotype dosages (matrix X) and the variants' unmeasured effect sizes (vector β), and is normally distributed with inferred variance σ^2. This can be stated as the formula $y \sim N(X\beta, \sigma^2 I)$. Ordinary least squares regression with multiple predictors uses this model too, as do many PRS methods. But unlike those methods, which fit β in a relatively unconstrained manner, CAVIAR and its successors enforce priors that makes β sparse, i.e. force β to contain few non-zero elements. These statistical fine-mapping methods then fit this model to genetic data, perform calculations to infer the posterior probability that any given entry in β is non-zero, and interpret that as the probability that the corresponding genetic variant is causal. I note that while the model above involves the genotypes of individuals, given by X, most current statistical fine-mapping methods can be fit to just GWAS summary statistics and a matrix describing the LD of variants in a region to one another, and so can be run without access to privileged information about individuals. Under this one overarching methodology set forth by CAVIAR, statistical fine-mappers differ in which priors they use to enforce the sparsity of β, how they explore space of possible combinations of causal configurations (that is, which elements of β are non-zero), and how they summarize that exploration.

Fine-mappers report their results as *PIPs* and *credible sets*. PIPs (posterior inclusion probabilities) are numbers between 0 and 1 assigned to each variant that summarize the fine-mapper's posterior belief that the variant is causal. This is the same information that the ABF method first reported, and in doing so formalized the use of GWAS for causal inference, though ABF did not use the term PIP and made simplifying assumptions to come to this information.

While the meaning of PIPs has remained stable throughout recent statistical fine-mapping history, the meaning of the term *credible set* has changed over successive publications. Now *credible set* is commonly used to refer to a collection of variants where the fine-mapper guarantees with some preset probability that at least one variant in the collection is causal[174], though even these guarantees differ between fine-mappers. Generally, fine-mappers return one credible set for each independent signal they identify in a GWAS region. A credible set containing only one variant can be interpreted to mean that that variant is likely to be causal, and a credible set containing many variants indicates those variants could not be sufficiently distinguished from one another due to confounding LD.

CAVIAR was the tool we chose to use for our paper in Chapter 1, being state of the art at the time we performed those analyses. However, while CAVIAR introduced the modern fine-mapping framework, it took the very simplistic approach of attempting to model every possible configuration of causal variants with a preset maximum number of causal variants. It would calculate a posterior likelihood based on how well each such configuration fit the data, and then calculate a PIP for each variant as the sum of the likelihoods of each configuration in which that variant was causal. This brute-force approach made CAVIAR very slow, and unable to consider more than two simultaneously causal variants for many genomic signals.

CAVIARBF[175], a tool based on CAVIAR by different authors, improved on CAVIAR's theoretical framework in showing the similarities between CAVIAR's model and the model of ABF. CAVIARBF also increased the speed of CAVIAR's posterior probability calculations. However, CAVIARBF still tried to enumerate all possible configurations, which increases exponentially with the number of causal variants allowed. Thus CAVIARBF could only reasonably allow for examining up to three simultaneously causal variants.

FINEMAP[176] in 2016 improved upon CAVIARBF by implementing a stochastic method for searching what it considers to be plausible causal variant configurations, instead of examining all such configurations. For this reason, FINEMAP is able to consider an effectively unbounded number of causal variants and still runs incomparably faster than CAVIARBF. This has made FINEMAP a common choice among current statistical fine-mapping methods. However, I note that both data presented in Chapter 3 and unpublished correspondence with FINEMAP's author suggest that, at a relatively small percentage of loci, FINEMAP's predictions may differ dramatically across repeated runs. This suggests that FINEMAP's speed may come at the cost of marginal, but significant and unstated losses in replicability.

SuSiE[174,177], published in 2020 and incorporated alongside FINEMAP in our work in Chapter 3, uses a similar overarching model to CAVIAR's, but importantly uses a different prior and exploration method than the CAVIAR family of tools. It enforces a prior that each GWAS region is composed of multiple independent signals. In this prior, each signal contains exactly one causal variant, though the prior allows for uncertainty as to what that causal variant is. SuSiE fits the distribution of uncertainty in each signal one signal at a time, fitting against the residual of the previously fit signals. In this way, each variant is assigned a chance of being the causal for each signal, though interpretable SuSiE signals generally only contain a few variants whose chance of being causal is non-negligible. Once all signals have been fit, SuSiE restarts the fitting procedure by dropping the fit of the first signal and refitting it against the residual of all the remaining signals, and proceeds to refit each signal in this manner, multiple times over, until the overall fit converges. While this is a stepwise method, SuSiE attempts to avoid the pitfalls of forward stepwise methods both by incorporating uncertainty and through mandatory reassessment of already-fit signals. Yet there is some marginal evidence that SuSiE may be slightly less precise than FINEMAP, possibly due to its stepwise approach[168].

The benefit SuSiE gains from its methodology is that it can estimate multiple credible sets independently, in that the probability of causality assigned to variants in one credible set is mostly independent from the choice of causal variants from other credible sets. In contrast, FINEMAP's credible sets are all reported with the assumption that every causal variant for each other signal in the region has been identified with certainty, and the only uncertainty is which variant is causal for the current signal. The clarity in its model has led SuSiE to be another commonly used statistical fine-mapping tool today alongside FINEMAP.

Validation of Statistical Fine-Mappers

Before I discuss other developments in statistical fine-mapping, it must be said that current statistical fine-mapping tools are largely unvalidated, and so each tool's claims of accuracy must be treated carefully. This lack of validation is because validating statistical fine-mapping algorithms is inherently difficult – they give inferences about the causality of variants in situations where such inferences cannot be readily confirmed by other means.

To skirt this issue, the majority of statistical fine-mapping developers use simulations as a means of quantifying their algorithms' efficacies[165,174–176,178,179]. These researchers attempt to mimic real conditions by building simulations off of variant dosages drawn from real genetic databases. They then decide which of those variants will be simulated as causal, providing ground truth data which fine-mapping results can be evaluated against. From there they simulate phenotype data from those causal variants, including a healthy dose of external noise in those simulations, run their fine-mapping tools on the phenotype and genotype data, and compare their tools' results to the ground truth. These comparisons are often used to show that a new statistical fine-mapper's credible sets are smaller than preceding algorithms', and that they contain the causal variant(s) more frequently.

However, many simplifying assumptions are made in these simulations. They universally assume that variant associations with the outcomes are truly linear and that there is no interaction between variants. These assumptions bias simulations to unknown extents, reducing their credibility as sources of validation, and fine-mapping methods papers rarely attempt to quantify the sensitivity of their methods to violations of these assumptions. (Note that I do not take issue with fine-mapping *models* making linearity assumptions – model misspecification may be acceptable if the model's output avoids large numbers of false-positives, but simulation misspecification is problematic because it is purporting to quantify the level of misspecification). Another difficult to justify assumption often made by statistical fine-mapping simulations is that no rare variants are causal[174,176,178].

Seemingly to provide orthogonal validation to simulations, many statistical fine-mapping publications run their algorithms genome-wide against real phenotype data and summarize how their algorithms behave[178,179], often pointing to their new tool's increased precision compared to competing methods. However, there are no external datasets to validate these genome-wide summaries against, and increased precision does not necessarily correlate with increased accuracy.

An alternative approach is to identify regions with some amount of experimental evidence suggesting which variants in the region are causal, and to validate statistical fine-mapping tools on those regions. Unfortunately, due to the difficulty in curating such data, this is much less common in the literature, and papers which perform this type of validation do so at small scale[165,176].

While statistical fine-mapping is widely used, our paper in Chapter 3 and a recent publication from the same month[168] both demonstrate that statistical fine-mappers are less reliable than they purport to be. To my knowledge, these are the first publications on that topic. I go into more detail on this in Chapter 3 and the overall thesis Discussion. A separate recent work has shown that statistical fine-mapping can be highly unreliable specifically when applied to summary statistics from meta-analyses of multiple GWAS[180]. In the Discussion I also describe opportunities for building a benchmark for statistical fine-mapping tools and for predicting scenarios where statistical fine-mapping is unreliable. Both projects could help allay the current lack of validation in the field of statistical fine-mapping.

Lastly, it should be stated that all the statistical fine-mapping tools described here always attempt to identify causal variants from among the tested variants. This approach cannot succeed if the causal variants have not been included in and tested by the GWAS providing the summary statistics, and statistical fine-mapping simulations rarely take this into account. The inclusion of complex variation in GWAS, include my work on STRs, attempts to address that problem.

Statistical Fine-Mapping with Varying Datasets and Data Types

Researchers attempting to fine-map GWAS results often bring in other sources of data, whether that data consists of genetic annotations or signals from other related phenotypes. Statistical fine-mapping algorithms have similarly developed to incorporate these sources of data to help deconvolute LD. In this section I describe those methods and their caveats. Unfortunately, these approaches suffer the same validation issues as the field of statistical fine-mapping as a whole. So while publications introducing these methods tend to claim their methods have greater power than methods which don't incorporate outside data sources, those claims should be weighed carefully.

Perhaps the most common extension of statistical fine-mapping is to incorporate genetic annotations, called *functionally informed* statistical fine-mapping. The intuition behind this is that variants overlapping annotations of known genetic functionality are more likely to be causal than the average variant which has no prior functional evidence, so statistical fine-mappers could combine the information from GWAS summary statistics and LD matrices with the information provided by annotations. Functionally informed statistical fine-mapping methods include SparsePro[181], CARMA[182], EMS[183], PolyFun[184], BFMAP[185] and fastPAINTOR[178,186,187]. The central challenge these methods all tackle differently is how they learn to weight information from different classes of annotations.

A drawback to functionally informed statistical fine-mapping is that the process is no longer hypothesis-free, and becomes biased towards identifying causal variants whose mechanisms are at least partially documented and biased against identifying variants whose mechanisms are unknown. This did not suit the purposes of my thesis, whose goal was to identify understudied causal non-coding STRs. Whether researchers choose to utilize annotations in statistical fine-mapping will depend on whether a hypothesis-free or hypothesis-driven search better fits their goals. Possibly a two-step approach, comparing annotation-free with annotation-driven statistical fine-mapping results, would allow researchers maximum insight into which data sources are driving their results at each locus. This is demonstrated by a type 2 diabetes GWAS[70], though that effort uses conditional regressions instead of more modern statistical fine-mapping approaches.

Multi-trait statistical fine-mapping, also called *multi-outcome* statistical fine-mapping or *colocalization*, attempts to run statistical fine-mapping on multiple traits simultaneously so as to jointly determine which variants are causal for which traits. Multi-trait statistical fine-mapping is desirable as it can identify whether distinct traits share etiology at a region or not (this goal is often called colocalization). Multi-trait statistical fine-mapping can also have greater power to fine-map a causal variant if the different phenotypes being jointly fine-mapped share a causal variant but have different sources of noise. Applications of multi-trait statistical fine-mapping include jointly analyzing traits with partially shared genetics (e.g. different types of irritable bowel disease), jointly analyzing the same trait measured in different settings (e.g. expression data for a gene measured in multiple different tissues), or jointly analyzing traits at different levels of granularity to build mechanistic hypotheses (e.g. colocalizing QTL data with GWAS of organism-level traits).

Many colocalization methods were developed in the early 2010s under the assumption of a single causal variant per locus. Even after method development efforts moved away from that assumption, many multi-trait fine-mappers were limited by the assumptions they made regarding the sharing of causal variants between traits. PAINTOR[178] required that all causal variants be shared between the traits, MFM[68], flashfm[179] and coloc with SuSiE[188] require the user to specify a prior likelihood of shared effects between traits, though such a choice is often difficult to motivate, and SuSiE^2 ("SuSiE squared", by different authors than SuSiE)[189] fine-maps two traits, assuming that the signal for the second trait is caused by the signal for the first trait (e.g. assuming a molecular QTL signal is the underlying mechanism for a organism-level trait signal). Most biological systems are complicated, making these sorts of assumptions and priors difficult to justify, and users of these methods should at least demonstrate that their conclusions are not overly sensitive to these assumptions. Alternatively, one limited set of scenarios where these methods may be of particular value is when it is known that the studied traits must share etiology and the goal of multi-trait fine-mapping is simply to use multiple datasets to improve statistical power to identify causal variants.

More recently, multi-trait statistical fine-mappers mvSuSiE[190] and CAFEH[191] have been developed to learn rates of causal sharing between traits from the data they are being trained on. I have limited exposure to these methods, but they look promising as attempts to move beyond limiting assumptions around causal variant sharing rates. I discuss the tie in between mvSuSiE and our work in the Chapter 1 Forward. Note that care must be taken to check whether multi-variate statistical fine-mapping methods require all traits to be measured on all individuals (e.g. mvSuSiE, possibly CAFEH) which precludes them from being used to jointly fine-map eQTL and GWAS signals, whether they require that traits be measured on separate cohorts (PAINTOR) or if they allow arbitrary sharing of individuals between cohorts (e.g. flashfm).

Lastly, there has been plenty of recent work on *multi-ethnic* statistical fine-mapping, also called *trans-ancestry* statistical fine-mapping, which is designed specifically to identify causal variants for traits studied in multiple distinct human populations. Due to differences in LD between populations, these efforts can have important gains in power over fine-mapping efforts performed in homogeneous populations. Unfortunately, a review of such methods is beyond the scope of this thesis.

In sum, GWAS is a hypothesis-free method for interrogating genetic contributions to human traits. Complex variants such as STRs have been omitted from most GWAS studies but causally effect many phenotypes. And statistical fine-mapping is a main tool by which causal variants with unknown mechanisms, such as STRs, can be identified at GWAS loci.

Contributed Research

Upcoming are Chapters 1-3 which contain full reprints of papers I have contributed to, in publication order, and which constitute the bulk of my doctoral work. In the first I contribute to an effort led by Stephanie Fotsing which provides evidence for the causal contribution of STRs towards gene expression levels, showing that common variation in STRs likely contributes to organism-level phenotypes. In the second Nima Mousavi and I coauthor a tool to ease the inclusion of STRs in bioinformatics analyses and pipelines. In the last, I lead the effort where we use extensive fine-mapping to suggest that common length variation of STRs across the genome is causally involved in a wide variety of blood traits and biomarkers in humans.

In each chapter I provide a forward. In the forwards I do not attempt to fully restate the results of each paper; the papers are apt records of their own results framed from the time points at which they were published. Rather I use the forwards to reflect on the works and attempt to place them within the context of my dissertation and the way this field has changed over time. I encourage the reader to read these forward sections in tandem with the abstracts, introductions and conclusions of the papers included.

Chapter 1: The Impact of Short Tandem Repeat Variation on Gene Expression

Forward to the Reprint

This chapter contains a full reprint of the paper *The Impact of Short Tandem Repeat Variation on Gene Expression* which was first authored by Stephanie Fotsing, to whom I was second author. In it we indicate that over a thousand STRs influence gene expression through the tools of association testing and statistical fine-mapping, using data from the Genotype Tissue Expression project (GTEx)[30]. We further demonstrate that many of these associations plausibly drive signals previously identified by GWAS which omitted STRs. (As an aside: in this paper we use the term fine-mapping to refer specifically to statistical fine-mapping).

This paper was conceived at a time when genetic data in large biobanks was still based on array data, not sequencing, and our lab had yet to complete its first set of analyses demonstrating that an STR reference panel could be used to accurately impute STR genotypes into array data[83]. Rather, GTEx was a relatively new resource that provided a valuable source of sequencing data in which STRs could be genotyped and a large source of gene expression data against which STR hypotheses could be tested. If performing large-scale STR GWAS was currently out of reach, then showing that STRs were causal for changes in gene expression, and thus would be likely to influence the traits those genes were causal for, was an important steppingstone towards that goal.

This is remarkably distinct from the current research landscape. GTEx is likely still the largest research-accessible biobank of a wide variety of healthy human tissues from a range of individuals that have already been assayed for gene expressions. But GTEx is limited in the tissues it assays, by the small numbers of individuals it assays, its lack of tissues sampled during exposure to important environmental conditions and its lack of diseased tissues. Further, GTEx only has tissues from adults. Sampling tissues from children at differing developmental time points is necessary to be able to study effects which may potentially be only visible during development, though collecting such tissues is clearly very challenging and is the subject of an ongoing effort[192]. Thus, though it is expected that a majority of causal non-coding variant effects are mediated through gene expression, it can be expected that GTEx will only identify some of those effects. Further, this suggests that not finding expression modulation evidence in GTEx is not sufficient to refute expression modulation as a mechanistic hypothesis.

Further, since we began this project, population-level biobanks have become huge[53] relative to the size of GTEx. We have also developed well tested reference panels from which tandem repeats can be imputed[83,84] into array data in those biobanks, to say nothing of biobanks which already have short tandem repeats called from whole genome sequencing[133]. Thus I expect GWAS of organism-level traits to be relatively more conclusive than eQTL analyses for the foreseeable future.

Nonetheless, this paper fundamentally succeeded. It is one of the earliest efforts which developed causal evidence of the effects of STR lengths on gene expression across the genome. It laid a roadmap for connecting such links to GWAS hits. And this paper provided a list of putatively causal STRs to be further studied for mechanistic insights.

Having identified STRs statistically fine-mapped to impact gene expression, this paper attempted to identify trends among those STRs. It most successfully showed that CG-rich repeats in 5’ UTR and promoter regions are likely to influence expression through stabilizing non-canonical DNA secondary structures. It inferred a few other trends statistically, such as nucleosome positioning signals and strand biases in AT repeats. However, this paper could not leverage those trends to infer how changes in the lengths of individual STRs might mechanistically impact gene expressions. This challenge is a fundamental limitation of all papers studying non-coding variation across the genome, and is magnified by the many distinct mechanisms an STR, or any other non-coding variant, may act by.

I joined this project after it had already been conceived and drafted by Stephanie, Melissa, Alon and the other co-authors. I thoroughly updated the paper with Melissa and Alon in response to reviewer comments and performed the mash[193] analysis to improve our cross-tissue analyses. I found the redrafting process to be a wonderful introduction to these research areas as it required me to understand the totality of the paper at a detailed and authoritative level, and I thank Melissa and Alon for introducing me to the field in this way.

While our initial paper draft identified a lack of shared STR expression effects across tissues, the mash analysis led us to reevaluate those results. Specifically, we concluded that STR expression effects are in fact commonly shared across tissue-clusters (Chapter 1 Figure 1d, Extended Data Figure 4, and Supplementary Figures 12 and 13), and that the lack of sharing noted in the first draft of the paper was likely due to small sample sizes in each individual tissue leading to large false-negative rates, and not lack of shared biology.

While the reexamination of effect sizes through mash was a success, we did not incorporate those results into our statistical fine-mapping analysis, instead running that on the per-tissue effect sizes. This meant that statistical fine-mapping results could not take advantage of the increased power from mash-derived effect sizes, and we likely fine-mapped fewer expression-associated STRs because of that.

Our subsequent paper on STR causality in blood traits (Chapter 3) again ran statistical fine-mapping multiple times when the goal was to identify a single causal STR. There, the different runs were not in different tissues but in highly related traits (e.g. red blood cell count and percentage of red blood cells among all blood cells). The dataset used in Chapter 3 was sufficiently large that the power concerns of the Chapter 1 analyses no longer applied. However, in Chapter 3, fine-mapping results that differed between very similar traits were difficult to interpret. For instance, if an STR causally increases red blood cell count, then it should also causally increase red blood cell percentage. So it is difficult to reconcile cases where fine-mapping marks it as causal for red blood cell count but not for red blood cell percentage.

The solution to both issues, the desire for increased power by sharing data, either sharing eQTL data across tissue or GWAS signal data across traits, and the desire for a consistent fine-mapping result across those tissues/traits, is multi-trait statistical fine-mapping. However, as discussed in the fine-mapping section of the thesis Introduction, up until recently all multi-trait fine-mapping methods have required priors for how often causal variants are shared across tissues/traits, and the choice of these priors can be very difficult to justify. Thus it is heartening to see this corner of GWAS analysis come full circle with mvSuSiE[194], which incorporates mash output as a prior for input to SuSiE, giving a principled approach to setting of priors to multi-trait statistical fine-mapping and potentially mitigating these issues for future GWAS and eQTL studies.

ANALYSIS
https://doi.org/10.1038/s41588-019-0521-9
nature genetics

The impact of short tandem repeat variation on gene expression

Stephanie Feupe Fotsing[1,2,6], Jonathan Margoliash[3,4], Catherine Wang[5], Shubham Saini[3], Richard Yanicky[4], Sharona Shleizer-Burko[4], Alon Goren[4*] and Melissa Gymrek[3,4*]

Short tandem repeats (STRs) have been implicated in a variety of complex traits in humans. However, genome-wide studies of the effects of STRs on gene expression thus far have had limited power to detect associations and provide insights into putative mechanisms. Here, we leverage whole-genome sequencing and expression data for 17 tissues from the Genotype–Tissue Expression Project to identify more than 28,000 STRs for which repeat number is associated with expression of nearby genes (eSTRs). We use fine-mapping to quantify the probability that each eSTR is causal and characterize the top 1,400 fine-mapped eSTRs. We identify hundreds of eSTRs linked with published genome-wide association study signals and implicate specific eSTRs in complex traits, including height, schizophrenia, inflammatory bowel disease and intelligence. Overall, our results support the hypothesis that eSTRs contribute to a range of human phenotypes, and our data should serve as a valuable resource for future studies of complex traits.

Expression quantitative trait loci (eQTL) studies attempt to link genetic variation to gene expression changes as potential molecular intermediates that drive disease and variation in complex traits. Recent studies have identified tens of thousands of eQTLs (genetic variants associated with expression of nearby genes) across multiple human tissue types[1,2]. Most of these have focused on biallelic SNPs or short indels. Yet multiple studies dissecting genome-wide association study (GWAS) loci have found repetitive[3,4] and structural variants[5–7] to be the underlying causal variants, highlighting the need to consider additional variant classes beyond SNPs.

Short tandem repeats, consisting of consecutively repeated units of 1–6 base pairs (bp), represent a large source of genetic variation. STR mutation rates are orders of magnitude higher than those of SNPs[8] and short indels[9], and each individual is estimated to harbor around 100 de novo mutations in STRs[10]. Expansions at several dozen STRs have been known for decades to cause mendelian disorders[11], including Huntington's disease and hereditary ataxias. Importantly, these pathogenic STRs represent a small minority of the more than 1.5 million STRs in the human genome[12]. Due to bioinformatics challenges of analyzing repetitive regions, many STRs are often filtered from genome-wide studies[13]. However, increasing evidence supports a widespread role of common variation at STRs in complex traits, such as gene expression[14–17].

STRs may regulate gene expression through a variety of mechanisms[18]. For example, the CCG repeat implicated in fragile X syndrome was shown to disrupt DNA methylation, altering expression of *FMR1* (ref. [19]). Yeast studies have demonstrated that homopolymer repeats act as nucleosome positioning signals with downstream regulatory effects[20,21]. Dinucleotide repeats may alter affinity of nearby DNA-binding sites[22]. Furthermore, certain STR repeat units may form noncanonical DNA and RNA secondary structures such as G-quadruplexes[23], R-loops[24] and Z-DNA[25].

We previously identified more than 2,000 STRs for which the number of repeats was associated with the expression of nearby genes[14], termed expression STRs (eSTRs). However, the quality of the datasets available for that study reduced our power to detect associations and prevented accurate fine-mapping of individual signals. STR genotypes were based on low coverage (4–6×) whole-genome sequencing data performed using short reads (50–100 bp), which are unable to span many STRs. As a result, STR genotype calls exhibited poor quality with less than 50% genotyping accuracy[12]. Additionally, the study used a single cell type (lymphoblastoid cell lines) with potentially limited relevance to most complex traits[26]. While our study and others[14,16] demonstrated that eSTRs explain a sizable portion (10–15%) of the *cis* heritability of gene expression, the resulting eSTR catalogs were not powered to robustly implicate eSTRs over other nearby variants.

Here, we leverage deep whole-genome sequencing (WGS) and gene expression data collected by the Genotype–Tissue Expression Project (GTEx)[1] to identify more than 28,000 eSTRs in 17 tissues. We employ fine-mapping to quantify the probability of causality of each eSTR and characterize the top 1,400 (top 5%) fine-mapped eSTRs. We additionally identify hundreds of eSTRs that are in strong linkage disequilibrium (LD) with published GWAS signals and implicate specific eSTRs in height, schizophrenia, inflammatory bowel disease and intelligence. To further validate our findings, we demonstrate evidence of a causal link between height and an eSTR for the gene *RFT1* and use a reporter assay to experimentally validate an effect of this STR on expression. Finally, our eSTR catalog is publicly available as a resource for future studies of complex traits.

Results

Profiling expression STRs across 17 human tissues. We performed a genome-wide analysis to identify associations between the number of repeats at each STR and expression of nearby genes (expression STRs, or 'eSTRs', which we use to refer to a unique STR by gene association). We focused on 652 individuals from the GTEx[1] dataset for which both high-coverage WGS and

[1]Biomedical Informatics and Systems Biology, University of California San Diego, La Jolla, CA, USA. [2]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA. [3]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [4]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [5]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [6]Present address: La Jolla Institute of Immunology, La Jolla, CA, USA. *e-mail: agoren@ucsd.edu; mgymrek@ucsd.edu

RNA-sequencing of multiple tissues were available (Fig. 1a). We used HipSTR[27] to genotype STRs in each sample. After filtering low quality calls (Methods), 175,226 STRs remained for downstream analysis. To identify eSTRs, for each gene and for each STR within 100 kilobases (kb) of that gene, we performed a linear regression between the average length of the STR in each person and normalized expression of the gene, controlling for sex, population structure and technical covariates (Methods and Supplementary Figs. 1–3). Analysis was restricted to 17 tissues where we had data for at least 100 samples (Supplementary Table 1 and Methods) and to genes with median reads per kilobase of transcript, per million mapped reads (RPKM) greater than 0. Altogether, we performed an average of 262,593 STR–gene tests across 15,840 protein-coding genes per tissue.

Using this approach, we identified 28,375 unique eSTRs associated with 12,494 genes in at least one tissue at a gene-level false discovery rate (FDR) of 10% (Fig. 1b, Supplementary Table 1 and Supplementary Data 1). The number of eSTRs detected per tissue correlated with sample size as expected (Pearson $r = 0.75$; $P = 0.00059$; $n = 17$), with the smallest number of eSTRs detected in the two brain tissues, presumably due to their low sample sizes (Extended Data Fig. 1). eSTR effect sizes previously measured in lymphoblastoid cell lines were significantly correlated with effect sizes in all GTEx tissues ($P < 0.01$ for all tissues, mean Pearson $r = 0.45$). We additionally examined previously reported eSTRs[28–35] that were mostly identified using in vitro constructs. Six of eight examples were significant eSTRs in GTEx ($P < 0.01$) in at least one tissue analyzed (Supplementary Table 2).

eSTRs identified above could potentially be explained by their tagging nearby causal variants. To prioritize potentially causal eSTRs we employed CAVIAR[36], a statistical fine-mapping framework. CAVIAR models the relationship between LD structure and association statistics of local variants to quantify the posterior probability of causality for each variant (which we refer to as the CAVIAR score). We used CAVIAR to fine-map eSTRs against all SNPs nominally associated ($P < 0.05$) with each gene under our model (Methods and Fig. 1a). On average across tissues, 12.2% of eSTRs had the highest causality scores of all variants tested.

We ranked eSTRs by their best CAVIAR score across tissues and chose the top 5% for downstream analysis (1,420 unique eSTRs with best CAVIAR score >0.3). We hereby refer to these as fine-mapped eSTRs (FM-eSTRs) (Supplementary Table 1 and Supplementary Data 2). Expected gene annotations are more strongly enriched in this subset compared to the entire set (Extended Data Fig. 2), and stricter thresholds reduced the power to detect eSTR-enriched features described below. Of the FM-eSTRs in each tissue, on average 78% explained gene expression variation beyond that explained by the best SNP (ANOVA $q < 0.1$). Furthermore, on average, each FM-eSTR had a CAVIAR score 0.41 higher (41% higher posterior probability) than the top-scoring SNP (Supplementary Fig. 4). Multiple STRs with known disease implications[35,37–40] were captured by this list (Fig. 1c). In many cases, FM-eSTRs show clear relationships between the number of repeats and gene expression across a wide range of repeat lengths (Extended Data Fig. 3).

To minimize power differences across tissues and enable cross-tissue comparisons of eSTR effects, we applied multivariate adaptive shrinkage (mash[41]) (Fig. 1a). Mash takes the per-tissue effect sizes and standard errors computed above as input and recomputes posterior estimates for each, while considering cross-tissue effect-size correlations. We compared FM-eSTR mash effect sizes across all pairs of tissues (Fig. 1d) and recovered previously observed relationships[41]. Tissues with similar origins (for example, adipose-visceral/adipose-subcutaneous) are highly concordant, whereas whole blood effects are less correlated with other tissues. These tissue sharing patterns are similar to those obtained using unadjusted effect sizes of single-tissue eSTRs (Supplementary Fig. 5). We further examined tissue sharing of FM-eSTRs by counting, for each FM-eSTR, the number of tissues for which mash computed a posterior Z-score with an absolute value >4. Most eSTRs are either shared across all tissues analyzed or are shared by only a small number of tissues (Extended Data Fig. 4), again similar to previously reported SNP analyses in this cohort[1].

FM-eSTRs demonstrate unique genomic characteristics. We next sought to characterize properties of STRs that might provide insights into their biological function. We reasoned that genomic characteristics that distinguish FM-eSTRs from all analyzed STRs would support the hypothesis that a subset of them are acting as causal variants. While results below are presented for FM-eSTRs as defined above (CAVIAR score >0.3), we also provide results recomputed using a range of score thresholds in the Supplementary Information. These results show that the major characteristics of FM-eSTRs identified below are robust to the precise threshold used.

We first considered whether the localization of FM-eSTRs differs from that of STRs overall (Fig. 2a,b and Extended Data Fig. 5). Overall, the majority of FM-eSTRs occur in intronic or intergenic regions, and only 11 FM-eSTRs fall in coding exons (Supplementary Table 3). However, compared to all STRs, those closest to transcription start sites and near DNase I hypersensitive (HS) sites are more likely to be FM-eSTRs (Fig. 2c,d and Extended Data Fig. 6). FM-eSTRs are strongly enriched at 5′ UTRs (odds ratio (OR) = 5.0; Fisher's two-sided $P = 4.9 \times 10^{-13}$), 3′ UTRs (OR = 2.78; $P = 5.85 \times 10^{-10}$) and within 3 kb of transcription start sites (OR = 3.39; $P = 3.94 \times 10^{-70}$). These enrichments are considerably stronger for FM-eSTRs compared to all eSTRs (Supplementary Table 4), suggesting, as expected, that FM-eSTRs are more likely to be causal.

We next examined nucleosome occupancy in the lymphoblastoid cell line GM12878 and DNA accessibility (measured by DNase-seq) in a variety of cell and tissue types within 500 bp of FM-eSTRs (Extended Data Fig. 7). As expected from previous studies[42], regions near homopolymer repeats are strongly nucleosome depleted. STRs with other repeat lengths also show distinct patterns of nucleosome positioning (Extended Data Fig. 7a–c). Nucleosome occupancy is broadly similar for FM-eSTRs compared to all STRs. Yet FM-eSTRs are generally located in regions with higher DNase-seq read count compared to non-eSTRs (Mann–Whitney U-test two-sided $P = 3.9 \times 10^{-37}$ in GM12878; Extended Data Fig. 7d–f). DNase I HS signal around homopolymer FM-eSTRs shows a periodic pattern in multiple cell and tissue types, with peaks located at multiples of 147 bp upstream and downstream from the STR (Extended Data Fig. 7d). Given that 147 bp is the length of DNA typically wrapped around a single nucleosome[42], we hypothesize that a subset of homopolymer FM-eSTRs may act by shifting nucleosome positions and thus modulating the accessibility of adjacent sites.

Next, we compared the sequence characteristics of FM-eSTRs with all STRs. We find that the total lengths of FM-eSTRs are significantly higher (Mann–Whitney U-test two-sided $P = 0.00032$ and $P = 2.4 \times 10^{-10}$ when comparing total repeat number and total length in bp, respectively, based on the sequence present in hg19). We tested FM-eSTRs combined across all tissues for enrichment of each canonical STR repeat unit (defined lexicographically, see Methods) and found that FM-eSTRs are most strongly enriched for repeats with GC-rich repeat units (Fig. 2e, Supplementary Table 5 and Supplementary Fig. 6). For example, the canonical repeat units CCCCGG, CCCCCG and CCG are 22-, 13- and 7-fold enriched in FM-eSTRs compared to all STRs, respectively. During transcription, these GC-rich repeat units have been shown to form highly stable secondary structures, such as G4 quadruplexes in single-stranded DNA[43] or RNA[44], that may be involved in regulation of gene expression. We found that, in general, higher repeat numbers at GC-rich eSTRs are associated with greater DNA or RNA stability

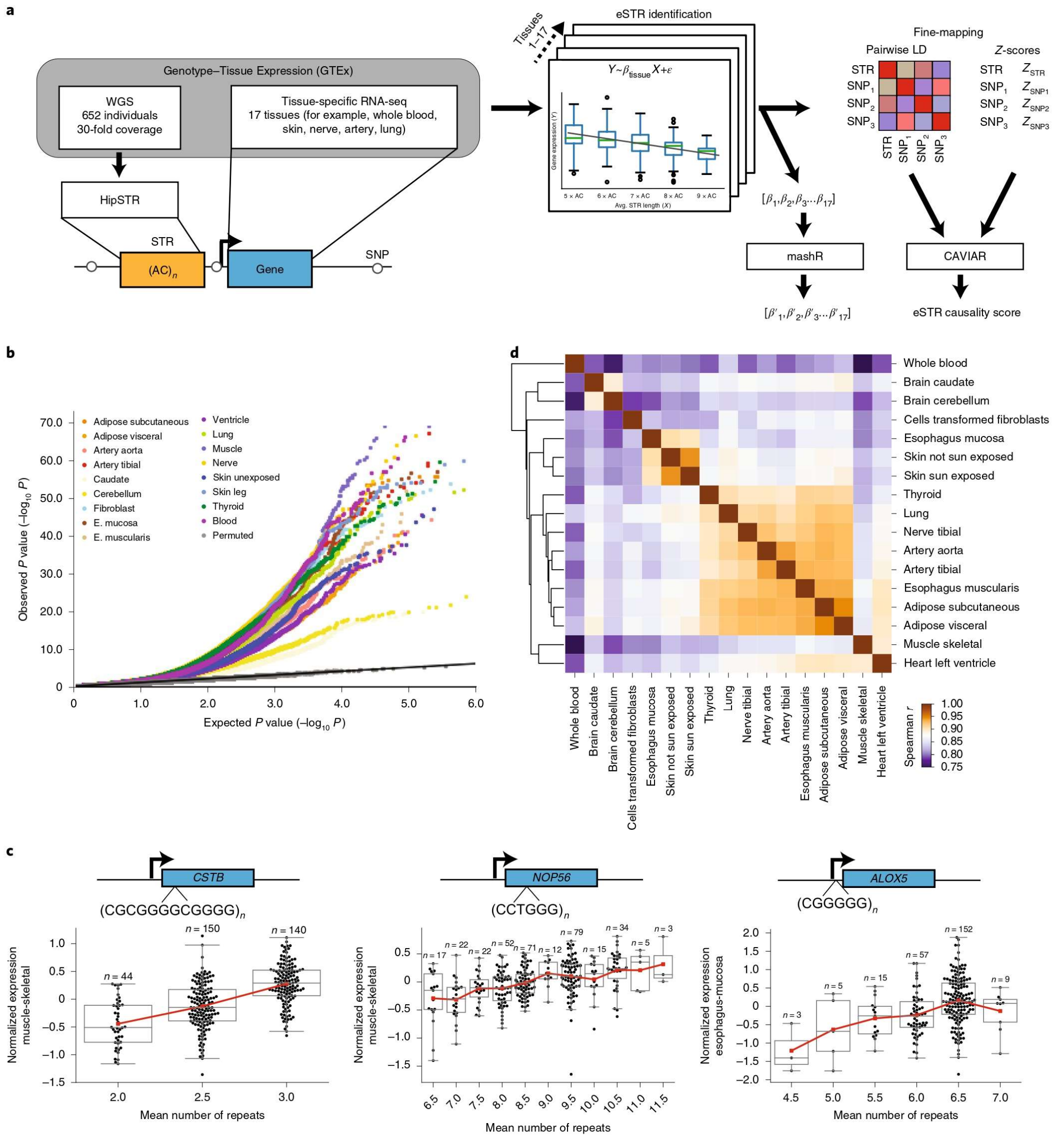


Fig. 1 | Multitissue identification of eSTRs. a, Schematic of eSTR discovery pipeline. We analyzed eSTRs using RNA-seq from 17 tissues and STR genotypes obtained from deep WGS for 652 individuals from the GTEx Project. **b**, eSTR association results. The quantile–quantile plot compares observed *P* values for each STR by gene test versus the expected uniform distribution for each tissue. Gray dots denote permutation controls (*n* = 336). Supplementary Table 1 gives the number of tests performed in each tissue. **c**, Example eSTRs previously implicated in disease. Example FM-eSTRs previously implicated in myoclonus epilepsy (left), spinocerebellar ataxia 36 (middle) and reduced lung function and cardiovascular disease (right) are shown. Black points represent single individuals. For each plot, the *x* axis represents the mean number of repeats in each individual and the *y* axis represents normalized expression in a representative tissue. Box plots summarize the distribution of expression values. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1 − 1.5 × IQR (bottom) and Q3 + 1.5 × IQR (top), where IQR is the interquartile range (Q3–Q1). The red line shows the mean expression for each *x* axis value. Gene diagrams not drawn to scale. **d**, eSTR correlations across tissues. Each cell shows the Spearman correlation between mashR FM-eSTR effect sizes for each pair of tissues. Only eSTRs with CAVIAR score >0.3 (FM-eSTRs) in at least one of the two tissues were included in each correlation. Supplementary Table 1 gives the number of FM-eSTRs identified in each tissue. Rows and columns were clustered using hierarchical clustering (Methods).

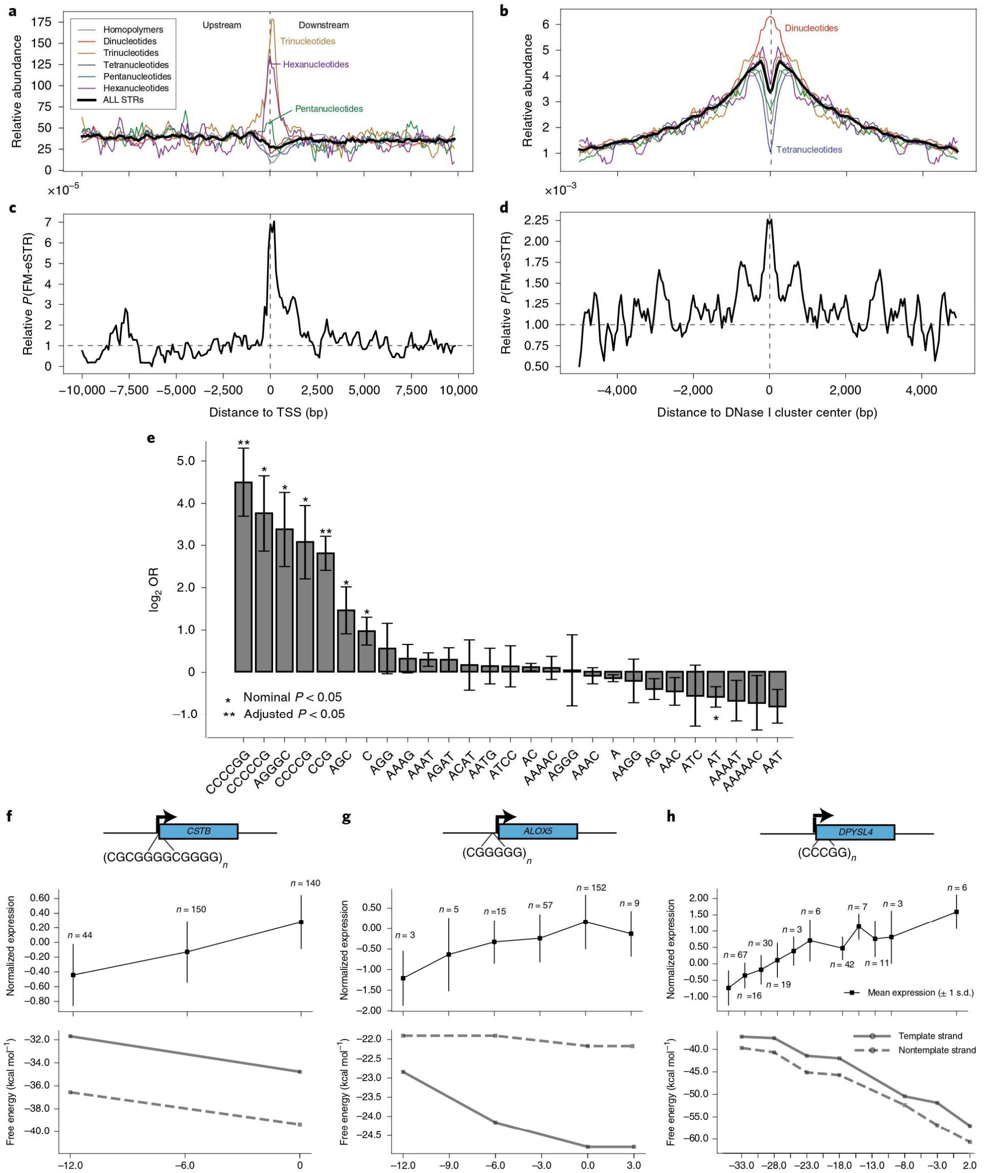


Fig. 2 | Characterization of FM-eSTRs. **a**, Density of all STRs around transcription start sites (TSS). The *y* axis shows the fraction of STRs with each repeat unit type located in each 100-bp bin around the TSS. **b**, Density of all STRs around DNase I HS sites. Plots are centered at ENCODE DNase I HS clusters and represent the fraction of STRs with each repeat unit type located in each 50-bp bin. **c**, Relative probability to be an FM-eSTR around TSSs. **d**, Relative probability to be an FM-eSTR around DNase I HS clusters. Values were smoothed using a sliding average of each four consecutive bins (**a**–**d**). **e**, Repeat unit enrichment at FM-eSTRs. The *x* axis shows all repeat units for which there are at least three FM-eSTRs across all tissues. The *y* axis center values denote the $\log_2$ OR comparing FM-eSTRs to all STRs. Error bars represent ±1 s.e.m. Asterisks denote repeat units that are significantly enriched or depleted in FM-eSTRs (based on two-sided Fisher's exact *P* value). Per repeat unit sample sizes and Fisher exact statistics are provided in Supplementary Table 5. **f**–**h**, Example GC-rich FM-eSTRs in promoters predicted to modulate secondary structure: example eSTRs are shown from skeletal muscle (**f**); esophagus mucosa (**g**) and transformed fibroblasts (**h**). Top plots show mean expression across all individuals with each mean STR length. Vertical bars represent ±1 s.d. Bottom plots show the free energy computed for each allele based on template (solid) and nontemplate (dashed) strands. The *x* axis shows STR lengths relative to hg19 (bp). Gene diagrams are not drawn to scale.

and increased expression of nearby genes (Supplementary Note, Fig. 2f–h and Supplementary Figs. 7–10).

We next examined effect-size biases in FM-eSTR associations. Overall, FM-eSTRs are equally likely to show positive versus negative correlations between repeat length and gene expression (Supplementary Fig. 11; binomial two-sided $P=0.94$). We additionally observe that FM-eSTRs with repeat units of the form (A_nC/G_nT) show strand-specific effects when in or near transcribed regions. Transcribed FM-eSTRs are more likely to have the T-rich version of the repeat unit on the template strand (binomial two-sided $P=0.0015$). These T-rich FM-eSTRs tend to have more positive effect sizes, with the most notable differences for AC versus GT repeats. These patterns are observed in transcribed regions across multiple distinct repeat types (A/T, AC/GT, AAC/GTT, AAAC/GGGT) but are not present in intergenic regions (Extended Data Fig. 8).

Finally, we wondered whether eSTRs might exhibit distinct characteristics in different tissues. We clustered tissue-specific Z-scores (absolute value) for each FM-eSTR calculated jointly across tissues by mash (Methods), to identify eight categories of FM-eSTRs (Supplementary Figs. 12 and 13). These include two clusters of FM-eSTRs present across many tissues (clusters 2 and 8) as well as several more tissue-specific clusters (for example, thyroid for cluster 1). Notably, clusters do not necessarily imply tissue specificity, but rather enrich for FM-eSTRs with particularly strong effects in one or more tissues (Supplementary Fig. 13). Clusters show similar repeat unit enrichment to all FM-eSTRs and do not exhibit distinct enriched repeat units (Supplementary Fig. 14). Similar results were achieved using different numbers of clusters. Overall, our results suggest that the majority of eSTRs act by global mechanisms and do not implicate tissue-specific characteristics of FM-eSTRs. However, low numbers of tissue-specific effects limit the power to detect differences.

eSTRs are potential drivers of published GWAS signals. We wondered whether our eSTR catalog could identify STRs affecting complex traits in humans. We first leveraged the NHGRI/EBI GWAS catalog[45] to identify FM-eSTRs that are nearby and in LD with published GWAS signals. Overall, 1,380 unique FM-eSTRs are within 1 megabase (Mb) of GWAS hits (Methods and Supplementary Data 3). Of these, 847 are in moderate LD ($r^2>0.1$) and 65 are in strong LD ($r^2>0.8$) with the lead SNP. When considering a more stringent set of FM-eSTRs, with a CAVIAR score >0.5, 403 and 26 are in moderate and strong LD with a GWAS hit, respectively.

We next sought to determine whether specific published GWAS signals could be driven by changes in expression due to an underlying but previously unobserved FM-eSTR. We reasoned that such loci would exhibit the following properties: (1) strong similarity in association statistics across variants for both the GWAS trait and expression of a particular gene, indicating the signals may be colocalized, that is driven, by the same causal variant; and (2) strong evidence that the FM-eSTR causes variation in expression of that gene (Fig. 3a). Colocalization analysis requires high-resolution summary statistic data. Thus, we focused on several example complex traits (height[46], schizophrenia[47], inflammatory bowel disease (IBD)[48] and intelligence[49]) for which detailed summary statistics computed on cohorts of tens of thousands of individuals, or more, are publicly available (Methods).

For each trait, we identified FM-eSTRs within 1 Mb of published GWAS signals from Supplementary Data 3. We then used coloc[50] to compute the probability that the FM-eSTR signals we derived from GTEx and the GWAS signals derived from other cohorts are colocalized. The coloc tool compares association statistics at each SNP in a region for expression and the trait of interest and returns a posterior probability that the signals are colocalized. We used coloc to test a total of 276 gene × trait pairs (138, 45, 29 and 64 for height, intelligence, IBD and schizophrenia, respectively). In total, we identified 62 GWAS loci with (1) an FM-eSTR in at least moderate LD ($r^2>0.1$) with a nearby SNP for that trait in the GWAS catalog, and (2) colocalization posterior probability between the target gene and the trait >50%, meaning colocalization of the eQTL and GWAS signals is the most probable model (Extended Data Figs. 9 and 10). Out of the 62 FM-eSTRs colocalized with GWAS signals, 40 have CAVIAR scores >0.5. Results of all colocalization tests are provided in Supplementary Table 6.

A top example is an FM-eSTR for *RFT1*, a gene encoding an enzyme involved in the N-glycosylation of proteins[51], which has 97.8% colocalization probability with a GWAS signal for height (Fig. 3b,c). The lead SNP in the NHGRI catalog (rs2336725:C>T) is in high LD ($r^2=0.85$) with an AC repeat that is a significant eSTR in 15 tissues. This STR falls in a cluster of transcription factor and chromatin regulator binding regions identified by ENCODE near the 3′ end of the gene (Fig. 3d) and exhibits a positive correlation with expression.

To more directly test for association between this FM-eSTR and height, we used our recently developed STR–SNP reference haplotype panel[52] to impute STR genotypes into available GWAS data. We focused on the eMERGE cohort (Methods) for which imputed genotype array data and height measurements are available. We tested for association between height and SNPs, as well as for height and AC repeat number, after excluding samples with low STR imputation quality (Methods). Imputed AC repeat number is significantly associated with height in the eMERGE cohort ($P=0.00328$; $\beta=0.010$; $n=6{,}393$, where β is the effect size), although with a slightly weaker P value compared to the top SNP (Fig. 3e). Notably, even in the case where the STR is the causal variant, power is likely to be reduced due to the lower quality of the imputed STR genotypes. Notably, AC repeat number shows a strong positive relationship with height across a range of repeat lengths (Fig. 3f), similar to the relationship between repeat number and *RFT1* expression.

To further investigate whether the FM-eSTR for *RFT1* could be a causal driver of gene expression variation, we devised a dual reporter assay in HEK293T cells to test for an effect of the number of repeats on gene expression (0, 5, 10 or 12 repeats, plus approximately 170 bp of genomic sequence context on either side) (Supplementary Table 7 and Methods). We observed a positive linear relationship between the number of AC repeats and reporter expression, as predicted (Fig. 3g) (Pearson $r=0.97$; $P=0.013$). Furthermore, all pairs of constructs with consecutive repeat numbers showed significantly different expression (one-sided t-test $P<0.01$) with the exception of 10 versus 12 repeats. Overall, these results further support the hypothesis that eSTRs may act as causal drivers of gene expression.

Discussion

Here we present the most comprehensive resource of eSTRs to date, which reveals more than 28,000 associations between the number of repeats at STRs and expression of nearby genes across 17 tissues. We performed fine-mapping to quantify the probability that each eSTR causally effects gene expression and characterized the top fine-mapped eSTRs. The eSTRs analyzed here consist of a large spectrum of repeat classes with a variety of repeat unit lengths and sequences. Based on the diverse characteristics of eSTRs, we hypothesize that different repeat classes work by distinct regulatory effects (Fig. 4). While we explored several potential mechanisms, including nucleosome positioning and the formation of noncanonical DNA or RNA secondary structures, our results do not rule out other potential mechanisms.

We leveraged our resource to provide evidence that FM-eSTRs may drive a subset of published GWAS associations for a variety of complex traits. STRs have a unique ability, compared with biallelic SNPs, to drive phenotypic variation along a spectrum of multiple alleles. In multiple examples, eSTRs show a linear trend between

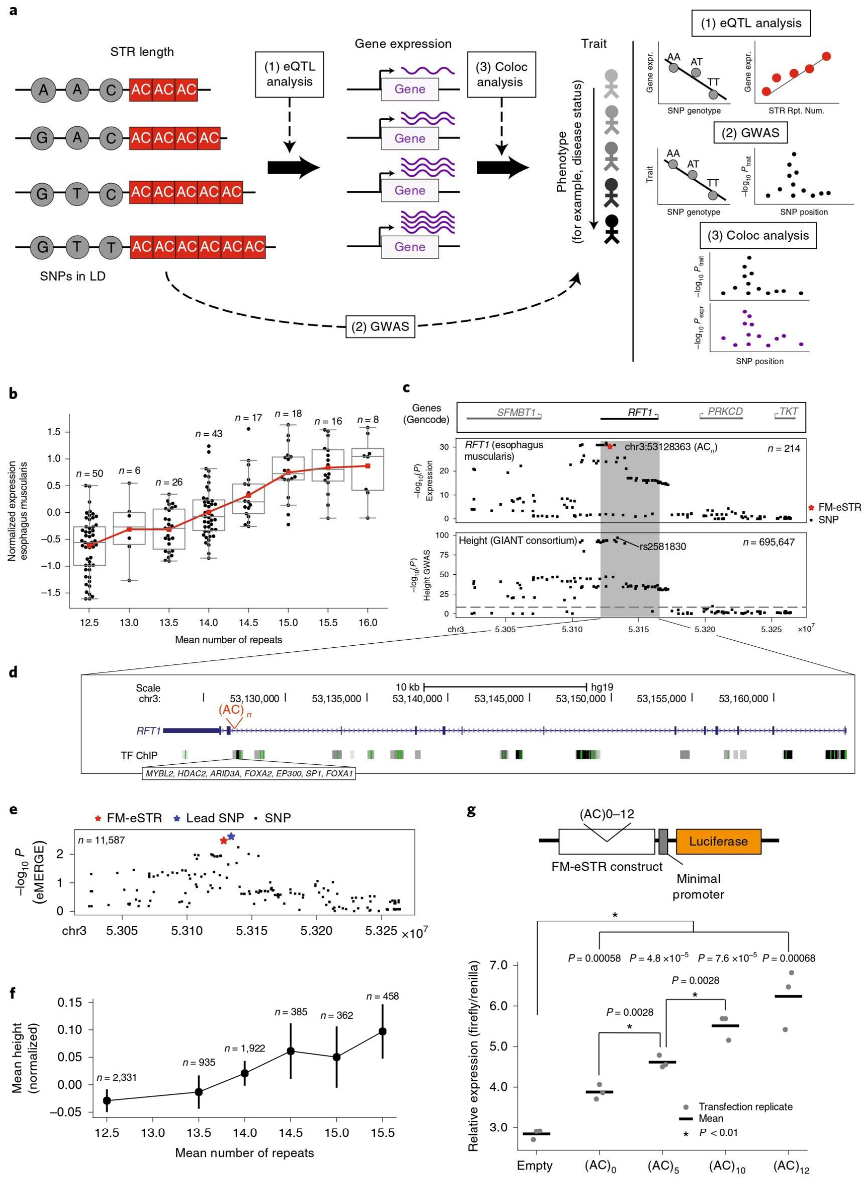


Fig. 3 | FM-eSTRs colocalize with GWAS signals. a, Overview of analyses to identify FM-eSTRs involved in complex traits. We assumed a model where variation in STR repeat number alters gene expression, which in turn affects the value of a particular complex trait. **b**, eSTR association for *RFT1*. The *x* axis shows STR genotype as the mean number of AC repeats and the *y* axis gives normalized *RFT1* expression, defined as in Fig. 1c. **c**, Summary statistics for *RFT1* expression and height. The $-\log_{10}$ *P* values of association between each variant and *RFT1* expression are shown in the middle panel. The $-\log_{10}$ *P* values of association for each variant with height are shown in the bottom panel. Black dots, SNPs; red star, FM-eSTR; gray dashed line, genome-wide significance threshold. **d**, Genomic view of the *RFT1* locus. **e**, eSTR and SNP associations with height in the eMERGE cohort. The *y* axis denotes association *P* values for each variant. Black dots, SNPs; red star, imputed FM-eSTR; blue star, top eMERGE SNP. **f**, Imputed *RFT1* repeat number is correlated with height. The *x* axis shows the mean number of AC repeats. The *y* axis shows the mean normalized height for all samples included in the analysis with a given genotype. Error bars show ±1 s.e.m. **g**, Reporter assay testing repeat number versus expression. A variable number of AC repeats plus genomic context were introduced upstream of a reporter gene. Gray dots show the value for each of $n = 3$ transfections, each averaged across three technical replicates. Black lines show the mean across the three transfections.

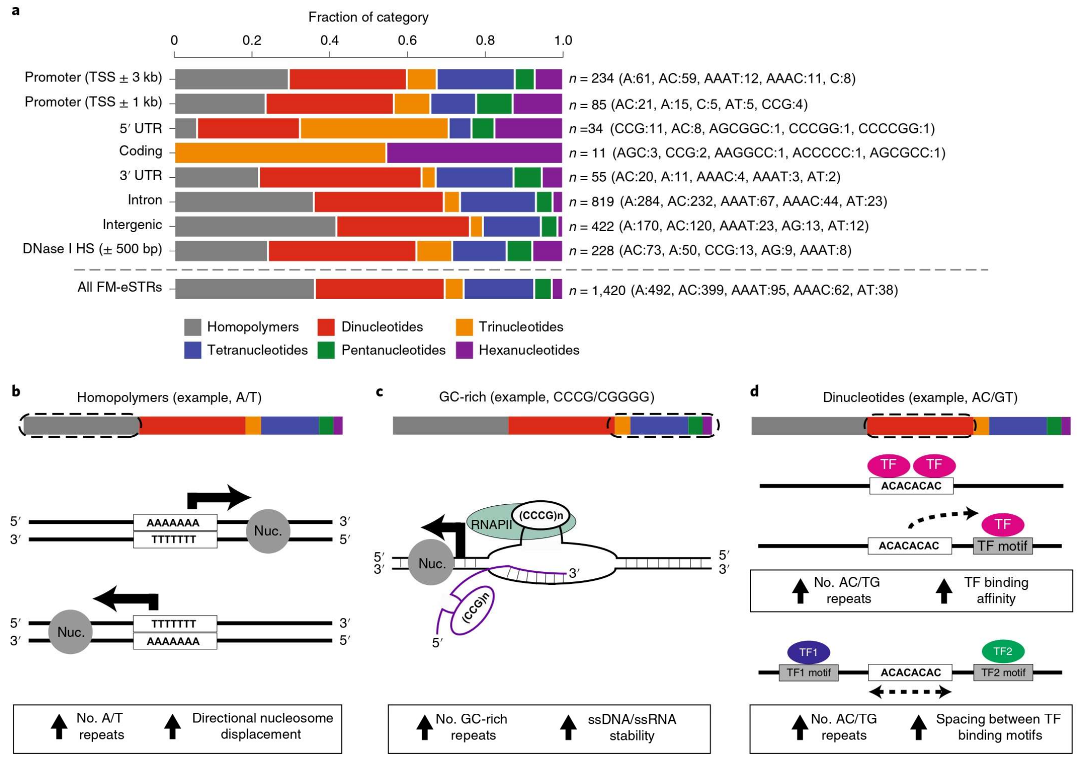


Fig. 4 | Summary of FM-eSTRs classes and potential regulatory mechanisms. a, Distribution of FM-eSTR classes across genomic annotations. Each bar shows the fraction of FM-eSTRs falling in each annotation consisting of homopolymer (gray), dinucleotide (red), trinucleotide (orange), tetranucleotide (blue), pentanucleotide (green) or hexanucleotide (purple) repeats. The total number of FM-eSTRs and the top five most common repeat units in each category are shown on the right. Of note, FM-eSTRs may be counted in more than one category. **b**, Homopolymer A/T STRs are predicted to modulate nucleosome positioning. Homopolymer repeats are depleted of nucleosomes (Nuc., gray circles) and may modulate expression changes in nearby genes through altering nucleosome positioning. **c**, GC-rich STRs form DNA and RNA secondary structures during transcription. Highly stable secondary structures such as G4 quadruplexes may act by expelling nucleosomes (gray circle) or stabilizing RNAPII (light green circle). These structures may form in DNA (black) or RNA (purple). The stability of the structure can depend on the number of repeats. **d**, Dinucleotide STRs can alter transcription factor (TF) binding. Dinucleotides are prevalent in putative enhancer regions. They may potentially alter transcription factor binding by forming binding sites themselves (top), changing affinity of nearby binding sites (middle) or modulating spacing between nearby binding sites (bottom). Text and arrows in the white boxes provide a summary of the predicted eSTR mechanism depicted in each panel (**b**–**d**).

repeat length and expression across a range of repeat numbers, a signal that cannot be easily explained by tagging nearby biallelic variants. Notably, our analysis is based only on signals that could be detected by standard SNP-based GWAS, which are underpowered to detect underlying multiallelic associations from STRs[52]. Further work to directly test for associations between STRs and phenotypes may reveal a widespread role for repeat number variation in complex traits.

Our study faced several limitations. (1) While we applied stringent fine-mapping approaches to find eSTRs whose signals are probably not explained by nearby SNPs in LD, some signals could plausibly be explained by other variant classes, such as structural variants[53] or *Alu* elements[54] that were not considered. Furthermore, our fine-mapping procedure may be vulnerable to false negatives for STRs in strong or perfect LD with nearby SNPs, or false positives due to noise present with small sample sizes. (2) Our study was limited to tissues available from GTEx with sufficient sample sizes. While this greatly expanded on the single tissue used in our previous eSTR analysis, some tissues, such as brain, were not well represented. Further, due to overwhelming sharing of eSTRs across tissues, we were unable to identify tissue-specific characteristics of eSTRs. (3) Despite strong evidence that the FM-eSTRs for *RFT1* and other genes may drive published GWAS signals, we have not definitively proved causality. Additional work is needed to validate effects on expression and evaluate the impact of these STRs in trait-relevant cell types.

Altogether, our eSTR catalog provides a valuable resource for studying the role of STRs in complex traits. Example applications of this resource include further analysis of the genetic architecture of gene expression by quantifying the contribution of different variant classes, genome-wide analyses to confirm or refute hypotheses about eSTR mechanisms and integration of eSTRs into GWAS fine-mapping to identify candidate variants not identified by SNP-based analyses. To facilitate these and other studies, all summary-level eSTR data are publicly available at http://webstr.ucsd.edu/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0521-9.

Received: 13 February 2019; Accepted: 25 September 2019;
Published online: 1 November 2019

References

1. GTEx Consortium Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
2. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
3. Grünewald, T. G. P. et al. Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. *Nat. Genet.* **47**, 1073–1078 (2015).
4. Song, J. H. T., Lowe, C. B. & Kingsley, D. M. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. *Am. J. Hum. Genet.* **103**, 421–430 (2018).
5. Boettger, L. M. et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48**, 359–366 (2016).
6. Leffler, E. M. et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**, eaam6393 (2017).
7. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
8. Sun, J. X. et al. A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).
9. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
10. Willems, T. et al. Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *Am. J. Hum. Genet.* **98**, 919–933 (2016).
11. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
12. Willems, T. et al. The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
13. Li, H. Towards better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
14. Gymrek, M. et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
15. Nasrallah, M. P. et al. Differential effects of a polyalanine tract expansion in Arx on neural development and gene expression. *Hum. Mol. Genet.* **21**, 1090–1098 (2012).
16. Quilez, J. et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44**, 3750–3762 (2016).
17. Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–1216 (2009).
18. Gemayel, R., Vinces, M. D., Legendre, M. & Verstrepen, K. J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
19. Liu, X. S. et al. Rescue of fragile X syndrome neurons by DNA methylation editing of the FMR1 gene. *Cell* **172**, 979–992.e6 (2018).
20. Raveh-Sadka, T. et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **44**, 743–750 (2012).
21. Suter, B., Schnappauf, G. & Thoma, F. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res.* **28**, 4083–4089 (2000).
22. Afek, A., Schipper, J. L., Horton, J., Gordan, R. & Lukatsky, D. B. Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl Acad. Sci. USA* **111**, 17140–17145 (2014).
23. Conlon, E. G. et al. The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *eLife* **5**, e17820 (2016).
24. Lin, Y., Dent, S. Y., Wilson, J. H., Wells, R. D. & Napierala, M. R loops stimulate genetic instability of CTG.CAG repeats. *Proc. Natl Acad. Sci. USA* **107**, 692–697 (2010).
25. Rothenburg, S., Koch-Nolte, F., Rich, A. & Haag, F. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl Acad. Sci. USA* **98**, 8985–8990 (2001).
26. Min, J. L. et al. The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PLoS ONE* **6**, e22070 (2011).
27. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods***14**, 590–59 (2017).
28. Borel, C. et al. Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. *Hum. Mutat.* **33**, 1302–1309 (2012).
29. Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30**, 315–320 (2002).
30. Gebhardt, F., Zänker, K. S. & Brandt, B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**, 13176–13180 (1999).
31. Johnson, A. D. et al. Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.* **18**, 2700–2710 (2009).
32. Matsuzono, K. et al. Antisense oligonucleotides reduce RNA foci in spinocerebellar ataxia 36 patient iPSCs. *Mol. Ther. Nucleic Acids* **8**, 211–219 (2017).
33. Saha, A. et al. Functional IFNG polymorphism in intron 1 in association with an increased risk to promote sporadic breast cancer. *Immunogenetics* **57**, 165–171 (2005).
34. Shimajiri, S. et al. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett.* **455**, 70–74 (1999).
35. Vikman, S. et al. Functional analysis of 5-lipoxygenase promoter repeat variants. *Hum. Mol. Genet.* **18**, 4521–4529 (2009).
36. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics.* **198**, 497–508 (2014).
37. Kobayashi, H. et al. Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.* **89**, 121–130 (2011).
38. Lalioti, M. D. et al. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386**, 847–851 (1997).
39. Mougey, E. et al. ALOX5 polymorphism associates with increased leukotriene production and reduced lung function and asthma control in children with poorly controlled asthma. *Clin. Exp. Allergy* **43**, 512–520 (2013).
40. Stephensen, C. B. et al. ALOX5 gene variants affect eicosanoid production and response to fish oil supplementation. *J. Lipid Res.* **52**, 991–1003 (2011).
41. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
42. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* **10**, 161–172 (2009).
43. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–780 (2012).
44. Ciesiolka, A., Jazurek, M., Drazkowska, K. & Krzyzosiak, W. J. Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. *Front. Cell. Neurosci.* **11**, 97 (2017).
45. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
46. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
47. Schizophrenia Working Group of the Psychiatric Genomics Consortium Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
48. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
49. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
50. Guo, H. et al. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum. Mol. Genet.* **24**, 3305–3313 (2015).
51. Haeuptle, M. A. et al. Human RFT1 deficiency leads to a disorder of N-linked glycosylation. *Am. J. Hum. Genet.* **82**, 600–606 (2008).
52. Saini, S., Mitra, I., Mousavi, N., Fotsing, S. F. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).
53. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
54. Hasler, J. & Strub, K. Alu elements as regulators of gene expression. *Nucleic Acids Res.* **34**, 5491–5497 (2006).

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

Methods

Dataset and preprocessing. Next-generation sequencing data were obtained from GTEx through dbGaP under phs000424.v7.p2. This included high-coverage (30×) Illumina WGS data and expression data from 652 unrelated individuals (Supplementary Fig. 1). The WGS cohort consisted of 561 individuals with reported European ancestry, 75 of African ancestry, and 8, 3 and 5 of Asian, Amerindian and unknown ancestry, respectively. For each sample, we downloaded BAM files containing read alignments to the hg19 reference genome and VCF files containing SNP genotype calls.

STRs were genotyped using HipSTR[27] v.0.5, which returns the maximum likelihood diploid STR allele sequences for each sample based on aligned reads as input. Samples were genotyped separately with nondefault parameters, --min-reads 5 and --def-stutter-model. VCFs were filtered using the filter_vcf.py script available from HipSTR, using recommended settings for high-coverage data (--min-call-qual 0.9, --max-call-flank-indel 0.15 and --max-call-stutter 0.15). VCFs were merged across all samples and further filtered to exclude STRs meeting the following criteria: call rate <80%; STRs overlapping segmental duplications (UCSC Genome Browser[55] hg19.genomicSuperDups table); penta- and hexamer STRs containing homopolymer runs of at least five or six nucleotides, respectively, in the hg19 reference genome, since we previously found these STRs to have high error rates due to indels in homopolymer regions[52]; and STRs whose frequencies did not meet the percentage of homozygous versus heterozygous calls expected under Hardy–Weinberg equilibrium (binomial two-sided $P<0.05$). Additionally, to restrict to polymorphic STRs, we filtered STRs with heterozygosity <0.1. Altogether, 175,226 STRs remained for downstream analysis.

We additionally obtained gene-level RPKM values for each tissue from dbGaP project phs000424.v7.p2. We focused on 15 tissues with at least 200 samples, and included two brain tissues with slightly more than 100 samples available (Supplementary Table 1). Genes with median RPKM of 0 were excluded, and expression values for remaining genes were quantile-normalized separately per tissue to a standard normal distribution. Analysis was restricted to protein-coding genes based on GENCODE v.19 (Ensembl 74) annotation.

Before downstream analyses, expression values were adjusted separately for each tissue to control for sex, population structure and technical variation in expression as covariates. For population structure, we used the top ten principal components (PCs) resulting from performing principal components analysis (PCA) on the matrix of SNP genotypes from each sample. PCA was performed jointly on GTEx samples and 1000 Genomes Project[56] samples genotyped using Omni 2.5 SNP genotyping arrays (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapeit2_scaffolds/hd_chip_scaffolds/). Analysis was restricted to biallelic SNPs present in the Omni 2.5 data and resulting loci were LD-pruned using plink[57] v.1.90b3.44 with option --indep 50 5 2. PCA on resulting SNP genotypes was performed using smartpca[58,59] v.13050. To control for technical variation in expression, we applied PEER factor correction[60]. Based on an analysis of number of PEER factors versus number of eSTRs identified per tissue (Supplementary Fig. 2), we determined an optimal number of $N/10$ PEER factors as covariates for each tissue, where N is the sample size. PEER factors were correlated with covariates reported previously for GTEx samples (Supplementary Fig. 3), such as ischemic time.

eSTR and expression SNP identification. For each STR within 100 kb of a gene, we performed a linear regression between STR lengths and adjusted expression values

$$Y' = \beta X + \epsilon$$

where X denotes STR genotypes, Y' denotes expression values adjusted for the covariates described above, β denotes the effect size and ϵ is the error term. A separate regression analysis was performed for each STR–gene pair in each tissue. For STR genotypes, we used the average repeat length of the two alleles for each individual, where repeat length was computed as a length difference from the hg19 reference, with 0 representing the reference allele. Linear regressions were performed using the OLS function from the Python statsmodels.api module[61] (https://www.statsmodels.org, v.0.8.0), which returns estimated regression coefficients computed using ordinary least squares and two-sided P values for each regression coefficient testing the null hypothesis $\beta=0$ computed from t-statistics for each coefficient. As a control, for each STR–gene pair, we performed a permutation analysis in which sample identifiers were shuffled.

Samples with missing genotypes or expression values were removed from each regression analysis. To reduce the effect of outlier STR genotypes, we removed samples with genotypes observed in fewer than three samples. If, after filtering samples, there were fewer than three unique genotypes, the STR was excluded from analysis. Adjusted expression values and STR genotypes for remaining samples were then Z-scaled to have mean 0 and variance 1, before performing each regression. This step forces resulting effect sizes to be between −1 and 1.

We used a gene-level FDR threshold (described previously[14]) of 10% to identify significant STR–gene pairs. We assume most genes have, at most, a single causal eSTR. For each gene, we determined the STR association with the strongest P value. This P value was adjusted using a Bonferroni correction for the number of STRs tested per gene, to give a P value for observing a single eSTR association for each gene. We then used the list of adjusted P values (one per gene) as input to the fdrcorrection function in the statsmodels.stats.multitest module to obtain a q-value for the best eSTR for each gene. FDR analysis was performed separately for each tissue.

Expressions SNPs (eSNPs) were identified using the same model covariates and normalization procedures, but using SNP dosages (0, 1 or 2) rather than STR lengths. Similar to the STR analysis, we removed samples with genotypes occurring in fewer than three samples and removed SNPs with fewer than three unique genotypes remaining after filtering. On average, we tested 17 STRs and 533 SNPs per gene.

Fine-mapping eSTRs. We used model comparison as an orthogonal validation to CAVIAR findings to determine whether the best eSTR for each gene explained variation in gene expression beyond a model consisting of the best eSNP. For each gene with an eSTR we determined the eSNP with the strongest P value. We then compared two linear models: $Y' \approx$ eSNP (SNP-only model) versus $Y' \approx$ eSNP + eSTR (SNP + STR model) using the anova_lm function in the Python statsmodels.api.stats module. Q-values were obtained using the fdrcorrection function in the statsmodels.stats.multitest module. On average across tissues, 17.4% of eSTRs tested improved the model over the best eSNP for the target gene (10% FDR). When restricting to FM-eSTRs, 78% improved the model (10% FDR).

We used CAVIAR[36] v.2.2 to further fine-map eSTR signals against all nominally significant eSNPs ($P<0.05$) within 100 kb of each gene. On average, 121 SNPs per gene passed this threshold and were included in the CAVIAR analysis. Pairwise LD between the eSTR and eSNPs was estimated using the Pearson correlation between SNP dosages (0, 1 or 2) and STR genotypes (average of the two STR allele lengths) across all samples. CAVIAR was run with parameters -f 1 -c 2 to model up to two independent causal variants per locus. In some cases, initial association statistics for SNPs and STRs might have been computed using different sets of samples if some were filtered due to outlier genotypes. To provide a fair comparison between eSTRs and eSNPs, for each CAVIAR analysis we recomputed Z-scores for eSTRs and eSNPs using the same set of samples before running CAVIAR.

Multitissue eSTR analysis. We used an R implementation of mash[41] (mashR) v.0.2.21 to compute posterior estimates of eSTR effect sizes and standard errors across tissues (https://stephenslab.github.io/mashr/articles/intro_mash_dd.html). Briefly, mashR takes, as input, effect sizes and standard error measurements per tissue, learns various covariance matrices of effect sizes between tissues and outputs posterior estimates of effect sizes and standard errors accounting for global patterns of effect-size sharing. We used all eSTRs with a nominal P value of $<1\times10^{-5}$ in at least one tissue as a set of strong signals to compute covariance matrices. eSTRs that were not analyzed in all tissues were excluded from this step. We included 'canonical' covariance matrices (identity matrix and matrices representing condition-specific effects) and matrices learned by extreme deconvolution initialized using PCA with five components, as suggested by mashR documentation. After learning covariance matrices, we applied mashR to estimate posterior effect sizes and standard errors for each eSTR in each tissue. For eSTRs that were filtered from one or more tissues in the initial regression analysis, we set input effect sizes to 0 and standard errors to 10 in those tissues to reflect high uncertainty in effect-size estimates at those eSTRs. For Fig. 1d, rows and columns of the effect-size correlation matrix were clustered using default parameters from the clustermap function in the Python seaborn library (https://seaborn.pydata.org/, v.0.9.0).

Canonical repeat units. For each STR, we defined the canonical repeat unit as the lexicographically first repeat unit when considering all rotations and strand orientations of the repeat sequence. For example, the canonical repeat unit for the repeat sequence CAGCAGCAGCAG would be AGC.

Enrichment analyses. Enrichment analyses were performed using a two-sided Fisher's exact test as implemented in the fisher_exact function of the Python package scipy.stats (https://docs.scipy.org/doc/scipy/reference/stats.html, v.1.2.1). Overlapping STRs with each annotation was performed using the intersectBed tool of the BEDTools[62] suite v.2.28.0. Genomic annotations were obtained by downloading custom tables using the UCSC Genome Browser[55] table browser tool to select either coding regions, introns, 5′ UTRs or 3′ UTRs. An STR could be assigned to more than one category in the case of overlapping transcripts. STRs not assigned to one of those categories were labeled as intergenic. ENCODE DNase I HS clusters were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz). Analysis was restricted to DNase I HS clusters annotated in at least 20 cell types. The distance between each STR and the center of the nearest DNase I HS cluster was computed using the closestBed tool from the BEDTools suite.

Analysis of DNase-seq, ChIP–seq and nucleosome occupancy. Genome-wide nucleosome occupancy signal in GM12878 was downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/

wgEncodeSydhNsome/wgEncodeSydhNsomeGm12878Sig.bigWig). ChIP–seq reads for RNA polymerase II (RNAPII) and DNase-seq reads were downloaded from the ENCODE Project website (https://www.encodeproject.org) (accessions GM12878 RNAPII: ENCFF000OBB, heart RNAPII: ENCFF643EGO, lung RNAPII: ENCSR033NHF, tibial nerve RNAPII: ENCFF750HDH, human embryonic stem cells RNAPII: ENCFF526YGE; GM12878 DNase I: ENCFF775ZJX, fat DNase I: ENCFF880CAD, tibial nerve DNase I: ENCFF226ZCG, skin DNase I: ENCFF238BRB). Histograms of aggregate read densities and heat maps for individual STR regions were generated using the annoatePeaks.pl tool of HOMER[63] v.4.10. For nucleosome occupancy and DNase I analyses on all STRs, we used parameters -size 1000 -hist 1. For analysis of GC-rich repeats in promoters, we used parameters -size 10000 -hist 5.

Characterization of tissue-specific eSTRs. We clustered FM-eSTRs based on Z-scores computed by mash for each eSTR in each tissue. We first created a tissue by FM-eSTR matrix of the absolute value of the Z-scores. We then Z-normalized the Z-scores for each FM-eSTR to have mean 0 and variance 1. We used the KMeans class from the Python sklearn.cluster module to perform K-means clustering with $K=8$ (https://scikit-learn.org/stable/, v.0.20.3). The number of clusters was chosen by visualizing the sum of squared distances from centroids for values of K ranging from 1 to 20 and choosing a value of K based on the 'elbow method'. Using different values of K produced similar groups. We tested for nonuniform distributions of FM-eSTR repeat units across clusters using a chi-squared test implemented in the scipy.stats chi2_contingency function.

Analysis of DNA and RNA secondary structure. For each STR, we extracted the repeat plus 50-bp flanking sequencing from the hg19 reference genome. We additionally created sequences containing each common allele for each STR. Common alleles were defined as those seen at least five times in a previously generated deep catalog of STR variation in 1,916 samples[52]. For each sequence and its reverse complement, we ran mfold[64] v.3.6 on the DNA and corresponding RNA sequences, with mfold arguments NA = DNA and NA = RNA, respectively, and otherwise default parameters to estimate the free energy of each single-stranded sequence. Mann–Whitney U-tests were performed using the mannwhitneyu function of the scipy.stats Python package.

Colocalization of FM-eSTRs with published GWAS signals. Published GWAS associations were obtained from the NHGRI/EBI GWAS catalog available from the UCSC Genome Browser Table Browser (table hg19.gwasCatalog) downloaded on 24 July 2019. Height GWAS summary statistics were downloaded from the GIANT Consortium website (https://portals.broadinstitute.org/collaboration/giant/images/0/0f/Meta-analysis_Locke_et_al%2BUKBiobank_2018.txt.gz). Schizophrenia GWAS summary statistics were downloaded from the Psychiatric Genomics Consortium website (https://www.med.unc.edu/pgc/results-and-downloads).

IBD summary statistics were downloaded from the International Inflammatory Bowel Disease Genetics Consortium website. We used the file EUR.IBD.gwas_info03_filtered.assoc with summary statistics in Europeans (ftp://ftp.sanger.ac.uk/pub/consortia/ibdgenetics/iibdgc-trans-ancestry-filtered-summary-stats.tgz). Intelligence summary statistics were downloaded from the Complex Trait Genomics laboratory website (https://ctg.cncr.nl/documents/p1651/SavageJansen_IntMeta_sumstats.zip). LD between STRs and SNPs was computed by taking the squared Pearson correlation between STR lengths and SNP dosages in GTEx samples for each STR–SNP pair. STR genotypes seen less than three times were filtered from LD calculations.

Colocalization analysis of eQTL and GWAS signals was performed using the coloc.abf function of the coloc[50] package. For all traits, dataset 1 was specified as type = "quant" and consisted of SNP effect sizes and their variances as input. We specified sdY = 1 since expression was quantile normalized to a standard normal distribution. Dataset 2 was specified differently for height and schizophrenia to reflect quantitative versus case-control analyses. For height and intelligence, we specified type = "quant" and used effect sizes and their variances as input. We additionally specified minor allele frequencies listed in the published summary statistics file and the total sample size of $N=695{,}647$ and $N=269{,}720$ for height and intelligence, respectively. For schizophrenia and IBD, we specified type = "CC" and used effect sizes and their variances as input. We additionally specified the fraction of cases as 33%.

Capture Hi-C interactions (Extended Data Fig. 10) were visualized using the 3D Genome Browser[65]. The visualization depicts interactions profiled in GM12878 (ref. [66]) and only shows interactions overlapping the STR of interest.

Association analysis in the eMERGE cohort. We obtained SNP genotype array data and imputed genotypes from dbGaP accessions phs000360.v3.p1 and phs000888.v1.p1 from consent groups c1 (Health/Medical/Biomedical), c3 (Health/Medical/Biomedical-Genetic Studies Only-No Insurance Companies) and c4 (Health/Medical/Biomedical-Genetic Studies Only). Height data were available for samples in cohorts c1 (phs000888.v1.pht004680.v1.p1.c1), c3 (phs000888.v1.pht004680.v1.p1.c3) and c4 (phs000888.v1.pht004680.v1.p1.c4). We removed samples without age information listed. If height was collected at multiple times for the same sample, we used the first data point listed.

Genotype data were available for 7,190, 6,100 and 3,755 samples from the c1, c3 and c4 cohorts, respectively (dbGaP study phs000360.v3.p1). We performed PCA on the genotypes to infer ancestry of each individual. We used plink to restrict to SNPs with minor allele frequency at least 10% and with genotype frequencies expected under Hardy–Weinberg equilibrium ($P > 1\times10^{-4}$). We performed LD pruning using the plink option --indep 50 5 1.5 and used pruned SNPs as input to PCA analysis. We visualized the top two PCs and identified a cluster of 14,147 individuals overlapping samples with annotated European ancestry. We performed a separate PCA using only the identified European samples and used the top ten PCs as covariates in association tests.

A total of 11,587 individuals with inferred European ancestry had both imputed SNP genotypes and height and age data available. Samples originated from cohorts at Marshfield Clinic, Group Health Cooperative, Northwestern University, Vanderbilt University and the Mayo Clinic. We adjusted height values by regressing on top ten ancestry PCs, age and cohort. Residuals were inverse normalized to a standard normal distribution. Adjustment was performed separately for males and females.

Imputed genotypes (from dbGaP study phs000888.v1.p1) were converted from IMPUTE2 (ref. [67]) to plink's binary format using plink, which marks calls with uncertainty >0.1 (score < 0.9) as missing. SNP associations were performed using plink with imputed genotypes as input and with the 'linear' option with analysis restricted to the region chr3:53022501–53264470.

The *RFT1* FM-eSTR was imputed into the imputed SNP genotypes using Beagle 5 (ref. [68]) with option gp = true and using our SNP–STR reference haplotype panel[52]. We previously estimated imputation concordance of 97% at this STR in a separate European cohort. Samples with imputed genotype probabilities of less than 0.9 were removed from the STR analysis. We additionally restricted analysis to STR genotypes present in at least 100 samples to minimize the effect of outlier genotypes. We regressed STR genotype (defined above as the average of an individual's two repeat lengths) on residualized height values for the remaining 6,393 samples using the Python statsmodels.regression.linear_model.OLS function (https://www.statsmodels.org).

Dual luciferase reporter assay. Constructs for 0, 5 or 10 copies of AC at the FM-eSTR for *RFT1* (chr3:53128363–53128413) plus approximately 170-bp genomic context on either side (RFT1_0rpt, RFT1_5rpt, RFT1_10rpt in Supplementary Table 7) were ordered as gBlocks from Integrated DNA Technologies. Each construct additionally contained homology arms for cloning into pGL4.27 (below). Additionally we amplified, using PCR, the region from genomic DNA for sample NA12878 with 12 copies of AC (NIGMS Human Genetic Repository, Coriell) using PrimeSTAR max DNA Polymerase (Clontech, catalog no. R045B) and primers RFT1eSTR_F and RFT1eSTR_R (Supplementary Table 7), which included the same homology arms.

Constructs were cloned into plasmid pGL4.27 (Promega, catalog no. E8451), which contains the firefly luciferase coding sequence and a minimal promoter. The plasmid was linearized using EcoRV (New England Biolabs, catalog no. R3195) and purified from agarose gel (Zymo Research, catalog no. D4001). Constructs were cloned into the linearized vector using In-Fusion (Clontech, catalog no. 638910). Sanger sequencing of isolated clones for each plasmid validated expected repeat numbers in each construct.

Plasmids were transfected into the human embryonic kidney 293 cell line (HEK293T; ATCC CRL-3216) and grown in DMEM media (Gibco, catalog no. 10566-016), supplemented with 10% fetal bovine serum (Gibco, catalog no. 10438-026), 2 mM glutamine (Gibco, catalog no. A2916801), 100 units ml^{-1} of penicillin, 100 µg ml^{-1} of streptomycin and 0.25 µg ml^{-1} amphotericin B (Gibco, Antibiotic-Antimycotic, catalog no. 15240062). Cells were maintained at 37 °C in a 5% CO$_2$ incubator. HEK293T cells (2×10^5) were plated onto each well of a 25 µg ml^{-1} poly-D-lysine (EMD Millipore, catalog no. A-003-E)-coated 24-well plate, the day before transfection. On the day of the transfection, medium was changed to Opti-MEM. We conducted cotransfection experiments to test expression of each construct. Empty pGL4.27 vector (100 ng) (Promega, E8451) or 100 ng of each one of the pGL4.27 derivatives, was mixed with 5 ng of the reference plasmid pGL4.73 (Promega, catalog no. E6911), harboring an SV40 promoter upstream of *Renilla* luciferase, and added to the cells in the presence of Lipofectamine 3000 (Invitrogen, catalog no. L3000015), according to the manufacturer's instructions. Cells were incubated for 24 h at 37 °C, washed once with PBS and then incubated in fresh completed medium for an additional 24 h.

Forty-eight hours after transfection the HEK293T cells were washed three times with PBS and lysed in 100 µl of Passive Lysis Buffer (Promega, catalog no. E1910). Firefly luciferase and *Renilla* luciferase activities were measured in 10 µl of HEK293T cell lysate using the dual luciferase reporter assay system (Promega, catalog no. E1910) in a Veritas Microplate Luminometer. Relative activity was defined as the ratio of firefly luciferase activity to *Renilla* luciferase activity. For each plasmid, transfection and the expression assay were done in triplicate using three wells of cultured cells that were independently transfected (biological repeats), and three individually prepared aliquots of each transfection reaction (technical repeats). Values from each technical replicate were averaged to get one ratio for each biological repeat.

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

Data availability

All eSTR summary statistics are available for download on WebSTR http://webstr.ucsd.edu/downloads.

Code availability

Code for performing analyses and generating figures is available at http://github.com/gymreklab/gtex-estrs-paper.

References

55. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
56. The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
57. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
58. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
59. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
60. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
61. Seabold, S. P. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 57–61 (SCIPY, 2010).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
63. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
64. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
65. Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
66. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
67. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
68. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

Acknowledgements

Research reported in this publication was supported in part by the Office of The Director, National Institutes of Health under Award Number DP5OD024577 (M.G.). We thank V. Bafna, E. Mendenhall, J. Gleeson and Y. Liu for helpful comments. See the Supplementary Note for additional acknowledgements.

Author contributions

S.F.F. performed all eSTR and SNP mapping, helped to perform downstream analyses and helped to draft the manuscript. J.M. performed multitissue analysis using mashR and helped to revise the manuscript. C.W. optimized and performed the reporter assay. S.S. participated in the design of the STR imputation analysis. S.S.-B. lead, designed and analyzed data from the reporter assay. R.Y. implemented the WebSTR web application. A.G. conceived and planned analyses and validation experiments of regulatory effects of eSTRs and wrote the manuscript. M.G. conceived the study, designed and performed analyses and wrote the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

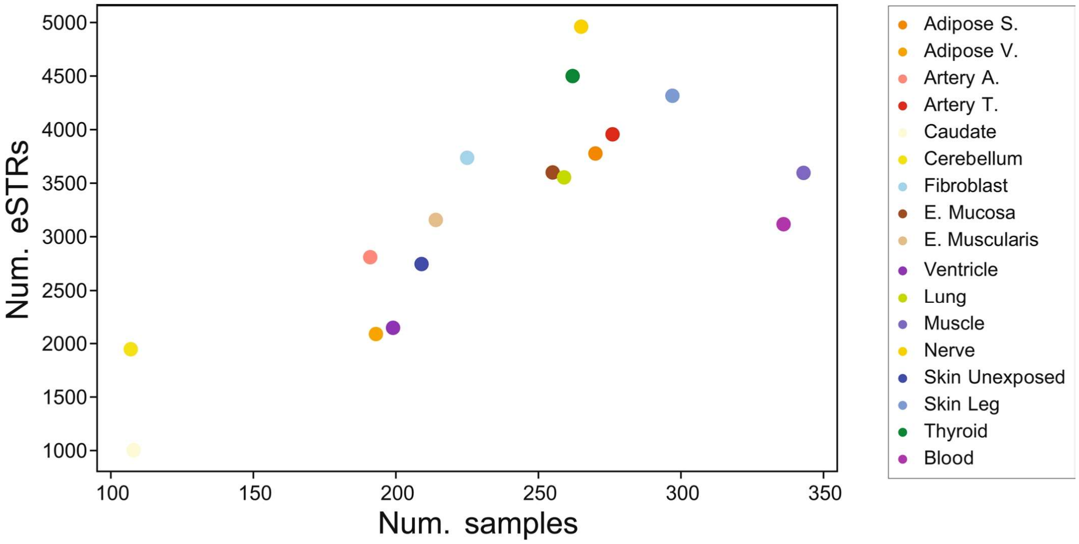
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-019-0521-9.

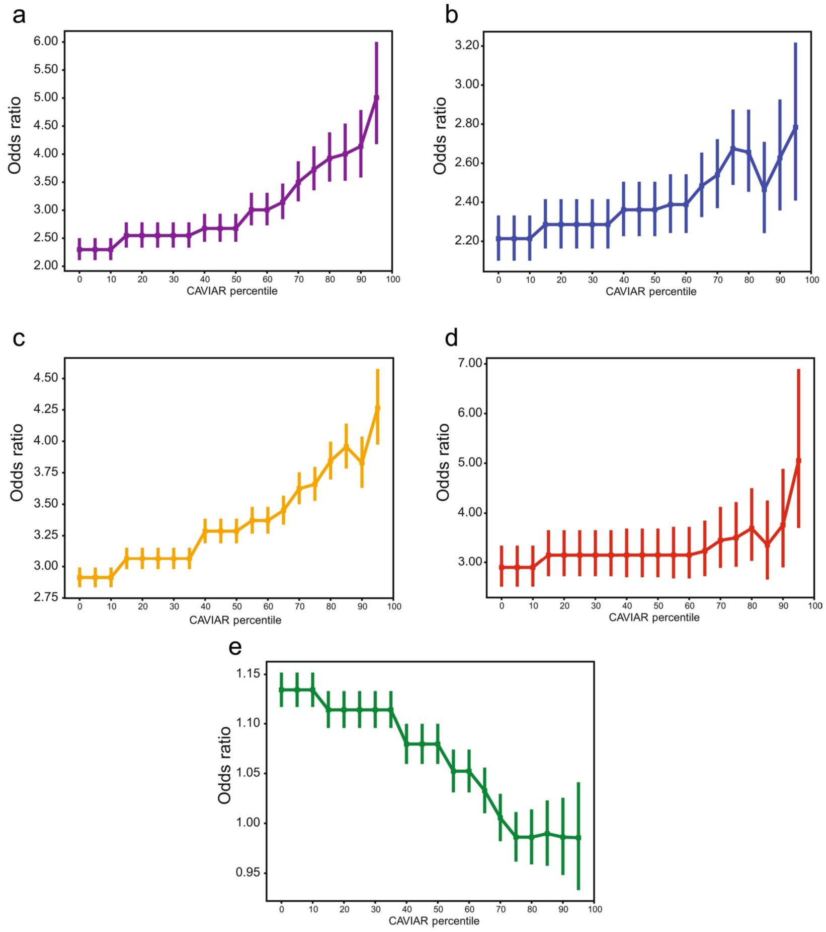
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0521-9.

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

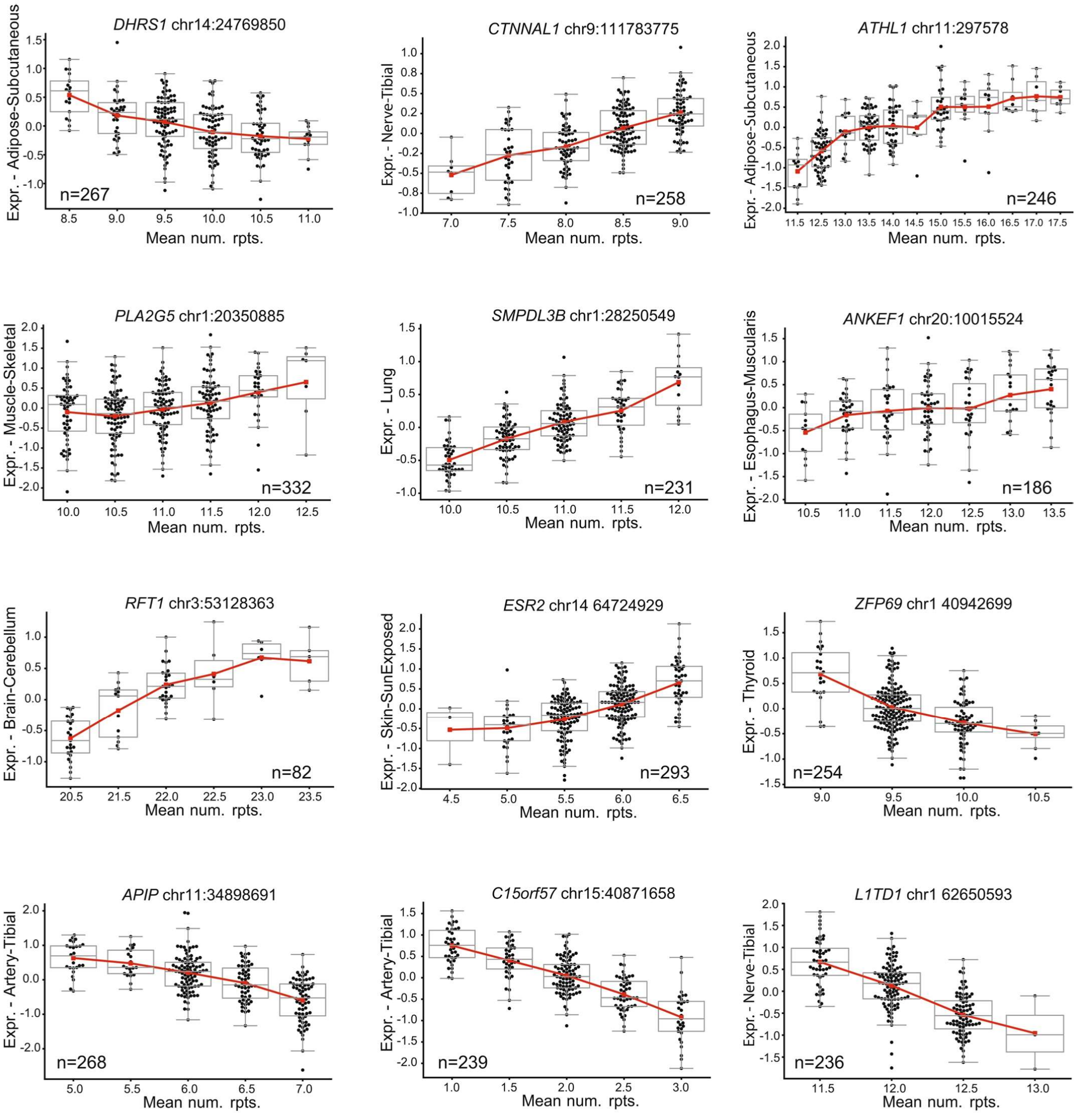
Reprints and permissions information is available at www.nature.com/reprints.



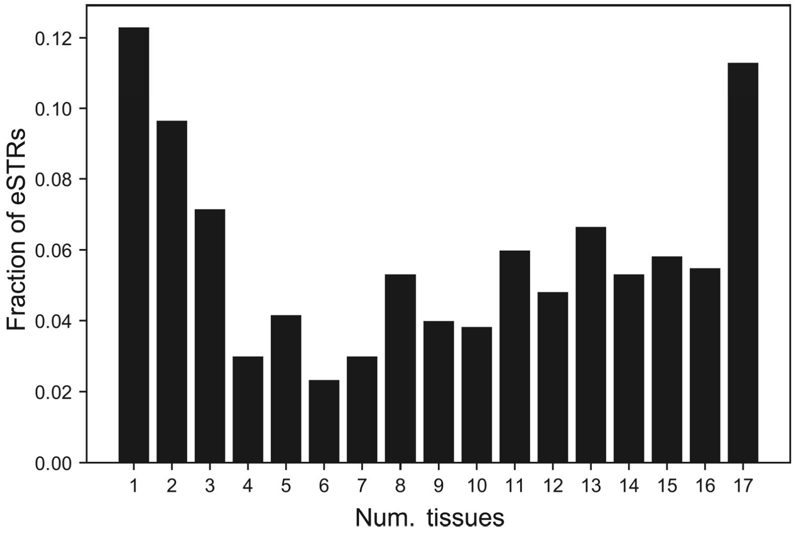
Extended Data Fig. 1 | Relationship between sample size and number of eSTRs detected. The x-axis shows the number of samples per tissue. The y-axis shows the number of eSTRs (gene-level FDR<10%) detected in each tissue. Each dot represents a single tissue, using the same colors as shown in Fig. 1 in the main text (see box on the right). Notably, although whole blood and skeletal muscle had the highest number of samples, we identified fewer eSTRs in those tissues than in others with lower sample sizes. This is concordant with previous results for SNPs in the GTEx cohort and may reflect higher cell-type heterogeneity in these tissue samples.



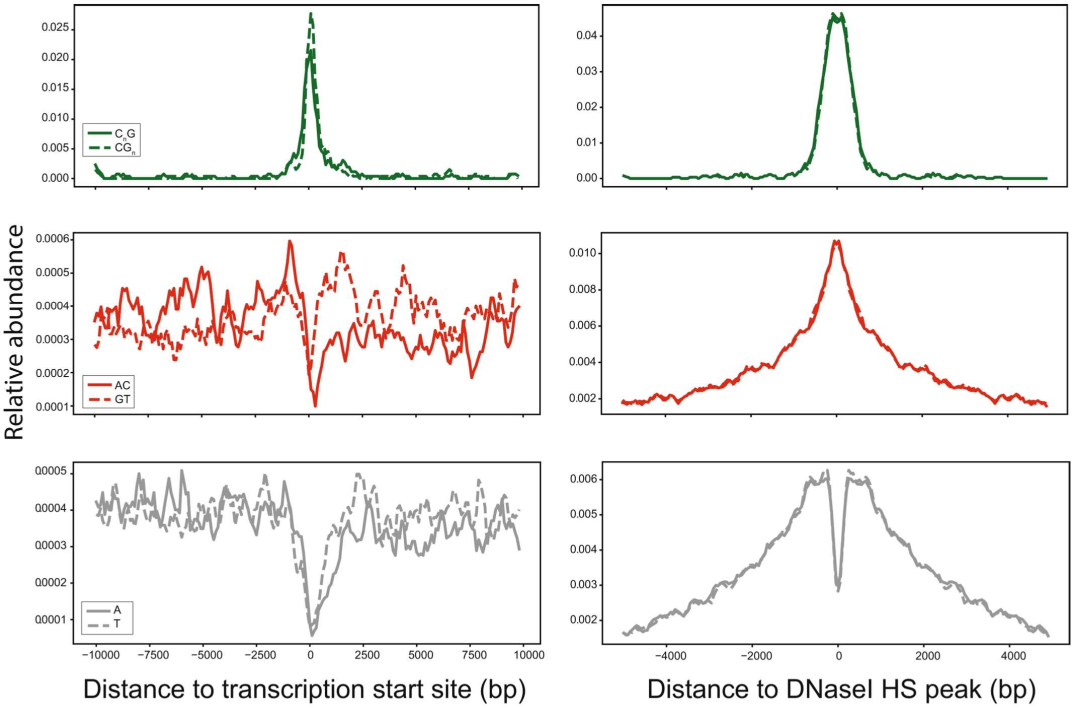
Extended Data Fig. 2 | Enrichment of genomic annotations as a function of CAVIAR threshold. The x-axis represents CAVIAR thresholds in terms of the percentile (percentage of all 28,375 eSTRs excluded by those thresholds). The y-axis represents the odds ratio for enrichment in eSTRs above each percentile threshold in each of these categories: **a**. 5′UTRs (purple); **b**. 3′UTRs (blue); **c**. promoters (orange; TSS +/- 3kb); **d**. Coding regions (red) and **e**. Introns (green). The y-axis center values denote the log$_2$ odds ratios comparing eSTRs passing each threshold to all STRs. Error bars represent +/−1 s.e.



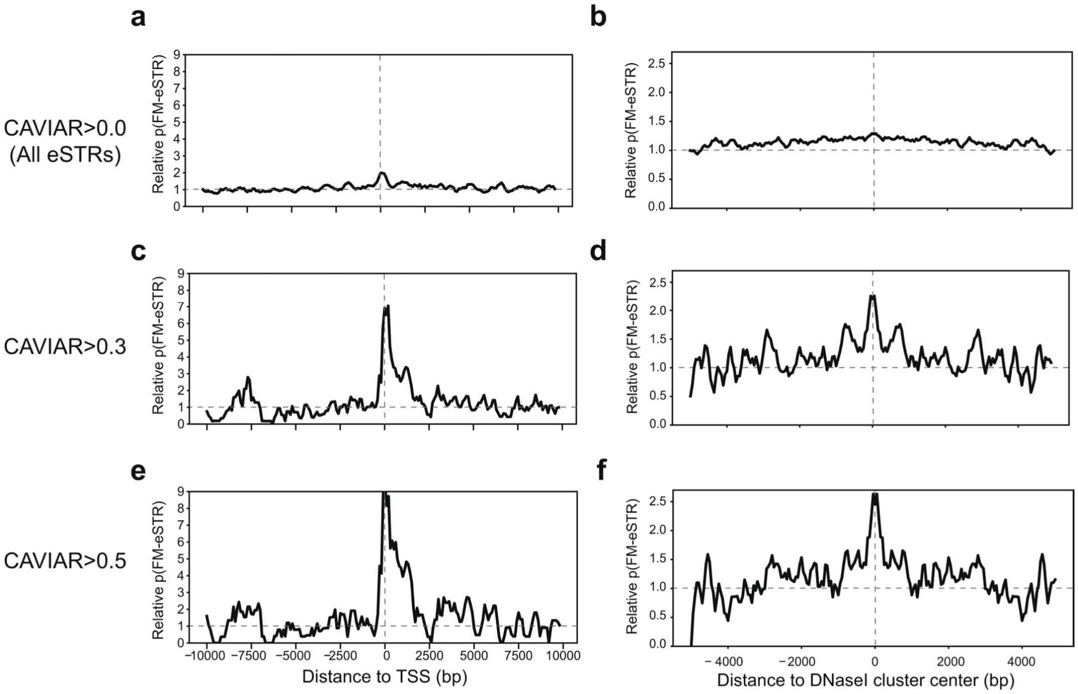
Extended Data Fig. 3 | Example multi-allelic FM-eSTRs. For each plot, the x-axis represents the mean number of repeats in each individual and the y-axis represents normalized expression in the tissue for which the eSTR was most significant. Boxplots summarize the distribution of expression values for each genotype. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (bottom) and Q3+1.5*IQR (top), where IQR gives the interquartile range (Q3-Q1). The red line shows the mean expression for each x-axis value.



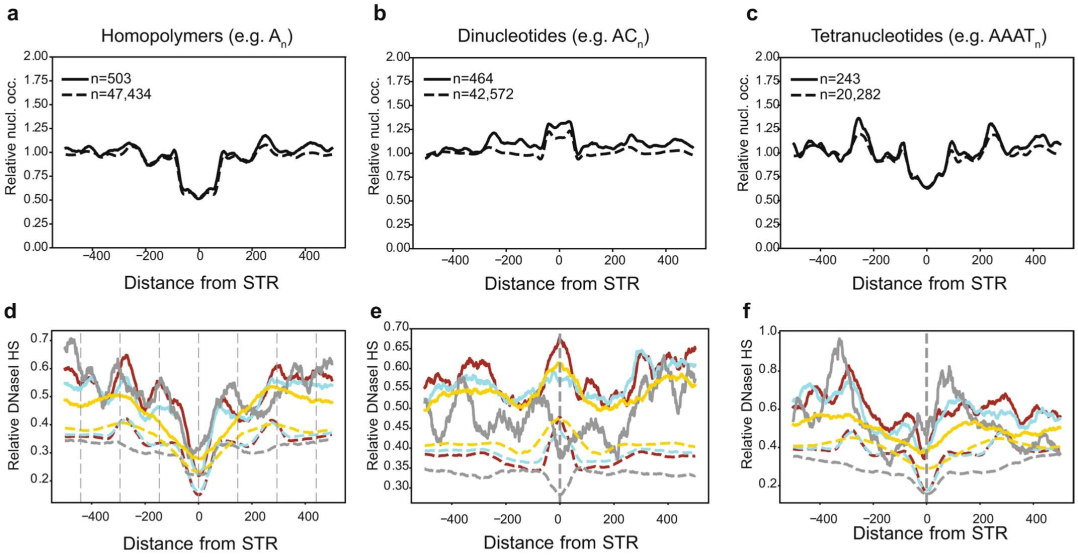
Extended Data Fig. 4 | Sharing of eSTRs across tissues. The x-axis represents the number of tissues that share a given eSTR (absolute value of mashR Z-score >4). The y-axis represents the number of eSTRs shared across a given number of tissues.



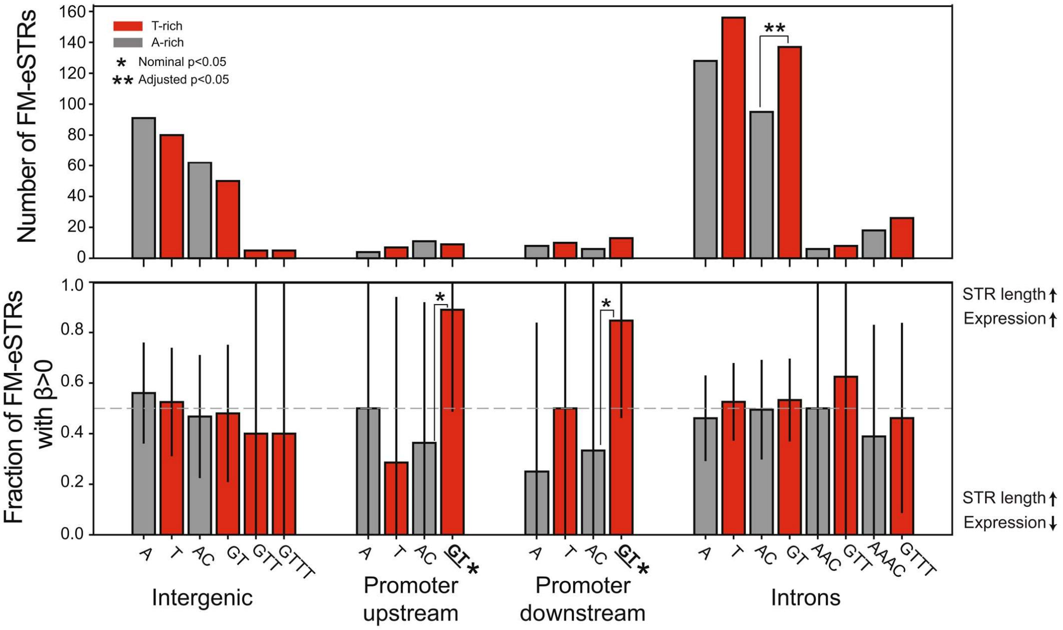
Extended Data Fig. 5 | Localization of all STRs around putative regulatory regions. Left and right plots show localization around transcription start sites and DNaseI HS clusters, respectively. The y-axis denotes the fraction of STRs of each type in each bin. For promoters, the x-axis is divided into 100bp bins. For DNaseI HS sites, the x-axis is divided into 50bp bins. In each plot, values were smoothed by taking a sliding average of each four consecutive bins. Only STR-gene pairs included in our analysis are considered. Each plot compares localization of the two possible sequences of a given repeat unit on the coding strand. Top plots compare repeat units of the form C_nG vs. their reverse complement on the opposite strand, middle plots compare AC vs. GT repeats, and bottom plots compare A vs. T repeats. The strand of each STR was determined based on the coding strand of each target gene.



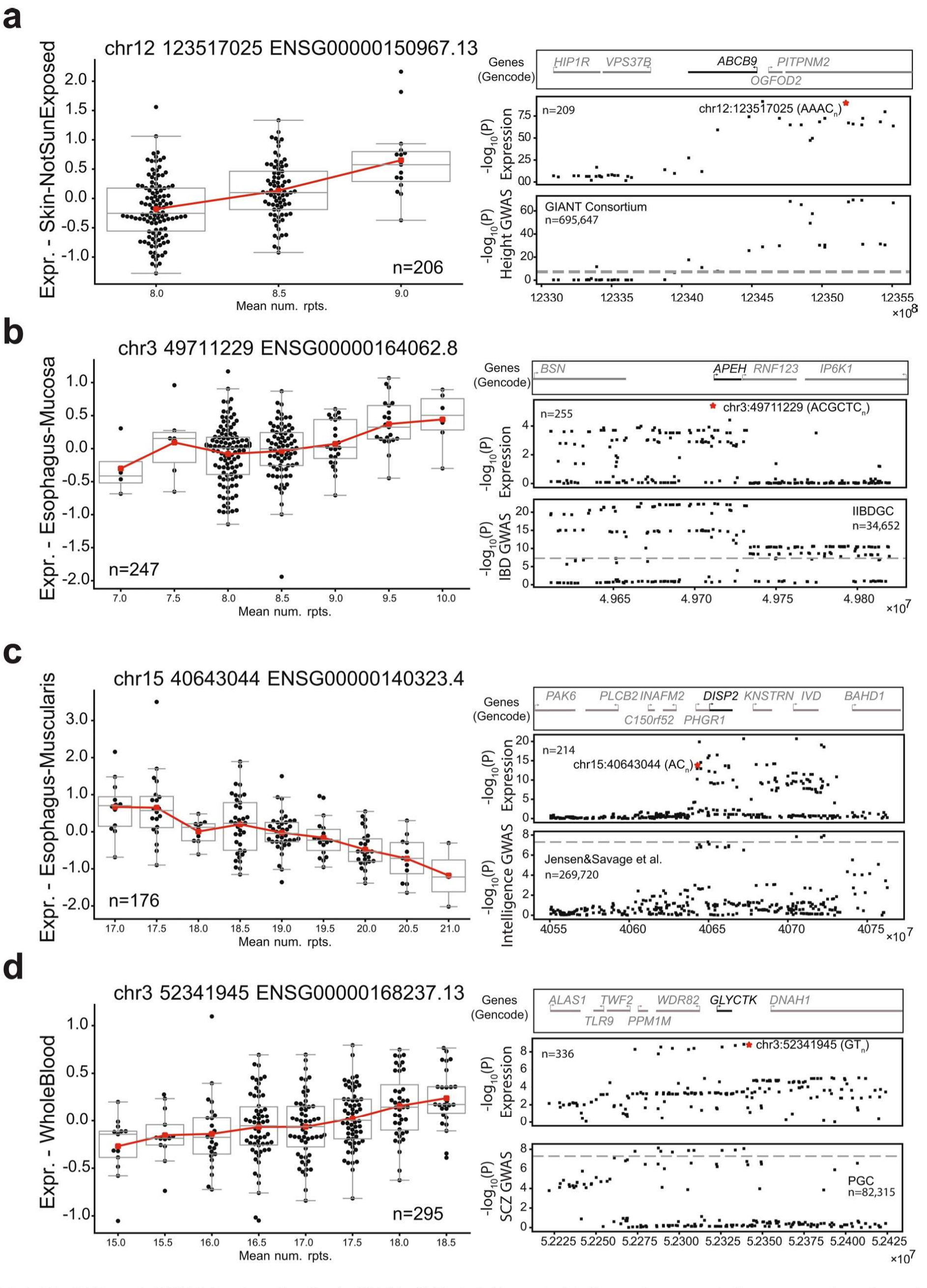
Extended Data Fig. 6 | Relative probability of eSTRs around TSSs and DNaseI HS sites for a range of CAVIAR scores. Plots are shown for FM-eSTRs defined using multiple CAVIAR thresholds (0, corresponding to all eSTRs, 0.3, as used in the main text, or 0.5). **a**., **c**., and **e**. show the relative probability of an STR to be an FM-eSTR around TSSs. The black lines represent the probability of an STR in each bin to be an FM-eSTR. Values were scaled relative to the genome-wide average. **b**., **d**., and **f**. show the relative probability of an STR to be an FM-eSTR around DNaseI HS clusters. Values were smoothed by taking a sliding average of each four consecutive bins.



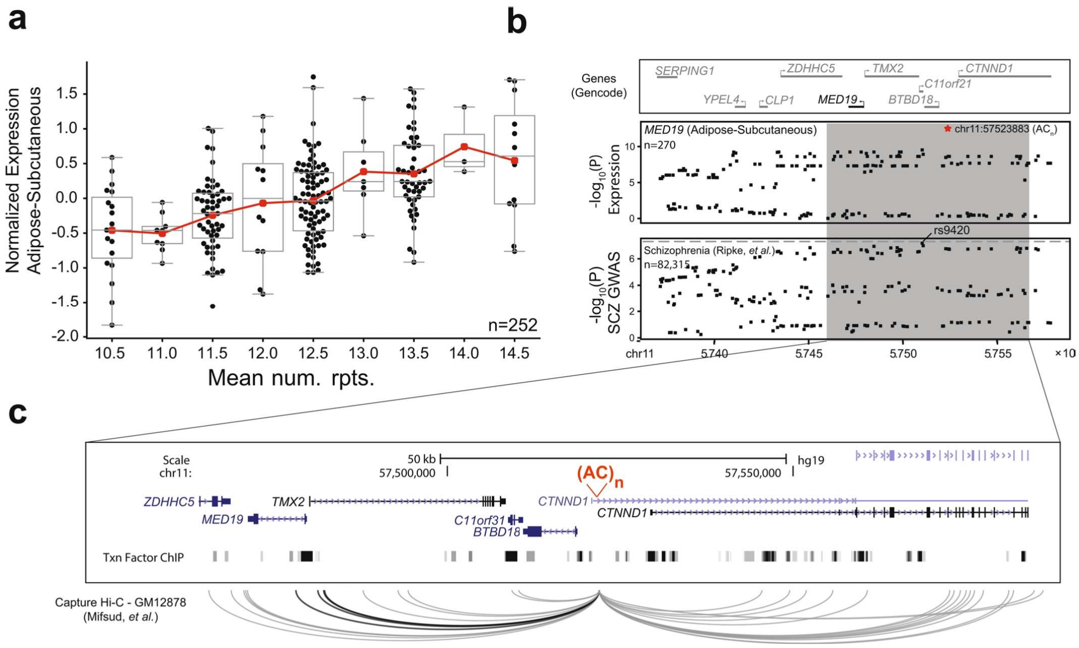
Extended Data Fig. 7 | Nucleosome occupancy and DNaseI hypersensitivity show distinct patterns around eSTRs. a-c. Nucleosome density around STRs with different repeat unit lengths. Nucleosome density in GM12878 in 5bp windows is averaged across all STRs analyzed (dashed) and FM-eSTRs (solid) relative to the center of the STR. **b**. **DNaseI HS density around STRs with different repeat unit lengths**. The number of DNaseI HS reads in GM12878 (gray), fat (red), tibial nerve (yellow), and skin (cyan) is averaged across all STRs in each category. Solid lines show FM-eSTRs. Dashed lines show all STRs. Left=homopolymers, middle=dinucleotides, right=tetranucleotides. Other repeat unit lengths were excluded since they have low numbers of FM-eSTRs (see Fig. 4a). Dashed vertical lines in (**d**) show the STR position +/- 147bp.



Extended Data Fig. 8 | Strand-biased characteristics of FM-eSTRs. Top panel: the y-axis shows the number of FM-eSTRs with each repeat unit on the template strand. Bottom panel: the y-axis shows the percentage of FM-eSTRs with each repeat unit on the template strand that have positive effect sizes. Gray bars denote A-rich repeat units (A/AC/AAC/AAAC) and red bars denote T-rich repeat units (T/GT/GTT/GTTT). Single asterisks denote repeat units nominally enriched or depleted (two-sided binomial p<0.05). Double asterisks denote repeat units significantly enriched after controlling for multiple hypothesis testing (Bonferroni adjusted p<0.05). Asterisks above brackets show significant differences between repeat unit pairs. Asterisks on x-axis labels denote departure from the 50% positive effect sizes expected by chance. Error bars give 95% confidence intervals.



Extended Data Fig. 9 | Example GWAS signals co-localized with FM-eSTRs. Left: For each plot, the x-axis represents the mean number of repeats in each individual and the y-axis represents normalized expression in the tissue with the most significant eSTR signal at each locus. Boxplots summarize the distribution of expression values for each genotype. Box plots are as defined in Fig. 1c. The red line shows the mean expression for each x-axis value. Right: Top panels give genes in each region. The target gene for the eQTL associations is shown in black. Middle panels give the -log$_{10}$ p-values of association of the effect-size between each SNP (black points) and the expression of the target gene. The FM-eSTR is denoted by a red star. Bottom panels give the -log$_{10}$ p-values of association between each SNP and the trait based on published GWAS summary statistics. P-values are two-sided and are based on t-statistics computed for effect sizes (β) (see Methods). Dashed gray horizontal lines give the genome-wide significance threshold of 5E-8.



Extended Data Fig. 10 | Example GWAS signal for schizophrenia potentially driven by an eSTR for *MED19* . a. eSTR association for *MED19*. The x-axis shows STR genotypes at an AC repeat (chr11:57523883) as the mean number of repeats in each individual and the y-axis shows normalized *MED19* expression in subcutaneous adipose. Each point represents a single individual. Red lines show the mean expression for each x-axis value. Boxplots are as defined in Fig. 1c. **b**. **Summary statistics for *MED19* expression and schizophrenia**. The top panel shows genes in the region around *MED19*. The middle panel shows the -log$_{10}$ p-values of association between each variant and *MED19* expression in subcutaneous adipose tissue in the GTEx cohort. The FM-eSTR is denoted by a red star. The bottom panel shows the -log$_{10}$ p-values of association for each variant with schizophrenia reported by the Psychiatric Genomics Consortium. The dashed gray horizontal line shows genome-wide significance threshold of 5E-8. **c**. **Detailed view of the *MED19* locus**. A UCSC genome browser screenshot is shown for the region in the gray box in (**b**). The FM-eSTR is shown in red. The bottom track shows transcription factor (TF) and chromatin regulator binding sites profiled by ENCODE. The bottom panel shows long-range interactions reported by Mifsud, *et al*. using Capture Hi-C on GM12878. Interactions shown in black include *MED19*. Interactions to loci outside of the window depicted are not shown.

Supplementary Note

GC-rich eSTRs are predicted to modulate DNA and RNA secondary structure

FM-eSTRs are most strongly enriched for repeats with high GC content (*e.g.*, canonical repeat units CCG, CCCCG, CCCCCG, AGGGC) (**Fig. 2e**, **Supplementary Table 5**) which are found almost exclusively in promoter regions (**Extended Data Fig. 5**). Given their strong enrichment, we decided to further explore potential biological mechanisms by which this subset of FM-eSTRs might be working. These GC-rich repeat units have been shown to form highly stable secondary structures during transcription such as G4 quadruplexes in single-stranded DNA[1] or RNA[2] that may regulate gene expression. We hypothesized that the effects of GC-rich eSTRs may be in part due to formation of non-canonical nucleic acid secondary structures that modulate DNA or RNA stability as a function of repeat number. We considered properties of two classes of GC-rich FM-eSTRs: (*i*) those following the standard G4 motif ($G_3N_{1\text{-}7}G_3N_{1\text{-}7}G_3N_{1\text{-}7}G_3$)[3] and (*ii*) repeats with canonical repeat unit CCG which does not meet the standard G4 definition. Notably, the majority of CCG FM-eSTRs (79%) occur in 5' UTRs compared to only 11% for G4 repeats. We observed that both classes of GC-rich repeats are associated with higher RNAPII (**Supplementary Fig. 7**) and lower nucleosome occupancy (**Supplementary Fig. 8**) compared to all STRs. This relationship with RNAPII was observed across a diverse range of cell and tissue types.

To evaluate whether GC-rich repeats could be modulating DNA or RNA secondary structure, we used mfold[4] to calculate the free energy of each STR and 50bp of its surrounding context in single stranded DNA or RNA. We considered all common allele lengths (number of repeats) observed at each STR (**Methods**) and computed energies for both the template and non-template strands. We then computed the correlation between the number of repeats and free energy at each STR region. Overall, both G4 and CGG STRs have lower mean free energy (greater stability) and more negative correlations between repeat number and free energy compared to all STRs (**Supplementary Fig. 9**; adjusted Mann-Whitney [MW] one-sided $p<0.05$). Compared to all STRs, FM-eSTRs tend to have lower free energy and more negative correlations with repeat number (MW $p<0.05$ in all categories except for CGG STRs). Notably, both metrics (mean free energy and correlation of repeat number vs. free energy) are significantly correlated with the total sequence length of the STR in all cases (Pearson correlation $p<0.01$). FM-eSTRs tend to be longer than STRs overall (see main text), which may partially explain the secondary structure trends observed.

Based on previous observations[5], we predicted that a higher number of repeat units at GC-rich eSTRs would result in greater DNA or RNA stability and in turn would increase expression of nearby genes. Three example FM-eSTRs following this trend are shown in **Fig. 2f-h**. We tested whether FM-eSTRs were biased toward negative vs. positive effect sizes. As described in the main text, overall FM-eSTRs show no bias in effect direction. However, when considering only repeats in promoter regions (TSS +/- 3kb), 59% of FM-eSTRs have positive effect sizes, significantly more than the 50% expected by chance (binomial two-sided p=0.04; n=137). This effect was stronger when considering only G4 FM-eSTRs (87% positive effect sizes; p=0.0074; n=15) but not significant for CCG FM-eSTRs (62% positive; n=13; p=0.58). This direction of effect bias was consistent across a range of CAVIAR thresholds used to define FM-eSTRs (**Supplementary Fig. 10**). Altogether, these results support a model in which higher repeat numbers at GC-rich eSTRs in promoter regions stabilize DNA secondary structures which promote transcription. Lastly, the contradictory results for CCG STRs may indicate that those repeats could act by distinct mechanisms compared to G4 STRs, but also may be due in part to limited power from a smaller sample size.

Acknowledgements

GTEx Project

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), GTEx Project March 5, 2014 version Page 5 of 8 Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the

University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v7.p2.

eMERGE Project - Mayo Clinic

Samples and associated genotype and phenotype data used in this study were provided by the Mayo Clinic. Funding support for the Mayo Clinic was provided through a cooperative agreement with the National Human Genome Research Institute (NHGRI), Grant #: UOIHG004599; and by grant HL75794 from the National Heart Lung and Blood Institute (NHLBI). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Marshfield Clinic

Funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS) The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Northwestern University

Samples and data used in this study were provided by the NUgene Project (www.nugene.org). Funding support for the NUgene Project was provided by the Northwestern University's Center for Genetic Medicine, Northwestern University, and Northwestern Memorial Hospital. Assistance with phenotype harmonization was provided by the eMERGE Coordinating Center (Grant number U01HG04603). This study was funded through the NIH, NHGRI eMERGE Network (U01HG004609). Funding support for genotyping, which was performed at The Broad Institute, was provided by the

NIH (U01HG004424). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

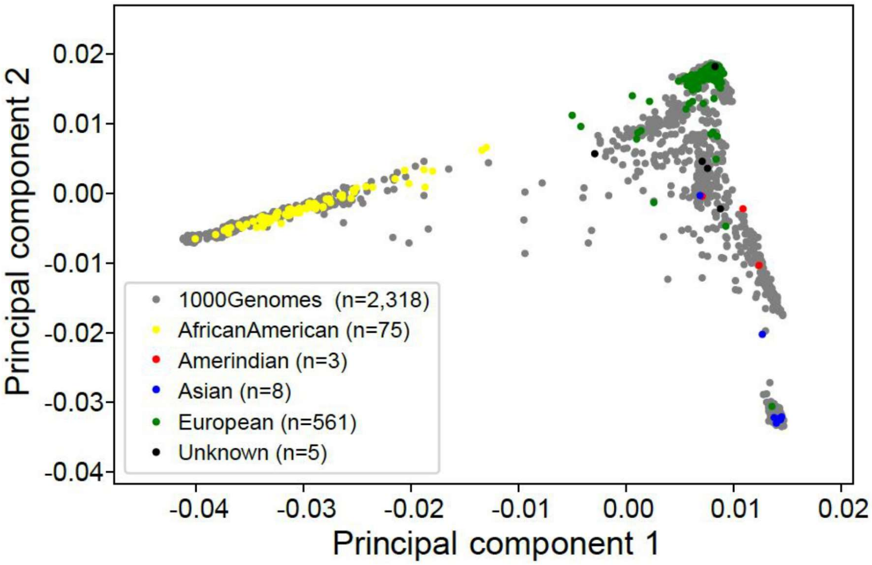
eMERGE Project - Vanderbilt University

Funding support for the Vanderbilt Genome-Electronic Records (VGER) project was provided through a cooperative agreement (U01HG004603) with the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences (NIGMS). The dataset and samples used for the VGER analyses were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant UL1RR024975 from NCRR/NIH. Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Group Health Cooperative

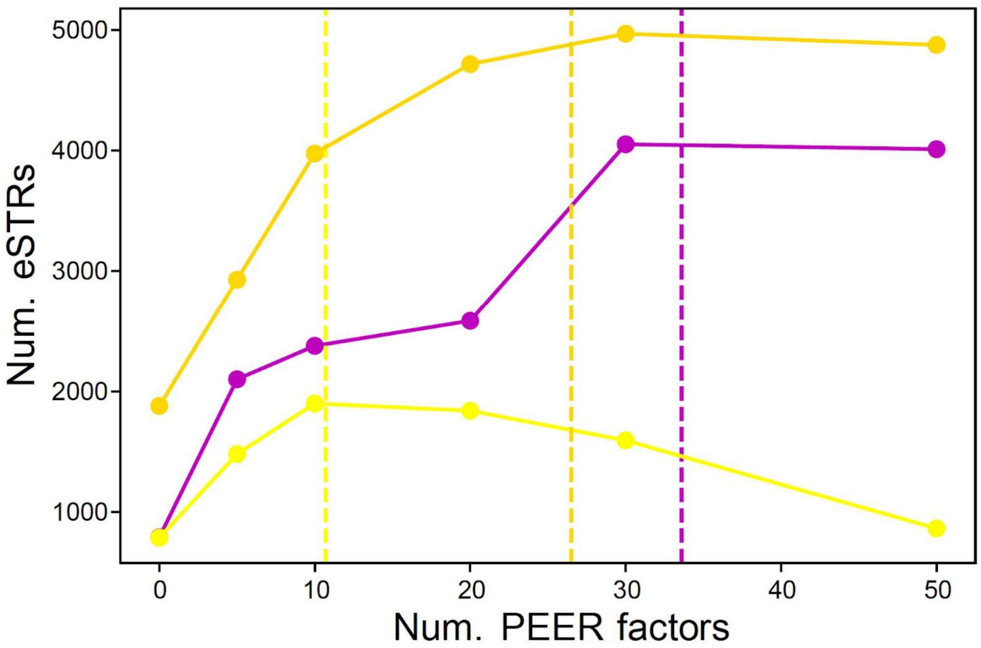
Funding support for Alzheimer's Disease Patient Registry (ADPR) and Adult Changes in Thought (ACT) study was provided by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). A gift from the 3M Corporation was used to expand the ACT cohort. DNA aliquots sufficient for GWAS from ADPR Probable AD cases, who had been enrolled in Genetic Differences in Alzheimer's Cases and Controls (Walter Kukull, PI, R01 AG007584) and obtained under that grant, were made available to eMERGE without charge. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Genome-wide association analyses were supported through a Cooperative Agreement from the National Human Genome Research Institute, U01HG004610 (Eric B. Larson, PI). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

Supplementary Figure 1: Analysis of GTEx population structure



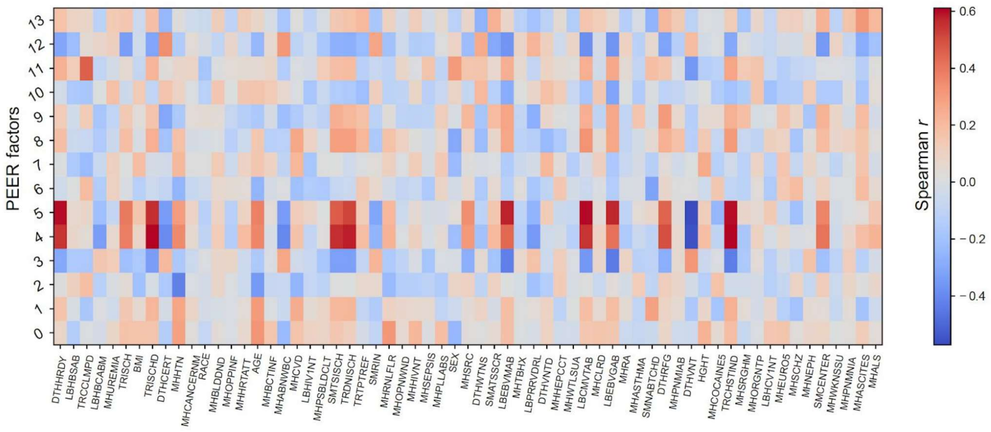
Principal component analysis was performed using SNP genotypes from the GTEx and 1000 Genomes cohorts. Samples from the 1000 Genomes project are shown in gray and GTEx samples are shown as colored dots based on ethnicity provided for each sample (yellow=African American; red=Amerindian; blue=Asian; green=European, black=Unknown).

Supplementary Figure 2: Effect of varying numbers of PEER factors on power to detect eSTRs



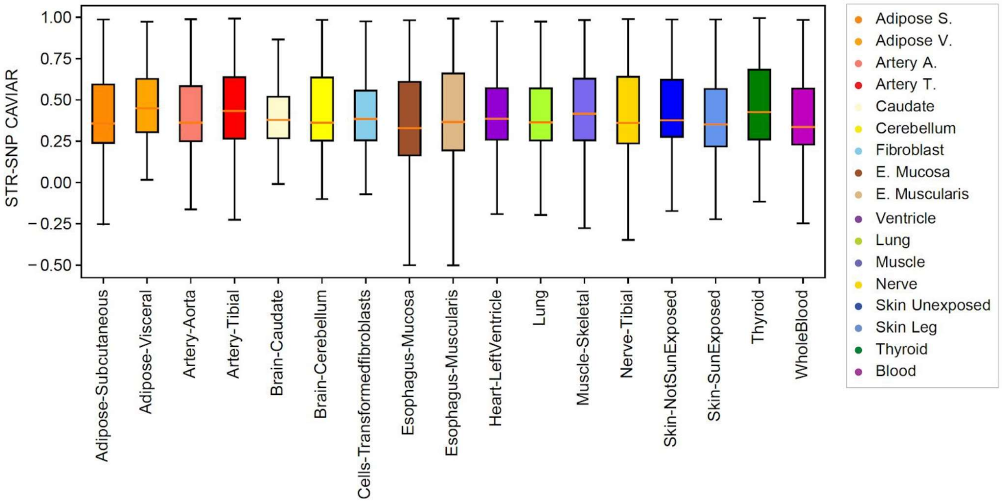
eSTRs (gene-level FDR 10%) were computed for each tissue after adjusting for a number of PEER factors ranging from 0 to 50. The numbers of STRxgene tests performed for each tissue are reported in **Supplementary Table 1**. The x-axis shows the number of PEER factors adjusted for. The y-axis shows the number of significant eSTRs at gene-level FDR of 10% (see **Methods**). Dashed vertical lines show the number of PEER factors equal to N/10, where N is the number of samples analyzed for each tissue. Purple=whole blood, yellow=Brain-Cerebellum, gold=Nerve-Tibial.

Supplementary Figure 3: Correlation of sample metadata with PEER factors



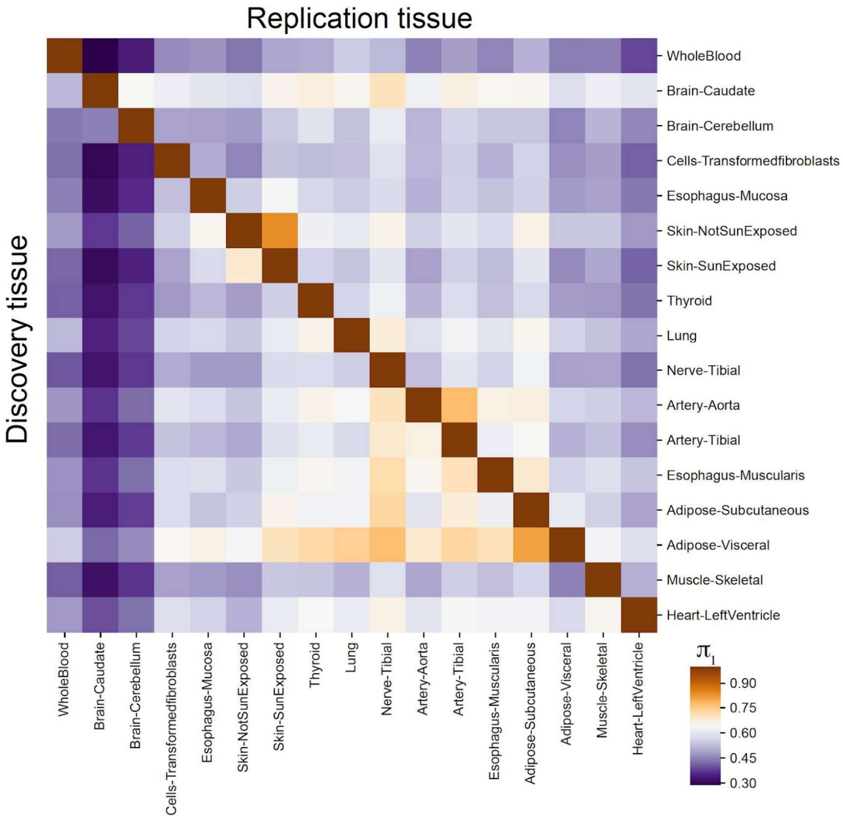
Each cell in the matrix shows the Spearman correlation of each PEER factor with data processing covariates. The x-axis represents each variable as defined for the GTEx cohort in dbGaP study phs000424.v7.p2. For example, covariates most strongly associated with PEER factors included DTHHRDY (Hardy scale for death classification) and TRISCHD (ischemic time). The y-axis represents factors obtained from PEER analysis of gene expression from Adipose-subcutaneous tissue (n=270 samples). Similar correlations were observed for other tissues.

Supplementary Figure 4: Difference in CAVIAR score between the top eSTR and top eSNP for each gene



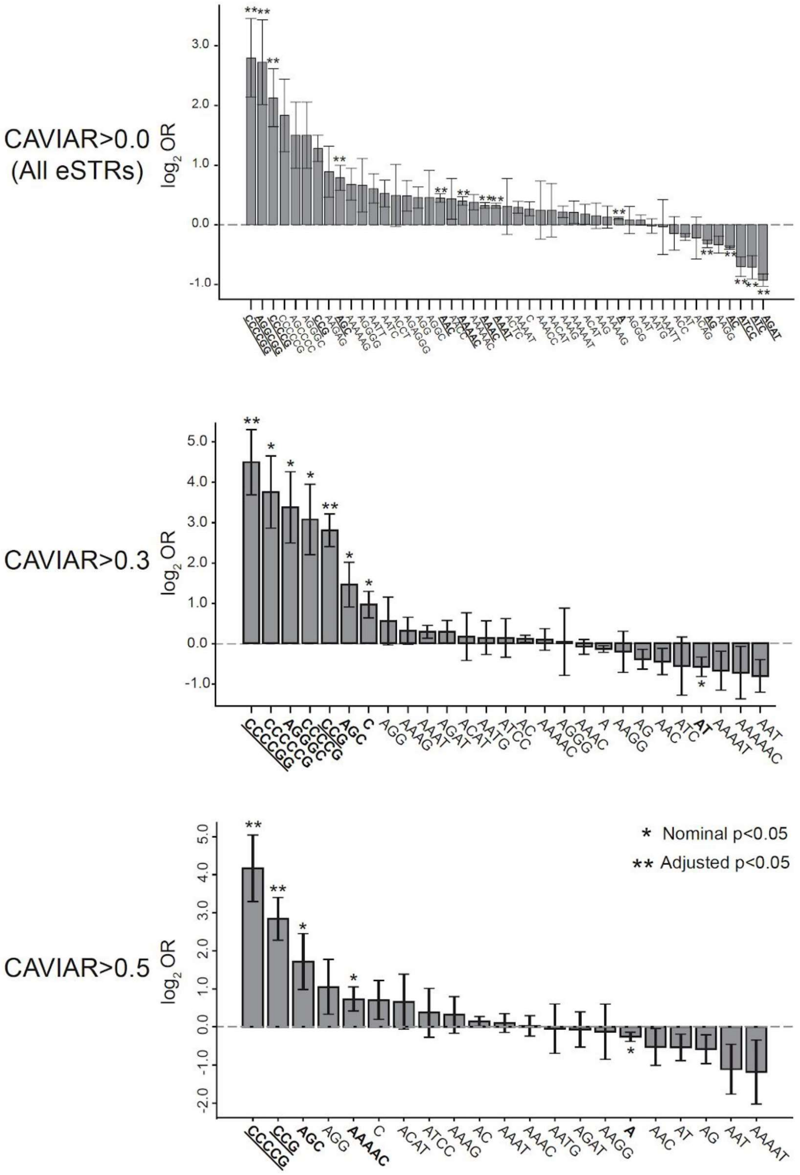
For each tissue, the boxplot shows the distribution of differences between the CAVIAR posterior score for the best STR and the best SNP for each gene. Data is only shown for genes with FM-eSTRs in each tissue. The numbers of FM-eSTRs identified in each tissue are shown in **Supplementary Table 1**. The colors of each box correspond to the different tissues (see legend on the right). Box plots are defined as in **Fig 1c.**

Supplementary Figure 5: Pairwise sharing of effect sizes across tissues



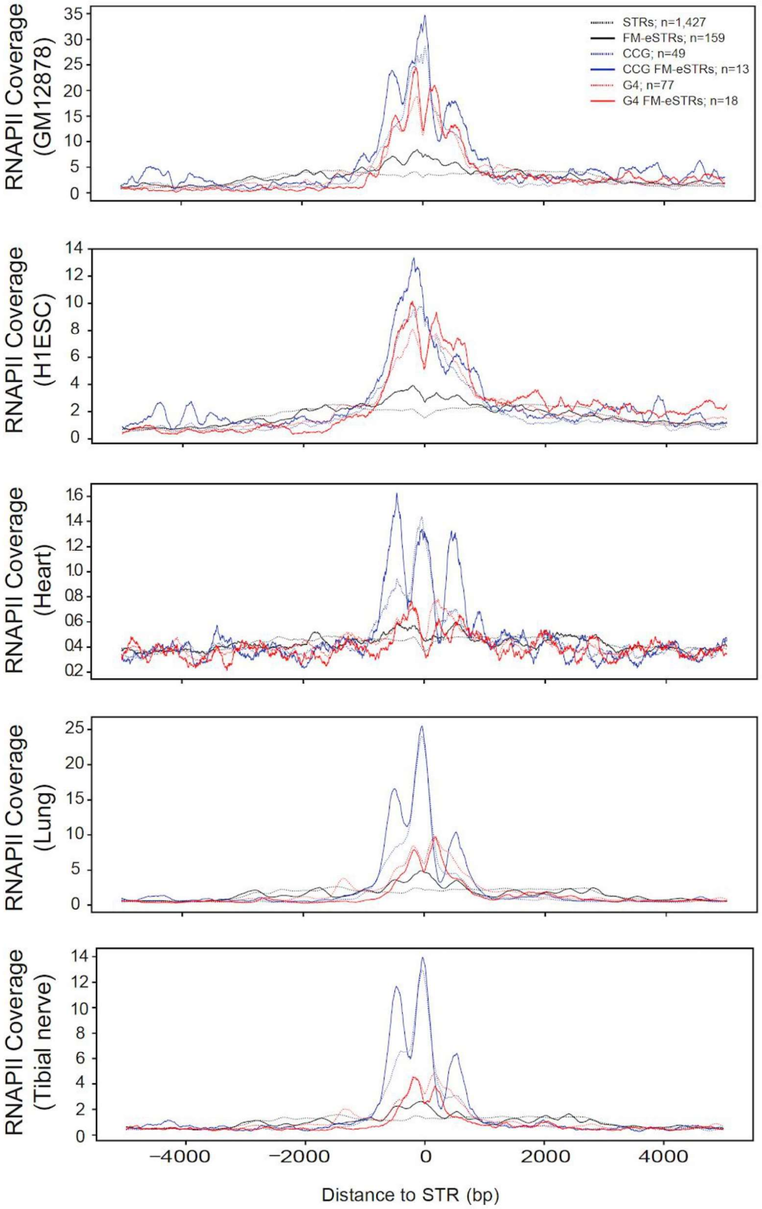
For each discovery tissue (rows), all eSTRs (gene-level FDR<10%) were tested for association in each other (replication) tissue (columns). The value in each cell gives the percent of eSTRs that were replicated (π_1). An eSTR was considered to be replicated if the eSTR measured effect size (β as defined in **Methods**) was nominally significantly different from 0 in the replication tissue (two-sided $p<0.05$). The numbers of eSTRs in each tissue are shown in **Supplementary Table 1**.

Supplementary Figure 6: Repeat unit enrichment at FM-eSTRs across all tissues for various CAVIAR thresholds.



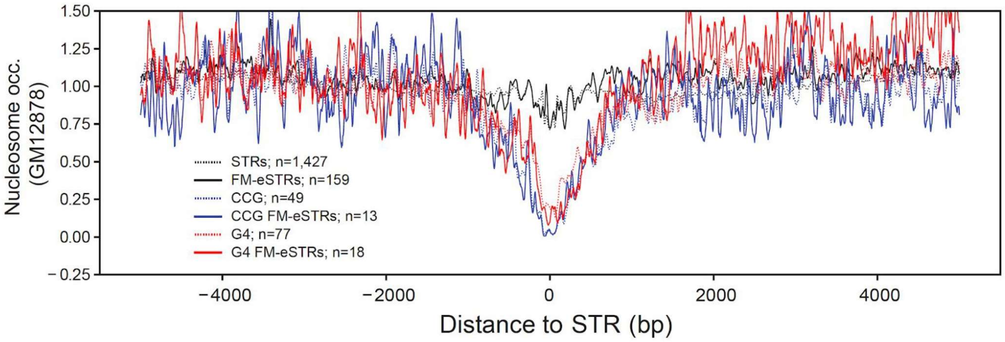
Repeat unit enrichment is shown for FM-eSTRs defined using multiple thresholds for CAVIAR scores (>0, corresponding to all eSTRs, >0.3, as used in the main text, or >0.5). The x-axis shows all repeat units for which there are at least 3 FM-eSTRs across all tissues. The y-axis denotes the $\log_2$ odds ratios (OR) from performing a Fisher's exact test comparing FM-eSTRs to all STRs. Center values are the $\log_2$ odds ratios, error bars represent +/-1 s.e. Single asterisks and bolded text denote repeat units nominally enriched or depleted (two-sided Fisher exact test $p<0.05$). Double asterisks and bolded-underlined text denote repeat units significantly enriched after controlling for the number of repeat units tested (Bonferroni adjusted $p<0.05$). Nominally significant repeat units are not denoted in the top plot due to difficulty in visualizing them for the large number of repeat units shown. The top plot only includes repeat units for which there are at least 10 FM-eSTRs across all tissues. Per repeat unit sample sizes and Fisher exact statistics are provided in **Supplementary Table 5**.

Supplementary Figure 7: Density of RNAPII localization around STRs



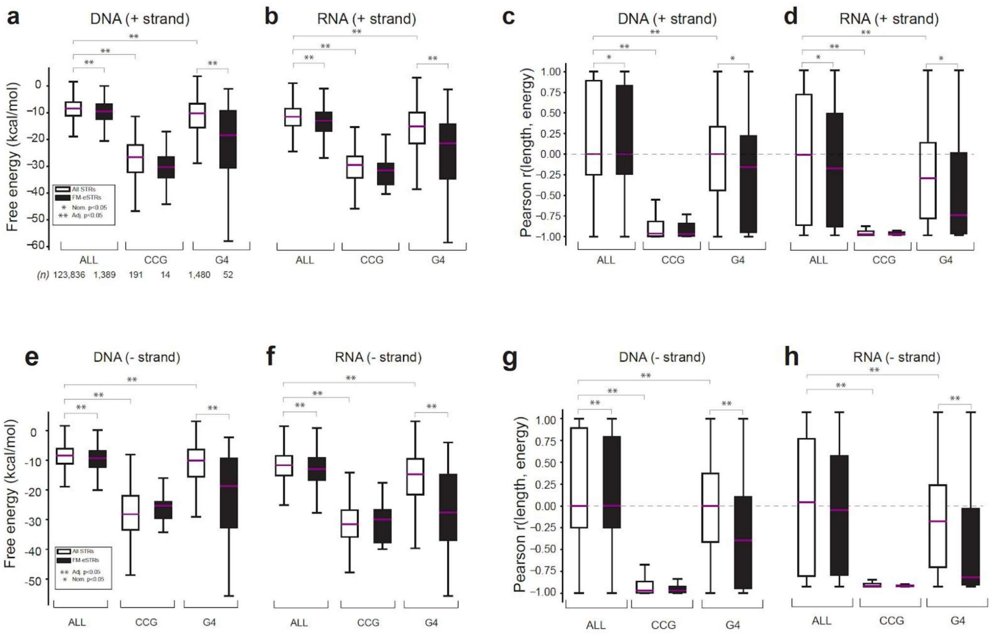
The y-axis shows the average number of ChIP-seq reads for RNA Polymerase II in 5bp bins centered at STRs within 3kb of any TSS. Black lines denote all STRs, blue lines denote CCG STRs, and red lines denote STRs matching the canonical G4 motif. Dashed lines represent all STRs of each class and solid lines represent FM-eSTRs. Plots show read counts in different cell and tissue types. From top to bottom: GM12878, human embryonic stem cells, heart, lung, and tibial nerve.

Supplementary Figure 8: Nucleosome occupancy around STRs.



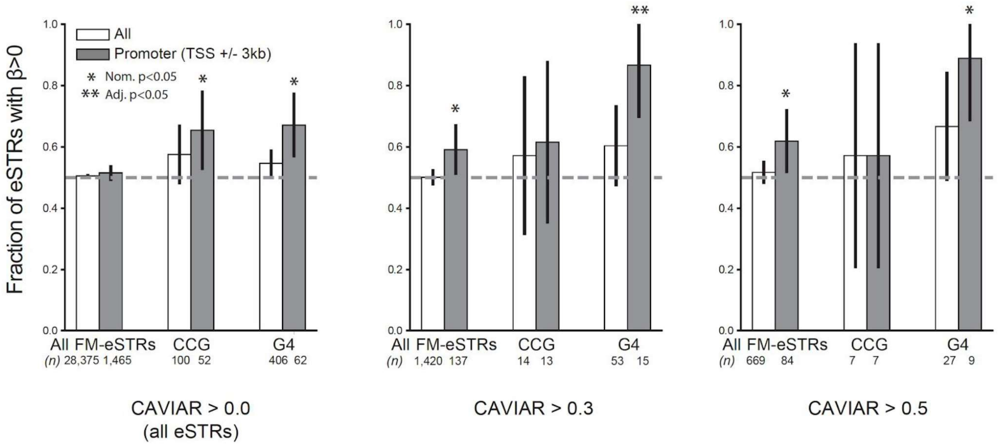
The y-axis denotes the average nucleosome occupancy in 5bp bins centered at STRs in GM12878. Black lines denote all STRs found within 3kb of any TSS, blue lines denote CCG STRs, and red lines denote STRs matching the canonical G4 motif. Dashed lines represent all STRs of each class and solid lines represent FM-eSTRs. Only STRs within 3kb of a TSS are included.

Supplementary Figure 9: GC-rich eSTRs are predicted to modulate DNA secondary structure.



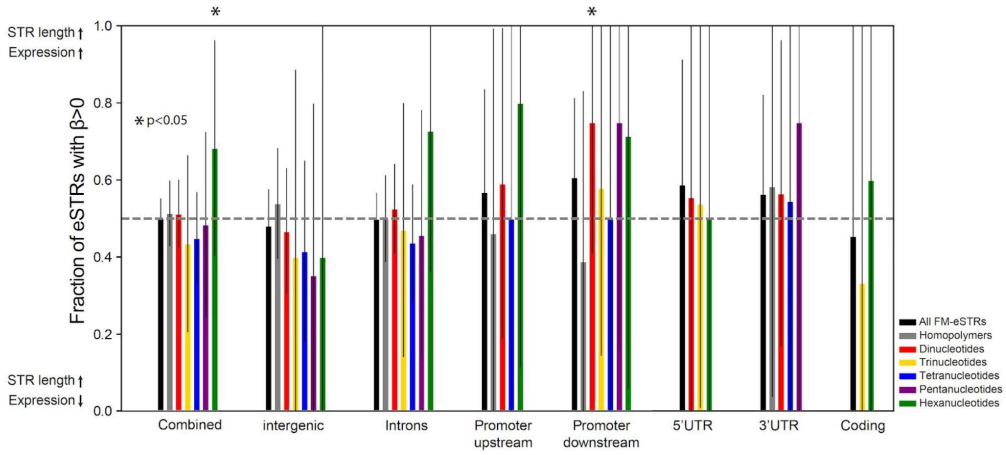
(**a-b, e-f) Free energy of STR regions.** Boxplots denote the distribution of free energy for each STR +/- 50bp of context sequence, computed as the average across all alleles at each STR. (**c-d, g-h) Pearson correlation between STR length and free energy.** Correlations were computed separately for each STR, and plots show the distribution of correlation coefficients across all STRs. The dashed horizontal line denotes 0 correlation as expected by chance. **(a)** and **(c)** show results computed using the template strand for DNA. **(b)** and **(d)** show results computed using the template strand for RNA. **(e)** and **(g)** show results computed using the non-template strand for DNA. **(f)** and **(h)** show results computed using the non-template strand for RNA. Nominally significant (Mann Whitney one-sided p<0.05) differences between distributions are denoted with a single asterisk. Differences significant after controlling for multiple hypotheses are denoted with double asterisks. For each category (free energy and Pearson correlation), we used a Bonferroni correction to control for 20 total comparisons: comparing all vs. FM-eSTRs separately in each category, comparing CCG vs. all STRs, and comparing G4 vs. all STRs, in four conditions (DNA +/- and RNA +/-). Box plots as in **Fig. 1c**. The numbers of FM-eSTRs in each distribution are shown beneath panel **a.** Numbers are identical for **b-h**.

Supplementary Figure 10: Bias in the direction of GC-rich eSTR effect sizes



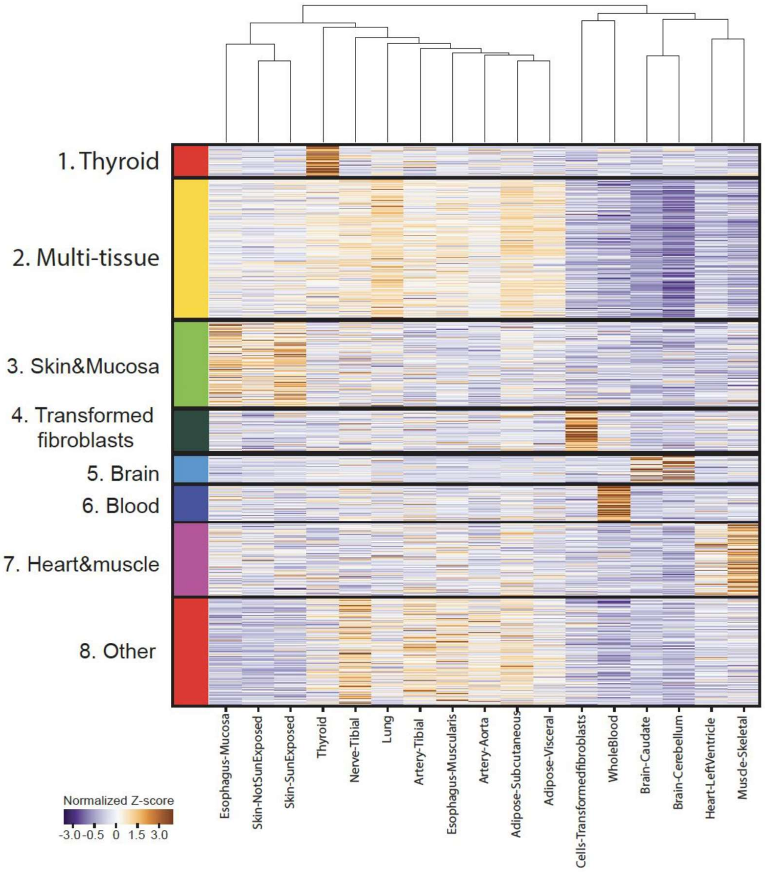
The y-axis shows the percentage of FM-eSTRs in each category with positive effect sizes, meaning a positive correlation between STR length and expression. White bars denote all STRs in each category. Gray bars denote STRs falling within 3kb of the TSS of the gene whose expression it is correlated with. Error bars give 95% confidence intervals. Results are shown for multiple thresholds for CAVIAR scores (>0, corresponding to all eSTRs, >0.3, as used in the main text, or >0.5).

Supplementary Figure 11: Bias in the direction of eSTR effect sizes



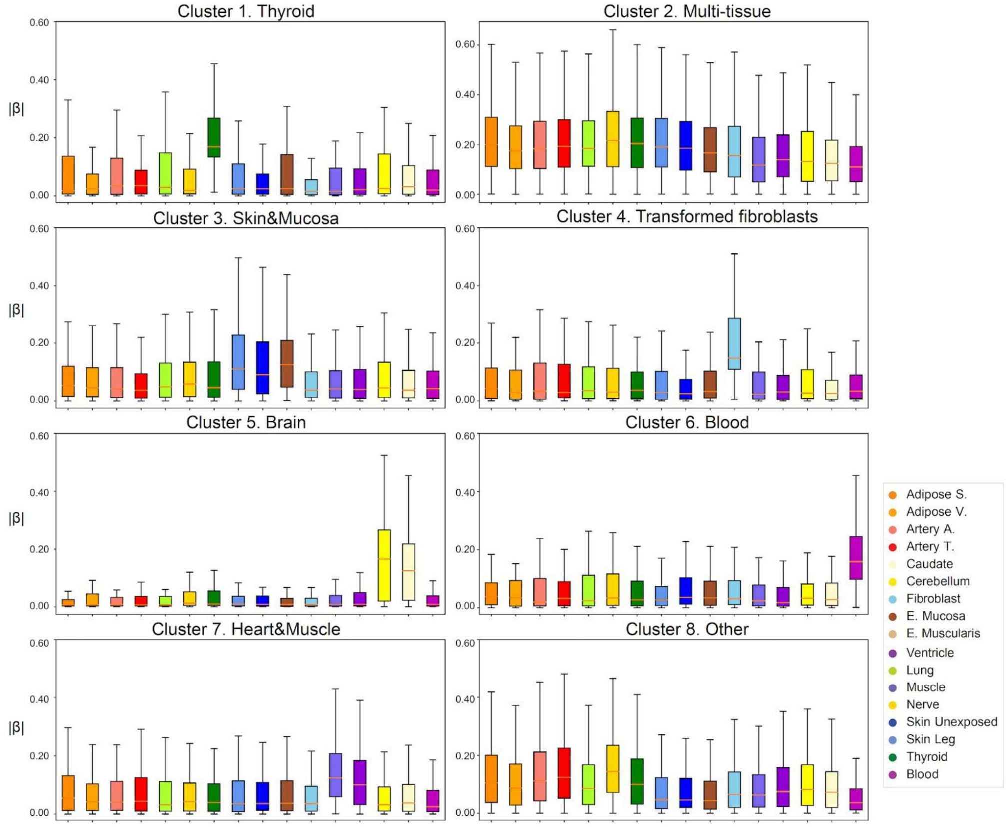
The y-axis shows the percentage of FM-eSTRs (n=1,420) in each category with positive effect sizes, meaning a positive correlation between STR length and expression. Colored bars represent different repeat unit lengths (black=all FM-eSTRs; gray=homopolymers; red=dinucleotides; gold=trinucleotides; blue=tetranucleotides; purple=pentanucleotides; green=hexanucleotides). Error bars show 95% confidence intervals. Asterisks denote categories that are nominally significant (binomial two-sided p<0.05) for having significantly more or less positive effect sizes than expected by chance (50%). No category was significant after accounting for multiple hypothesis testing. The number of FM-eSTRs in each category and breakdown by repeat unit length are shown in **Fig. 4a**.

Supplementary Figure 12: Characterization of tissue-specific FM-eSTRs



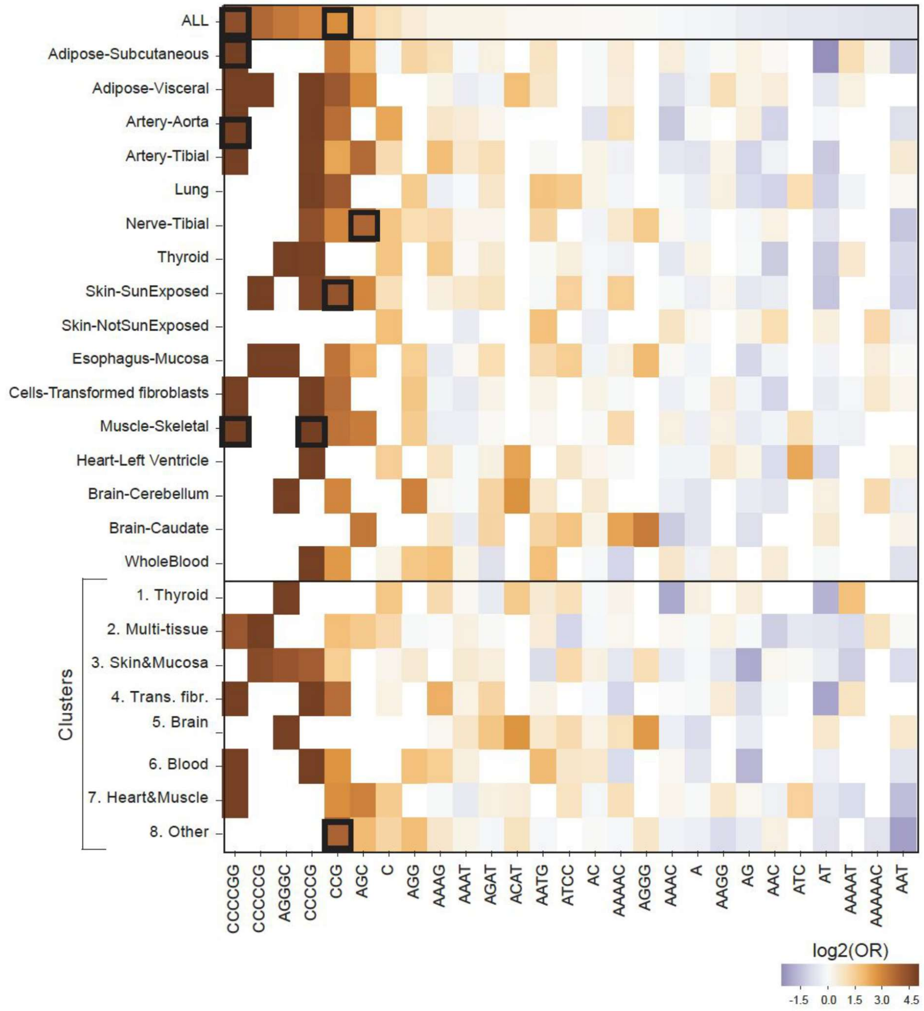
FM-eSTRs were clustered by absolute Z-scores computed by mash using K-means (**Methods**). The heatmap shows absolute values of Z-scores in each tissue, Z-normalized by row. (Number of genes in each cluster: Cluster 1=86, Cluster 2=360, Cluster 3=227; Cluster 4=106, Cluster 5=79, Cluster 6=98, Cluster 7=186, Cluster 8=278).

Supplementary Figure 13: Characterization of tissue-specific eSTR clusters



Each panel shows the distribution of the absolute value of posterior effect sizes computed by mash in each tissue for the set of FM-eSTRs in each cluster (see Supplementary **Fig. 12** above). Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (bottom) and Q3+1.5*IQR (top), where IQR gives the interquartile range (Q3-Q1). The red line shows the mean expression for each x-axis value. (Number of genes in each cluster: Cluster 1=86, Cluster 2=360, Cluster 3=227; Cluster 4=106, Cluster 5=79, Cluster 6=98, Cluster 7=186, Cluster 8=278).

Supplementary Figure 14: eSTR repeat unit enrichment



We evaluated repeat unit enrichment in multiple FM-eSTR groups: all FM-eSTRs combined across tissues (similar to **Fig. 2e**), FM-eSTRs identified per-tissue (see **Supplementary Table 1** for the total number of FM-eSTRs in each tissue), and FM-eSTRs belonging to each cluster (see **Supplementary Fig. 12** for the total number of FM-eSTRs in each cluster). For each group of FM-eSTRs, the heatmap shows the $\log_2$ of the odds ratio computed using a Fisher's Exact test (scipy.stats.fisher_exact). Columns are sorted from highest to lowest enrichment in all FM-eSTRs.

Bold boxes indicate enrichments statistically significant (adjusted $p<0.05$, adjusted separately per row for the number of repeat units tested).

Supplementary References

1. Bochman, M. L., Paeschke, K. &amp; Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet 13, 770-780, doi:10.1038/nrg3296 (2012).
2. Sawaya, S. et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One 8, e54710, doi:10.1371/journal.pone.0054710 (2013).
3. Todd, A. K., Johnston, M. &amp; Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res 33, 2901-2907, doi:10.1093/nar/gki553 (2005).
4. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406-3415, doi:10.1093/nar/gkg595 (2003).
5. Hansel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. Nat Genet 48, 1267-1272, doi:10.1038/ng.3662 (2016).

Reprinting Acknowledgements

This chapter contains a full reprint of the publication Fotsing, S. F.; Margoliash, J.; Wang, C.; Saini, S.; Yanicky, R.; Shleizer-Burko, S.; Goren, A.; Gymrek, M. The Impact of Short Tandem Repeat Variation on Gene Expression. *Nat. Genet.* **2019**, *51* (11), 1652–1659. https://doi.org/10.1038/s41588-019-0521-9.

Chapter 2: TRTools: A Toolkit for Genome-Wide Analysis of Tandem Repeats

Forward to the Reprint

This chapter contains a full reprint of the application note *TRTools: A Toolkit for Genome-Wide Analysis of Tandem Repeats* of which I was joint first author with Nima Mousavi. This application note announced the publication of the command line tool suite TRTools which accomplished two main tasks. Firstly, it provided a uniform interface for manipulating tandem repeat (TR) VCF files produced by five different TR callers, overcoming the different formats those tools use for representing TRs. Secondly, it created five command line tools for the processing and preliminary analysis of TR calls based on their lengths.

As a software package designed to support research, I judge the success of this effort by how much it assisted future projects both in our lab and the broader community. I weigh this against the considerable amount of time it took for our lab members to implement – at the time of developing TRTools, I was unsure if it was worth the effort we put in.

As of writing this thesis, Google Scholar suggests that TRTools has been cited 41 times in the less than 4 years since its release, most of which are citations from peer-reviewed publications. These citations suggest to me that TRTools was likely worth the effort from our lab. A particular highlight for me from our lab was the use of TRTools by Ziaei Jam et al.[84] in creating of a reference panel of TR genotypes from whole genome sequencing in individuals in the 1000 Genomes and H3Africa cohorts using multiple different TR callers. Highlights from outside the lab include resources such as TR calls from whole-genome sequencing of 4,000 Chinese individuals[195] and population-specific allele frequencies for 860,000 TRs from 340,000 individuals[196], as well as studies showing the potential contribution of TRs to Parkinson's disease[197] and the contribution of TRs to gene expression in colorectal cancer[198]. While TRTools only played a minor role in each of these studies, I am encouraged that it is making TR research simpler for researchers in the community. This project is also a success if its publication and existence have encouraged others to take up the study of tandem repeats, regardless of the extent to which it has featured in their publications. Further, I note personal correspondence from the authors of the long-read TR caller TRGT[199] has showed their interest in including TRGT as a supported caller in TRTools. This suggests that TRTools will continue to be useful to the research community going forward.

Google Scholar searches suggest that mergeSTR (used to merge separate TR call files produced by the same caller) and dumpSTR (used to *filter*, i.e. remove, low quality TR calls) have been the most used command line tools in TRTools. Searches also show that TRTools has been used to process VCF files produced by the AdVNTR, ExpansionHunter, GangSTR and HipSTR TR genotypers. In contrast, the command line tools qcSTR (used to produce quality control metrics for TR calls) and compareSTR (used to compare TR calls from different callers) have seen very little use, as has the ability of TRTools to take input from the popSTR TR caller. With greater foresight, we could have omitted supporting those features till they were more immediately valuable. Despite these small excess time sinks, the overall project remains a success in my mind. And as a nice side effect, tackling this project improved software development practices in our lab.

One of the biggest limitations of TRTools upon publication was its inability to perform testing for associations between traits and STR lengths. We remedied this by publishing the AssociaTR tool alongside our publication of the STR GWAS paper in Chapter 3. But AssociaTR was still at least an order of magnitude slower than standard GWAS tooling such as PLINK 2[33], and the added computational cost restricted its utility. Fortunately, following the advice of Manigbas et al.[200] who authored a recent STR GWAS, we are in the process of publishing a script within TRTools for recoding VCFs containing Beagle-imputed length dosages as PLINK 2 pgen files. This will allow the use of PLINK 2 for quickly running length-based STR GWAS, overcoming the computational limitation imposed by AssociaTR. As an added benefit, this will also allow PLINK 2 to quickly compute LD matrices between SNPs and the lengths of STRs, which is an essential preprocessing step before running statistical fine-mapping with STRs.

A limitation of TRTools that has yet to be addressed is that it is unable to process TR calls on sex chromosomes. Similarly, our analyses in Chapters 1 and 3 omitted study of the X chromosome. This omission is a casualty of the small but non-trivial effort to account for the different numbers of copies of sex chromosomes in women vs men. This omission must be fixed, both in updating TRTools, and also in including the X chromosome in our future STR GWAS and eQTL studies. Part of our responsibility when we demonstrate how to include oft-missing genetic variants in GWAS is not to omit full chromosomes from our analyses.

Bioinformatics, 37(5), 2021, 731–733
doi: 10.1093/bioinformatics/btaa736
Advance Access Publication Date: 17 August 2020
Applications Note

Genetics and population analysis

TRTools: a toolkit for genome-wide analysis of tandem repeats

Nima Mousavi[1,†], Jonathan Margoliash[2,†], Neha Pusarla[3], Shubham Saini[4], Richard Yanicky[2] and Melissa Gymrek[2,4,*]

[1]Department of Electrical and Computer Engineering, [2]Department of Medicine, [3]Department of Bioengineering and [4]Department of Computer Science and Engineering, University of California San Diego, La Jolla, 92093, USA

*To whom correspondence should be addressed.
[†]These authors contributed equally to this work.

Associate Editor: Russell Schwartz

Received on February 25, 2020; revised on August 7, 2020; editorial decision on August 10, 2020; accepted on August 12, 2020

Abstract

Summary: A rich set of tools have recently been developed for performing genome-wide genotyping of tandem repeats (TRs). However, standardized tools for downstream analysis of these results are lacking. To facilitate TR analysis applications, we present TRTools, a Python library and suite of command line tools for filtering, merging and quality control of TR genotype files. TRTools utilizes an internal harmonization module, making it compatible with outputs from a wide range of TR genotypers.

Availability and implementation: TRTools is freely available at https://github.com/gymreklab/TRTools. Detailed documentation is available at https://trtools.readthedocs.io.

Contact: mgymrek@eng.ucsd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Tandem repeats (TRs) represent one of the largest sources of human genetic variation and are well known to affect many human phenotypes (Hannan, 2018). Improvements in sequencing technology and bioinformatics algorithms have led to the recent development of a rich set of tools for performing genome-wide analysis of TR variation (Bakhtiari *et al.*, 2018; Dolzhenko *et al.*, 2017; Kristmundsdottir *et al.*, 2020; Mousavi *et al.*, 2019; Willems *et al.*, 2017). These tools take aligned sequencing reads as input and output Variant Call Format (VCF) files containing estimates of TR copy number at one or more genomic TRs. The resulting VCF files may be used for a wide variety of downstream applications. However, before doing so it is usually necessary to perform filtering, quality control (QC) and merging of files across samples. While utilities exist for performing such manipulations on VCF files containing SNP variants, these tools often do not handle multi-allelic TRs and are not designed to compute TR-specific statistics. Further, different TR genotypers use different allele annotations, complicating the use of downstream tools.

Here, we present TRTools, an open-source toolkit for performing analyses on TR genotypes. TRTools provides utilities for filtering, merging, comparing and performing QC on TR VCF files. It may be used to analyse either short tandem repeats (STRs; repeat units 1–6 bp) or variable number tandem repeats (VNTRs; repeat units >6 bp) collectively referred to here as TRs. It is currently compatible with five genotypers (GangSTR, HipSTR, ExpansionHunter, PopSTR2 and adVNTR, summarized in Supplementary Table S1) and can easily be extended to handle VCFs from additional tools.

2 Features and methods

TRTools consists of a suite of command-line utilities and a corresponding Python library for performing common operations on TR genotypes, including filtering, callset comparisons and other workflows. It parses VCF files using the PyVCF (Casbon, 2012) library and implements a 'TR harmonizer' module that converts VCF formats from each tool to a standardized representation (Supplementary Material). This harmonization step enables downstream operations to proceed agnostic of the original tool used to produce the genotypes. For all utilities described below, the –vcftype argument may be used to specify the genotyping tool used. If not specified, the type is automatically inferred. In the following sections, we summarize the current functionality available in TRTools. Utilities are summarized in Table 1. Each utility described below is available as a standalone command line tool within the TRTools package.

2.1 DumpSTR

dumpSTR is a tool for filtering TR VCF files. It performs call-level filtering (e.g. minimum call depth, minimum call quality) and locus-level filtering (e.g. minimum call rate or deviation from Hardy–Weinberg Equilibrium). dumpSTR is specially built to handle VCF FORMAT and INFO fields unique to TR genotypers. Unlike standard VCF filtering tools, it also computes locus-level metrics such as heterozygosity and Hardy–Weinberg Equilibrium based on TR allele lengths. It takes a VCF file as input and gives a new VCF with locus-level filters annotated in the FILTER column and call-level filters annotated in the FORMAT field for each call as output.

```
dumpSTR –vcf VCF –out OUTPREFIX \
  [-vcftype={eh|gangstr|hipstr|popstr|advntr}] \
  [filters]
```

2.2 MergeSTR

mergeSTR is a method for merging VCF files generated by TR genotyping methods. While methods for merging VCF files currently exist (Li, 2011), TR VCFs have unique characteristics that call for a specialized merging tool. TRs are often multi-allelic, and VCFs generated using different sample sets may contain different alternate allele sets. Further, existing tools may normalize TR alleles to remove redundant sequence when merging, which can interfere with downstream analysis of TR lengths. (For example, BCFtools (Li, 2011) normalizes REF=CAG, ALT=CAGCAG to REF=C, ALT=CAGC, which is not desirable in a TR analysis.) mergeSTR takes two or more VCFs generated by the same TR genotyper as input and a merged VCF file containing all of the samples included in the input VCFs as output.

```
mergeSTR –vcfs VCF1, VCF2[,...], VCFn \
   –out OUTPREFIX
```

Table 1. Summary of current TRTools utilities

Command	Description
dumpSTR	Filter a TR genotype dataset
mergeSTR	Merge two or more VCFs generated by a TR genotyper
statSTR	Generate per-locus statistics from a VCF of TR genotypes
compareSTR	Compare two TR genotype call sets
qcSTR	Output QC plots for a TR genotype call set

2.3 Statistics and QC utilities

TRTools provides a suite of statistics and QC utilities to allow fast high-level checks of TR runs.

statSTR allows users to compute locus-level statistics on multi-sample TR VCFs, such as the mean allele length, allele frequency distributions and call rate. It outputs a tab-delimited file listing user-specified statistics for each TR. statSTR can also output plots of allele frequency distributions at specific TRs (Fig. 1a).

```
statSTR –vcf VCF –out OUTPREFIX [statistics]
```

compareSTR allows users to compare calls from two VCF files. These can be generated by the same or different tools. This allows users to compare calls across platforms or for different runtime options. Figure 1b shows an example plot created by compareSTR comparing two call sets.

```
compareSTR –vcf1 VCF1 –vcf2 VCF2 \
   [-vcftype1 VCFTYPE] [-vcftype2 VCFTYPE] \
   –out OUTPREFIX [options] \
```

qcSTR automatically generates plots for performing quality control of TR genotype datasets. For example, Fig. 1c shows a plot demonstrating an expected deletion bias at long alleles based on popSTR2 genotypes.

```
qcSTR –vcf VCF –out OUTPREFIX [options]
```

Additional use cases for each utility using output from each supported TR genotyping tool are provided in the TRTools documentation.

2.4 Python library for data analysis

To enable researchers to leverage TRTools features in their own custom tools, we have packaged it as a Python library. The underlying functionality for operations such as harmonizing VCF records across TR genotypers or performing string manipulations on TR sequences can be accessed by importing the library into a Python script.

```
import vcf, trtools.utils.tr_harmonizer as trh
reader = vcf. Reader(open("my.vcf"))
vcftype = trh. InferVCFType(reader)
rec = reader.next()
trrecord = trh. HarmonizeRecord(vcftype, rec)
```

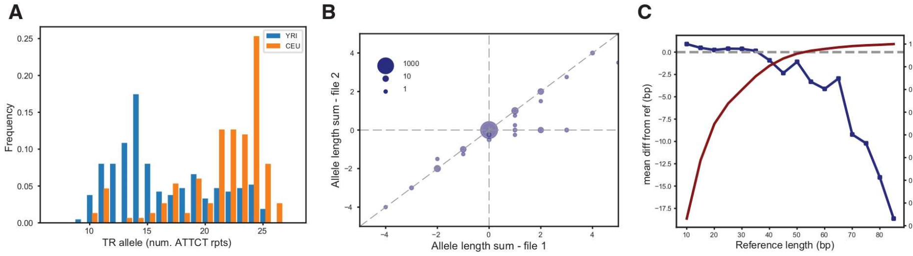


Fig. 1. TRTools visualizations. (**a**) Allele frequency distribution at an example pentanucleotide TR output by statSTR based on GangSTR genotypes for two sample sets (YRI population consisting of Yorubans from Nigeria and CEU population of Northwestern European descent). (**b**) Example TR genotype comparison output by compareSTR. The plot compares genotypes (in terms of number of repeats difference from hg19) from HipSTR (*x*-axis) to those from ExpansionHunter (*y*-axis) on 5000 tetranucleotide TRs. Bubble sizes give the number of calls included in each point. (**c**) Example reference bias plot output by qcSTR using popSTR2 genotypes. The plot shows the average deviation of TR alleles called versus the reference length of the TR (in bp). The red line shows the cumulative percentage of allele calls below each reference length threshold

```
trrecord.GetAlleleFrequencies(uselength=True)
# {10: 0.2, 15: 0.8} dict of num. rpts.->freq
```

3 Discussion

QC and filtering are crucial steps for nearly any genome- or population-scale analysis. TRTools meets a pressing need for standardized tools for performing these tasks on TR datasets, which are not handled well by mainstream tools. This toolkit currently supports five major TR genotypers. It can easily be extended to additional TR genotyping methods for either short or long reads as long as they are compatible with the VCF standard and report precise repeat copy numbers. Improved handling of imprecise repeat copy numbers and more complex repeat sequences reported by error-prone long reads is a topic for future development. Finally, TRTools can incorporate additional utilities as the community continues to develop standards for TR analysis.

Acknowledgements

The authors thank Vineet Bafna, Mehrdad Bakhtiari, Jonghun Park and Snædis Kristmundsdottir for helpful comments and sharing VCF files. Whole genome sequencing data for individuals NA12881, NA12892, and NA12878 was obtained from dbGaP (phs001224.v1.p1).

Funding

This work was supported by the National Institutes of Health [R01HG010149, DP5OD024577].

Conflict of Interest: none declared.

References

Bakhtiari,M. *et al.* (2018) Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.*, **28**, 1709–1719.

Casbon,J. (2012) *PyVCF – A Variant Call Format Parser for Python.* Available from https://pyvcf.readthedocs.io/

Dolzhenko,E. *et al.*; The US–Venezuela Collaborative Research Group. (2017) Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, **27**, 1895–1903.

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

Kristmundsdottir,S. *et al.* (2020) popSTR2 enables clinical and population-scale genotyping of microsatellites. *Bioinformatics*, **36**, 2269–2271.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Mousavi,N. *et al.* (2019) Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.*, **47**, e90–e90.

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

Supplementary Material - TRTools

Datasets

BAM files containing reads from high-coverage whole genome sequencing datasets for the 1000 Genomes Project (Consortium *et al.*, 2015) were accessed through the European Nucleotide Archive accession number `PRJEB31736`. They were processed using GangSTR (Mousavi *et al.*, 2019) v2.4.2.12 with non-default parameter `--grid-threshold 250` using the TR reference file `hg38_ver17.bed.gz` available on the GangSTR website (https://github.com/gymreklab/gangstr). Allele frequencies for a pentanucleotide repeat in the promoter of *RUNX1* (hg38 chr21:35348646-35348646) for samples from the YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah Residents with Northern and Western European Ancestry) populations are shown in Fig. 1a in the main text.

Whole genome sequencing (BAM file aligned to hg19) for Platinum Genomes sample NA12881 was downloaded from dbGaP (accession phs001224.v1.p1). A single chromosome (chromosome 10) was extracted using samtools (Li *et al.*, 2009). TRs were genotyped using ExpansionHunter (Dolzhenko *et al.*, 2017) v3.2.0 and HipSTR (Willems *et al.*, 2017) v0.6.2. For both, we used the GangSTR version 16 reference subsetted to the first 5,000 tetranucleotides as the input set of TRs. ExpansionHunter was run with parameter `-a path-aligner`. HipSTR was otherwise run with default parameters. We used dumpSTR to filter each callset and compareSTR to generate the bubble plot in Fig. 1b using the options shown below

A VCF file generated by popSTR2 (Kristmundsdottir *et al.*, 2019) on Platinum Genomes samples NA12891, NA12892, and NA12878 (dbGaP phs001224.v1.p1) was obtained from the PopSTR authors. This file was used to generate the reference bias plot shown in Fig. 1c in the main text.

TR Harmonizer implementation details

The TRHarmonizer Python library provides a uniform interface for accessing VCFs created by different tandem repeat (TR) genotypers. This library is the shared basis for all the command-line tools in the TRTools package. It is designed to cleanly handle differences in how different genotypers represent alleles, quality-scores and other metadata describing TR genotypes. This allows coding against a uniform interface while analyzing genetic variation at TRs regardless of which genotyper was used. The TRHarmonizer library also allows third parties to leverage the harmonization functionality outside of the command-line tools provided in TRTools.

A major challenge in analyzing TR genotypes is that alleles are represented differently in VCF outputs of different genotypers. The example below for chr21:47251618 (hg19) genotyped in Platinum Genomes sample NA12878 shows the different ways reference and alternate alleles are specified in VCFs by the genotypers which TRTools currently supports.

- adVNTR*, GangSTR, HipSTR
  - REF: AGTTAGTTAGTTAGTT
  - ALT: AGTTAGTTAGTTAGTTAGTT
- ExpansionHunter
  - REF: A
  - ALT: <STR5>
  - INFO: REF=4;RU=AGTT
- PopSTR
  - REF: AGTTAGTTAGTTAGTT

  - ALT: <5>
  - INFO: Motif=AGTT

* Note that while this TR was not called by AdVNTR because its motif is too short, AdVNTR output represents alleles in the same format as HipSTR and GangSTR.

Furthermore, consider the example at chr21:16402147:

- adVNTR, GangSTR, HipSTR
  - REF: AAATAAATAAATAAATAAAT
  - ALT: AAATAAATAAATAAAT
- ExpansionHunter
  - REF: A
  - ALT: <STR4>
  - INFO: REF=5;RU=AAAT
- PopSTR
  - REF: AAATAAATAAATAAATAAATAATAAA
  - ALT: <5.5>
  - INFO: Motif=AAAT

Here, popSTR's representation of alleles changes to specify impurities and partial repeats.

The key function of the TRHarmonizer module, `HarmonizeRecord`, takes as input a PyVCF Casbon (2012) record (a `PyVCF.model._Record object`) and a VCF type (one of: "advntr", "eh", "gangstr", "hipstr" or "popstr", corresponding to the supported genotypers) and outputs a TRRecord object (analogous to `PyVCF.model._Record`) storing alleles and other metadata in a standardized format. This allows downstream analyses to proceed agnostic of the genotyper which created the record. The TRRecord stores allele length genotypes as the number of copies of the motif corresponding to that length. This number is a float to allow for impurities and partial repeats. For genotypers which infer sequence alleles, the record additionally stores the sequence of the allele in all uppercase. In addition to alleles, a TRRecord also provides a uniform method for accessing the TR motif, per-sample quality scores and other metadata supplied by the underlying genotyper. The main text and TRTools documentation show examples of how to use the TRHarmonizer interface from Python.

TRHarmonizer is designed to be lightweight, and as such there are similar yet more complex use-cases that TRHarmonizer intentionally does not support. It does not have any insight into sequencing technologies which produce data that is later processed by TR genotypers into VCFs. As such it relies on the alleles, calls and associated quality scores output by the genotypers, each of which use their own models to compute quality scores. TRTools makes no attempt to modify those scores based on sequencing errors or other sources of error.

TRHarmonizer also does not handle differences in variant coordinates, whether due to differences in choice of variant reference set or differences between calling algorithms. Note that this is only relevant to compareSTR, as that is the only one of our tools designed to process TRs from multiple VCFs produced by different genotypers simultaneously. The types of differences related to variant coordinates that TRHarmonizer does not handle includes:

- Repeat regions which some callers choose to represent as a single variant and other callers represent as multiple variants
- Overlapping variants of different lengths due to decisions about whether to phase the repeat variant with other nearby variants
- Overlapping variants of different lengths due to different choices as to which parts of a locus constitute impure repeats and which constitute flanking regions

Rather, TRHarmonizer restricts itself to comparing variants called by different callers whose reference alleles start and end at the same base pairs. Handling different variant representations is a complex problem that has been the subject of significant work Cleary *et al.* (2015); Krusche (2010) and is best handled by haplotype comparison tools which have been tailored to the specific use-case at hand.

Finally, TRHarmonizer can be readily extended to support any TR genotyping tool built on top of any sequencing or genotyping technology as long as the tool produces a valid VCF file representing each TR as a distinct record in the VCF. Supporting additional tools simply requires adding a short function to the TRHarmonizer module converting records to the standardized format described above.

Commands for generating figures

The following code snippets show the commands used to generate the figures in the main text.

Listing 1: Code to generate Fig. 1a

```
#!/bin/bash
# YRIVCF and CEUVCF were generated by GangSTR v2.4.2.12
REGION=chr21:35348646-35348646
tabix --print-header $YRIVCF $REGION | bgzip -c > yri_runx1.vcf.gz
tabix --print-header $CEUVCF $REGION | bgzip -c > ceu_runx1.vcf.gz
tabix -p vcf yri_runx1.vcf.gz
tabix -p vcf ceu_runx1.vcf.gz

# Merge
mergeSTR --vcfs yri_runx1.vcf.gz,ceu_runx1.vcf.gz --out yri_ceu_runx1
bgzip -f yri_ceu_runx1.vcf
tabix -p vcf -f yri_ceu_runx1.vcf.gz

# Get sample lists
bcftools query -l yri_runx1.vcf.gz > yri_samples.txt
bcftools query -l ceu_runx1.vcf.gz > ceu_samples.txt

# StatSTR
# Compute stats separately on YRI and CEU samples
statSTR \
    --vcf yri_ceu_runx1.vcf.gz \
    --samples yri_samples.txt,ceu_samples.txt --sample-prefixes YRI,CEU \
    --region $REGION \
    --out yri_ceu_runx1 \
    --afreq --use-length --plot-afreq
# Output file yri_ceu_runx1-chr21-35348646.pdf shown in Fig. 1a
```

Listing 2: Code to generate Fig. 1b

```
#!/bin/bash

SAMPLE=NA12881
# $SAMPLE-hipstr.vcf.gz and $SAMPLE-eh-path.vcf.gz generated
# by calling HipSTR and ExpansionHunter on the same TR reference

# Filter
dumpSTR \
    --vcf $SAMPLE-hipstr.vcf.gz \
    --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 \
    --hipstr-min-supp-reads 2 \
    --out $SAMPLE-hipstr.filtered
cat $SAMPLE-hipstr.filtered.vcf | vcf-sort | \
        bgzip -c > $SAMPLE-hipstr.filtered.vcf.gz
tabix -p vcf $SAMPLE-hipstr.filtered.vcf.gz

dumpSTR \
    --vcf $SAMPLE-EH-path.vcf \
    --vcftype eh \
    --eh-min-call-LC 50 \
    --out $SAMPLE-eh-path.filtered
# Edit sample name to be same
cat $SAMPLE-eh-path.filtered.vcf | sed ’s/NA12881.chr10/NA12881/’ | \
        vcf-sort | bgzip -c > $SAMPLE-eh-path.filtered.vcf.gz
tabix -p vcf $SAMPLE-eh-path.filtered.vcf.gz

# Add contigs to EH
zcat $SAMPLE-hipstr.filtered.vcf.gz | grep congi > hg19_contigs.txt
bcftools annotate -h hg19_contigs.txt $SAMPLE-eh-path.filtered.vcf.gz | \
        bgzip -c > $SAMPLE-eh-path-contigs.filtered.vcf.gz
tabix -p vcf -f $SAMPLE-eh-path-contigs.filtered.vcf.gz

# Make bubbles plot and compare
compareSTR \
    --vcf1 $SAMPLE-eh-path-contigs.filtered.vcf.gz \
    --vcf2 $SAMPLE-hipstr.filtered.vcf.gz \
    --vcftype1 eh \
    --vcftype2 hipstr \
    --out eh-path-hipstr \
    --bubble-min -5 --bubble-max 5
# Output file eh-path-hipstr-bubble-periodALL.pdf shown in Fig. 1b
```

Listing 3: Code to generate Fig. 1c

```
#!/bin/bash
qcSTR --vcf $popSTR_vcf --out popstr_qc
# Output file popstr_qc-diffref-bias.pdf shown in Fig. 1c
```

Supplementary Tables

Supplementary Table 1: TR calling methods currently supported by TRTools

Method (Version tested)	Supported TR classes	Num. TRs in reference	Supported sequencing technologies	Use case notes
AdVNTR Bakhtiari *et al.* (2018) (v1.3.3)	Repeat unit length 6-100bp.	158,522 (genic hg19)	Illumina or PacBio	Designed for targeted genotyping of VNTRs on a single sample at a time. Only handles repeats shorter than the read length. Infers allele lengths by default. May alternatively identify putative frameshift mutations within VNTRs. May be run on large panels of TRs but is compute-intensive.
Exp. Hunter Dolzhenko *et al.* (2017) (v3.2.2)	Designed for STRs (typically with repeat unit length ≤6bp). Can handle complex repeat structures specified by regular expressions (e.g. (CAG)*(CCG)*).	25 (hg19)	PCR-free* Illumina	Designed for targeted genotyping of repeat expansions at known pathogenic TRs but may be run genome-wide on both short and expanded TRs using a custom TR panel. Can handle repeats with complex structures such as interruptions or nearby repeats.
GangSTR Mousavi *et al.* (2019) (v2.4.4)	Designed for STRs or VNTRs with repeat unit length 1-20bp.	829,233 (hg19_ver_13_1, excludes homopolymers)*	Paired-end PCR-free* Illumina	Designed for genome-wide genotyping of short or expanded TRs. Infers allele lengths. Allows multi-sample calling.
HipSTR Willems *et al.* (2017) (v0.6.2)	Repeat unit length 1-9bp.	1,620,030 (hg19)	Illumina	Designed for genome-wide genotyping of STRs shorter than the read length. Can phase repeats with SNPs. Allows multi-sample calling.
PopSTR2 Kristmundsdottir *et al.* (2019)	Repeat unit length 1-6bp.	540,1401 (hg38)	Illumina	Designed for genome-wide genotyping of short or expanded TRs. Allows multi-sample calling.

*These tools may be run on Illumina data that is not PCR-free, but may have reduced accuracy on those datasets.

Since each of these tools take as input a list of TRs to genotype, they can also be used on custom panels of TR loci. Tool information and reference panel numbers shown above are based on downloads from the github repository of each tool as of July 2, 2020.

References

Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018). Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.*, **28**(11), 1709–1719.

Casbon, J. (2012). *PyVCF - A Variant Call Format Parser for Python*.

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Rathod, M., Ware, D., Zook, J. M., Trigg, L., and De La Vega, F. M. (2015). Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv*.

Consortium, . G. P. *et al.* (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H. P., van Es, M. A., McLaughlin, R., Sproviero, W., Shatunov, A., Jones, A., Al Khleifat, A., Pittman, A., Morgan, S., Hardiman, O., Al-Chalabi, A., Shaw, C., Smith, B., Neo, E. J., Morrison, K., Shaw, P. J., Reeves, C., Winterkorn, L., Wexler, N. S., Housman, D. E., Ng, C. W., Li, A. L., Taft, R. J., van den Berg, L. H., Bentley, D. R., Veldink, J. H., and Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, **27**(11), 1895–1903.

Kristmundsdottir, S., Eggertsson, H. P., Arnadottir, G. A., and Halldorsson, B. V. (2019). popSTR2 enables clinical and population-scale genotyping of microsatellites. *Bioinformatics*.

Krusche, P. (2010). *Haplotype Comparison Tools*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.*, **47**(15), e90.

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**(6), 590–592.

Reprinting Acknowledgements

This chapter contains a full reprint of the publication Mousavi, N.; Margoliash, J.; Pusarla, N.; Saini, S.; Yanicky, R.; Gymrek, M. TRTools: A Toolkit for Genome-Wide Analysis of Tandem Repeats. *Bioinformatics* **2021**, *37* (5), 731–733. https://doi.org/10.1093/bioinformatics/btaa736.

Chapter 3: Polymorphic Short Tandem Repeats Make Widespread Contributions to Blood and Serum Traits

Forward to the Reprint

This chapter contains a full reprint of the paper *Polymorphic Short Tandem Repeats Make Widespread Contributions to Blood and Serum Traits* of which I was first author, and which was the subject of a complimentary perspective by a fellow researcher[201]. In this paper we provide evidence of widespread contribution of STRs to human blood cell counts and serum biomarkers, estimating that 5.2-7.6% of identifiable causal variants for these traits are STRs. We use statistical fine-mapping to identify many STR-trait associations with strong statistical evidence for causality and strong impacts on the studied trait. And we corroborate some of these candidate causal associations with plausible mechanistic hypotheses. (As an aside: in this paper we use the term fine-mapping to refer specifically to statistical fine-mapping).

A significant portion of the effort I put into this paper was devoted to simply getting the STR-length association testing pipeline to run. I describe the challenges I faced with bioinformatics pipelining tools more in the thesis Discussion. The software to perform the association tests themselves was developed both during this project (published as AssociaTR in TRTools) and has been updated since to utilize the speed of PLINK 2[33], see the Chapter 2 Forward for more information. I hope that the results here, coupled with the ease of our new integration of length-based testing with PLINK 2, will make length-based STR GWAS standard for GWAS practitioners going forward.

In this paper we caveat these length-based STR GWAS efforts saying that “no new strong peaks were identified only by STRs (Figure 1F), which is unsurprising, since the STRs were imputed from SNP genotypes.” In the time since writing that statement, I have come to a different perspective. Both the SNPs and indels we tested as well as the STRs we tested were imputed from the same phased microarray variants, yet there are many loci where SNP signal strengths outstrip STR signal strengths. Thus imputation cannot be the reason why conversely no STR signals stand out from the SNP signals. Instead, I now hypothesize that this is merely due to the quality of the reference panels used to impute these classes of variants. 45% of the variants imputed by the UK Biobank team[53] were sourced from the Haplotype Reference Consortium panel which contains ~32,500 individuals. By comparison, the Saini et al. SNP-STR reference panel[83] we used was itself imputed from less than one thousand individuals. All this is to say that there is decent possibility that length-based GWAS of STRs imputed from newer reference panels, such as the Ziaei Jam et al. panel[84], may in fact be able to detect some STR length signals whose strength stands out from other nearby associations. And of course, this probably will be the case for some associations with STRs called from WGS data.

Most of the other major challenges we faced in this project stemmed from statistical fine-mapping. We approached statistical fine-mapping from a straightforward perspective – we wanted to run the fine-mapping tools and use their results to highlight likely causal STRs. But as time passed and we analyzed more fine-mapping runs it became apparent that which tool we used, or which runs of each tool we drew results from had large impacts on which variants were being highlighted as causal. We eventually concluded that SuSiE and FINEMAP are frequently non-concordant with one another, that FINEMAP has a previously unacknowledged built-in level of instability, and that these inconsistencies, when they occur, are rather dramatic. While these conclusions were important to our results and were important to share with the research community, the process through which we came to these conclusions was long and indirect. It is exciting to see a recent paper by Cui et al. which addresses the challenge of fine-mapping inconsistencies more directly[168]. Further, these previously hidden fine-mapping inconsistencies have reinforced my belief that statistical fine-mapping tools need to be benchmarked. I consider benchmarking and the Cui et al. paper more in the thesis Discussion.

Another statistical fine-mapping challenge we ran into was how to incorporate orthogonal sources of data. Over the course of this project we measured trait associations for each STR in population groups other than the White British population, we measured STR length associations with expression of nearby genes, and we measured STR length associations with the methylation of nearby CpG bonds. However, it was unclear whether to integrate any of this data into our statistical fine-mapping efforts, as each of these other data sources had 30 to 1,000 times fewer individuals than the White British dataset.

If fine-mapping with data from other populations identified different causal variants than fine-mapping without that data, we would have struggled to distinguish if that was due to lack of power in the other population groups or truly different signal patterns in those groups. If the two fine-mapping results were the same, we would have struggled to determine if that was solely due to the overpowering amount of White British data or if the identified variants actually showed the same signal across populations. And it was unclear if our fine-mapping would become more unreliable due to model failures under this regime of massive data size disparities.

Further, while those struggles would have existed for the case of multi-ethnic fine-mapping, where we could expect causal variants to largely be shared across populations, integrating eQTL and mQTL data into a multi-trait fine-mapping effort would have introduced even more challenges. On top of the existing struggles with multi-ethnic fine-mapping, here we would additionally have had to set a prior for how frequently we expected causal GWAS signals to share or not share corresponding causal mQTL or eQTL signals. In the end, we decided to look for broad enrichments in replication rates across population groups (Chapter 3 Figure 3, Supplementary Figure 3 and Supplementary Table 10). We did not systematically account for rates of sharing between GWAS and eQTL and mQTL signals, and did not incorporate these additional sources of data into our statistical fine-mapping. It would be valuable if future projects could shed better light on how to perform such statistical fine-mapping with massively disparate dataset sizes.

Another fine-mapping challenge arose during when we came across an association between the length of an STR in the gene *PACSIN2* and the phenotype *platelet (size) distribution width*, i.e. the width of the distribution of the sizes of platelets in any one person. This was initially one of the most compelling signals we found, and we planned to highlight it in the results section of our paper. We even produced most of a preliminary figure for this finding.

(Continued on the following page …)

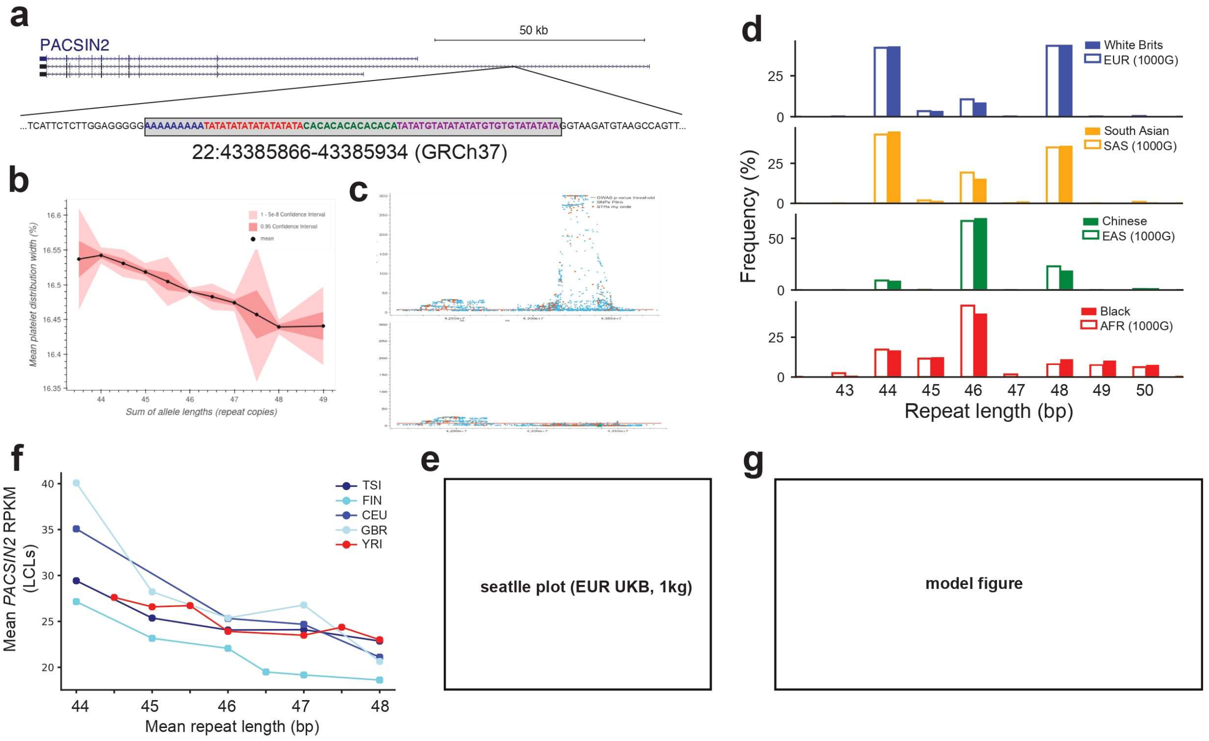


Figure 3.Forward.1: A preliminary draft of a figure showing that the total length of a compound STR in an intron of *PACSIN2* is strongly associated with platelet distribution width. (a) We highlight four consecutive STRs making up the compound STR in the *PACSIN2* region. This compound repeat falls in an intron of the longest of the displayed *PACSIN2* isoforms. Note that the purple repeat on the right is not a pure STR, as the base between the Ts alternates between As and Gs at different locations. Also note that there is a poly-G STR before the four highlighted STRs in this figure, making this a compound of five individual repeats. **(b)** Total length of the compound STR vs platelet distribution width among White British participants that passed quality control. This graph is unadjusted for covariates. Genotypes with total population dosage less than 0.1% of all alleles are omitted for clarity. **(c)** The *PACSIN2* association region, before and after conditioning on the total length of the compound STR. **(d)** The total length of the compound STR in different populations, in both the 1000 Genomes[13] cohort (outlined) and the UK Biobank cohort (solid) – see the paper methods for population definition specifics. **(e)** This was going to contain some version of Figure 3.Forward.2 below **(f)** Total length of the compound STR vs *PACSIN2* expression in different 1000 Genomes populations with Geuvadis[202] expression data. All populations represented are European except for YRI, which are Yoruba African individuals. RPKM: reads per kilobase million, a standard unit of gene expression. **(g)** We had planned to identify a potential mechanism of action for this association and would have highlighted it here.

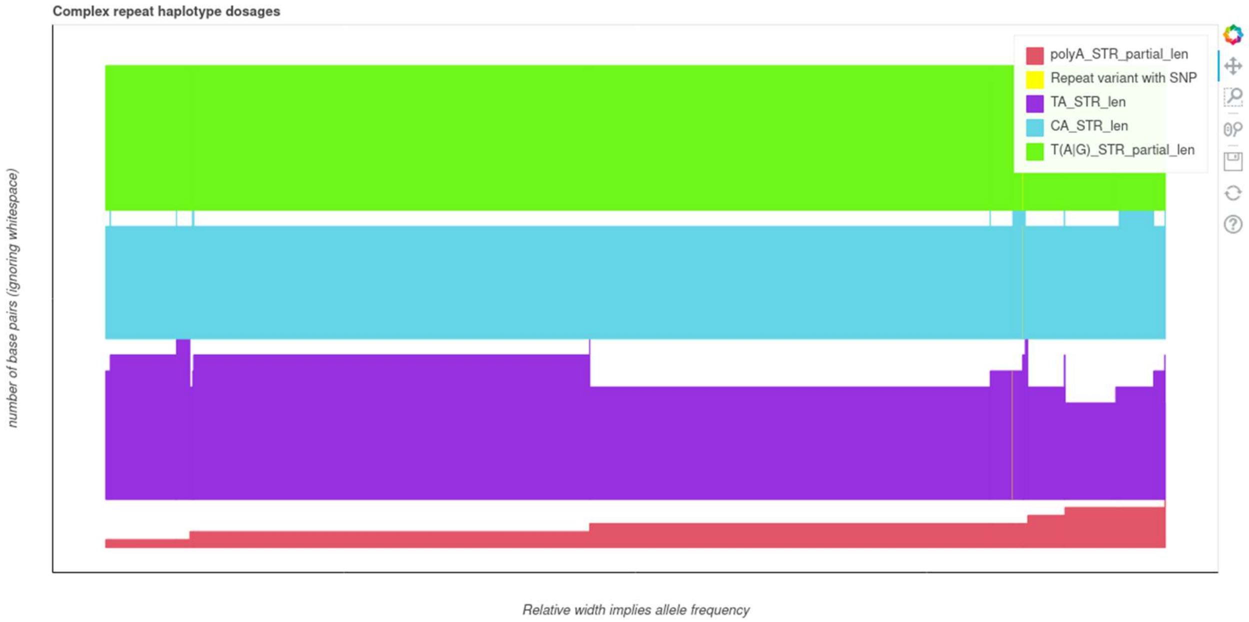


Figure 3.Forward.2: A preliminary draft of a figure showing the different lengths of the component STRs of the compound *PACSIN2* STR. Each column represents an individual in our cohort. The height of each colored section for an individual indicates the number of repeats of that STR in that individual, and the total height of all of these columns, ignoring whitespace indicates the total measured length of the repeat in that individual. The thin golden strips indicate individuals with impure SNP variation within the corresponding STRs. The reference panel from which we imputed STRs had calls for the full lengths of the TA and CA STRs, as well as flanking bases from the poly-A and T(A|G) STRs (read: T followed by an A or a G). It is likely, though not assured, that the TA and CA STR lengths reliable, and unclear if the flanking poly-A or T(A|G) lengths correspond to the relative lengths of those full STRs or not. Note that the T(A|G) lengths were uniform in this data.

The *PACSIN2* STR was one of the strongest STR associations we identified with good fine-mapping evidence (SuSiE CP=1 and FINEMAP CP=0.71; CP is defined in the paper), having $p < 1 \times 10^{-300}$ (beyond our pipeline's numeric precision; we had yet to modify our pipeline to work with z-scores to avoid this issue). The association was linear across a range of lengths (Chapter 3 Forward Figure 1b) and conditioning on the length of this STR completely accounted for the full association signal in the region (Chapter 3 Forward Figure 1c). The STR was also strongly associated with expression of *PACSIN2* in the Geuvadis cohort[202] across many European populations, as well as non-negligible association with *PACSIN2* in the Yoruba African Geuvadis population (Chapter 3 Forward Figure 1f).

However, a cursory examination of this locus showed that the STR we were associating with platelet distribution width was in fact a *compound* STR, consisting of five STRs back-to-back-to-back (Chapter 3 Forward Figure 1a). We were measuring the lengths of between two and four of these individual STRs (Chapter 3 Forward Figure 2; the STR reference panel had only designated the internal TA and CA STRs for calling, and the calls happened to contain some of the bases from the two flanking STRs). What we were associating with platelet distribution width was the total measured length of all these STRs.

While this total length was confidently fine-mapped and fully explained the signal, our first concerning result was that it took independent conditioning on the lengths of every individual STR to achieve similar signal reduction as conditioning on the top SNP in the region; no one STR's length explained a similar amount of signal. More concerning was we ran fine-mapping of these region with the lengths of the two internal STRs and the length of one of the flanking STRs as independent variants while excluding the total length variable, and none of the individual lengths had strong fine-mapping evidence. As fine-mapping was our only source of causal, non-correlative, evidence for this association, we deemed that inconsistent fine-mapping evidence made this story too unverified to present to our readers, and we cut the result from our paper.

We still remain uncertain as to whether this compound STR is causal and think it valuable for further follow-up. It is unclear what is the correct method for fine-mapping this region, and whether we should have accepted the fine-mapping evidence for the total length of the region or not. I delve into this type of issue more in the thesis Discussion. Further, it is relatively likely that attempting to jointly call the length of the compound STR introduced error that would have been avoided by calling the lengths of the individual units. That issue, at least, should be easy to fix in follow-up analyses that can currently be performed.

Overall, we chose to be quite stringent with our statistical fine-mapping approach in order to insulate our results, to the extent possible, from these statistical fine-mapping concerns.

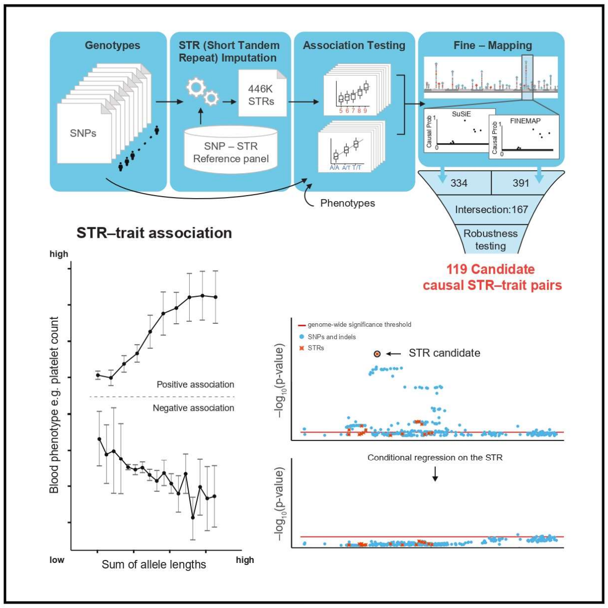
Much larger sets of putatively causal variants should be identifiable once there is better insight into how to rely on statistical fine-mapping methodologies. Yet despite our stringency, we identify 119 putatively causal STR signals across many traits and loci. This provides a wealth of results for further exploration and motivates the extension of these analyses to many more traits, which I reflect on in the thesis Discussion.

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 © 2023 The Author(s).
https://doi.org/10.1016/j.xgen.2023.100458

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.

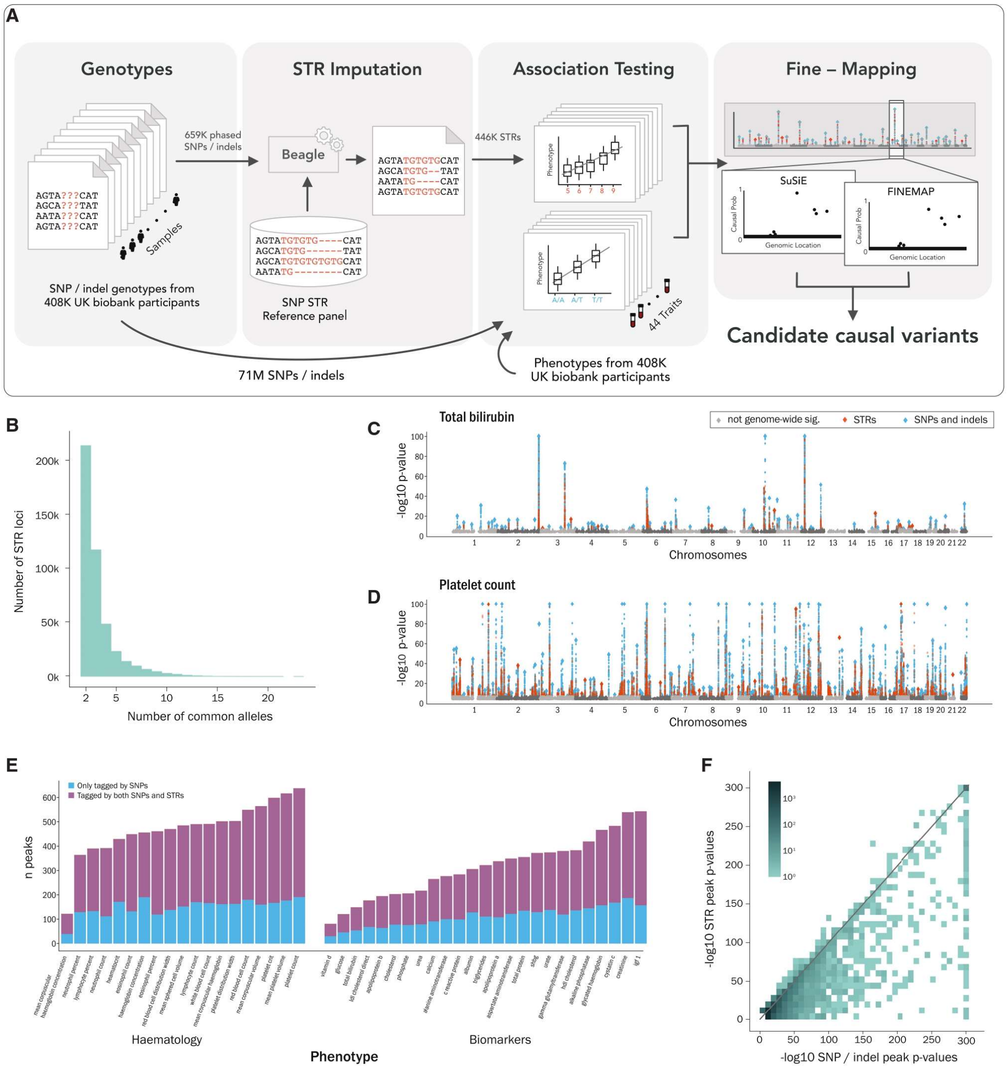


Figure 1. Genome-wide association tests identify STRs and SNPs associated with blood and biomarker traits in the UKB

(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.

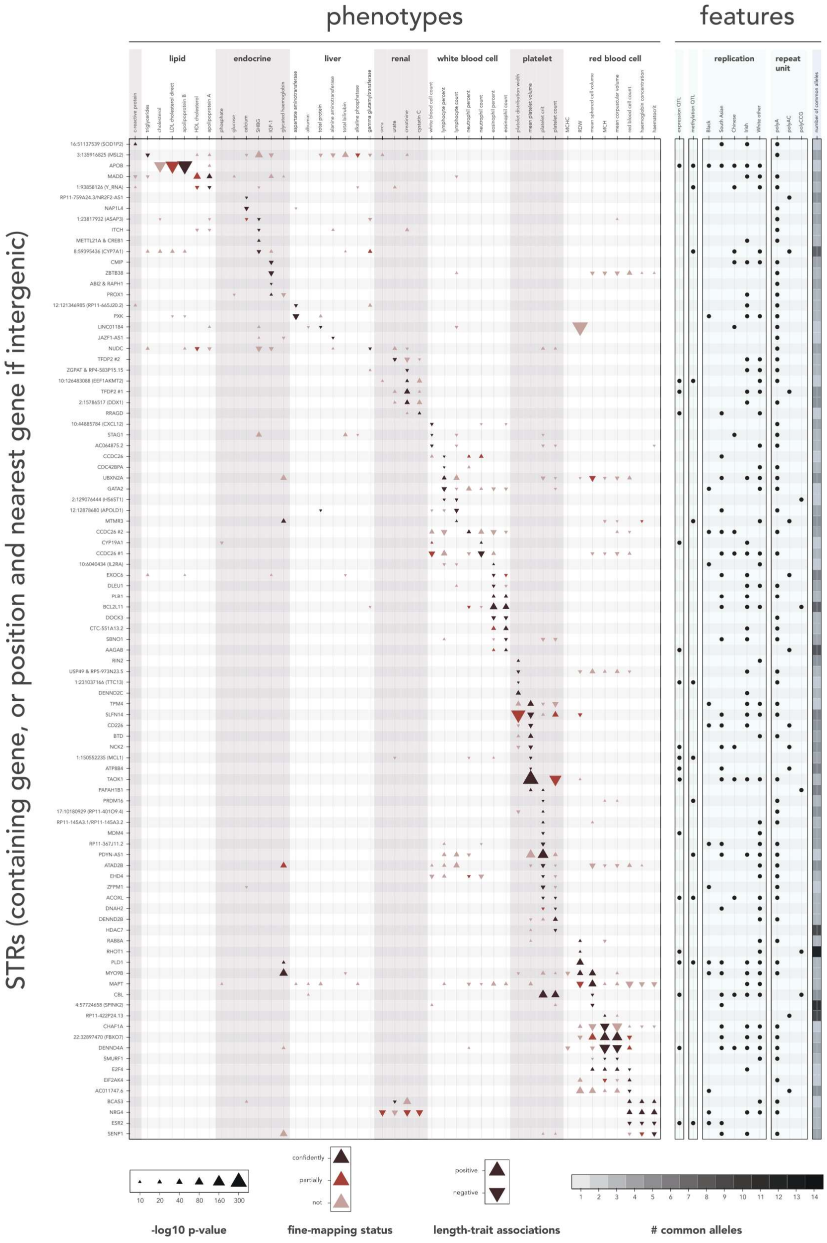
In simulations with randomly chosen causal SNPs, STRs comprised between 0% and 0.46% of genome-wide significant variants with $CP \geq 0.8$. In contrast, using phenotypes simulated from the second strategy, 1.4%–3.2% were STRs (Table S7). Of the total CP assigned by SuSiE or FINEMAP to genome-wide significant 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.

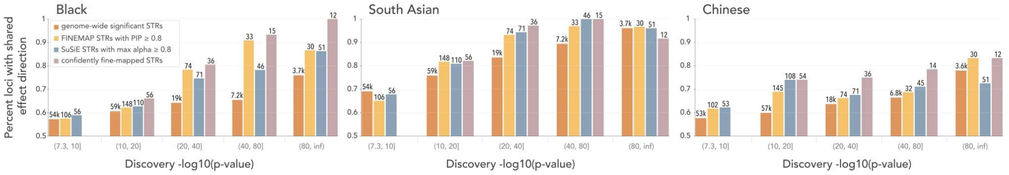


Figure 3. Concordance of White British STR effect directions in Black, South Asian, and Chinese populations

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	$(CAG)_6(CGCAGGCAG)[CGC(CAG)_2]_2CGC$	CTG (polyleucine)	apolipoprotein B	1.37e−279	−35.76	*APOB* (coding)	−
2:106510441	$(AC)_6GTG(CA)_{10}C(TA)_7T$	AC	mean platelet volume	6.93e−29	−11.15	*NCK2* (3′ UTR)	+
2:111878544	$(CGC)(CGCTGC)_2(CGC)_{13}C$	CCG	eosinophil count; 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.

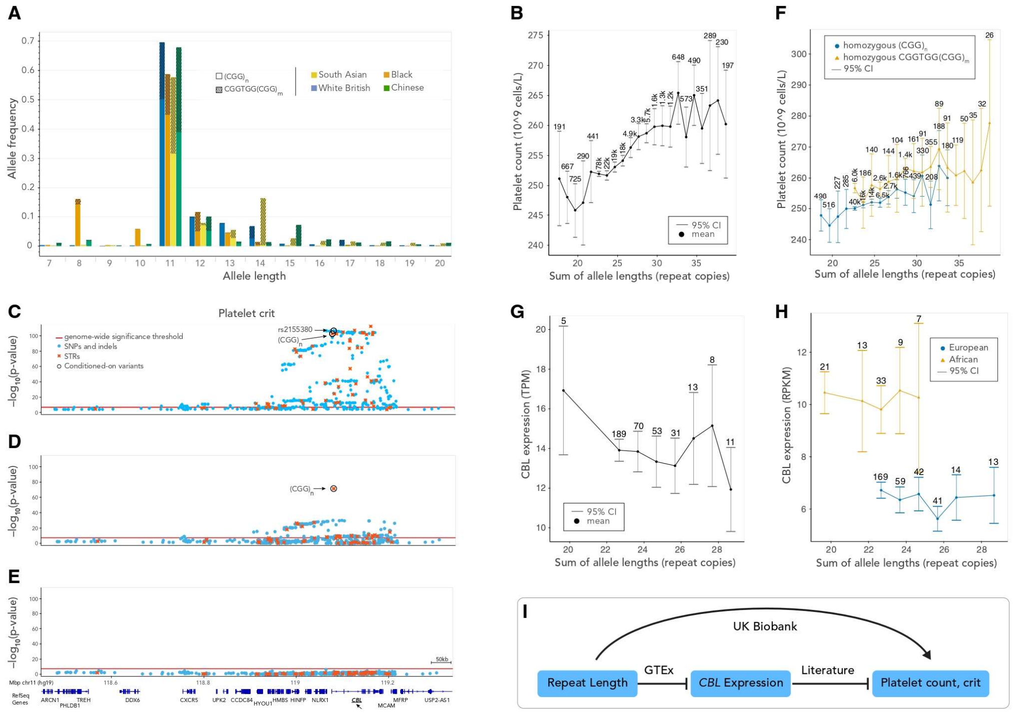


Figure 4. A polymorphic CGG repeat in the promoter of *CBL* influences platelet traits

(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 multi-allelic 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 $\overrightarrow{y} = \overrightarrow{g} * \beta_g + C * \overrightarrow{\beta_C} + \overrightarrow{\epsilon}$ where $\overrightarrow{y}$ is the vector of rank-inverse-normalized phenotype values per individual, $\overrightarrow{g}$ 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, $\overrightarrow{\beta_C}$ is the vector of covariate effect sizes, and $\overrightarrow{\epsilon}$ 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 $\overrightarrow{g}$ is the vector of dosages of the non-reference SNP or indel allele.

For conditional regressions, we fit the model $\overrightarrow{y} = \overrightarrow{g} * \beta_g + \overrightarrow{f} * \beta_f + C * \overrightarrow{\beta_C} + \overrightarrow{\epsilon}$ where all the terms are as described above, except $\overrightarrow{f}$ 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 ≤ 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 ≥ 0.2 or size ≤ 50). We reran those regions with the additional parameters L = 30 (maximum number of credible sets) and max_iter = 500. No regions failed to converge in under 500 iterations. We re-analyzed several loci that produced 30 plausible credible sets again with L = 50. No regions produced 50 plausible credible sets. SuSiE failed to finish for two regions (urate 4:8165642-11717761, total bilirubin 12:19976272-22524428) in under 48 hours; we excluded those regions from downstream analyses. A prior version of our pipeline had applied a custom filter to some SuSiE fine-mapping runs that caused SNPs with total minor allele dosage less than 20 across the entire population to be excluded. For consistency, any regions run with that filter which produced STRs included in our confidently fine-mapped set were rerun without that filter. Results from the rerun are reported in 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 ≥ 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 ≥ 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 ≥ 0.8 in the original runs and maintained CP ≥ 0.8 in a selected set of those alternate conditions.

For SuSiE, we evaluated using best-guess genotypes vs. genotype dosages as input. For FINEMAP, we tested varying the p value threshold, choice of non-major allele frequency threshold, effect size prior, number of causal variants per region, and stopping threshold. 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 $\overrightarrow{\epsilon} \sim \mathrm{N}(0, \mathrm{diag}(1 - \sum_{v \in V} \beta_v^2 Var[\overrightarrow{g_v}]))$ 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.

Cell Genomics, Volume *3*

Supplemental information

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

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

Notes

Note S1: Summary of fine-mapping models, related to STAR Methods

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

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

SuSiE reports a purity value for each signal, and we used that value to discard signals which were not well fine-mapped. SuSiE constructs 90%-credible sets for each signal so that the estimated probability of the credible set containing a variant causal for that signal is at least 90% (other values, such as 95%-credible sets, could be constructed similarly). SuSiE defines the purity of the credible set for each signal to be the minimum absolute correlation between any pair of variants in the set. The SuSiE manuscript suggests discarding signals with purity less than 0.5, but also states that the threshold is arbitrary. Looking at the distribution of credible set purities across all of our trait-regions (**Figure S3**) we decided to discard credible sets with purity less than 0.8, reasoning that the upper mode of the distribution is well above that threshold and that a signal containing two variants with correlation less than 0.8 has not been acceptably resolved.

SuSiE's PIPs are calculated across all credible sets regardless of purity, while we wished to conservatively only consider variants which had passed this added layer of scrutiny. Additionally, we saw that the PIP metric is sensitive to values of `L` (the number of signals fit per locus) – for one extreme example, in a locus with 57 variants, SuSiE run with `L=50` assigned each variant a PIP ≥ 0.5, which is unrealistic. So instead of using SuSiE's PIPs, we took each variant's highest alpha score from among credible sets with purity at least 0.8 as its SuSiE CP (or zero if it was in no such credible sets). This choice was uniformly conservative; CPs defined this way must be less than SuSiE's PIPs. We also found it to be less sensitive to `L` – we examine this more thoroughly in **Note S3** below, but in the above example we note that there was only one credible set containing less than 50 variants and it was not pure, so each variant in that region has a SuSiE CP of 0. We compared our SuSiE CP metric to SuSiE's PIP metric in **Figure S4** and saw that, among variants residing in pure credible sets, these two measures only strongly differed for variants whose contribution to any single pure signal was small. As our downstream analyses focused on variants with large alpha values in pure credible sets, this means that our use of alpha values instead of PIPs was not strongly impactful in analyzing those variants. The impact is that we conservatively restricted which variants we examined. Lastly, we note that for high purity thresholds such as the one we use, our metric should be very similar to calling SuSiE's `susie_get_pip` function with the flag `prune_by_cs=TRUE`, a method not examined in the SuSiE manuscript and one we did not encounter until after performing this work.

Note S2: Comparing results across fine-mapping methods, related to **STAR Methods**

To assess the reliability of our fine-mapping results, we measured how often the two fine-mapping methods agreed with one another, and how sensitive they were to model settings. First, we used SuSiE's credible sets as a proxy for the truly independent signals in our data. We observed that while SuSiE and FINEMAP were in agreement for most of the signals, their results were strongly discordant for a sizable number of signals (**Figure S8**). In particular, for 8.5% of 90%-credible sets returned by SuSiE (which by definition are assigned at least a 90% chance of containing a causal variant), the sum of FINEMAP's assigned CPs for all variants in each of those sets was less than 0.1, indicating that FINEMAP concluded those sets had a < 10% chance of containing a causal variant.

Second, we looked at the variant level and saw that for most variants, the CPs from FINEMAP and SuSiE were similar (**Figure S9**), with FINEMAP assigning slightly higher CPs overall (possibly due to our use of SuSiE alpha values per variant instead of the overall PIPs). However, we again saw that SuSiE and FINEMAP markedly disagree at a subset of loci. For instance, among all SNPs and indels which at least one fine-mapping method assigned a CP ≥ 0.95 and the other method was decisive about their causality (assigning either CP ≥ 0.95 or CP ≤ 0.05), 12.2% of those were assigned a CP ≥ 0.95 by one method and a CP ≤ 0.05 by the other. For STRs, the fine-mapping methods disagreed at nearly half of the loci (43.5%) that were assigned CP ≥ 0.95 by one method and decisively scored by the other, suggesting the CPs for STRs are even less reliable. This highlights the need for additional quality control before stating that variants assigned a high posterior probability by a single fine-mapper are likely to be causal. Without any prior on which fine-mapper to believe when the two disagreed, we focused only on the 167 trait-STR associations for which association p-values were well below the genome-wide significance threshold (p-value<1e-10) and both fine-mappers assigned high CPs (CPs≥0.8) (**Figure S10a;** the 167 associations can be extracted from **Table S4**).

Note S3: Assessing robustness of fine-mapping results, related to **STAR Methods**

We further assessed how robust our fine-mapping results were to fine-mapping instability and differences in the fine-mapping conditions, data filtering thresholds and algorithm metaparameters used. For SuSiE, we modified the inputs (1) `scaled_prior_variance`, (2) `tol`, (3) `residual_variance`, and (4) `L`, and also (5) changed the input genotypes from dosage genotypes to best-guess genotypes and (6) changed the prior to favor SNPs and indels over STRs as causal variants. For FINEMAP, we modified the inputs (1) `–prior-std` and (2) `–prob-conv-sss-tol` and also (3) filtered input variants with total non-major allele dosage less than 100, (4) filtered variants with p-value ≤ 5e-4, (5) set the prior on the number of causal variants per region to 4, and (6) changed the prior to favor SNPs and indels over STRs as causal variants. Further, we tested the ability to reproduce each fine-mappers results under the same conditions, and found that while SuSiE produced identical results when run on identical initial conditions, FINEMAP did not, so (7) we ran FINEMAP a second time with the same initial conditions.

We tested a few of the SuSiE settings on a subset of mean platelet volume fine-mapping regions before broader testing. We were encouraged that these settings had minimal impact on the results in this initial run and so did not include these conditions in our downstream tests for identifying confidently fine-mapped STRs. These settings were:

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

- `tol` – This determines what amount of change in the objective function between optimization rounds is small enough to cause SuSiE to terminate. We reduced this from the default of 1e-3 to 1e-4 and saw only miniscule changes in the results (**Figure S11a**).

- `residual_variance` – This is the initial value for the estimation of the residual variance of the phenotype after controlling for all effects at the locus. By default, the `residual_variance` is initialized to the full variance of the phenotype, which in our study was slightly less than 1 (rank-inverse normalization set it to 1, and then regressing covariates out of the phenotype before running SuSiE reduced it slightly). We ran SuSiE with alternate `residual_variance` values of 0.95 and 0.8 and saw small changes in the results (**Figure S11b,c**), while noting that a residual variance value of 0.8 would be unrealistic for the large majority of fine-mapping regions in traits we studied.

- `L` – This is the number of signals SuSiE fits in a region, or equivalently, the upper bound on the number of causal variants SuSiE attempts to find (**Figure S11d**). In our original fine-mapping runs, we ran SuSiE with `L=10`, and only increased `L` in a region if needed (first to 30, and then to 50 if still needed, **STAR Methods**). Below, we compared SuSiE runs with `L=10` in every region to runs with `L=50`. The SuSiE manuscript[1] states that inflated `L` values should not adversely impact model fitting because extraneous signals contribute small probabilities dispersed over many variants, thus not strongly changing any single variant's prediction, and also the learned effect sizes of these extra signals are shrunk towards zero. We see in our comparison that this only induces a large change in CP for a small fraction of variants. Of those, almost all of them are variants with non-zero CP values under the `L=10` case and zero CP in the `L=50` case. Thus, if they have any effect, this indicates that in most cases inflated values of `L` should lead to more conservative fine-mapping results. While overestimating `L` does not seem to harm our CP estimates, we found that when using SuSiE's standard PIP metric, overestimates of `L` may indeed lead to poor performance in some cases (see the discussion in **Note S1** above).

On the other hand, the fine-mapping conditions we document below did impact the end results. For each of these conditions, we ran fine-mapping on the trait-regions of the 167 STR-trait associations above, (in the same manner as the fine-mapping section of the main **STAR Methods**), and present supplemental figures showing how the CPs of variants changed under those conditions (**Figures S12, S13a-f**). Due to our lack of confidence in signals that were not robust to these choices, we restricted our set of confidently fine-mapped STR associations to the 119 associations that had CP ≥ 0.8 under each of those conditions (**Table S5**). While our focus here was to find confidently fine-mapped STRs, and while the set of trait-regions used for running these tests was chosen for that purpose, **Figures S12, S13a-f** identify similar trends for SNPs and indels in those regions. Thus, we hypothesize that these comparisons are relevant for fine-mapping of all variant types.

SuSiE with best-guess genotypes vs dosage genotypes

We ran SuSiE with the best-guess genotypes from our imputation pipeline instead of the dosage genotypes from that pipeline (**Figure S12**). Discrepancies in best-guess vs. dosages reflect imputation uncertainty, and contrasting runs under those two conditions allowed us to discard loci where this uncertainty strongly impacted the results.

FINEMAP under identical conditions

We reran FINEMAP on the same data with the same conditions and compared the CPs of the two runs (**Figure S13a**).

FINEMAP with alternative p-value thresholds

By default, we chose to filter as few variants as possible from our fine-mapping runs while still controlling for computational costs, which meant filtering variants with p>5e-2 from our FINEMAP runs and variants with p>5e-4 from our SuSiE runs, as FINEMAP was less computationally intensive. To check if this difference impacted the fine-mappers' results we ran FINEMAP having filtered all variants with p>5e-4 and compared it to our default FINEMAP runs (**Figure S13b**).

FINEMAP with alternative choice of non-major allele frequency threshold

To test whether FINEMAP results were strongly influenced by rare variants, we excluded all variants with total non-major allele dosage < 100 (population frequency less than approximately 0.015%) on top of the filter excluding variants with p-value ≥ 0.05 (**Figure S13c**). (Note that variants with total non-major allele dosage < 20 were excluded from association testing and thus from all fine-mapping runs).

Inadvertantly, when running FINEMAP with the alternative non-major allele frequency threshold, we only applied the threshold to SNPs and indels whose reference allele was the major allele, thus failing to filter out such variants whose reference allele had total dosage < 100 in the tested population. This reduces the useful interpretation of the results from this particular run with FINEMAP. Additionally, as a minor note, in this run we inadvertantly included the few variants with association p-value exactly equal to 0.05, in addition to including all variants with p-value < 0.05 as normal.

FINEMAP with alternative choice of effect size prior

FINEMAP's default `–prior-std` value is 0.05 which gives causal variants a default effect size of 0.25% of phenotypic variance. We modified this to `–prior-std 0.0224` to reflect published expected effect sizes for GWAS variants of about 0.05%[3] (**Figure S13d**).

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

We ran FINEMAP with the prior of four causal variants per trait-region instead of one (**Figure S13e**). We did this by adding a column prob to the FINEMAP input Z file which contained the value 4/n for each variant, where n was the number of variants in the trait-region, and by running FINEMAP with the `–prior-snps` flag.

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

We ran FINEMAP with the flag `-prob-conv-sss-tol` 0.0001 (reduced from the default of 0.001) (**Figure S13f**). This reduced what amount of change in the objective function over the last 100 rounds of optimization would be considered small enough to cause FINEMAP to terminate.

Summary

To conclude, 48 (28.3%) of the 167 STR-trait associations failed to replicate in one of the above alternate fine-mapping runs. The dosages vs best-guess genotypes choice when running SuSiE was the most impactful of these alternate conditions, accounting for 29 of those 48 cases, 24 of which did not fail any other alternate conditions (**Figure S12**), suggesting that imputation uncertainty has a sizable effect on downstream inferences.

Surprisingly, 7 (4.1%) of the 167 STR associations failed to replicate in the FINEMAP run under identical conditions, indicating that FINEMAP is moderately unstable (**Figure S13a**). This is a concern as FINEMAP has no seed parameter to allow for study reproducibility. Overall, 24 associations failed to replicate in at least one of the 6 FINEMAP alternate runs, 19 of which did not additionally fail the SuSiE best-guess condition. At these rates, it is difficult to distinguish if this is driven by FINEMAP's underlying instability, or if any of these other conditions are impactful by themselves. We suspect that FINEMAP's instability only appears at specific loci due to some nature of their LD patterns and do not currently hypothesize that FINEMAP results at all loci are subject to such instability.

For STR associations which failed to replicate in any of the above conditions, meaning that in the alternate run they had CP < 0.8, the associations rarely failed to replicate because of slight decreases in CP. Rather, there was an average decrease of 0.64 CP across failed replications, suggesting that when fine-mappers are sensitive to modeling conditions (or their own instability), those conditions strongly impact the end results.

Encouragingly, we saw that these comparisons agreed more frequently in regions containing variants which both the default SuSiE and FINEMAP runs agreed had high CPs (both CPs ≥ 0.8, **Figures S12, S13**) than in regions where the default SuSiE and FINEMAP runs disagreed (data not shown). This suggests that concordance between different fine-mapping algorithms may be able to provide security against the instability in the results of any single algorithm. While we focused on fine-mapping results for STRs, which generally showed lower concordance across methods than SNPs, our results suggest similar robustness checks should be performed when fine-mapping SNPs and other variant types.

There were several fine-mapping conditions we tested that strongly impacted the resulting CPs but that we did not use as filters when selecting our causal STR candidates since they represent unrealistic parameter choices. We report their values in **Table S4**. Those conditions were:

- We ran FINEMAP with the flag `-prior-std 0.005`, corresponding to an expected effect size 0.0025% (**Figure S13g**). We concluded that this was much lower than the effect sizes we were hoping to detect.

- Both SuSiE and FINEMAP have the default assumption that each variant is as likely to be causal as any other variant (regardless of allele frequency). We instead conservatively ran SuSiE and FINEMAP with the prior assumption that SNPs and indels were 4x more likely to be causal than STRs. For this, we set the prior probability of causality for each SNP or indel to 4/(4*n_SNPs_indels + n_STRs) and for each STR to 1/(4*n_SNPs_indels + n_STRs). For SuSiE we did this by setting the `prior_weights` input to an array containing those probabilities. For FINEMAP we did this by adding a column `prob` to the FINEMAP input Z file which contained those probabilities, and by running FINEMAP with the `-prior-snps` flag. As expected, this resulted in overall decreased STR CPs (**Figure S14**). While we did not filter our candidate STRs based on this setting, we were encouraged to see that a majority of the strongest hits replicated despite this conservative setting.

Finally, we note there are other parameters which were not tested here but that could be tested for robustness. This includes whether FINEMAP results are sensitive to imputation uncertainty or overestimating `--n-causal-snps`, whether SuSiE results are sensitive to a non-major allele frequency threshold, or to increasing p-value thresholds higher than is computationally necessary and testing if either method's results are sensitive to the size of the trait-regions being fine-mapped.

Note S4: Additional details for coding fine-mapped STRs, related to **Figure 2** which displays these and other confidently fine-mapped associations

Coding trinucleotide repeat in APOB: This repeat did not initially appear in our list of confidently fine-mapped STRs due to limitations in our process for filtering indels which are STR alleles (**STAR Methods**). We did not filter an indel imputed by the UKB team that corresponded exactly to the short allele of the STR imputed from our reference panel, since the indel/short allele corresponds to the deletion of an imperfect repeat sequence (GCCAGCAGC for a CAG repeat), and we cautiously only performed automated filtering for indels without imperfections. The presence of this indel alongside the STR during fine-mapping caused SuSiE's (but not FINEMAP's) results to, in some cases, show low confidence as to which of the two variants were causal. Specifically, FINEMAP assigned a CP of 1 to the STR for both traits apolipoprotein B and LDL cholesterol under each FINEMAP run used for filtering down to the confidently fine-mapped set. For the original run for the apolipoprotein B trait and for the best-guess run for both traits, SuSiE created a credible set containing both the indel and the STR and assigned each a CP of less than 0.8, causing the association not to pass our filters for confidently fine-mapped STRs. However, if we sum the SuSiE CPs of both variants in those runs we get a CP of over 0.97 in each case, making the apolipoprotein B association pass our confidently fine-mapped thresholds. Thus, we added this association to our confidently fine-mapped set. We note that the original SuSiE run for the LDL trait assigned low CPs to both the STR and the indel. While that was the only fine-mapping of the eight runs used for filtering that did not assign the pair of variants a combined CP ≥ 0.8 for LDL, it precludes us from adding the LDL association to the confidently fine-mapped set. For both the apolipoprotein B and LDL cholesterol associations, we updated the CPs in **Tables S4** and S**5** to reflect the combined CPs for both variants.

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

We used AlphaFold[4] to investigate whether the two common repeat alleles (referred to in previous literature as SP24 and SP27[5]) at *APOB* might have an effect on protein structure. Although analysis of the impact of small mutations has not yet been validated using AlphaFold[6], it could help generate hypotheses regarding the impact of protein-coding repeats on protein function. We restricted analysis to the first 600 amino acids as analysis of the full-length APOB protein (~4500) was computationally prohibitive. The two alleles did not induce any notable changes in predicted structure. Previous work on this repeat[7] suggests this variant, which resides in the signal peptide of the protein, affects secretion efficiency. We hypothesize that this reported change in secretion efficiency, rather than large changes in protein structure, drives this particular signal.

Coding trinucleotide repeat in E2F4: We identified a protein-coding (poly-serine) repeat in *E2F4* confidently fine-mapped to multiple red blood cell traits. We similarly used AlphaFold[4] to assess the impact of varying the number of serine repeats on E2F4's protein structure, which did not result in any noticeable difference (**Figure S16a**). However E2F4 is known to form a complex with RBL2, in which RBL2 stabilizes E2F4 and protects it from degradation via the ubiquitin-proteasome pathway[8]. This prompted us to explore the joint structure of E2F4 with RBL2. In addition to not being validated for predicting the effect of mutations, the standard AlphaFold software is also not designed or validated for predicting the structure of protein complexes[6]. Nonetheless, it can be used for such prediction, and that may generate useful hypotheses. In this case, AlphaFold results suggest E2F4 and RBL2 are more tightly bound when E2F4 contains a short vs. long poly-serine track (**Figure S16b**). We therefore hypothesize that long poly-serine alleles destabilize the complex and lead to faster degradation of E2F4. While existing proteomics datasets[9,10] did not allow us to test this hypothesis, it does have the expected direction of association: previous knockout[11] and knockdown[12] experiments show that presence of E2F4 is positively associated with counts of other cell types, and our GWAS shows a negative association between this STR's length and red blood cell counts.

Methods for 3D structure simulation and analysis using AlphaFold: We performed protein 3D structure simulations using AlphaFold v2.2.0[4]. The simulations were conducted using 1 NVIDIA V100 SMX2 GPU. We used the settings `--max_template_date=2020-05-14` `--use_gpu_relax=True` `--`

`model_preset=monomer` . The simulated protein structures were aligned and visualized using PyMOL 2.5.4 (https://pymol.org/2/).

The `NP_000375` fasta file for APOB was downloaded from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein). Due to the large size of the full APOB protein (4563 amino acids) and the focus on the N terminal STR variants, only the first 600 amino acids were used in the structure simulation.

The `NP_001941` fasta file for E2F4 was downloaded from the NCBI protein database. The full length of the E2F4 protein was used for the structure simulation. Among all the poly-serine variants, we selected the shortest (containing a 12 unit repeat) and longest variants (containing a 19 unit repeat, the reference) for comparison. Note that the poly-serine sequence in E2F4 is 2 amino acids longer than the repeat as the last two serines use different codings than the repeat unit AGC. (e.g. the reference is SSSSSSSSSSSSSNSNSSSSS which is 21 amino acids, while the reference repeat is 19 units long).

For the E2F4_RBL2 complex, we downloaded the `NP_001310537` fasta file for RBL2 from the NCBI protein database. The E2F4 sequences were linked with the RBL2 sequence using a flexible 70 amino acid linker $(GGGGS)_{14}$.

The protein sequences used for the simulations are listed below:

APOB longer STR allele
```
MDPPRPALLALLALPALLLLLLAGARAEEEMLENVSLVCPKDATRFKHLRKYTYNYEAESSSGVPGTADSRSATRINCKVELEVPQLCSFILKTSQCTLKEVYGFNPEGKA
LLKKTKNSEEFAAAMSRYELKLAIPEGKQVFLYPEKDEPTYILNIKRGIISALLVPPETEEAKQVLFLDTVYGNCSTHFTVKTRKGNVATEISTERDLGQCDRFKPIRTGI
SPLALIKGMTRPLSTLISSSQSCQYTLDAKRKHVAEAICKEQHLFLPFSYKNKYGMVAQVTQTLKLEDTPKINSRFFGEGTKKMGLAFESTKSTSPPKQAEAVLKTLQELK
KLTISEQNIQRANLFNKLVTELRGLSDEAVTSLLPQLIEVSSPITLQALVQCGQPQCSTHILQWLKRVHANPLLIDVVTYLVALIPEPSAQQLREIFNMARDQRSRATLYA
LSHAVNNYHKTNPTGTQELLDIANYLMEQIQDDCTGDEDYTYLILRVIGNMGQTMEQLTPELKSSILKCVQSTKPSLMIQKAAIQALRKMEPKDKDQEVLLQTFLDDASPG
DKRLAAYLMLMRSPSQADINKIVQILPWEQNEQVKNFVASHIANI
```

APOB shorter STR allele
```
MDPPRPALLALPALLLLLLAGARAEEEMLENVSLVCPKDATRFKHLRKYTYNYEAESSSGVPGTADSRSATRINCKVELEVPQLCSFILKTSQCTLKEVYGFNPEGKALLK
KTKNSEEFAAAMSRYELKLAIPEGKQVFLYPEKDEPTYILNIKRGIISALLVPPETEEAKQVLFLDTVYGNCSTHFTVKTRKGNVATEISTERDLGQCDRFKPIRTGISPL
ALIKGMTRPLSTLISSSQSCQYTLDAKRKHVAEAICKEQHLFLPFSYKNKYGMVAQVTQTLKLEDTPKINSRFFGEGTKKMGLAFESTKSTSPPKQAEAVLKTLQELKKLT
ISEQNIQRANLFNKLVTELRGLSDEAVTSLLPQLIEVSSPITLQALVQCGQPQCSTHILQWLKRVHANPLLIDVVTYLVALIPEPSAQQLREIFNMARDQRSRATLYALSH
AVNNYHKTNPTGTQELLDIANYLMEQIQDDCTGDEDYTYLILRVIGNMGQTMEQLTPELKSSILKCVQSTKPSLMIQKAAIQALRKMEPKDKDQEVLLQTFLDDASPGDKR
LAAYLMLMRSPSQADINKIVQILPWEQNEQVKNFVASHIANILNS
```

E2F4 longer STR allele
```
MAEAGPQAPPPPGTPSRHEKSLGLLTTKFVSLLQEAKDGVLDLKLAADTLAVRQKRRIYDITNVLEGIGLIEKKSKNSIQWKGVGPGCNTREIADKLIELKAEIEELQQRE
QELDQHKVWVQQSIRNVTEDVQNSCLAYVTHEDICRCFAGDTLLAIRAPSGTSLEVPIPEGLNGQKKYQIHLKSVSGPIEVLLVNKEAWSSPPVAVPVPPPEDLLQSPSAV
STPPPLPKPALAQSQEASRPNSPQLTPTAVPGSAEVQGMAGPAAEITVSGGPGTDSKDSGELSSLPLGPTTLDTRPLQSSALLDSSSSSSSSSSSSSNSNSSSSSGPNPST
SFEPIKADPTGVLELPKELSEIFDPTRECMSSELLEELMSSEVFAPLLRLSPPPGDHDYIYNLDESEGVCDLFDVPVLNL
```

E2F4 shorter STR allele
```
MAEAGPQAPPPPGTPSRHEKSLGLLTTKFVSLLQEAKDGVLDLKLAADTLAVRQKRRIYDITNVLEGIGLIEKKSKNSIQWKGVGPGCNTREIADKLIELKAEIEELQQRE
QELDQHKVWVQQSIRNVTEDVQNSCLAYVTHEDICRCFAGDTLLAIRAPSGTSLEVPIPEGLNGQKKYQIHLKSVSGPIEVLLVNKEAWSSPPVAVPVPPPEDLLQSPSAV
STPPPLPKPALAQSQEASRPNSPQLTPTAVPGSAEVQGMAGPAAEITVSGGPGTDSKDSGELSSLPLGPTTLDTRPLQSSALLDSSSSSSNSNSSSSSGPNPSTSFEPIKA
DPTGVLELPKELSEIFDPTRECMSSELLEELMSSEVFAPLLRLSPPPGDHDYIYNLDESEGVCDLFDVPVLNL
```

E2F4-RBL2 longer STR allele
```
MAEAGPQAPPPPGTPSRHEKSLGLLTTKFVSLLQEAKDGVLDLKLAADTLAVRQKRRIYDITNVLEGIGLIEKKSKNSIQWKGVGPGCNTREIADKLIELKAEIEELQQRE
QELDQHKVWVQQSIRNVTEDVQNSCLAYVTHEDICRCFAGDTLLAIRAPSGTSLEVPIPEGLNGQKKYQIHLKSVSGPIEVLLVNKEAWSSPPVAVPVPPPEDLLQSPSAV
STPPPLPKPALAQSQEASRPNSPQLTPTAVPGSAEVQGMAGPAAEITVSGGPGTDSKDSGELSSLPLGPTTLDTRPLQSSALLDSSSSSSSSSSSSSNSNSSSSSGPNPST
SFEPIKADPTGVLELPKELSEIFDPTRECMSSELLEELMSSEVFAPLLRLSPPPGDHDYIYNLDESEGVCDLFDVPVLNLGGGGSGGGGSGGGGSGGGGSGGGGSGGGGSG
GGGSGGGGSGGGGSGGGGSGGGGSGGGGSGGGGSGGGGSMPSGGDQSPPPPPPPPAAAASDEEEEDDGEAEDAAPPAESPTPQIQQRFDELCSRLNMDEAARAEAWDSYRS
MSESYTLEGNDLHWLACALYVACRKSVPTVSKGTVEGNYVSLTRILKCSEQSLIEFFNKMKKWEDMANLPPHFRERTERLERNFTVSAVIFKKYEPIFQDIFKYPQEEQPR
QQRGRKQRRQPCTVSEIFHFCWVLFIYAKGNFPMISDDLVNSYHLLLCALDLVYGNALQCSNRKELVNPNFKGLSEDFHAKDSKPSSDPPCIIEKLCSLHDGLVLEAKGIK
EHFWKPYIRKLYEKKLLKGKEENLTGFLEPGNFGESFKAINKAYEEYVLSVGNLDERIFLGEDAEEEIGTLSRCLNAGSGTETAERVQMKNILQQHFDKSKALRISTPLTG
VRYIKENSPCVTPVSTATHSLSRLHTMLTGLRNAPSEKLEQILRTCSRDPTQAIANRLKEMFEIYSQHFQPDEDFSNCAKEIASKHFRFAEMLYYKVLESVIEQEQKRLGD
MDLSGILEQDAFHRSLLACCLEVVTFSYKPPGNFPFITEIFDVPLYHFYKVIEVFIRAEDGLCREVVKHLNQIEEQILDHLAWKPESPLWEKIRDNENRVPTCEEVMPPQN
LERADEICIAGSPLTPRRVTEVRADTGGLGRSITSPTTLYDRYSSPPASTTRRRLFVENDSPSDGGTPGRMPPQPLVNAVPVQNVSGETVSVTPVPGQTLVTMATATVTAN
NGQTVTIPVQGIANENGGITFFPVQVNVGGQAQAVTGSIQPLSAQALAGSLSSQQVTGTTLQVPGQVAIQQISPGGQQQKQGQSVTSSSNRPRKTSSLSLFFRKVYHLAAV
RLRDLCAKLDISDELRKKIWTCFEFSIIQCPELMMDRHLDQLLMCAIYVMAKVTKEDKSFQNIMRCYRTQPQARSQVYRSVLIKGKRKRRNSGSSDSRSHQNSPTELNKDR
TSRDSSPVMRSSSTLPVPQPSSAPPTPTRLTGANSDMEEEERGDLIQFYNNIYIKQIKTFAMKYSQANMDAPPLSPYPFVRTGSPRRIQLSQNHPVYISPHKNETMLSPRE
KIFYYFSNSPSKRLREINSMIRTGETPTKKRGILLEDGSESPAKRICPENHSALLRRLQDVANDRGSH
```

E2F4-RBL2 shorter STR allele

```
MAEAGPQAPPPPGTPSRHEKSLGLLTTKFVSLLQEAKDGVLDLKLAADTLAVRQKRRIYDITNVLEGIGLIEKKSKNSIQWKGVGPGCNTREIADKLIELKAEIEELQQRE
QELDQHKVWVQQSIRNVTEDVQNSCLAYVTHEDICRCFAGDTLLAIRAPSGTSLEVPIPEGLNGQKKYQIHLKSVSGPIEVLLVNKEAWSSPPVAVPVPPPEDLLQSPSAV
STPPPLPKPALAQSQEASRPNSPQLTPTAVPGSAEVQGMAGPAAEITVSGGPGTDSKDSGELSSLPLGPTTLDTRPLQSSALLDSSSSSSNSNSSSSSGPNPSTSFEPIKA
DPTGVLELPKELSEIFDPTRECMSSELLEELMSSEVFAPLLRLSPPPGDHDYIYNLDESEGVCDLFDVPVLNLGGGGSGGGGSGGGGSGGGGSGGGGSGGGGSGGGGSGGG
GSGGGGSGGGGSGGGGSGGGGSGGGGSGGGGSMPSGGDQSPPPPPPPPAAAASDEEEEDDGEAEDAAPPAESPTPQIQQRFDELCSRLNMDEAARAEAWDSYRSMSESYTL
EGNDLHWLACALYVACRKSVPTVSKGTVEGNYVSLTRILKCSEQSLIEFFNKMKKWEDMANLPPHFRERTERLERNFTVSAVIFKKYEPIFQDIFKYPQEEQPRQQRGRKQ
RRQPCTVSEIFHFCWVLFIYAKGNFPMISDDLVNSYHLLLCALDLVYGNALQCSNRKELVNPNFKGLSEDFHAKDSKPSSDPPCIIEKLCSLHDGLVLEAKGIKEHFWKPY
IRKLYEKKLLKGKEENLTGFLEPGNFGESFKAINKAYEEYVLSVGNLDERIFLGEDAEEEIGTLSRCLNAGSGTETAERVQMKNILQQHFDKSKALRISTPLTGVRYIKEN
SPCVTPVSTATHSLSRLHTMLTGLRNAPSEKLEQILRTCSRDPTQAIANRLKEMFEIYSQHFQPDEDFSNCAKEIASKHFRFAEMLYYKVLESVIEQEQKRLGDMDLSGIL
EQDAFHRSLLACCLEVVTFSYKPPGNFPFITEIFDVPLYHFYKVIEVFIRAEDGLCREVVKHLNQIEEQILDHLAWKPESPLWEKIRDNENRVPTCEEVMPPQNLERADEI
CIAGSPLTPRRVTEVRADTGGLGRSITSPTTLYDRYSSPPASTTRRRLFVENDSPSDGGTPGRMPPQPLVNAVPVQNVSGETVSVTPVPGQTLVTMATATVTANNGQTVTI
PVQGIANENGGITFFPVQVNVGGQAQAVTGSIQPLSAQALAGSLSSQQVTGTTLQVPGQVAIQQISPGGQQQKQGQSVTSSSNRPRKTSSLSLFFRKVYHLAAVRLRDLCA
KLDISDELRKKIWTCFEFSIIQCPELMMDRHLDQLLMCAIYVMAKVTKEDKSFQNIMRCYRTQPQARSQVYRSVLIKGKRKRRNSGSSDSRSHQNSPTELNKDRTSRDSSP
VMRSSSTLPVPQPSSAPPTPTRLTGANSDMEEEERGDLIQFYNNIYIKQIKTFAMKYSQANMDAPPLSPYPFVRTGSPRRIQLSQNHPVYISPHKNETMLSPREKIFYYFS
NSPSKRLREINSMIRTGETPTKKRGILLEDGSESPAKRICPENHSALLRRLQDVANDRGSH
```

Note S5: Additional details for non-coding fine-mapped STRs, related to **Figure 2** which displays these and other confidently fine-mapped associations

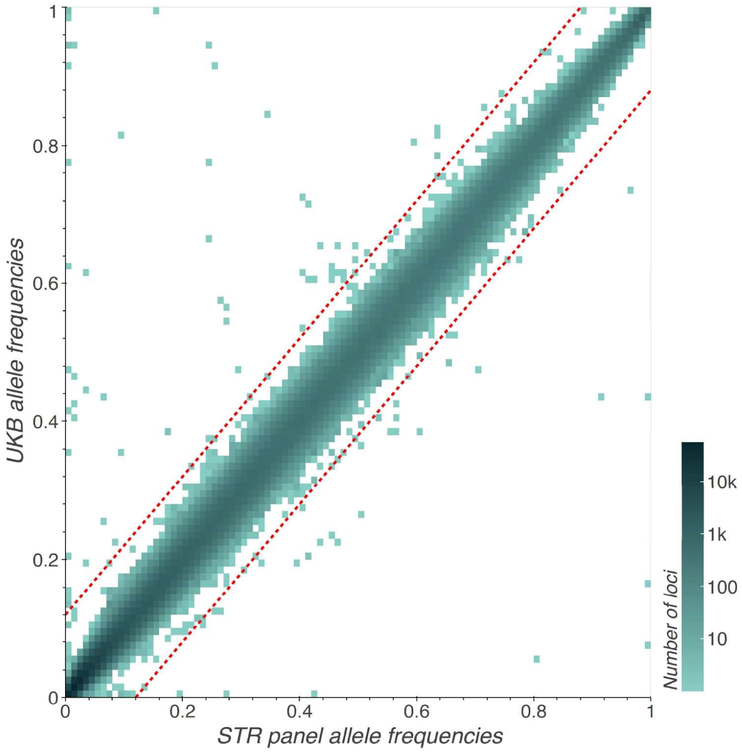
Dinucleotide repeat in SLC2A2 (GLUT2): We identified a dinucleotide repeat immediately upstream of exon 4 of *SLC2A2* as a confidently fine-mapped STR for bilirubin. While *SLC2A2* has not previously been causally linked to bilirubin levels, *SLC2A2* mediates glucose transport to hepatocytes, where glucose is stored in the form of glycogen[13]. Glycogen degradation produces intermediates that are substrates in the process that regulates bilirubin conjugation and excretion[14,15] and thus could potentially impact bilirubin levels in the blood. This effect of *SLC2A2* on bilirubin levels may be partially corroborated by a large cohort study on babies with congenital hyperinsulinemic hypoglycemia, a condition that inhibits glycogen breakdown, which reported elevated bilirubin in that population[16].

Tetranucleotide repeat in ESR2: We identified a GTTT repeat in an intron of *ESR2* whose length is negatively associated with haemoglobin concentration, red blood cell count, and haematocrit. *ESR2* is known to regulate red blood cell production. Studies conducted in populations chronically exposed to high altitude hypoxia, a driver of erythrocytosis (excess red blood cell production), demonstrated inhibition of erythrocytosis through activation of estrogen beta signaling in ex vivo models[17]. These observations are corroborated by a study of rat models under hypoxia, where beta-estrogen treatment reduced circulating levels of erythropoietin, a kidney-derived factor that stimulates red blood cell production[18].

We additionally identified a negative association between length of this STR and *ESR2* expression. However, the expected direction of association between *ESR2* expression and red blood cell count is unclear. Multiple *ESR2* isoforms exist, either as a result of alternative splicing of the last coding exons (exon 8 and exon 9, respectively), deletion of one or more coding exons, or alternative usage of untranslated exons in the 5′ region[19]. One of the five isoforms found in humans even has an undetectable affinity to estrogen, and instead was found to antagonize estrogen-alpha receptor signaling[20]. Thus any change in overall *ESR2* expression would need to be understood in the context of the isoforms whose expressions are changing and which tissues those isoforms are common in, complicating any mechanistic predictions.

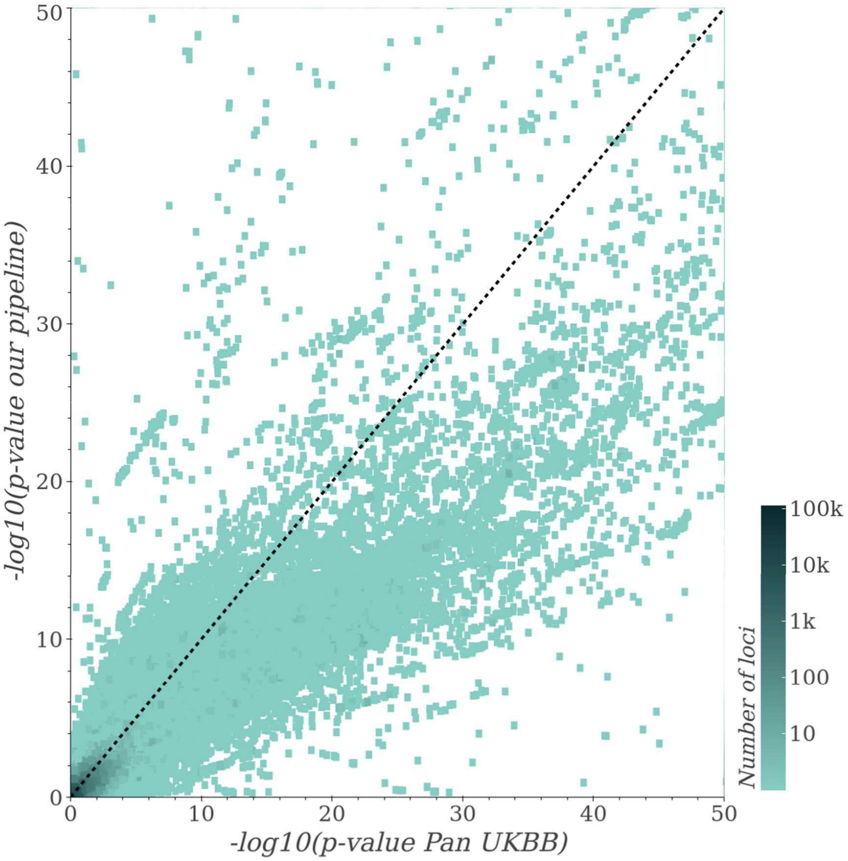
Figures

Figure S1: Comparison of SNP alternate allele frequencies between our SNP-STR reference panel and UKB phased hard-called variants, related to **STAR Methods**



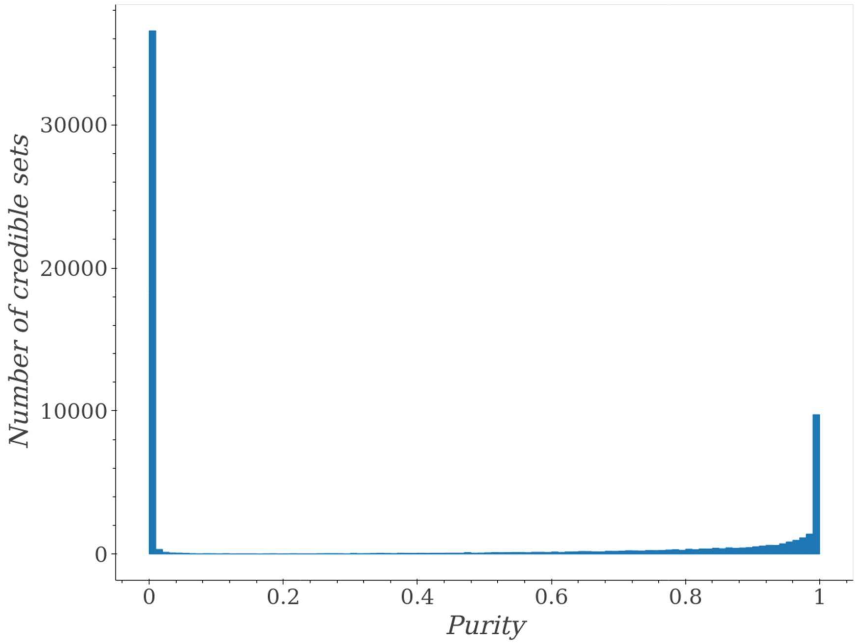
The x-axis indicates the alternate allele frequency of variants calculated from the European individuals in our SNP-STR reference panel[21] (see **Key Resources Table**). The y-axis indicates their alternate allele frequency in unrelated participants in the White British population in the UKB. We filtered variants with more than a 12% difference in alternate allele frequency (indicated by the red diagonal lines). The color gradient represents the number of variants ($\log_{10}$ scale) whose p-values fall in each region. White regions contain no variants.

Figure S2: Comparison of association p-values between our pipeline and summary statistics published by Pan UKBB, related to **STAR Methods** which details other quality control steps



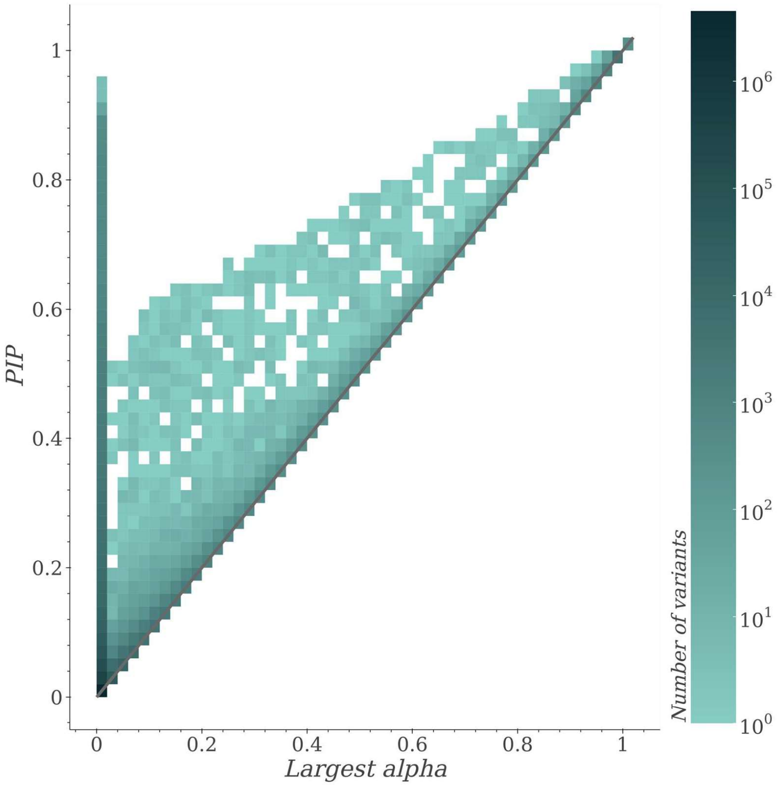
Heatmap of -log$_{10}$ p-values obtained from the Pan UKBB[22] study of UKB data (x-axis) vs. from our study (y-axis) for total bilirubin associations with SNPs and indels. The color gradient represents the number of variants (log$_{10}$ scale) whose p-values fall in each region. White regions contain no variants. P-values less than 1e-50 are truncated. Our pipeline's p-values are highly correlated with Pan UKBB's but are overall more conservative, which may be attributable to differences in models used (linear mixed model for Pan-UKB vs. linear model used here).

Figure S3: Distribution of SuSiE 90%-credible set purities, related to **STAR Methods**, **Note S1**



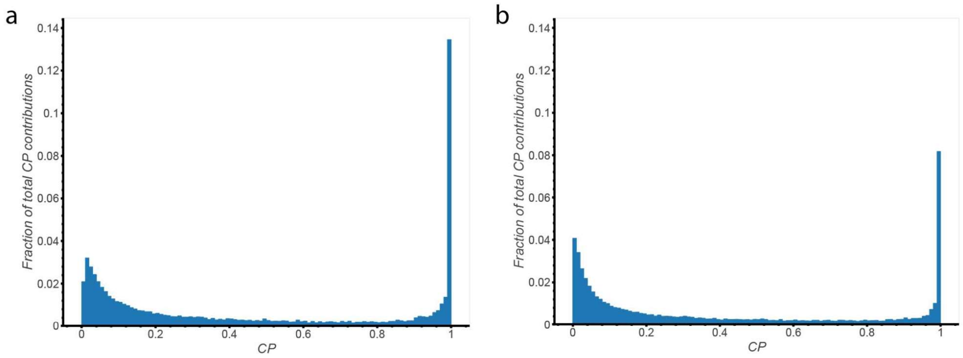
Distribution of SuSiE 90%-credible set purities across all trait-regions. (The rightmost bin is inclusive, containing SuSiE credible sets with purity up to and including 1, e.g. those that consist of a single variant.) Purity is defined as the minimum absolute correlation between any pair of variants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

Figure S4: PIP vs alpha values assigned by SuSiE, related to **STAR Methods**, **Note S1**



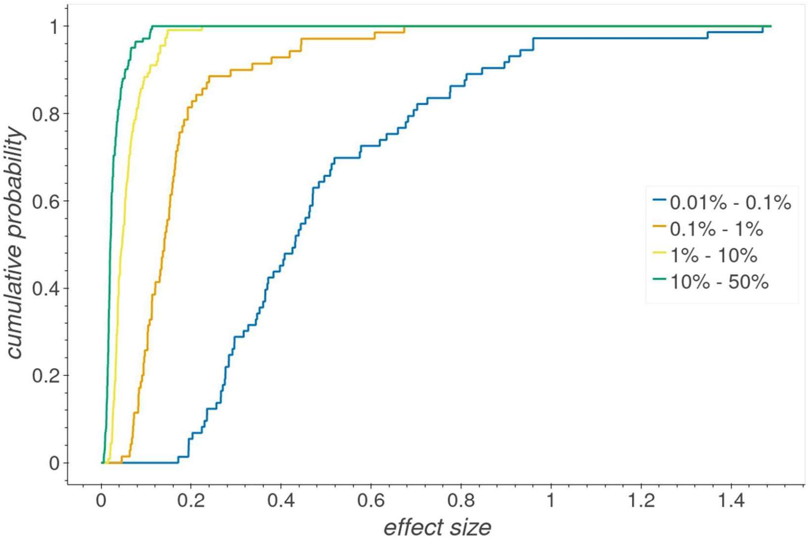
Largest alpha value across pure credible sets (x-axis) vs. PIP (y-axis) for all variants across all trait regions. Color ($\log_{10}$ scale) indicates the number of data points falling in each bin.

Figure S5: Contribution of variants to signals genome-wide by variant CP, related to **STAR Methods** which contains other fine-mapping details



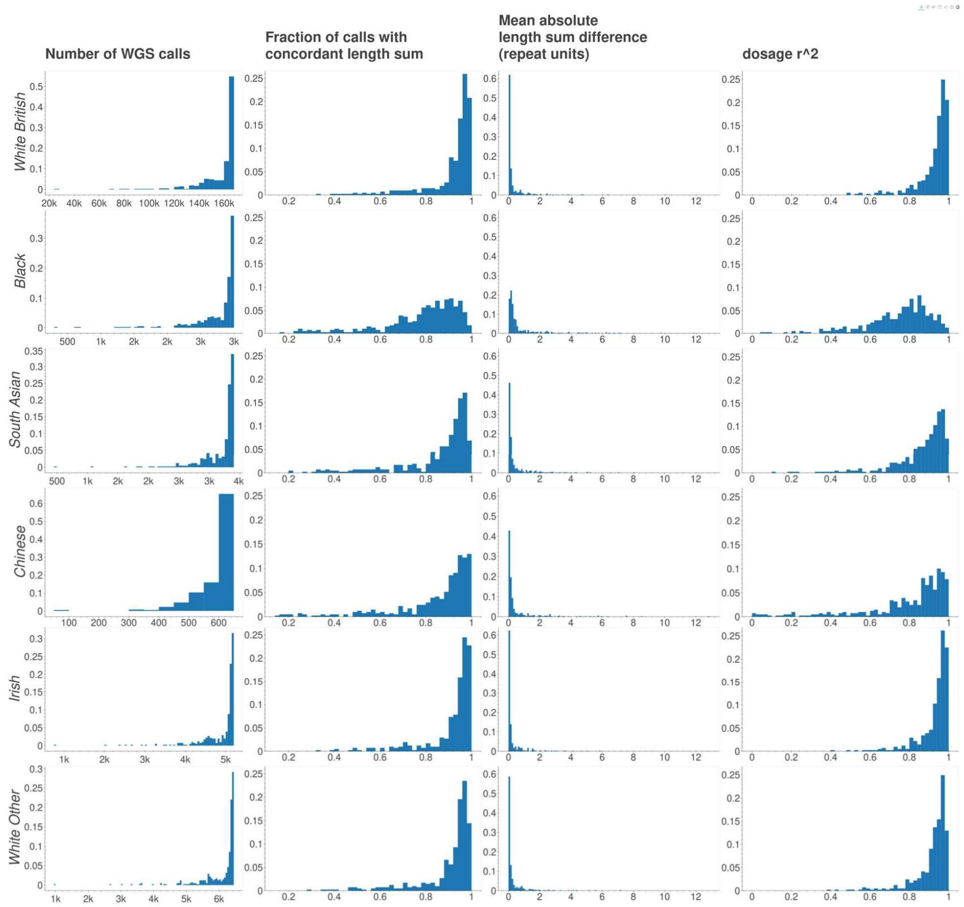
Summed contribution of genome-wide significant variants across all regions binned by variant CP as a fraction of the total CP of all genome-wide significant variants across all regions for **(a)** SuSiE and **(b)** FINEMAP. (The rightmost bin for each graph is inclusive, containing variants with CPs up to and including 1.) The total contribution of all variants across all regions with CP < 0.1 was 29.3% for SuSiE and 35.1% for FINEMAP.

Figure S6: Effect sizes for causal variants in simulation strategy 1, related to **STAR Methods**, **Tables S6-S8**



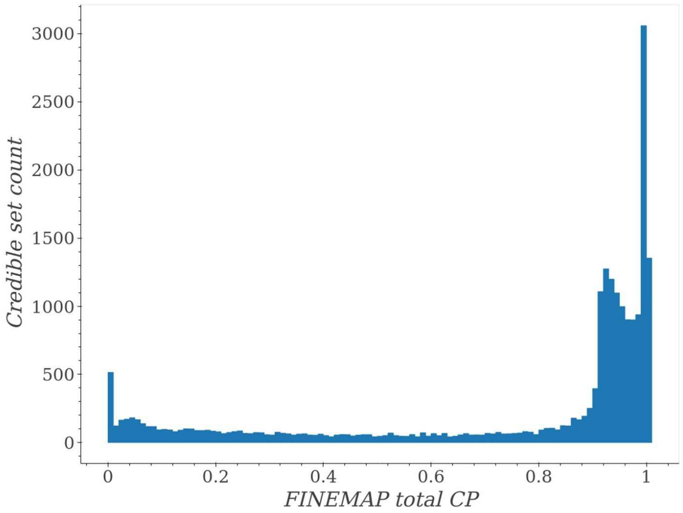
Cumulative distribution functions of the discrete effect size distributions for each minor allele frequency bin for simulation strategy 1. Per the **STAR Methods**, these effect sizes were drawn from all SNPs/indels in all platelet count regions that had either FINEMAP CP ≥ 0.5 or SuSiE CP ≥ 0.5. SNPs/indels chosen to be causal for strategy 1 simulations had their effect sizes drawn from the bin their minor allele frequency corresponded to. As expected, more common variants tend to have smaller effect sizes than rarer variants.

Figure S7: Concordance between STR length imputation and WGS calls, related to **STAR Methods** which details how these metrics were computed



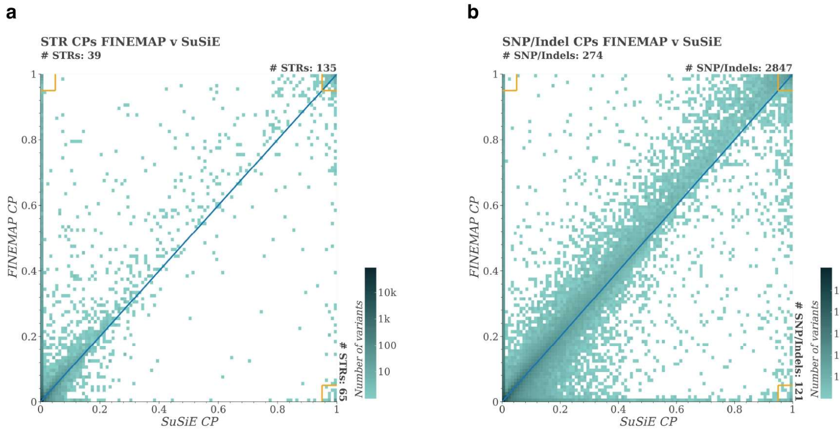
Each histogram is a distribution over the 409 distinct STRs with an association with either FINEMAP or SuSiE CP ≥ 0.8 (**Table S4**). Each row represents a different ethnicity group of QCed (potentially related) individuals. The first column displays the number of WGS calls per locus for each group. Each subsequent column is a different measurement of per-locus concordance between STR length calls from imputation and WGS: the fraction of concordant length sums, the mean absolute difference between length sums and the correlation between length sum dosages (**STAR Methods**). For sake of comparison, all histograms in the same column share the same x-axis and the same y-axis bounds, excepting the first column.

Figure S8: Total CPs assigned to SuSiE credible sets by FINEMAP, related to **Note S2**, **STAR methods**



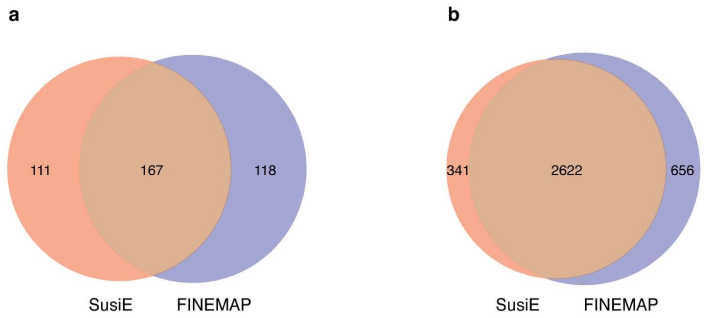
SuSiE 90%-credible sets across all trait-regions (with purity ≥ 0.8) were each binned by the total CP FINEMAP assigned to all variants in that set. Sets in the rightmost bin have FINEMAP total CP between 1 and 1.01 (i.e. FINEMAP predicts them to contain on average between 1 and 1.01 causal variants). FINEMAP assigned 6 SuSiE credible sets to have total CP greater than 1.01 (none of which attained total CP greater than 1.13); those 6 are omitted from the figure. By definition, SuSiE has estimated each 90%-credible set to have between a 90% and 100% chance of containing a single causal variant.

Figure S9: Discordance between SuSiE and FINEMAP CPs, related to **STAR Methods**, **Note S2**



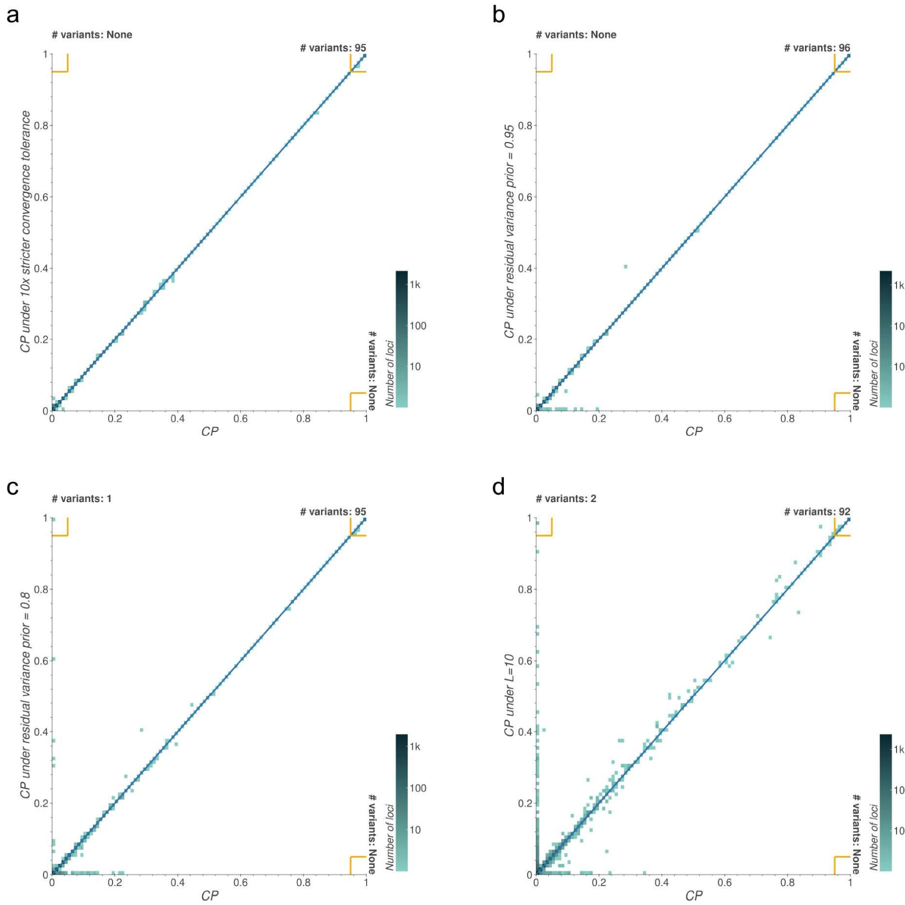
Comparison of CPs across all trait-regions between SuSiE (x-axis) and FINEMAP (y-axis) for genome-wide significant STRs **(a)** and SNPs and indels **(b)**. The blue line denotes equal CP. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes.

Figure S10: Discordance between fine-mappers, related to **STAR Methods**, **Note S2**



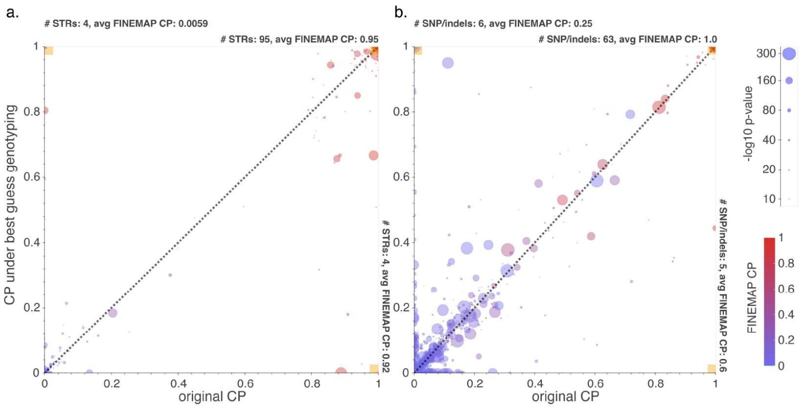
The number of **(a)** STRs and **(b)** SNPs/indels with p-value < 1e-10 assigned a CP ≥ 0.8 by only SuSiE (red), only FINEMAP (purple), or both (brown).

Figure S11: SuSiE settings not used for filtering, related to **STAR Methods**, **Note S3**



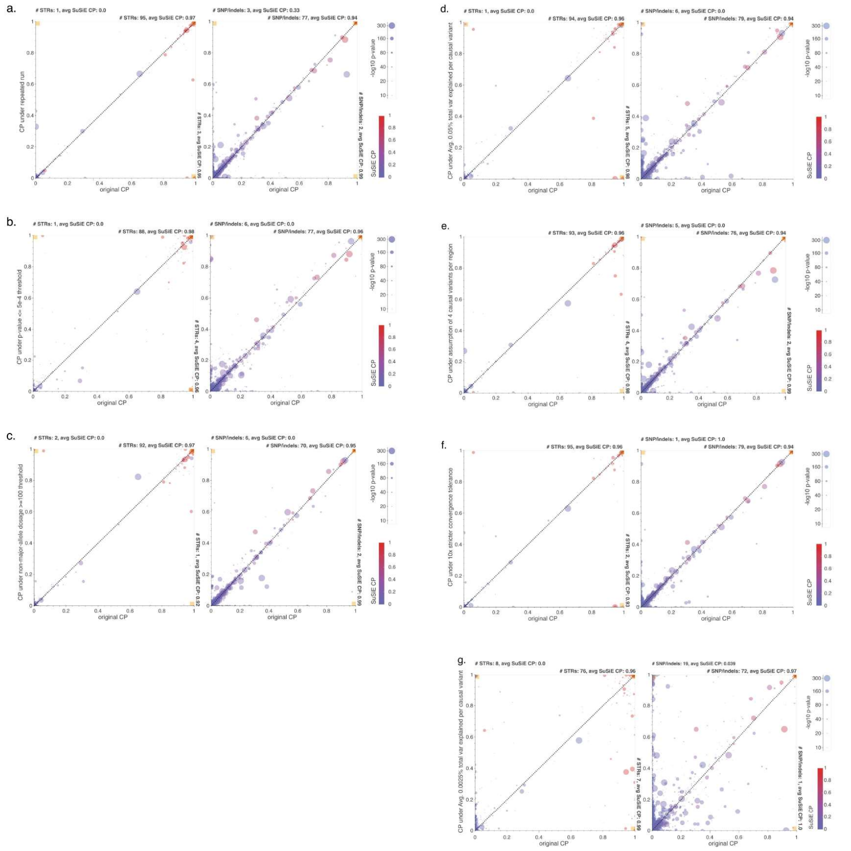
Concordance between SuSiE CPs for all genome-wide significant variants across most small-to-medium sized mean platelet volume fine-mapping regions under default settings on the x-axis (`tol=1e-3`, `residual_variance` slightly less than 1, and `L=50`) vs. a single alternate setting on the y-axis **(a)**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

Figure S12: Effect of best-guess genotypes on SuSiE results, related to **STAR Methods**, **Note S3**



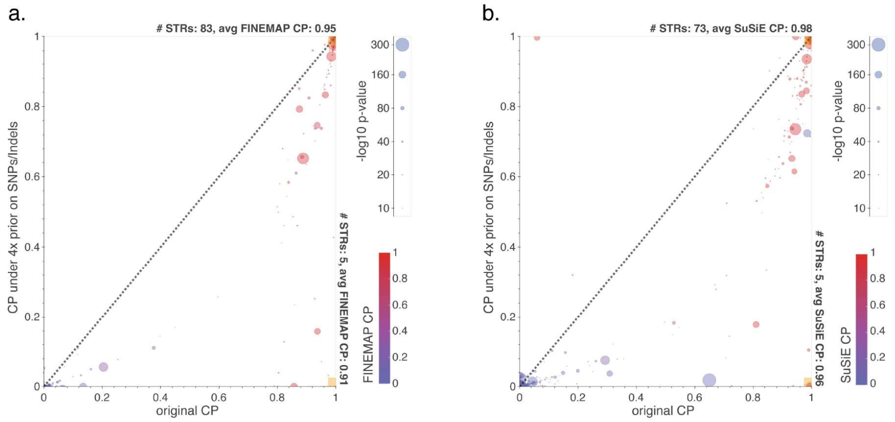
Discordance between SuSiE CPs for variants with p-value < 1e-10 when run with dosage genotypes (x-axis) vs best-guess genotypes (y-axis) among STRs **(a)** and SNPs and indels **(b)**. These data points are taken from running SuSiE on the trait-regions containing the 167 STR-trait associations with p-value < 1e-10 and with both SuSiE and FINEMAP CPs ≥ 0.8. Black lines denote equal CP. Larger circle sizes denote larger variant -log$_{10}$ association p-values. Circle color denotes the CP of that variant from our default FINEMAP run. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes and the average FINEMAP CP value of those variants.

Figure S13: Effect of alternate settings on FINEMAP results, related to **STAR Methods**, **Note S3**



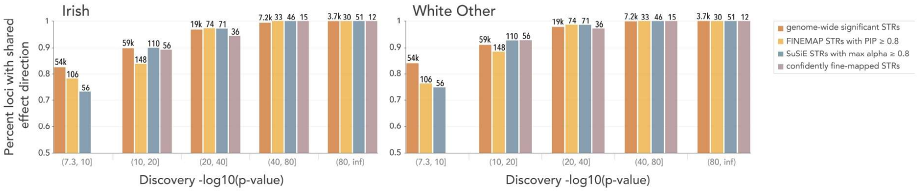
Discordance between FINEMAP CPs for variants with p-value < 1e-10 . x-axis values correspond to CPs from FINEMAP runs using our standard settings: with a p-value > 5e-2 filter, `--prior-std 0.05`, prior of one causal variant per trait-region and `--prob-conv-sss-tol 0.001` . y-axis values correspond to CPs from FINEMAP runs when **(a)** rerun with the same default settings on the same data (to account for random variation in the algorithm) or **(b-g)** run on the same data with a single alternate setting. Those alternate settings were: **(b)** p-value > 5e-4 filter **(c)** additionally filtering those variants with total non-major-allele dosage < 100 **(d)** `–prior-std 0.0224` **(e)** prior of four causal variants per trait region **(f)** `--prob-conv-sss-tol 0.0001` and **(g)** `--prior-std=0.005`. The variants in these plots are all those with p-value < 1e-10 in the trait-regions containing the 167 STR-trait associations with p-value < 1e-10 and with both SuSiE and FINEMAP CPs ≥ 0.8. CP discordance among STRs is plotted on the left, and among SNPs and indels is plotted on the right. Black lines denote equal CP. Larger circle sizes denote larger variant -$\log_{10}$ association p-values. Circle color denotes the CP of that variant from our default SuSiE run. Yellow boxes in the three extreme corners are summarized by the number of variants residing in those boxes and the average SuSiE CP value of those variants.

Figure S14: Effect of conservative prior favoring SNPs and indels on estimated causality of STR variants, related to **STAR Methods**, **Note S3**



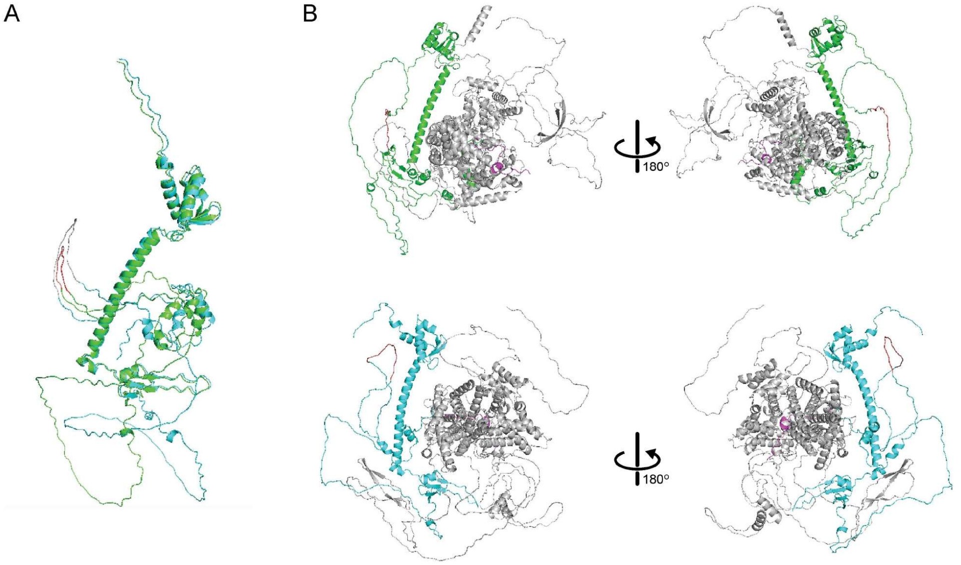
Discordance between CPs for STRs with p-value < 1e-10 when run under default settings (x-axis) vs with a 4x prior of causality for SNPs and indels as compared to STRs (y-axis) in **(a)** SuSiE and **(b)** FINEMAP. These data points are taken from running SuSiE and FINEMAP on the trait-regions containing the 167 STR-trait associations with p-value < 1e-10 and with both SuSiE and FINEMAP CPs ≥ 0.8. Black lines denote equal CP. Larger circle sizes denote larger variant -$\log_{10}$ association p-values. Circle color denotes the CP of that variant from the other fine-mapper's default run. Yellow boxes in the extreme corners are summarized by the number of variants residing in those boxes and the average SuSiE CP value of those variants.

Figure S15: Replication of White British STR associations in other White populations, related to **Figure 3**, **Table S10**



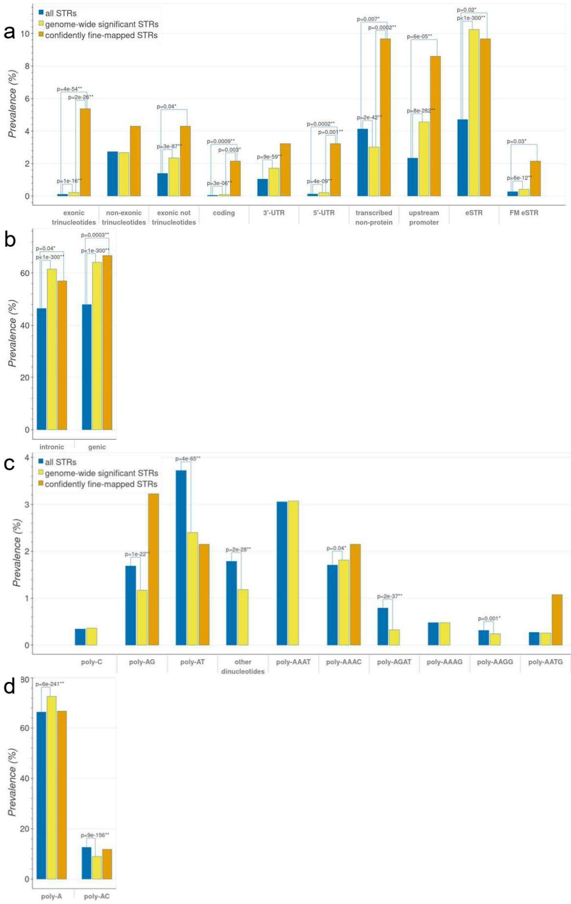
The y-axis gives the fraction of STR associations measured in the discovery cohort that have the same direction of effect when measured in the replication population regardless of p-value (left=Irish, right=White Other, see **Figure 3** for the non-White populations). Brackets beneath the x-axis denote the binning of discovery $-\log_{10}$ p-values. Brown=genome-wide significant associations (discovery p<5e-8), orange=FINEMAP fine-mapped STR associations (discovery p<5e-8 and FINEMAP CP ≥ 0.8), teal=SuSiE fine-mapped STR associations (discovery p<5e-8 and SuSiE CP ≥ 0.8) and purple=confidently fine-mapped STR associations. Annotations above each bar indicate the number of STR-trait associations considered. We required confidently fine-mapped STR associations to have p-value < 1e-10, thus they do not appear in the left-most bin. The trends in these figures are somewhat sensitive to the choice of p-value bin boundaries so we additionally analyze this data using logistic regression models (**Table S10**).

Figure S16: Evaluation of E2F4 protein structure using AlphaFold, related to **Note S4**, **Figure 2**



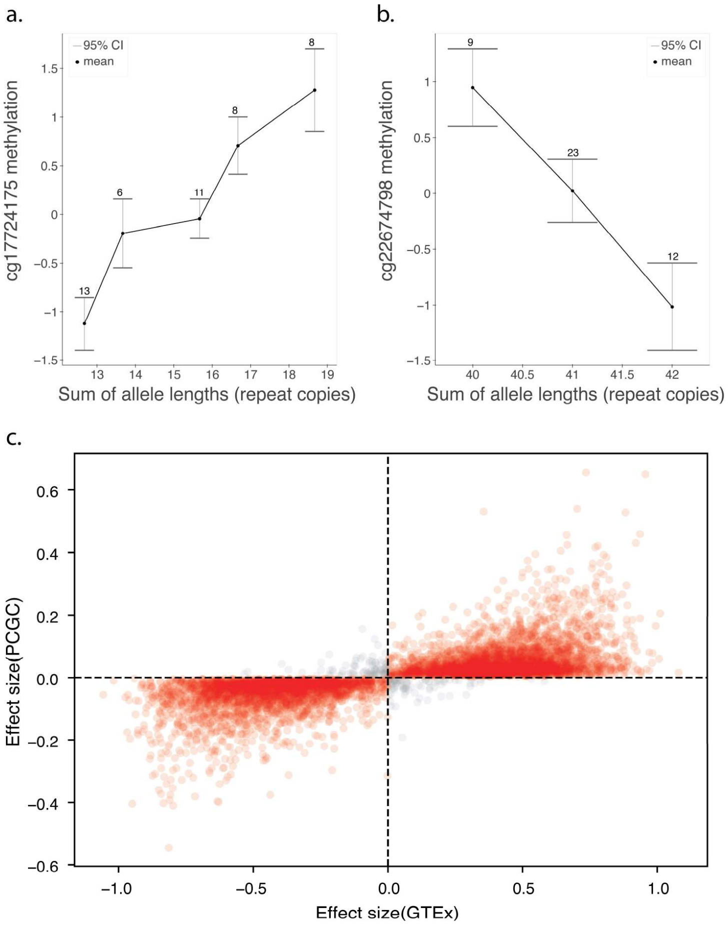
(a) depicts the E2F4 protein, while **(b)** depicts the E2F4-RBL2 complex. In both, green denotes the E2F4 protein variant with the shorter STR allele (12 units) while cyan represents the E2FF4 protein containing longer, reference STR allele (19 units). In **(a)** the shorter variant's poly S region coded by the STR is highlighted in red, and the longer variant's poly S region is highlighted in grey. In **(b)**, both poly S regions are highlighted in red, and RBL2 is in grey. Notice in **(a)** that the structure of the two E2F4 variants are highly similar but that in **(b)** the distance between E2F4 and RBL2 is smaller for the shorter-allele variant than the longer-allele variant.

Figure S17: Prevalence of STR features, related to **STAR Methods** which details how this was calculated



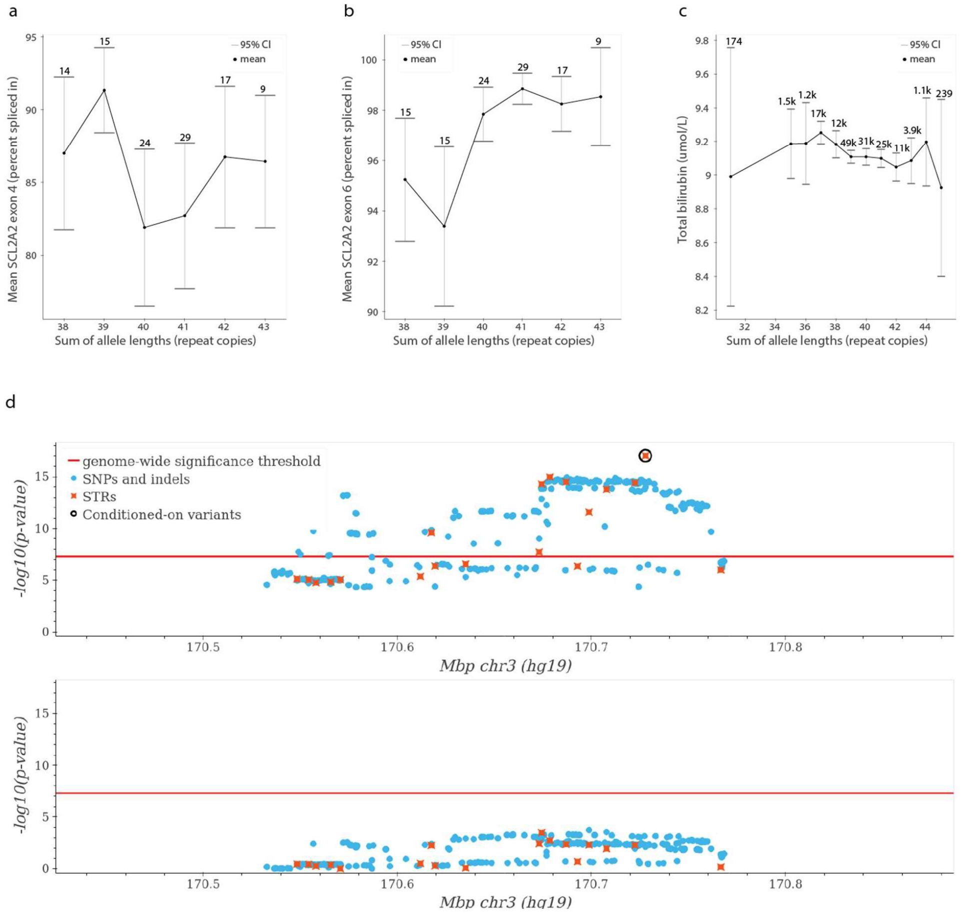
Genomic annotation **(a-b)** and repeat unit **(c-d)** prevalences are shown for different categories of STRs. (Blue=all STRs in our imputation panel, yellow=genome-wide significant STRs for at least one trait, orange=confidently fine-mapped STRs). **(a,c)** show low prevalence categories, **(b,d)** show high prevalence categories. In **(a)**, “upstream promoter” is defined as the region 3kb upstream of a transcription start site, and “eSTR” and “FM eSTR” are categories from our previous study to identify STR-gene expression associations in the Genotype Tissue Expression (GTEx) cohort[23]. **(c-d)** contains all repeat units represented by at least one thousand STRs in our reference panel, except for trinucleotide STR repeat units, as enrichments for those could not be distinguished from the enrichment for exonic trinucleotide STRs as a whole. See the **STAR Methods** for more details. P-values from two-sided tests of difference between proportions are only displayed when p≤0.05. Note that strong p-values for differences between the all STRs and genome-wide significant STRs categories could often be due to restricting to phenotypically-important genomic regions and not necessarily due to enrichment for causal variants.

Figure S18: Confidently fine-mapped STRs influencing DNA methylation, related to **Figure 2**, **Table S14** which contain other methylation associations with confidently fine-mapped STRs



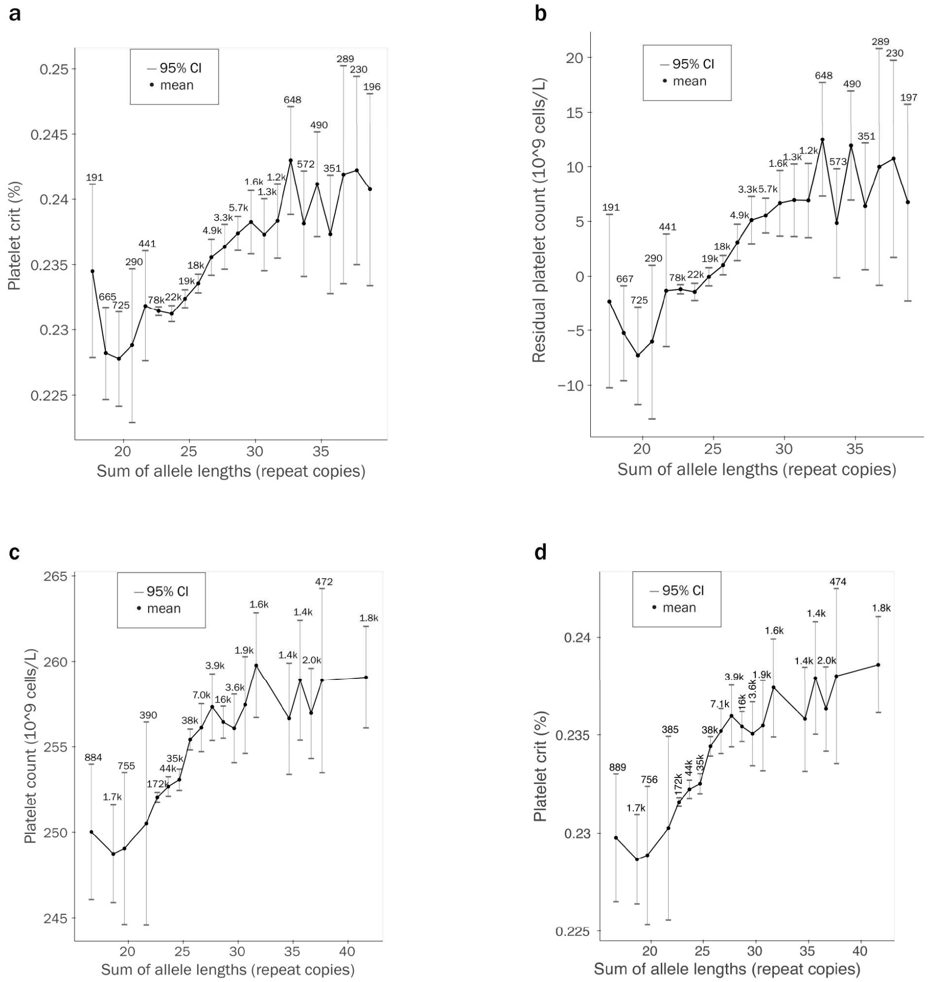
(a) STR 1:150579759-150579814 (hg38), residing less than 200bp upstream from the transcription start site of the gene *MCL1* and confidently fine-mapped to mean platelet volume (p=3e-16), is associated with methylation of a CpG site 500bp farther upstream (cg17724175). **(b)** STR 1:3170058-3170078 (hg38), residing in a conserved region of an intron of the gene *PRDM16* and confidently fine-mapped to platelet crit (p=4e-12), is associated with the methylation of a CpG site ~10kb away that resides in the same intron (cg22674798). Summed allele lengths are on the x-axis, and inverse normalized methylation levels are on the y-axis. Box plots indicate 25th, median and 75th percentiles, with whiskers extending up to 1.5 times the interquartile range beyond the 25th and 75th percentiles. Circles denote the mean methylation values for each allele length sum. **(c)** The effect sizes from STR length vs CpG methylation associations as measured here in the GTEx cohort compared to those measured by Martin-Trujillo et al.[24] in the Pediatric Cardiac Genomics Consortium (PCGC) cohort. Each of the n=7661 circles represents one of the associations measured in both cohorts. Red/grey circles denote matching/opposite effect directions in the two cohorts. We see that effects are broadly consistent (r=0.73, $p<10^{-200}$) between the two studies.

Figure S19: Association of an STR in *SLC2A2* with bilirubin levels, related to **Figure 2** which contains other fine-mapped examples



Scatter plots showing the association between the GT repeat at chr3:170727702 (hg19) and the splicing (percent spliced in) of exon 4 (**a**; linear association p=0.77) and exon 6 (**b**; linear association p=8.7e-07) of *SLC2A2* in Liver samples from GTEx[25]. For each plot, the x-axis represents the sum of repeat copies of STR in each individual and the y-axis represents percent spliced in for the indicated exon. The solid line gives the mean percent spliced in. Population counts are displayed for each length sum, only length sums with a count of at least 5 are displayed. © Association between length (from whole genome sequencing) of the GT repeat and eosinophil percentage in the UKB. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only allele length sums with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum. **(d)** Association of variants at the *SLC2A2* locus and total bilirubin levels, before (top) and after (bottom) conditioning on the GT repeat. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(GT)_n$ STR.

Figure S20: Associations of an STR in *CBL* with platelet crit and residual platelet volume, related to **Figure 4**

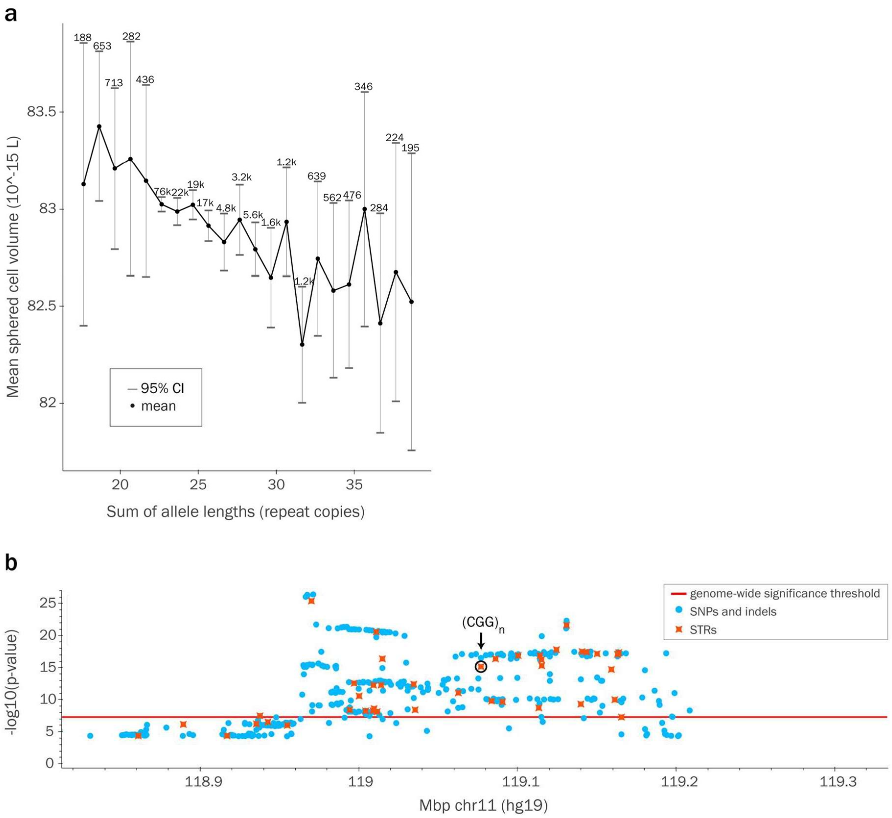


Summed STR length (from whole genome sequencing) vs platelet crit **(a)** and residual platelet count **(b)**. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. Rounded population-wide counts are displayed for each sum. Only allele length sums with a population frequency of 0.1% or greater are displayed.

The trends in **(a-b)** are nearly identical to those in **Figure 4b** for unadjusted platelet count. For **(b)** we calculated residuals by linearly regressing out the same covariates that were used in association p-value calculations (**STAR Methods**), including sex, age, population principal components and categorical covariates for batch effects. We then calculated the mean residual for each allele length sum. Note that in our association pipeline, covariates are included as we test STRs for association with rank inverse normalized phenotypes, while for **(b)** here we did not rank inverse normalize the phenotype values.

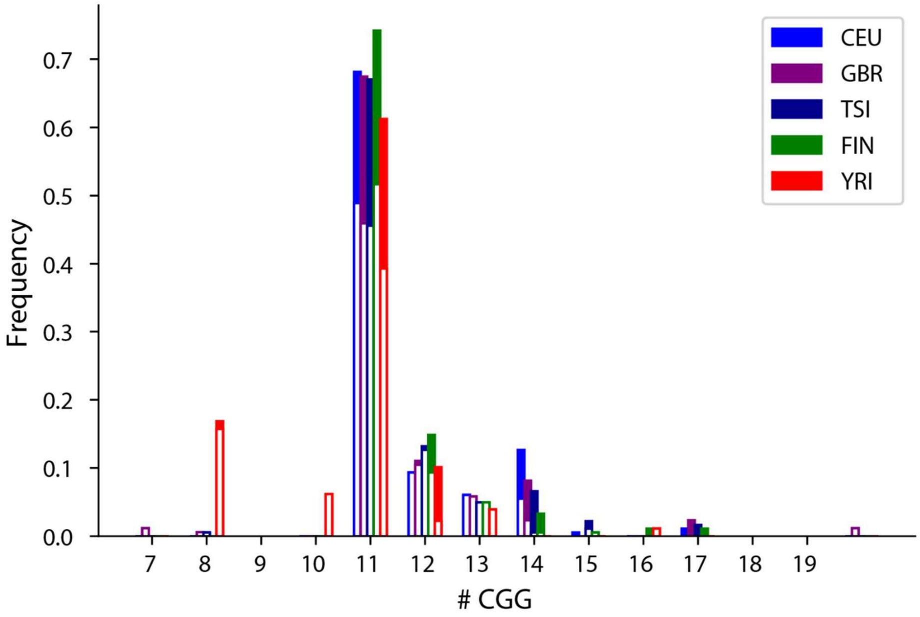
(c-d) display the associations with platelet count and platelet crit, respectively, of STR lengths derived from imputation. These trends are overall similar to those with genotypes derived from WGS data (**Figure 4b**, part **(a)** of this figure). For **(c-d)** we calculate the mean trait value for each allele length sum across QCed, unrelated White British participants, where each participant's contribution to each allele length sum's mean is weighted by that participant's imputed likelihood of having that allele length sum genotype. 95% confidence intervals were calculated similarly**.**

Figure S21: Associations of an STR in *CBL* with mean sphered cell volume, related to **Figure 4**



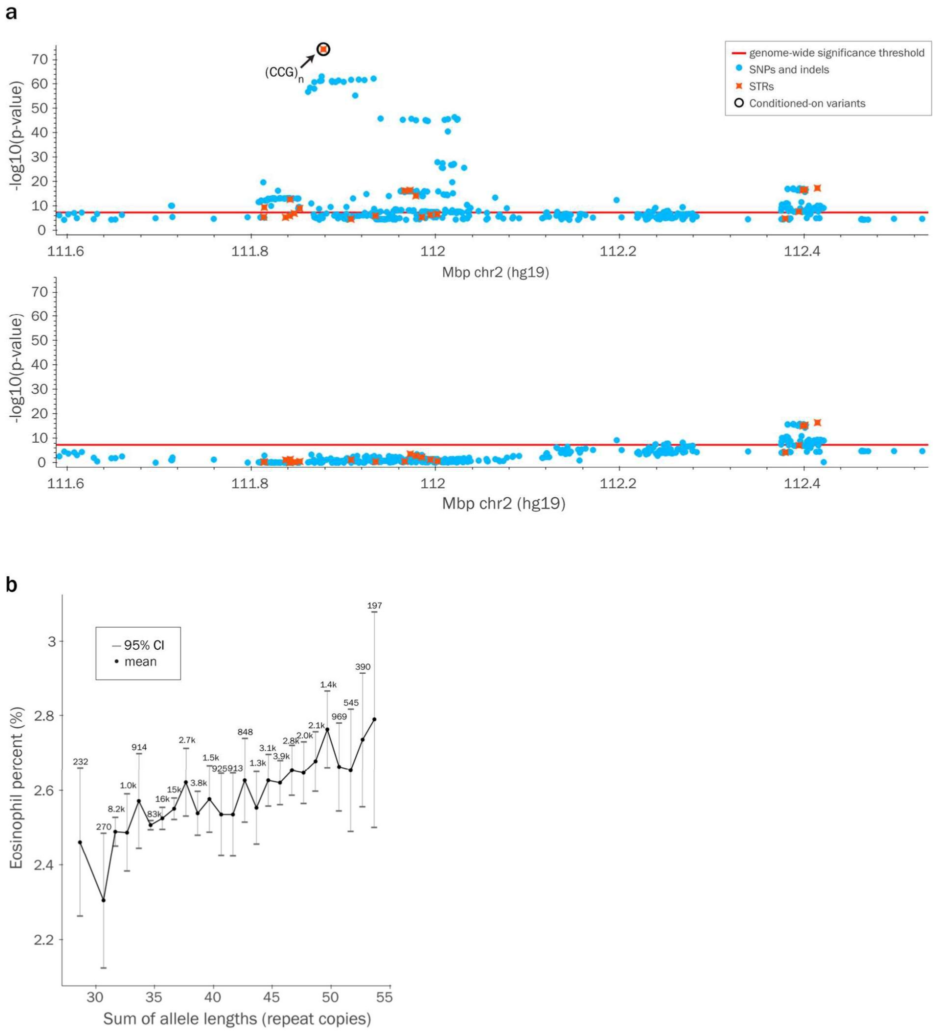
(a) Association between length (from whole genome sequencing) of the CGG repeat at chr11:119077000 (hg19) and mean sphered cell volume. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only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

Figure S22: Distribution of alleles of an STR in *CBL* across 1000 Genomes populations, related to **Figure 4**



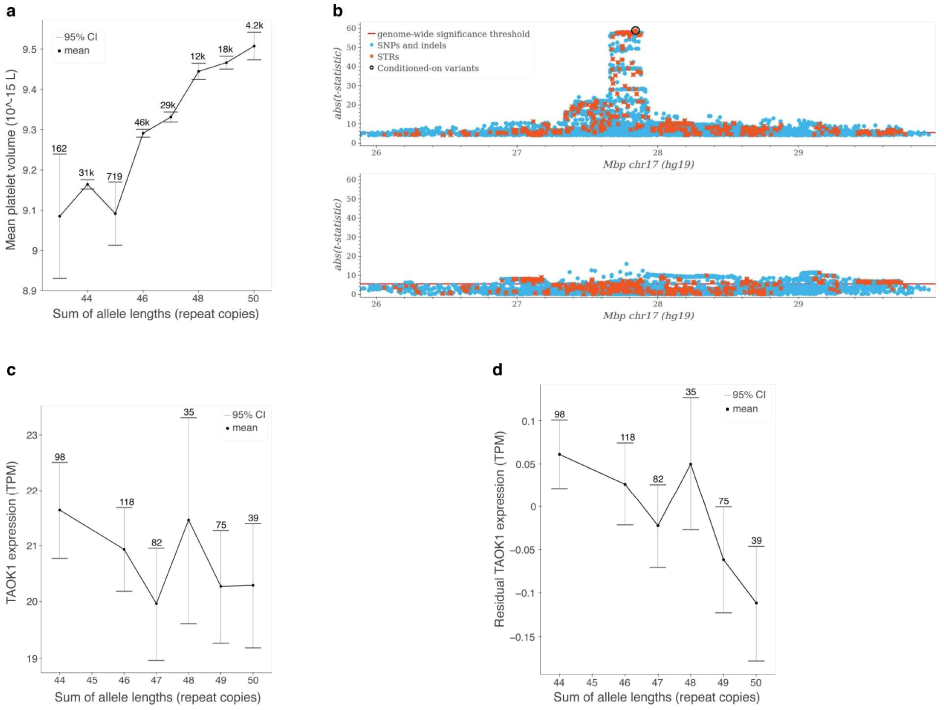
The x-axis gives STR length (number of repeat units) and y-axis gives the population frequency. The solid portion of each bar corresponds to the alleles of that length that include a “TGG” imperfection at the second repeat (rs7108857). Colors denote 1000 Genomes populations that were included in the Geuvadis cohort[26].

Figure S23: Association of an STR in *BCL2L11*, related to **Figure 2** which contains this and other confidently fine-mapped associations



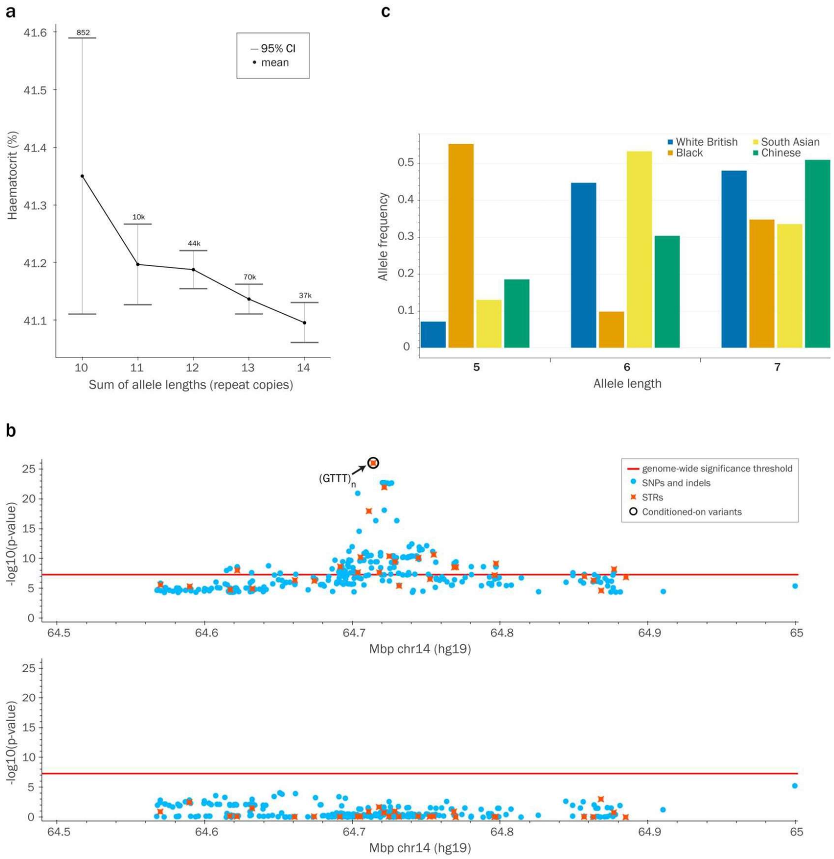
(a) Association of variants at the *BCL2L11* locus and eosinophil percent, before (top) and after (bottom) conditioning on the CCG repeat at chr2:111878544 (hg19). Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(CCG)_n$ STR. **(b)** Association between length (from whole genome sequencing) of the CCG repeat and eosinophil percentage. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only allele length sums with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum.

Figure S24: Association of an STR in *TAOK1*, related to **Figure 2** which contains this and other confidently fine-mapped associations



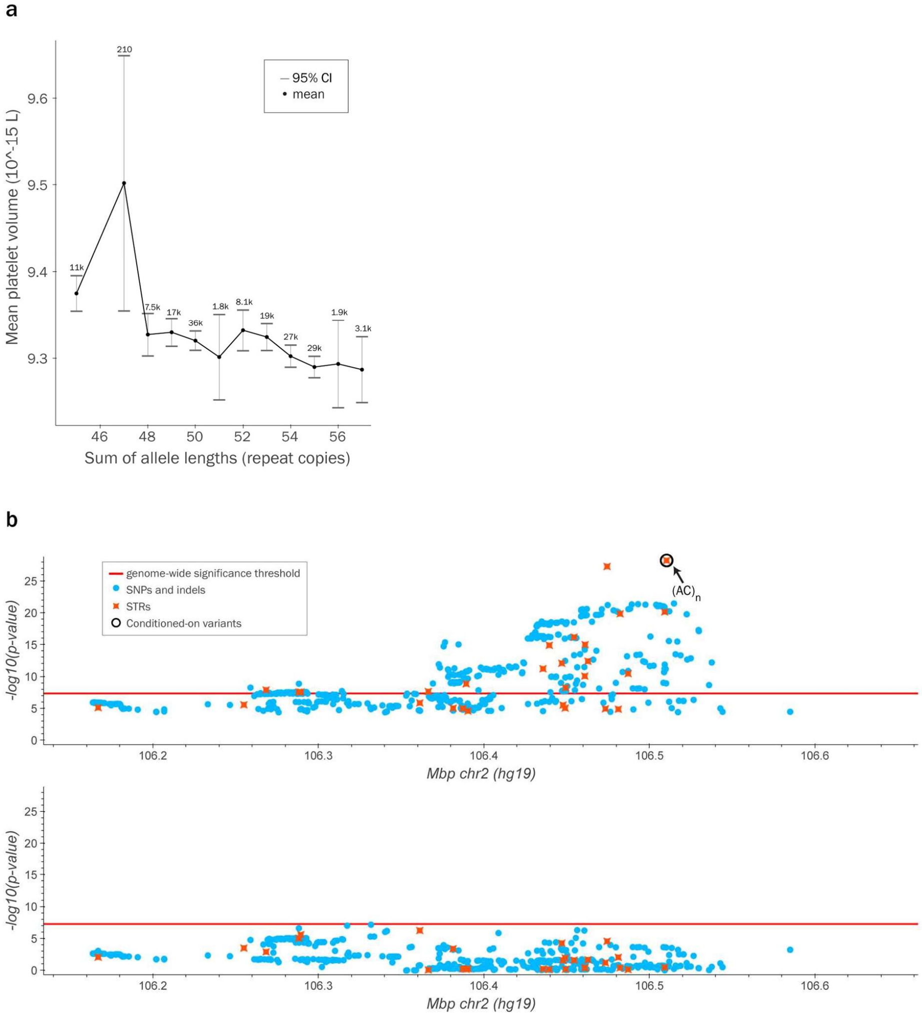
(a) Association between length (from whole genome sequencing) of the A repeat at chr17:27842016 (hg19) and mean platelet volume. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only allele length sums with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum. **(b)** Association of variants at the *TAOK1* locus and mean platelet volume before (top) and after (bottom) conditioning on the STR. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(A)_n$ STR. In this Manhattan plot we display associations according to the absolute value of their t-statistics instead of their -log10 p-values as those p-values exceeded the precision of our software (<1e-300). **(c-d)** Association between imputed best-guess genotypes of the repeat and *TAOK1* gene expression in thyroid tissue in the GTEx cohort[25]. Population-wide counts are displayed for each allele length sum. **(c)** displays the association with *TAOK1* expression TPM, while **(d)** displays the association with residual TPM values obtained after regressing out genetic principal components and PEER factors.

Figure S25: Association and allele distribution of an STR in *ESR2*, related to **Figure 2** which contains this and other confidently fine-mapped associations



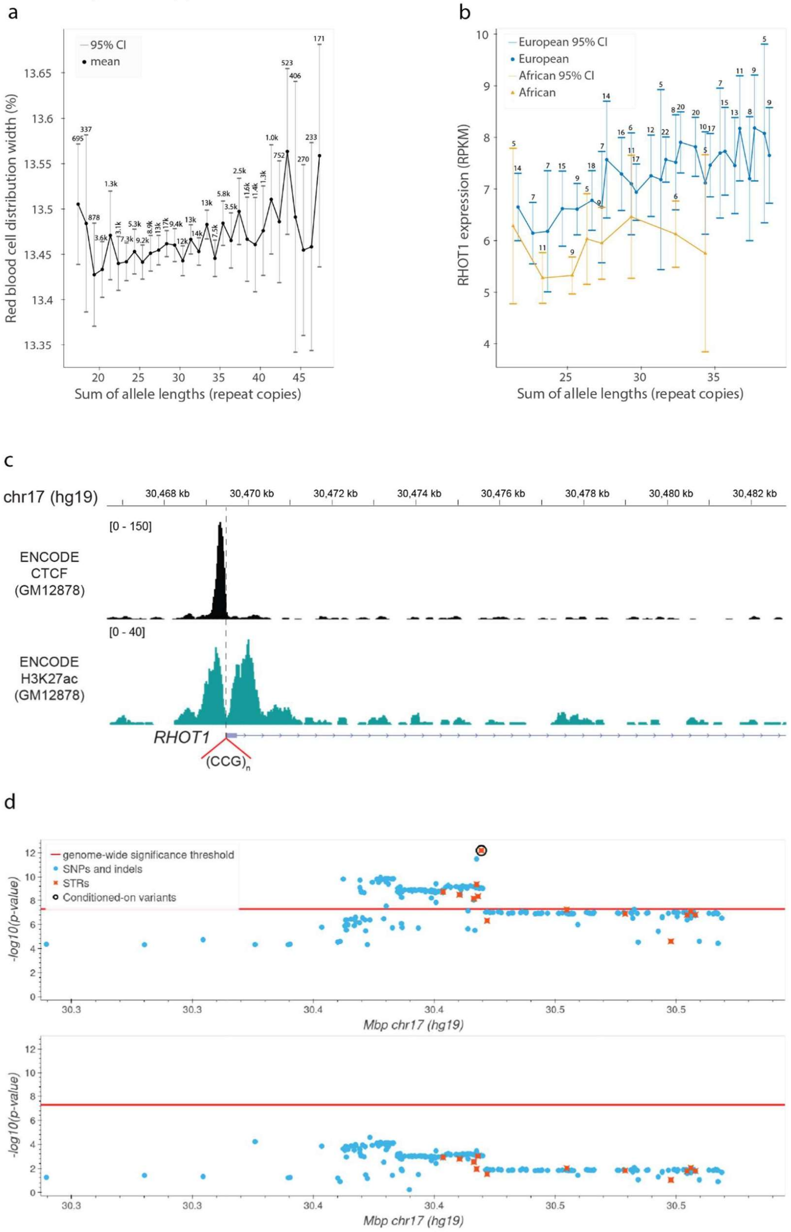
(a) Association between length (from whole genome sequencing) of the GTTT repeat at chr14:64714051 (hg19) and haematocrit. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only allele length sums with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum. **(b)** Association of variants at the *ESR2* locus and haematocrit. Conditioning on the repeat fully accounts for the signal seen in this region. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(GTTT)_n$ STR. **(c)** Differing distributions of STR length alleles from whole genome sequencing in different populations (blue=White British, orange=Black, yellow=South Asian; green=Chinese). Length alleles with frequency < 0.1% in all populations have been omitted.

Figure S26: Association of an STR in *NCK2*, related to **Figure 2** which contains this and other confidently fine-mapped associations



(a) Association between length (from whole genome sequencing) of the AC repeat at chr2:106510441 (hg19) and mean platelet volume. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only allele length sums with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum. **(b)** Association of variants at the *NCK2* locus and mean platelet volume. Conditioning on the repeat fully accounts for the signal seen in this region. Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(AC)_n$ STR.

Figure S27: Associations of an STR in *RHOT1*, related to **Figure 2** which contains this and other confidently fine-mapped associations



(a) Association between length (from UKB whole genome sequencing) of the CCG repeat at chr17:30469471 (hg19) and red blood cell distribution width. The mean trait value for each sum of STR allele lengths was calculated across QCed White British participants. 95% confidence intervals were calculated similarly. Only allele length sums with a population frequency of 0.1% or greater are displayed. Rounded population-wide counts are displayed for each sum. **(b)** Association between dosage of the repeat and *RHOT1* gene expression in the Geuvadis cohort[26] (LCLs; n=447). Solid lines give mean expression values for each STR dosage bin with at least 5% frequency in each group. Dosages were binned into groups spanning 3 repeat copies each since individually each genotype was relatively rare at this locus. **(c)** Positioning of the CCG repeat relative to the H3K27ac signal (note the localization within the nadir of the signal, which indicates a nucleosome depleted region) and a CTCF binding site at the 5' UTR of *RHOT1*. The visualization was generated using the Integrative Genomics Viewer[27] loading the ENCODE[28] data for GM12878 LCLs. The image does not display the gene NR_136413 that also overlaps the STR as it is not expressed in LCLs. **(d)** Association of variants at the *RHOT1* locus and red blood cell distribution width, before (top) and after (bottom) conditioning on the CCG repeat at chr17:30469471 (hg19). Light blue=SNPs and indels; orange=STRs. Red line=significance threshold, black circle=the $(CCG)_n$ STR.

Table S3: **Percentage of genetic variation attributable to STR lengths**, related to **STAR Methods** which provides SNP and STR processing details

Ethnicity	*White British*	*Black*	*South Asian*	*Chinese*	*Irish*	*White Other*
Non-major autosomal alleles per person	3,685,859.9	4,616,247.8	3,812,516.6	3,488,214.6	3,680,401.1	3,719,650.5
Non-major autosomal base pairs per person	5,046,660.5	6,300,734.3	5,216,199.4	4,770,162.0	5,039,114.7	5,092,602.0
STR percentage of nonmajor autosomal alleles per person	7.7%	7.2%	7.6%	7.5%	7.7%	7.7%
STR percentage of nonmajor autosomal base pairs per person	18.6%	18.1%	18.5%	18.3%	18.6%	18.6%

Here we calculate the average number of non-major alleles per participant from variants included in our GWAS, broken down by ethnicity. Variants include both SNPs/Indels, as well as STRs. We did not include indels which we deemed to be STR alleles in these calculations (**STAR Methods**). Both sets of variants are imputed; all counts in the table are derived from imputed dosages and so are fractional. In addition to the number of non-major alleles, we calculate the number of non-major base pairs per participant. We weighted non-major SNP alleles as 1 base pair, and non-major indel or STR alleles according to the length in base pairs of the insertion or deletion relative to the major allele in that ethnicity. There were no variants in the SNP/Indel dataset that could not be completely described as a single location with a SNP, simple insertion, or simple deletion. We observe that the percentage of non-major alleles and non-major base pairs attributable to STRs is substantially more consistent across populations than the total number of non-major alleles and non-major base pairs (which was greatest in the Black populations and least in the Chinese population). Caveat: these percentages only measure the total contribution of STR lengths to genetic variation, and thus we expect them to modestly underestimate the total contribution of STRs to genetic variation, which should include sequence variation in STR regions as well as length changes. We also note that our procedure for identifying indels as STR alleles was conservative, and so these numbers would likely increase to some extent with a more precisely permissive identification and filtering of such indels.

Table S6: Simulation strategy 1 minor allele frequency bins, related to **STAR Methods**, **Figure S6, Tables S7, S8**

min minor allele frequency (%)	**max minor allele frequency (%)**	**num.**	**% of total**	**num. causal**	**% of total causal**	**per variant weight**
0.01	0.1	147043	27.2%	73	18.4%	1.25E-06
0.1	1	81196	15.0%	70	17.6%	2.17E-06
1	10	115288	21.3%	112	28.2%	2.45E-06
10	50	197643	36.5%	142	35.8%	1.81E-06

The four bins we used for weighted random sampling for causal variants for simulation strategy 1.

min minor allele frequency (%)	The minimum minor allele frequency for variants in this bin
max minor allele frequency (%)	The maximum minor allele frequency for variants in this bin
num.	Total number of variants in this bin across all platelet count fine-mapping regions
% of total	Percent of all variants in all bins that fall in this bin
num. causal	Total number of variants with FINEMAP CP ≥ 0.5 or SuSiE CP ≥ 0.5 across all platelet count fine-mapping regions that fall in this bin
% of total causal	Percent of all causal variants in all bins that fall in this bin
per variant weight	Unnormalized probability, per variant, of being drawn as causal for simulation strategy 1.

See **Figure S6** for effect size distributions per minor allele frequency bin.

Table S8: Total CP attributed to SNPs/indels vs STRs in simulation by fine-mapper and simulation strategy, related to **STAR Methods**, **Figure S6, Tables S6, S7**

strategy	**n variants with FINEMAP CP > 0**	**fraction total FINEMAP CP to STRs**	**n variants with SuSiE CP > 0**	**fraction total SuSiE CP to STRs**
random_one_var	38797	0.007112815	6974	0.005032121
random_two_var	79982	0.008894848	14463	0.007882322
random_three_var	143480	0.009476106	19419	0.007620056
susie_no_strs	146316	0.030871123	24256	0.031883727

For this table, we only considered variants with p-value < 5e-8. For this table, we do not present results for the simulation strategy which included causal STRs as it did not simulate all regions genome wide and so this statistic would not be meaningful.

strategy	Fine-mapping strategy
n variants with FINEMAP CP > 0	Total across all simulations for this strategy
Fraction total FINEMAP CP to STRs	Of the CP assigned to those variants, the percent of that CP that was assigned to STRs by FINEMAP
n variants with SuSiE CP > 0	Total across all simulations for this strategy
Fraction total SuSiE CP to STRs	Of the CP assigned to those variants, the percent of that CP that was assigned to STRs by SuSiE

Table S10: Replication in Non-White British populations, related to **Figures 3, S15**

	Black	South Asian	Chinese	Irish	Other White
FINEMAP vs all gwsig	0.00146	0.00354	0.017	0.506	0.585
SuSiE vs all gwsig	0.00575	0.00333	0.0181	0.914	0.392
Confidently fine-mapped vs all gwsig	0.00124	0.015	0.00346	0.733	0.498
Confidently fine-mapped vs FINEMAP	0.0635	0.258	0.0196	0.194	0.19
Confidently fine-mapped vs SuSiE	0.0119	0.277	0.078	0.715	0.883

The table details effect of fine-mapping on the rate of replication of White British STR associations in other populations. The table is constructed as follows: each column represents a different population, and each row represents a comparison between different categories of STRs. For example, for the top left cell in the table, the replication population is the Black population, the model considers all gwsig STR-trait associations and the target category is all STR-trait associations selected by FINEMAP. The cells contain p-values how much more likely STR associations in the target category were to replicate than all STRs being considered in that test (further detailed in **STAR Methods**). The p-values highlighted in yellow are nominally significant ($p < 0.05$). It is unsurprising that we do not see that fine-mapping improves replication rates in White populations, as we expect nearly identical LD patterns in those populations and replication rates will only differ at loci with different ancestral LD structures.

Table S13: Sample sizes of the GTEx tissues, related to **STAR Methods** which provides other GTEx analysis details

Tissue	**Sample_n**
Adipose_Subcutaneous	663
Adipose_Visceral_Omentum	541
AdrenalGland	258
Artery_Aorta	432
Artery_Coronary	240
Artery_Tibial	663
Brain_Amygdala	152
Brain_Anteriorcingulatecortex_BA24	176
Brain_Caudate_basalganglia	246
Brain_CerebellarHemisphere	215
Brain_Cerebellum	241
Brain_Cortex	255
Brain_FrontalCortex_BA9	209
Brain_Hippocampus	197
Brain_Hypothalamus	202
Brain_Nucleusaccumbens_basalganglia	246
Brain_Putamen_basalganglia	205
Brain_Spinalcord_cervicalc_1	159
Breast_MammaryTissue	459
Cells_Culturedfibroblasts	504
Cells_EBV_transformedlymphocytes	174
Colon_Sigmoid	373
Colon_Transverse	406
Esophagus_GastroesophagealJunction	375
Esophagus_Mucosa	555
Esophagus_Muscularis	515
Heart_AtrialAppendage	429
Heart_LeftVentricle	432
Liver	226
Lung	578
Muscle_Skeletal	803
Nerve_Tibial	619
Pancreas	328
Pituitary	283
Skin_NotSunExposed_Suprapubic	604
Skin_SunExposed_Lowerleg	701
SmallIntestine_TerminalIleum	187
Spleen	241
Stomach	359
Thyroid	653
WholeBlood	755

The table provides the number of individuals analyzed for each tissue in GTEx for the expression analysis.
(For the methylation analysis, only whole blood was studied, and samples from 47 individuals were used).

References

1. 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. Ser. B Stat. Methodol. *82*, 1273–1300. 10.1111/rssb.12388.
2. 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. 10.1093/bioinformatics/btw018.
3. Flint, J. (2013). GWAS. Curr. Biol. *23*, R265–R266. 10.1016/j.cub.2013.01.040.
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. 10.1038/s41586-021-03819-2.
5. Wu, J.H., Wen, M.S., Lo, S.K., and Chern, M.S. (1994). Increased frequency of apolipoprotein B signal peptide sp24/24 in patients with coronary artery disease. General allele survey in the population of Taiwan and comparison with Caucasians. Clin. Genet. *45*, 250–254. 10.1111/j.1399-0004.1994.tb04150.x.
6. AlphaFold Protein Structure Database FAQ https://alphafold.ebi.ac.uk/faq.
7. Sturley, S.L., Talmud, P.J., Brasseur, R., Culbertson, M.R., Humphries, S.E., and Attie, A.D. (1994). Human apolipoprotein B signal sequence variants confer a secretion-defective phenotype when expressed in yeast. J. Biol. Chem. *269*, 21670–21675. 10.1016/S0021-9258(17)31858-6.
8. Hateboer, G., Kerkhoven, R.M., Shvarts, A., Bernards, R., and Beijersbergen, R.L. (1996). Degradation of E2F by the ubiquitin-proteasome pathway: regulation by retinoblastoma family proteins and adenovirus transforming proteins. Genes Dev. *10*, 2960–2970. 10.1101/gad.10.23.2960.
9. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. Nat. Genet. *53*, 1712–1721. 10.1038/s41588-021-00978-w.
10. Jiang, L., Wang, M., Lin, S., Jian, R., Li, X., Chan, J., Dong, G., Fang, H., Robinson, A.E., Aguet, F., et al. (2020). A Quantitative Proteome Map of the Human Body. Cell *183*, 269-283.e19. 10.1016/j.cell.2020.08.036.

11. Hsu, J., Arand, J., Chaikovsky, A., Mooney, N.A., Demeter, J., Brison, C.M., Oliverio, R., Vogel, H., Rubin, S.M., Jackson, P.K., et al. (2019). E2F4 regulates transcriptional activation in mouse embryonic stem cells independently of the RB family. Nat. Commun. *10*, 2939. 10.1038/s41467-019-10901-x.

12. Liu, J., Xia, L., Wang, S., Cai, X., Wu, X., Zou, C., Shan, B., Luo, M., and Wang, D. (2021). E2F4 Promotes the Proliferation of Hepatocellular Carcinoma Cells through Upregulation of CDCA3. J. Cancer *12*, 5173–5180. 10.7150/jca.53708.

13. Thorens, B. (2015). GLUT2, glucose sensing and glucose homeostasis. Diabetologia *58*, 221–232. 10.1007/s00125-014-3451-1.

14. Bánhegyi, G., Garzó, T., Antoni, F., and Mandl, J. (1988). Glycogenolysis - and not gluconeogenesis - is the source of UDP-glucuronic acid for glucuronidation. Biochim. Biophys. Acta BBA - Gen. Subj. *967*, 429–435. 10.1016/0304-4165(88)90106-7.

15. Meech, R., Hu, D.G., McKinnon, R.A., Mubarokah, S.N., Haines, A.Z., Nair, P.C., Rowland, A., and Mackenzie, P.I. (2019). The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms. Physiol. Rev. *99*, 1153–1222. 10.1152/physrev.00058.2017.

16. Edwards, M., Falzone, N., and Harrington, J. (2021). Conjugated hyperbilirubinemia among infants with hyperinsulinemic hypoglycemia. Eur. J. Pediatr. *180*, 1653–1657. 10.1007/s00431-021-03944-0.

17. 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. 10.1038/s12276-020-00550-2.

18. 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. 10.1152/ajpregu.00573.2001.

19. Lewandowski, S., Kalita, K., and Kaczmarek, L. (2002). Estrogen receptor β. FEBS Lett. *524*, 1–5. 10.1016/S0014-5793(02)03015-6.

20. Zhao, C., Dahlman-Wright, K., and Gustafsson, J.-Å. (2008). Estrogen Receptor β: An Overview and Update. Nucl. Recept. Signal. *6*, nrs.06003. 10.1621/nrs.06003.

21. 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. 10.1038/s41467-018-06694-0.

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

23. 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. 10.1038/s41588-019-0521-9.

24. 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. 10.1101/gr.277057.122.

25. THE GTEX CONSORTIUM (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. 10.1126/science.aaz1776.

26. 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. 10.1038/nature12531.

27. 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. 10.1038/nbt.1754.

28. Dunham, I., 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. 10.1038/nature11247.

Reprinting Acknowledgements

This chapter contains a full reprint of the publication Margoliash, J.; Fuchs, S.; Li, Y.; Zhang, X.; Massarat, A.; Goren, A.; Gymrek, M. Polymorphic Short Tandem Repeats Make Widespread Contributions to Blood and Serum Traits. *Cell Genomics* **2023**, *3* (12), 100458. https://doi.org/10.1016/j.xgen.2023.100458.

Discussion

Over the previous three chapters, I, my mentors Professor Melissa Gymrek and Professor Alon Goren, and our collaborators have demonstrated the ability to conduct genome-wide analyses of the involvement of STR lengths in human traits. We have shown that STR contributions are widespread, used compelling statistical evidence and plausible mechanistic hypotheses to identify likely-causal STRs which show effects both at the gene and phenotypic levels, and estimated that STRs perhaps represent 5% of all causal variants.

Broadening the pool of identifiably causal genetic variants to include STRs, or any other source of missing variation, contributes to the goals of population genetics research at large. Primarily, causal variant identification is an important contributor to fine-mapping GWAS signals. A variant causal for a GWAS signal can resolve the gene that signal acts through via overlap with functional annotations, can identify which cell type(s) that gene acts in if those annotations are cell type specific, and can illuminate which biomolecular mechanisms drive that signal. Even for signals where such insights are more difficult to come by, causal variant identification can motivate targeted experimental follow-up. When pools of candidate causal STRs with unidentified mechanisms are made available for study, that can motivate the discovery of new mechanisms that whole classes of STRs may act by. And even for the many GWAS signals where a causal variant cannot be easily identified, broadening the pool of variants tested to include more variant types slightly increases the chance that follow-up studies will identify the underlying causal variants.

We further hypothesize that identifying causal STRs will improve PRS. We expect this to be true in training populations in cases when LD between tagging and causal variants is lacking, but even more so when transferring PRS to target populations, where LD patterns can shift dramatically. We hope this is one of many contributions that can ease the inequity in current PRS performance across populations.

I spend the rest of this discussion outlining the current advances, challenges and available research projects in three areas of study related to my work: the study of common, genome-wide variation in STRs, the study of statistical fine-mapping, and the development of bioinformatics pipelines at terabyte scale. Afterwards, I provide a summary of the future projects I have proposed.

Prospects and Challenges in the Study of Common Variation in Short Tandem Repeats

The Need to Expand Existing STR Analyses

While there are many methodological improvements that would aid the study of causal common variation in STRs and I touch on a some of those below, our work has already demonstrated that current GWAS approaches, careful statistical fine-mapping, and cross referencing existing genetic annotations are sufficient for identifying many likely causal STRs. Thus I believe the most important next step in the study of causal common STR variation is the simplest one: expand the use of existing pipelines.

In our paper in Chapter 3 we studied blood traits, which being highly polygenic as well as reliably and widely phenotyped, served well for demonstrating that we could identify STRs causal for human traits. We should expand that work to the broad analysis of many medically relevant disease phenotypes. We know that causal STR signals which are identifiable by our current pipeline are relatively sparse, and that some of the identifiable signals are relatively weak. Thus it is unlikely a priori that causal common variation in STRs strongly contributes to a specific disease, but there is good evidence that such variation will causally contribute to some diseases, and we should be able to identify those diseases through sufficiently broad testing. Broad testing is similarly important for developing STR-strengthened PRS – we will either want to demonstrate that PRS built on top of STRs show small improvements compared to existing PRS across a wide swath of traits, and/or identify a small subset of traits whose PRS are substantially improved by the inclusion of STRs. Either goal will require the analysis of many traits.

In our paper in Chapter 1 we studied gene expression, and I discuss the limitations with existing gene expression data sets in the Chapter 1 Forward. Recently, blood proteome data was released for ~3,000 proteins in ~55,000 individuals in the UK Biobank[93]; we should expand our study of STR mechanisms to that dataset. Protein abundance measurements in blood are farther removed from transcriptional mechanisms than gene expression measurements. But protein levels are more closely related to many organism-wide traits than expression measurements, and thus this dataset potentially will provide more insight into STR contributions to disease. And while this data only encompasses up to 3,000 unique proteins, it includes many more individuals than existing eQTL datasets and will have much greater power to detect *cis* effects of STRs near the genes which code for those proteins. In short, this is a strongly powered dataset where it easier than normal to infer the target gene of any identified effect.

Further, we should extend our GWAS work to many more populations. The UK Biobank was a wonderful dataset for our initial STR GWAS due to its superior depth, breadth and quality of data. Yet it behooves us to extend our work to more diverse biobanks, such as All of Us[203] and the Million Veterans Program[204], as well as the large East Asian biobanks, so that any insights we garner will be more equitably applicable to the broader population. I acknowledge, however, that we have experienced barriers to working with the large US biobanks; I touch on those more below.

I note that recent research has been highly successful in causally linking common, multiallelic, coding VNTRs[148] and copy number repeats[82] to a variety of GWAS signals. While much of the work and benefit of our pipeline has been to identify likely-causal non-coding STRs for follow-up study, and while these common non-coding STRs in some cases explain as much phenotypic variation as any signal for the studied trait, coding or non-coding, it still makes sense to prioritize high-impact categories of STRs for further analysis. Focusing on high-impact STR categories would increase the rate at which we uncover maximally interpretable STR signals and would not preclude us from studying of non-coding variation afterward.

In particular, our work in Chapter 1 demonstrated that CG-rich STRs increase the relative stability of DNA during transcription as their length increases, and identified multiple statistically fine-mapped examples of this in promoter regions. Our work in Chapter 3 demonstrated that, among all STRs residing in GWAS signals, statistically fine-mapped STRs were enriched in 5’ UTR regions of genes. Chapter 3 also provided multiple examples of candidate causal CG-rich STRs in such regions. In line with this evidence, I believe this category of CG-rich STRs in 5’ UTR and promoter regions is the category of non-coding STRs with the most compelling evidence for widespread causality and strong effect sizes. This category also possesses a clear mechanistic hypothesis for explaining STR effects based on their lengths. So I believe a study focused on the STRs most likely to impact disease traits should focus not only on coding STRs but also CG-rich STRs in 5’ UTRs and promoter regions.

A recent preprint[200] has attempted just such a study. That study called ~36,000 TRs in WGS data for ~170,000 individuals in the UK Biobank cohort, focusing on TRs with mechanistic priors, including most of the coding STRs and most of the CG-rich STRs in promoters and 5’ UTRs, and tested them for associations with ~30,000 traits. Through this they discover many compelling associations. However, I note that there are a few aspects of the statistical fine-mapping performed by this study which may potentially have limited its power, suggesting that there may be many more such STR signals to uncover. First, they identify STRs as causal if they are conditionally independent from the top SNP in the region, which could result in loss of causal STRs that are well tagged by SNPs but distinguishable from such SNPs via statistical fine-mapping methods. Second, they do not mark STRs given a large PIP by statistical fine-mapping as causal if statistical fine-mapping puts a higher PIP on another variant in the region, which could result in the loss of causal STRs in regions with multiple independently causal variants.

As evidence, if one compares their fine-mapping results to our results from Chapter 3, one sees that both studies identify roughly the same number of causal transcribed STRs whose motifs only contain Cs and Gs (jointly identified by both studies: 2; their study only: 4; our study only: 2). But their study tested >650 times as many traits as ours. This suggests that their methodology may be missing many causal STRs, and that another study could complement their identified causal STRs with further discoveries.

Challenges Characterizing STR Mechanisms

There are still many limitations to detecting causal STRs. In my opinion, the biggest such limitation is the difficulty of assigning causal mechanisms to putatively causal STRs. Our work does suggest that CG-rich STRs in 5' UTR and promoter regions are likely to act through stabilizing non-canonical secondary structures. Yet aside from that, our results in Chapters 1 and 3 only suggest broad enrichment patterns within statistically fine-mapped STRs, which we have yet to find strong mechanistic interpretations for. In Chapter 1 we demonstrate that transcribed AC/GT STRs are more likely to be causally fine-mapped to gene expression if the T-rich repeat unit is on the template strand. In Chapter 3 we identify that transcribed non-coding STRs are enriched among statistically fine-mapped STRs. But neither of these findings suggests mechanisms for individual STRs that are dependent on STR length. Similarly, recent work by others[158] has shown that STR lengths can be causally linked to changes in local DNA methylation, but this work does not explain the mechanism by which this happens or provide the ability to a priori predict which STRs might contribute in this manner.

One of the most promising recent developments in the mechanistic interpretation of STRs is a publication by Horton et al.[123]. This study showed, in vitro, that many transcription factors preferentially bind specific STRs compared to random DNA sequences. Further, they show that when an STR preferred by a transcription factor is situated near a copy of that factor's canonical binding motif, the STR may increase the rate at which the factor recognizes and binds to that canonical motif.

Follow up study is needed to identify STRs in the human genome which may act through this mechanism. Specifically, Horton et al. identified 63 human transcription factors with preferential binding to at least one homopolymer, dinucleotide or tetranucleotide STR compared to random DNA sequences. This list of pairs of transcription factors and their preferred STRs should be made available so that the genome can be scanned for loci where the canonical motifs of these transcription factors are near to STRs which they prefer. Perhaps such loci could be added to the list of STRs to prioritize for future study described above. Importantly, in figure 6G of their paper, Horton et al. demonstrate that different transcription factors from the same family with the same canonical binding motifs may prefer to bind to markedly different STRs. This means that STR preferences should be demonstrated at the level of individual transcription factors, and not for transcription factor families.

Yet, as mentioned in Chapter 1 Figure 4 and the section of the Introduction on STRs, there are many hypothesized mechanisms by which STRs act. Few of these mechanisms have in vitro evidence at similar scale to the Horton study, much less in vivo evidence at scale. Fundamentally, further experimental work is needed to characterize STR mechanisms.

Improvements in STR Genotyping and Imputation Datasets

Essential to the hunt for causal STRs are STR genotypes. These have grown immensely in scale and quality over the course of my thesis work. The paper I contributed to in Chapter 1, conceived before I joined the Gymrek and Goren labs in early 2019, decided to study STRs in the GTEx[30] cohort partly because it was a large source of WGS data at the time and because no validated STR imputation panel existed then. By the time that paper was published and I moved to a new project, our lab had validated and published the Saini et al.[83] STR imputation panel in the 1000 Genomes cohort[13] which allowed us to study imputed STRs at population scale in the UK Biobank, published in Chapter 3.

During the course of that project, multiple important new milestones have been reached. Firstly, in 2023 our lab published the Ziaei Jam et al.[84] TR reference panel in the 1000 Genomes cohort. This is, to my knowledge, currently the largest reference panel for imputing STRs genome-wide without biobank access controls. It represents the combined output of four different TR callers, each with different capabilities, leaving calls at many loci validated by multiple callers, and including twice as many repeat loci and more repeat alleles per locus as compared to the Saini panel. Further, high coverage WGS data did not exist in the 1000 Genomes dataset when the Saini reference panel was published, so the Saini 1000 Genomes panel itself had to be imputed from a predominantly European cohort for which WGS data existed. The Ziaei Jam panel improves upon this by calling TRs directly in 1000 Genomes from now-existing WGS data, which is especially important for the accuracy of STR calls for non-European individuals. Fundamentally, the Ziaei Jam reference panel should be a large boon to upcoming TR GWAS and the study of a wider set of TRs.

Also in 2023 was the release of WGS STR calls in 150,000 participants in the UK Biobank cohort[133]. This is, to my knowledge, currently the largest dedicated WGS STR call set available to the broader scientific community, and an important milestone in the study of common variation in STRs. It should be noted that there are a few caveats to this resource. First, it is only accessible on the UK Biobank cloud platform, which incurs significant research effort overhead; I discuss this more below. Secondly, from personal experience I can say that the dataset is not comprehensively documented. Lastly, the dataset was generated with the popSTR STR caller[205]. To my knowledge, popSTR's accuracy has yet to be recapitulated by independent researchers, and I hope to see that done in the future.

Still, I believe the genotypes produced by popSTR are likely of high quality and that these hurdles will likely be overcome, making the 150,000 person WGS popSTR call set an amazing resource for the community. Also to note is that the preprint discussing the release of WGS calls in the full 500,000 participants in the UKB cohort does not specifically mention STR calls[14]. I wonder if its authors believe their general purpose indel caller is sufficiently good that a STR-specific caller is not needed (and if so that should be validated), or if they intend to run popSTR on the new individuals at a later date, or if they intend to leave the 150,000 individual call set (which was then imputed into the rest of the cohort) as the final STR call set in the UK Biobank.

On a much smaller scale, but still important, is the 2023 publication of WGS STR genotypes in nearly 4,000 Chinese individuals in the NyuWa cohort[195]. This is, to my knowledge, one of the largest WGS STR datasets outside of the UKB that is available to the research community and likely the largest such WGS STR dataset in a Chinese population. As such, it is a valuable resource.

All this improvement in STR genotyping will lead to testing of more STR loci and more accurate estimates of their effect sizes. This will clearly improve our statistical fine-mapping results to some extent; I wonder how much improvement that will provide. It is possible, if STR genotyping was more accurate than our imputed calls in Chapter 3, that statistical fine-mapping could have implicated many more multiallelic STRs as causal, as being multiallelic leaves them with reduced LD to surrounding SNPs. Thus modern STR genotyping accuracies may lead to a greatly increased number of causally identifiable STRs. This is only a hypothesis, and one that I do not have evidence to claim is likely; perhaps causal STR identification will remain at roughly the same rate with modern genotyping.

Another important unresolved question around modern STR genotyping is whether the current generation of general purpose WGS indel callers are up to the task of calling STRs, or if STRs should still be called using purpose-built algorithms. Determining this is important for two reasons. Firstly, the freely available TOPMed imputation service contains indel calls for over 130,000 individuals with a focus on many disease cohorts[56]. That resource would easily be the largest freely available STR imputation panel if it was sufficiently effective at calling STR variants, but until that can be verified, the Ziaei Jam et al. reference panel with 3,000 participants must fill that role.

Secondly, many biobanks are producing WGS data for their cohorts and are using general purpose indel callers, not STR-specific genotypers, on those datasets. For example, the All of Us cohort contains WGS indel calls for over nearly 250,000 participants[203]. If indel calls are not sufficient for studying STR variation, then researchers studying STRs genome-wide in those cohorts will likely rely on imputing STRs despite the WGS data, as calling STRs in hundreds of thousands of individuals from the available WGS data can be financially prohibitive.

Lastly, while it is not the focus of my research, I would like to briefly mention that tools for calling TRs have improved nicely in recent years. The EnsembleTR method pioneered by the Ziaei Jam reference panel paper[84] enables improved STR calling from short read data by overlapping the results of multiple short read STR methods. Efforts by our lab and others to genotype STRs from long read data are progressing nicely[199,206]. And effort is being put into the calling of impurities in STRs[124].

Open Questions in Modeling STRs

While there are many questions to resolve using existing STR data and methods, there are also unresolved questions around how STRs should be represented and associated with traits.

Firstly, while our work has demonstrated the involvement of STR lengths with human traits, it is also known that impurities in STRs – variation within a repeat sequence other than gains or losses of full copies of the repeat unit – can fundamentally change the biology of STRs[80,123,124]. Yet I do not know of a methodology that has been developed to identify the impacts of impurities genome-wide.

Some impurities (it is unclear how many) will be called sufficiently accurately by SNP and indel callers and their effects will be detected by standard GWAS. However, I expect there are many impurities which are difficult to properly represent, and existing approaches gloss over that challenge. If the reference genome at a location is AAAAA, and there are three eight base alternate alleles AAAAAAAA, AAATAAAA and AAAATAAA, are the two T impurities the same variant or not? As in, should all individuals with any length-eight allele with a T impurity be grouped together when testing for association with the number of alternate alleles of this T impurity? Or should each alternate allele containing a T be tested separately? There seems to be no a priori correct answer here, those two alleles could behave similarly (and so should be tested together) or differently (as so should be tested independently) depending on the biological mechanism of the repeat. I suspect that current pipelines do not treat this question in a principled or uniform manner, and instead leave this choice to the vagaries of the underlying variant calling software.

More complicated models of STR variation could be used to account for this case. For example, the repeat above could be represented as three separate variables: the number of As before any impurity (if present), the presence or absence of the impurity, and then the number of As after the impurity (or zero if no impurity is present). However, this deceptively simple example poses many unsolved research problems.

First, building this multivariable model of an STR becomes increasingly complicated if the STR contains many distinct impurities across different alleles, and it may be unclear what the best representation of such an STR is. Second, it is quite possible that there are strong interaction effects between impurities and STR lengths. For instance, one can hypothesize a repeat whose function is modulated by its length but has no function if an impurity interrupts that length. Testing for such an effect would mean moving beyond the standard method of treating each variant as a separate additive component.

Third, it is unclear how to determine if the association of a multivariable representation of an STR is significant enough that it is unlikely caused just by chance. Should each variable in the model be compared to the standard genome-wide significance threshold? Or should there be a single such test which combines information from all the variables? If so, as more complicated models always fit training data better than a linear model, such a test will need a more stringent threshold to avoid false positives. It is unclear what that threshold should be, and if causal effects can still reliably be detected beyond that higher standard.

Lastly, and perhaps most important, is how to statistically fine-map such results. Should each variable be additively included in an existing statistical fine-mapping model? If so, what should be made of the tendency of statistical fine-mapping methods to favor parsimonious models – those with few causal variables – and how should the causality of an STR be interpreted if statistical fine-mapping prioritizes as causal some of the variables used to represent it but not others? Or should new statistical fine-mapping methods be created with the requirement that they simultaneously select or reject as causal all variables in a multivariable representation of an STR? (Fine-mapping interaction effects would already require a new method to be developed). Either way, how should false-positive rates be measured and controlled?

Similar issues crop up in the handling of *complex short tandem repeats*, by which I mean multiple STRs located back-to-back, or STRs with non-standard repetitive patterns (for instance, a region which consists of a sequence of Ts alternating with either As or Gs). The *PACSIN2* locus described in the Chapter 3 Forward exemplifies both of these cases, and that discussion demonstrates the challenges of reasoning about such loci. As with representing impurities, complex loci could be handled by representing the region as multiple variables, possibly with interaction terms between them. This poses similar research questions as struggles with representing impurities.

Finally, it is very likely that there exist STRs in the human genome with markedly non-linear effects on human traits even at common repeat lengths. For example, this has been demonstrated in STRs yeast[207], and in VNTRs in humans[148]. Such STR effects may be difficult to detect by statistical fine-mapping if the linear model does not fit adequately, but may even be missed completely by linear association testing if the effects are fully non-linear. Hypothesized effects include effects which plateau beyond a given repeat length, or effects which attain a maximum or minimum at an intermediate repeat length.

In theory quadratic or other non-linear association tests could fit some of these non-linear trends better. These pose similar research questions as the STR representation issues mentioned above – what are the correct ways to test such models for association, and to fine-map their results? And for all these issues – representing impurities in STRs, representing complex STRs, and testing STRs for non-linear associations – the choice a researcher makes in modelling these phenomena should be influenced by what effects the researcher expects to exist in the genome, and what models are best powered to detect such effects.

Perhaps a look at the techniques for non-linear SNP association tests – i.e. dominant/recessive testing – would prove beneficial. Those efforts also must account for non-linear terms in both association testing and statistical fine-mapping, are there is a chance that some solutions to these problems may already be explored.

Validating Statistical Fine-Mapping

There has been plenty of exciting work extending statistical fine-mapping to new use cases. This includes incorporating annotations into fine-mapping priors, fine-mapping multiple traits simultaneously, and fine-mapping effects in individuals from multiple ethnicities. However, as I have already discussed throughout this thesis, the biggest limitation to statistical fine-mapping is the concern that its results may not live up to their stated probabilistic guarantees. For our work, statistical fine-mapping was usually our only source of causal evidence that was not subject to LD confounding, making this concern especially poignant, and leading us to go to great lengths to attempt to shore up its weaknesses. Here I focus on two efforts to remedy these issues. First, I discuss the possibility of building a benchmark for statistical fine-mapping methods. Second, I envision how to predict at which loci statistical fine-mapping tools will perform inconsistently.

Building a Benchmark

No benchmarks for statistical fine-mapping tools currently exist. This is a fundamental problem for the field: it has allowed instabilities in fine-mapping outputs, described recently by us in Chapter 3 and by others[168], to go undiscovered for many years prior. My hunch is that the lack of benchmarking has only been allowed to persist because of the perceived difficulty in curating sufficiently large sets of variants known to be causal for specific traits, where both the variant and the trait are reasonably common. (No statistical fine-mapping tool is expected to have enough power to succeed at detecting a causal variant for a trait when either the variant or the trait is quite rare).

In 2022, Alsheikh et al. used automated and manual curation of genetics publications to identify 309 non-coding variants with some form of experimental validation of their causality[73]. I believe this should be used to build a statistical fine-mapping benchmark. In order to determine if a set of putatively causal variants is suitable for a statistical fine-mapping benchmark, it needs to pass four tests: first, there must be enough variants, second, the variants and traits must be sufficiently common, third, the evidence for causality must be strong, and fourth, the evidence used to nominate these variants as causal must be orthogonal to the evidence used by statistical fine-mappers. Before this publication, I had seen lists of many causal variants that all had some level of annotated functional evidence, but statistical fine-mapping or colocalization was always used to handle potential LD confounding, violating the fourth test and making this list not being suitable as a benchmark. Alsheikh et al.'s list is the first I have seen which has curated large quantities of wet-lab evidence that is likely independent of LD-based statistical fine-mapping approaches. Wet-lab evidence cannot be generated at sufficient scale by a single research lab, but this study overcomes that through aggregation of publications across the research community.

A note on confounding: it should be expected that many variants in the Alsheikh et al. list had existing functional annotations which encouraged the wet-lab validation. Thus existing functional annotations are likely overrepresented among variants in this list compared to the class of causal variants as a whole. Yet this should not impact the utility of this list for benchmarking LD-based statistical fine-mapping methods, as they are unaware of genetic annotations, and LD structure should be reasonably independent of those annotations. However, this does preclude benchmarking functionally informed statistical fine-mapping, which would be confounded by this bias.

In a similar vein, some of these variants may have been prioritized for wet-lab validation because their low LD with nearby variants made them easy to identify in GWAS signal visualizations, or because they were picked out as likely to be causal by prior statistical fine-mapping efforts. Both of these will cause benchmarking statistics generated on this list of variants to be biased upwards compared to hypothetical statistics generated on the set of all causal variation. Thus results from such a benchmark must be interpreted cautiously.

Lastly, I do not expect this list of causal variants to be perfect. The curation of such a large dataset from individual papers is bound to have some level of error. Further, the curation in that study allowed for gene expression, reporter assays, chromatin interaction and transcription factor binding all to be used as forms of causal evidence, in addition to stronger forms of evidence such as CRISPR gene editing. While all valuable, the extent to which those weaker forms of evidence imply function to an individual variant as opposed to an LD block, and the extent to which they demonstrate that function is the cause for the trait-association and is not just coincidental, will depend on the experimental set up of the individual papers surveyed. Some of these papers, despite the care with which they were reasoned, will have come to incorrect causal hypotheses.

Yet these sources of error are potentially acceptable. As long as the Alsheikh et al. list is accurate enough in aggregate, a benchmark built on it should be able to distinguish between successful and unsuccessful fine-mapping efforts. Alternatively, if one category of experimental evidence seems sufficiently unconvincing as a form of causality validation, papers relying solely on that form evidence could be excluded from this list, and there would still be many causal variants remaining.

As one step towards building a statistical fine-mapping benchmark, a dedicated researcher could also tailor the methodology used by Alsheikh et al. to collect additional causal variants (or alternatively, collaborate with Alsheikh et al. to perform this tailoring). Firstly, the Alsheikh et al. list is two years old, and could be regenerated to include new results. Secondly, Alsheikh et al. designed their list to exclude coding variation, rare variation, non-SNP variation, variation causally associated with human traits that are not disease traits, and variation not tagged by GWAS signals. All those restrictions are not of interest to a fine-mapping benchmark (even rare variation is important for such a benchmark so the benchmark can demonstrate at what rarities statistical fine-mapping is no longer applicable). Having decided upon the set of causal variants to be used in this benchmark, the researcher will need to identify datasets and the phenotypes measured within them that match the corresponding phenotype affected by each of these causal variants. These datasets ideally should be accessible to the broader research community so that future authors can replicate the results of the benchmark and test newly designed tools against it.

Even once the causal variants and corresponding datasets and phenotypes are identified, I still expect building this benchmarking to be a time-intensive effort. It will require running many different types of statistical fine-mappers: while some are currently more popular that others, e.g. FINEMAP[176] and SuSiE[174], as no statistical fine-mappers to this point have been properly validated, this benchmark should be run on a wide range of methodologies. Moreover, each tool should be run under a range of alternative settings and priors to see if it can be optimally calibrated.

Dissecting the results of such an analysis will also require some care. The researcher will undoubtedly utilize the PIPs different fine-mappers assign to different causal variants. But it is unclear if fine-mappers should be compared by the number of the causal variants passing specific PIP thresholds or by the average PIP of causal variants. It would also be interesting to develop a metric for how well credible sets capture causal variation, to account for cases when fine-mappers correctly prioritize causal LD blocks but cannot identify the causal variants themselves. Further, the success of statistical fine-mappers is likely to be significantly influenced by the proportion of phenotypic variation explained by the causal variant, by the causal variant’s minor allele frequency, by the number of other variants in high LD with the causal variant, by sample size, and by the strength of causal signal from other variants in the region (this last influence could potentially be estimated by looking at the peak p-value after conditioning on the causal variant). It would be valuable to see fine-mapping results stratified by those details, and interesting to see if different statistical fine-mappers performed better in different sets of conditions. As a last suggestion, it would be great if such a study would randomly down-sample existing sample sizes to see how fine-mapping accuracy varies according to sample size.

This is all to say that creating a fine-mapping benchmark is a time-intensive effort, but one that may now succeed, and would be a valuable addition to the field if it did. A benchmark would directly improve the study of causal complex non-coding variation by informing which statistical fine-mappers should be used and how much they can be trusted. And it would improve the development of future statistical fine-mappers by giving them a reliable manner to compare themselves to their competitors.

Predicting Statistical Fine-Mapping Non-Reproducibility

In our paper in Chapter 3, we showed that (a) FINEMAP and SuSiE could not both be as reliable as they claimed as they frequently strongly contradicted one another, and that (b) at a relatively small number of loci FINEMAP strongly contradicts itself across successive runs. A recent, valuable, contribution by Cui et al. helped formalize this problem[168], identifying that statistical fine-mappers do not reproduce their own results as often as they guarantee they will. Cui et al. also identified that statistical fine-mapper guarantees are reliable when the fine-mappers are run on data from ideal simulations but not when they are run on real data. This shows that statistical fine-mapping unreliability stems from reliance on simulations with overly simplistic assumptions which do not fully capture patterns observed in real data.

From their choice of simulations, it can be inferred that Cui et al. are proposing that the difference between previous simulations and real data which has led to these broken guarantees is that in fact at least 0.5% of all variants in the genome are causal for every trait, but with small effect sizes. Further, they seem to be proposing that the total effect of these small effect size variants corresponds to at least 1.5 times as much heritability as non-sparse large effect size variants. Cui et al. then build new fine-mapping tools which model the genome with those assumptions in mind. I am unclear if these new models are the best solution to the problem they insightfully identified. Firstly, I would wish to see more evidence that ubiquitously causal variants contribute the majority of the genetic heritability for most traits. I also wonder if other invalid simulation assumptions may contribute to fine-mapping irreproducibility. Secondly, while I am encouraged by the decrease in replication failure and the increase in PRS performance shown by Cui et al.'s new methods, these methods have 10% less power than existing methods and only reduce the excess replication failure rate by a moderate amount. I believe it is possible Cui et al.'s new methodology is the correct solution to this problem, but I would like to see more evidence to that point.

Nevertheless, I find this paper's clean identification of the problem of replication failure to be quite important. I hypothesize that through bootstrapping (resampling of individuals) or resampling of univariate effect sizes under a few normality assumptions (and thus bypassing the need for bootstrapping and rerunning GWAS), statistical fine-mapping could be run many times on a single locus as if there were many independent datasets. If this was done, the frequency at which statistical fine-mapping reproduced or failed to reproduce identical results could be measured at a per-locus level. This would identify some signals at which statistical fine-mapping could not be trusted, and thus ameliorate one current source of error in statistical fine-mapping. That would, in turn, aid our ability to identify causal non-coding variation.

Bioinformatics Struggles

Struggles with Pipelining Biobank-Scale Datasets

I spent a great deal of time in my PhD writing bioinformatics pipelines for running population-scale GWAS on terabytes of data. I estimate that I spent more time in my PhD struggling with errors from these pipelines than time waiting for those pipelines to run, than analyzing results and thinking about the biology they implicate, or than writing papers. These problems significantly impeded the progress of my research and I hope that documenting them here may lead to a greater recognition of their cost to the field.

Most of my computational PhD work was done on computing clusters at the University of California San Diego campus. Not knowing of pipelining tools such as Snakemake and WDL, I began my work writing a variant imputation pipeline by hand. This involved manually batching the imputation steps and launching the jobs, writing scripts to check their successful completion, identifying and rerunning jobs which had failed, and launching a series of downstream steps to gather the data.

A year and a half into my PhD I was informed of the existence of the Snakemake pipelining tool[208] and moved all of my work to that system. A year and a half later I moved away from Snakemake and rewrote my pipeline in the WDL (Workflow Description Language) pipeline configuration language[209], running the WDL config files on my compute cluster using its Cromwell execution engine[210]. Both Snakemake and WDL dramatically eased my ability to create and scale analyses compared to writing pipelines manually. Of particular importance was these tools' ability to automate job submission, to automate job failure detection and resubmission, to automate submission of downstream jobs once the jobs they depended on had succeeded, to automatically reuse intermediate results from previous pipeline invocations, and to provide partially self-documenting pipeline configurations.

Yet my experience with each tool was also seriously flawed. My impression from brief conversations with colleagues in different fields is that pipelining tools outside of bioinformatics, especially in commercial cloud platforms, are much less painful to work with. Such solutions should be evaluated for use in the bioinformatics community. For now I will describe my challenges using Snakemake and WDL, noting that the pain points I am about to discuss would likely only be faced by projects operating at large scale.

I had two main struggles with Snakemake that eventually prompted me to move my pipeline configuration to WDL, turning my three thousand lines of Snakemake configuration into what would end up being ten thousand lines of WDL configuration. Firstly, Snakemake only worked on compute clusters with shared filesystems and not in cloud environments, and we were anticipating moving our analyses to the UK Biobank Research Analysis Platform (UKB RAP) on DNA Nexus[211]. Since December 2023, Snakemake has supported plugins to enable computing and storage in a variety of cloud platforms[212], so this is potentially no longer an issue. I do not know if such support is possible or currently available in proprietary biobank clouds such as the UKB RAP or the All of Us Researcher Workbench.

The other debilitating issue I ran into using Snakemake is that Snakemake would hang for over an hour when asked to run a pipeline whose corresponding job graph was sufficiently deep and wide. As this also occurred when asking Snakemake to validate such a pipeline, this made developing large pipelines in Snakemake impossible. I do not know if Snakemake's hanging problem has been fixed in the last two years.

Snakemake uses file creation timestamps to determine which jobs need to be rerun: if a file output by a job has an earlier timestamp than one of its inputs then that job is marked for re-execution. Having since moved to WDL, I now recognize that this was a major drawback to my method for developing pipelines. Initially, I did not list my scripts as inputs to the Snakemake jobs which used them. But as I continually update my scripts, that required me to manually track which version of each script had been used to generate each output, creating significant overhead to using Snakemake.

So I instead marked each script as an input to the job which used it. That way, whenever a script was modified its output would be regenerated. This design pattern works in WDL, for if the new version of the script generates identical output to the old version (perhaps because a new feature was added to the script which was not used by the existing job, or a corner case bug was fixed which only impacted a small percentage of output files) then WDL would notice that the new and old outputs were identical and would not rerun any downstream analyses. Snakemake, on the other hand, would only notice that the output file had a new timestamp and thus would rerun all downstream analyses, potentially wasting a lot of time and money. This made me worried to open existing scripts in case I accidentally saved them and updated their timestamps, and discouraged me from cleaning, documenting and debugging my scripts. The last struggle with Snakemake I wish to mention is that Snakemake's design choice of implicitly linking the inputs and outputs of jobs based on their filename patterns made it very difficult to debug when these filename regexes were accidentally misaligned.

None of the above Snakemake issues apply to WDL. WDL is supported on the UKB RAP through dxCompiler and via Cromwell on the All of Us Researcher Workbench (though I discuss issues with dxCompiler below). Further, WDL requires that tasks be explicitly composed through consecutive function calls. This has the downside of making WDL incredibly verbose. But it circumvents the challenge of generating call graphs which caused Snakemake to hang and reduces the difficulty of debugging those call graphs when they are improperly composed. Since WDL avoids Snakemake's pitfalls, I wish I could say that working with it has been substantially easier than working with Snakemake. But that has not been true.

Writing control flow logic in WDL is a nightmare. WDL's authors have strictly limited the data manipulation functions within the configuration language (even going so far as to exclude if-else statements). Further, they have prevented users from writing their own functions to ease these limitations. To my knowledge, these choices have been made to prevent users from running compute intensive work within the workflow execution environment itself. Yet, by forcing users into this good workflow practice, WDL's authors have created a language where it routinely takes an hour or more fabricating convoluted uses of WDL primitives in order to perform simple control flow operations, operations which would take a few minutes to write in any other context. This problem is compounded by the fact that these workarounds often lead to type errors, and these type errors often lead to difficult to interpret compile time errors. Or worse, these errors escape compile time checks and cause pipelines to fail with obtuse error messages midway through execution.

Further, WDL's support from its creators is seemingly lacking. Cromwell only supports v1.0 of the language, despite v1.1 having existed for more than three years. This prevents users from taking advantage of the improvements to WDL's data manipulation functions that were built into v1.1. I worry if WDL will cease to be maintained in the long run.

Another challenge with WDL is that configuring Cromwell for use on a computer cluster is difficult and nonintuitive. This is especially true of its call caching feature. To wit, despite my having written a multipage piece of documentation on how to set up this configuration, I have had a lab member decide after a week of struggling that configuring the Cromwell call caching feature was not worth the effort.

Lastly, while I benefit from Cromwell's call caching system, the directory structure it stores intermediate results in is poorly organized. Finding intermediate results involves traversing a filesystem, often ten levels deep, where half the directories have hash-code names. Moreover, when I run a pipeline, if its call graph or call caching choices do not match my expectations, having to navigate this filesystem makes it difficult to figure out what went wrong. And it is rarely possible to identify which files will be pulled from cache and which are obsolete, making it impossible to delete only the intermediate results which are out-of-date.

Again, I would encourage any new student working with population-scale workflows to use one of these two tools, or find an alternative workflow execution tool, rather than rolling their own solution. But I sincerely hope that new students struggle less with their adopted pipelining tools than I did.

Struggles Working with Biobanks and their Clouds

While pipelining tools added a lot of struggle to my PhD work, these challenges are only exacerbated by the biobanks themselves. A few years ago, the UK Biobank moved all their new large-scale genetics data releases to their Research Analysis Platform (RAP) cloud, hosted by DNA Nexus[211]. The UK Biobank now charges research teams ~$10,000 to access this platform, where any computation costs additional money. Yet this new platform seems only designed for small scale analyses. It provides easy access to Jupyter notebooks and remote terminals on single compute nodes and provides GUIs for composing small pipelines of existing tools, but building new pipelines that are distributed across multiple nodes has proven excruciating.

The RAP documentation recommends building distributed pipelines in WDL using their dxCompiler system. But their GUI and command line interfaces obfuscate debugging such pipeline. And their storage system does not reasonably support call caching, which was an essential feature for my productivity when working on my university's computer cluster. Further, we have run into numerous issues where simple scatter-gather analyses run through dxCompiler fail in surprising ways due to being run on too many input files, even though the express purpose of the RAP is to enable researchers to examine all of this data. DNA Nexus support has frequently been unable to resolve these issues, either because the support staffer assigned to our case did not understand the dxCompiler system, or because they did not respond for weeks. We have even suffered weeks-long delays in DNA Nexus processing our payments so that we could run additional compute jobs.

All this is to say that I was much more productive working on the computer cluster we were previously allowed to process UK Biobank data with than on the UKB RAP. Moreover, in large part due to the difficulties with the UKB RAP, our lab has yet to successfully work with the WGS STR calls in 150k individuals in the UK Biobank, despite this data having been accessible for a full year and likely being the deepest STR call set in the world.

Lastly, I note anecdotes from coworkers have suggested that working with the US biobanks, specifically the All of Us Research Workbench and the Million Veterans Program, is even more of a struggle than working with the UKB RAP. The recent massive growth of available human genetic data has been a huge boon to population genetics researchers. It seems that infrastructure to support the use of this data has not grown at nearly the same rate.

Proposed Projects

All in all, it is an exciting time to be a researcher of common human genetic variation. There are many unsolved challenges in the study of STRs and the use of statistical fine-mapping that are waiting for a researcher to tackle them. I have proposed many project directions throughout this discussion; I reproduce them here for clarity. Those projects which I believe to be the most important are:

- Expand existing STR GWAS and statistical fine-mapping to disease traits, with the prospect of applications to PRS
- Expand existing STR GWAS and fine-mapping efforts to US and East Asian biobanks for improved equity, and other diverse cohorts as they become available.
- Focus on phenome-wide associations with coding STRs, and CG-rich STRs in 5' UTR and promoter regions
- Use Horton et al.[123] to identify possible loci with transcription factor-STR interactions
- Build a statistical fine-mapping benchmark from the Alsheikh et al.[73] list of experimentally validated causal variants
- Develop methods for predicting statistical fine-mapping non-reproducibility
- Incorporate sex chromosome analyses into TRTools

Other projects I propose include:

- Identify STR-gene interactions in the UK Biobank blood proteome dataset
- Evaluate the TOPMed[56] imputation panel as a source of STR genotypes
- Evaluate the All of Us[203] WGS indel call set as a source of STR genotypes
- Evaluate the 150k individual WGS STR calls in the UK Biobank[133] and modify pipelines to use this resource

- Develop models for genome-wide scans and statistical fine-mapping of STR impurity effects
- Develop models for association and statistical fine-mapping of complex short tandem repeat loci
- Develop models for association and statistical fine-mapping of non-linear STR effects
- Develop guidance for running statistical fine-mapping jointly on eQTL and GWAS data when those datasets have thousand-fold differences in sample size
- Attempt to identify alternative pipelining tools to Snakemake and WDL

REFERENCES

(1) *Genetics vs. Genomics Fact Sheet*. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics (accessed 2024-04-23).

(2) Aspden, J. L.; Wallace, E. W. J.; Whiffin, N. Not All Exons Are Protein Coding: Addressing a Common Misconception. *Cell Genomics* **2023**, *3* (4). https://doi.org/10.1016/j.xgen.2023.100296.

(3) Rogalska, M. E.; Vivori, C.; Valcárcel, J. Regulation of Pre-mRNA Splicing: Roles in Physiology and Disease, and Therapeutic Prospects. *Nat. Rev. Genet.* **2023**, *24* (4), 251–269. https://doi.org/10.1038/s41576-022-00556-8.

(4) Jia, L.; Mao, Y.; Ji, Q.; Dersh, D.; Yewdell, J. W.; Qian, S.-B. Decoding mRNA Translatability and Stability from the 5′ UTR. *Nat. Struct. Mol. Biol.* **2020**, *27* (9), 814–821. https://doi.org/10.1038/s41594-020-0465-x.

(5) Nurk, S.; Koren, S.; Rhie, A.; Rautiainen, M.; Bzikadze, A. V.; Mikheenko, A.; Vollger, M. R.; Altemose, N.; Uralsky, L.; Gershman, A.; Aganezov, S.; Hoyt, S. J.; Diekhans, M.; Logsdon, G. A.; Alonge, M.; Antonarakis, S. E.; Borchers, M.; Bouffard, G. G.; Brooks, S. Y.; Caldas, G. V.; Chen, N.-C.; Cheng, H.; Chin, C.-S.; Chow, W.; de Lima, L. G.; Dishuck, P. C.; Durbin, R.; Dvorkina, T.; Fiddes, I. T.; Formenti, G.; Fulton, R. S.; Fungtammasan, A.; Garrison, E.; Grady, P. G. S.; Graves-Lindsay, T. A.; Hall, I. M.; Hansen, N. F.; Hartley, G. A.; Haukness, M.; Howe, K.; Hunkapiller, M. W.; Jain, C.; Jain, M.; Jarvis, E. D.; Kerpedjiev, P.; Kirsche, M.; Kolmogorov, M.; Korlach, J.; Kremitzki, M.; Li, H.; Maduro, V. V.; Marschall, T.; McCartney, A. M.; McDaniel, J.; Miller, D. E.; Mullikin, J. C.; Myers, E. W.; Olson, N. D.; Paten, B.; Peluso, P.; Pevzner, P. A.; Porubsky, D.; Potapova, T.; Rogaev, E. I.; Rosenfeld, J. A.; Salzberg, S. L.; Schneider, V. A.; Sedlazeck, F. J.; Shafin, K.; Shew, C. J.; Shumate, A.; Sims, Y.; Smit, A. F. A.; Soto, D. C.; Sović, I.; Storer, J. M.; Streets, A.; Sullivan, B. A.; Thibaud-Nissen, F.; Torrance, J.; Wagner, J.; Walenz, B. P.; Wenger, A.; Wood, J. M. D.; Xiao, C.; Yan, S. M.; Young, A. C.; Zarate, S.; Surti, U.; McCoy, R. C.; Dennis, M. Y.; Alexandrov, I. A.; Gerton, J. L.; O'Neill, R. J.; Timp, W.; Zook, J. M.; Schatz, M. C.; Eichler, E. E.; Miga, K. H.; Phillippy, A. M. The Complete Sequence of a Human Genome. *Science* **2022**, *376* (6588), 44–53. https://doi.org/10.1126/science.abj6987.

(6) Diamantopoulos, M. A.; Tsiakanikas, P.; Scorilas, A. Non-Coding RNAs: The Riddle of the Transcriptome and Their Perspectives in Cancer. *Ann. Transl. Med.* **2018**, *6* (12), 241. https://doi.org/10.21037/atm.2018.06.10.

(7) Amaral, P.; Carbonell-Sala, S.; De La Vega, F. M.; Faial, T.; Frankish, A.; Gingeras, T.; Guigo, R.; Harrow, J. L.; Hatzigeorgiou, A. G.; Johnson, R.; Murphy, T. D.; Pertea, M.; Pruitt, K. D.; Pujar, S.; Takahashi, H.; Ulitsky, I.; Varabyou, A.; Wells, C. A.; Yandell, M.; Carninci, P.; Salzberg, S. L. The Status of the Human Gene Catalogue. *Nature* **2023**, *622* (7981), 41–47. https://doi.org/10.1038/s41586-023-06490-x.

(8) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Chromosomal DNA and Its Packaging in the Chromatin Fiber. In *Molecular Biology of the Cell. 4th edition*; Garland Science, 2002.

(9) Ernst, J.; Kellis, M. Chromatin-State Discovery and Genome Annotation with ChromHMM. *Nat. Protoc.* **2017**, *12* (12), 2478–2492. https://doi.org/10.1038/nprot.2017.124.

(10) *ENCODE Encyclopedia, Version 4 (Beta): Overview – ENCODE*. https://www.encodeproject.org/data/annotations/v4beta/ (accessed 2024-05-09).

(11) *ENCODE Encyclopedia, Version 4: Genomic Annotations – ENCODE*. https://www.encodeproject.org/data/annotations (accessed 2024-05-09).

(12) Moore, J. E.; Purcaro, M. J.; Pratt, H. E.; Epstein, C. B.; Shoresh, N.; Adrian, J.; Kawli, T.; Davis, C. A.; Dobin, A.; Kaul, R.; Halow, J.; Van Nostrand, E. L.; Freese, P.; Gorkin, D.

U.; Shen, Y.; He, Y.; Mackiewicz, M.; Pauli-Behn, F.; Williams, B. A.; Mortazavi, A.; Keller, C. A.; Zhang, X.-O.; Elhajjajy, S. I.; Huey, J.; Dickel, D. E.; Snetkova, V.; Wei, X.; Wang, X.; Rivera-Mulia, J. C.; Rozowsky, J.; Zhang, J.; Chhetri, S. B.; Zhang, J.; Victorsen, A.; White, K. P.; Visel, A.; Yeo, G. W.; Burge, C. B.; Lécuyer, E.; Gilbert, D. M.; Dekker, J.; Rinn, J.; Mendenhall, E. M.; Ecker, J. R.; Kellis, M.; Klein, R. J.; Noble, W. S.; Kundaje, A.; Guigó, R.; Farnham, P. J.; Cherry, J. M.; Myers, R. M.; Ren, B.; Graveley, B. R.; Gerstein, M. B.; Pennacchio, L. A.; Snyder, M. P.; Bernstein, B. E.; Wold, B.; Hardison, R. C.; Gingeras, T. R.; Stamatoyannopoulos, J. A.; Weng, Z. Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature* **2020**, *583* (7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4.

(13) 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.; Fairley, S.; Runnels, A.; Winterkorn, L.; Lowy, E.; Eichler, E. E.; Korbel, J. O.; Lee, C.; Marschall, T.; Devine, S. E.; Harvey, W. T.; Zhou, W.; Mills, R. E.; Rausch, T.; Kumar, S.; Alkan, C.; Hormozdiari, F.; Chong, Z.; Chen, Y.; Yang, X.; Lin, J.; Gerstein, M. B.; Kai, Y.; Zhu, Q.; Yilmaz, F.; Xiao, C.; Flicek, P.; Germer, S.; Brand, H.; Hall, I. M.; Talkowski, M. E.; Narzisi, G.; Zody, M. C. High-Coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. *Cell* **2022**, *185* (18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

(14) Li, S.; Carss, K. J.; Halldorsson, B. V.; Cortes, A.; Consortium, U. B. W.-G. S. Whole-Genome Sequencing of Half-a-Million UK Biobank Participants. medRxiv December 8, 2023, p 2023.12.06.23299426. https://doi.org/10.1101/2023.12.06.23299426.

(15) Chiang, C.; Scott, A. J.; Davis, J. R.; Tsang, E. K.; Li, X.; Kim, Y.; Hadzic, T.; Damani, F. N.; Ganel, L.; Montgomery, S. B.; Battle, A.; Conrad, D. F.; Hall, I. M. The Impact of Structural Variation on Human Gene Expression. *Nat. Genet.* **2017**, *49* (5), 692–699. https://doi.org/10.1038/ng.3834.

(16) Sudmant, P. H.; Rausch, T.; Gardner, E. J.; Handsaker, R. E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Hsi-Yang Fritz, M.; Konkel, M. K.; Malhotra, A.; Stütz, A. M.; Shi, X.; Paolo Casale, F.; Chen, J.; Hormozdiari, F.; Dayama, G.; Chen, K.; Malig, M.; Chaisson, M. J. P.; Walter, K.; Meiers, S.; Kashin, S.; Garrison, E.; Auton, A.; Lam, H. Y. K.; Jasmine Mu, X.; Alkan, C.; Antaki, D.; Bae, T.; Cerveira, E.; Chines, P.; Chong, Z.; Clarke, L.; Dal, E.; Ding, L.; Emery, S.; Fan, X.; Gujral, M.; Kahveci, F.; Kidd, J. M.; Kong, Y.; Lameijer, E.-W.; McCarthy, S.; Flicek, P.; Gibbs, R. A.; Marth, G.; Mason, C. E.; Menelaou, A.; Muzny, D. M.; Nelson, B. J.; Noor, A.; Parrish, N. F.; Pendleton, M.; Quitadamo, A.; Raeder, B.; Schadt, E. E.; Romanovitch, M.; Schlattl, A.; Sebra, R.; Shabalin, A. A.; Untergasser, A.; Walker, J. A.; Wang, M.; Yu, F.; Zhang, C.; Zhang, J.; Zheng-Bradley, X.; Zhou, W.; Zichner, T.; Sebat, J.; Batzer, M. A.; McCarroll, S. A.; Mills, R. E.; Gerstein, M. B.; Bashir, A.; Stegle, O.; Devine, S. E.; Lee, C.; Eichler, E. E.; Korbel, J. O. An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* **2015**, *526* (7571), 75–81. https://doi.org/10.1038/nature15394.

(17) Fan, J.-B.; Chee, M. S.; Gunderson, K. L. Highly Parallel Genomic Assays. *Nat. Rev. Genet.* **2006**, *7* (8), 632–644. https://doi.org/10.1038/nrg1901.

(18) *All of Us Research Program Makes Nearly 250,000 Whole Genome Sequences Available to Advance Precision Medicine*. All of Us Research Program | NIH. https://allofus.nih.gov/news-events/announcements/all-us-research-program-makes-nearly-250000-whole-genome-sequences-available-advance-precision-medicine (accessed 2024-05-21).

(19) Beyter, D.; Ingimundardottir, H.; Oddsson, A.; Eggertsson, H. P.; Bjornsson, E.; Jonsson, H.; Atlason, B. A.; Kristmundsdottir, S.; Mehringer, S.; Hardarson, M. T.; Gudjonsson, S. A.; Magnusdottir, D. N.; Jonasdottir, A.; Jonasdottir, A.; Kristjansson, R. P.; Sverrisson, S. T.; Holley, G.; Palsson, G.; Stefansson, O. A.; Eyjolfsson, G.; Olafsson, I.;

Sigurdardottir, O.; Torfason, B.; Masson, G.; Helgason, A.; Thorsteinsdottir, U.; Holm, H.; Gudbjartsson, D. F.; Sulem, P.; Magnusson, O. T.; Halldorsson, B. V.; Stefansson, K. Long-Read Sequencing of 3,622 Icelanders Provides Insight into the Role of Structural Variants in Human Diseases and Other Traits. *Nat. Genet.* **2021**, *53* (6), 779–786. https://doi.org/10.1038/s41588-021-00865-4.

(20) Mars, N.; Widén, E.; Kerminen, S.; Meretoja, T.; Pirinen, M.; della Briotta Parolo, P.; Palta, P.; Palotie, A.; Kaprio, J.; Joensuu, H.; Daly, M.; Ripatti, S. The Role of Polygenic Risk and Susceptibility Genes in Breast Cancer over the Course of Life. *Nat. Commun.* **2020**, *11* (1), 6383. https://doi.org/10.1038/s41467-020-19966-5.

(21) Visscher, P. M.; Wray, N. R.; Zhang, Q.; Sklar, P.; McCarthy, M. I.; Brown, M. A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101* (1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

(22) Omenn, G. S.; Lane, L.; Overall, C. M.; Pineau, C.; Packer, N. H.; Cristea, I. M.; Lindskog, C.; Weintraub, S. T.; Orchard, S.; Roehrl, M. H. A.; Nice, E.; Liu, S.; Bandeira, N.; Chen, Y.-J.; Guo, T.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. *J. Proteome Res.* **2023**, *22* (4), 1024–1042. https://doi.org/10.1021/acs.jproteome.2c00498.

(23) The Gene Ontology Consortium; Aleksander, S. A.; Balhoff, J.; Carbon, S.; Cherry, J. M.; Drabkin, H. J.; Ebert, D.; Feuermann, M.; Gaudet, P.; Harris, N. L.; Hill, D. P.; Lee, R.; Mi, H.; Moxon, S.; Mungall, C. J.; Muruganugan, A.; Mushayahama, T.; Sternberg, P. W.; Thomas, P. D.; Van Auken, K.; Ramsey, J.; Siegele, D. A.; Chisholm, R. L.; Fey, P.; Aspromonte, M. C.; Nugnes, M. V.; Quaglia, F.; Tosatto, S.; Giglio, M.; Nadendla, S.; Antonazzo, G.; Attrill, H.; dos Santos, G.; Marygold, S.; Strelets, V.; Tabone, C. J.; Thurmond, J.; Zhou, P.; Ahmed, S. H.; Asanitthong, P.; Luna Buitrago, D.; Erdol, M. N.; Gage, M. C.; Ali Kadhum, M.; Li, K. Y. C.; Long, M.; Michalak, A.; Pesala, A.; Pritazahra, A.; Saverimuttu, S. C. C.; Su, R.; Thurlow, K. E.; Lovering, R. C.; Logie, C.; Oliferenko, S.; Blake, J.; Christie, K.; Corbani, L.; Dolan, M. E.; Drabkin, H. J.; Hill, D. P.; Ni, L.; Sitnikov, D.; Smith, C.; Cuzick, A.; Seager, J.; Cooper, L.; Elser, J.; Jaiswal, P.; Gupta, P.; Jaiswal, P.; Naithani, S.; Lera-Ramirez, M.; Rutherford, K.; Wood, V.; De Pons, J. L.; Dwinell, M. R.; Hayman, G. T.; Kaldunski, M. L.; Kwitek, A. E.; Laulederkind, S. J. F.; Tutaj, M. A.; Vedi, M.; Wang, S.-J.; D'Eustachio, P.; Aimo, L.; Axelsen, K.; Bridge, A.; Hyka-Nouspikel, N.; Morgat, A.; Aleksander, S. A.; Cherry, J. M.; Engel, S. R.; Karra, K.; Miyasato, S. R.; Nash, R. S.; Skrzypek, M. S.; Weng, S.; Wong, E. D.; Bakker, E.; Berardini, T. Z.; Reiser, L.; Auchincloss, A.; Axelsen, K.; Argoud-Puy, G.; Blatter, M.-C.; Boutet, E.; Breuza, L.; Bridge, A.; Casals-Casas, C.; Coudert, E.; Estreicher, A.; Livia Famiglietti, M.; Feuermann, M.; Gos, A.; Gruaz-Gumowski, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Pedruzzi, I.; Pourcel, L.; Poux, S.; Rivoire, C.; Sundaram, S.; Bateman, A.; Bowler-Barnett, E.; Bye-A-Jee, H.; Denny, P.; Ignatchenko, A.; Ishtiaq, R.; Lock, A.; Lussi, Y.; Magrane, M.; Martin, M. J.; Orchard, S.; Raposo, P.; Speretta, E.; Tyagi, N.; Warner, K.; Zaru, R.; Diehl, A. D.; Lee, R.; Chan, J.; Diamantakis, S.; Raciti, D.; Zarowiecki, M.; Fisher, M.; James-Zorn, C.; Ponferrada, V.; Zorn, A.; Ramachandran, S.; Ruzicka, L.; Westerfield, M. The Gene Ontology Knowledgebase in 2023. *Genetics* **2023**, *224* (1), iyad031. https://doi.org/10.1093/genetics/iyad031.

(24) Milacic, M.; Beavers, D.; Conley, P.; Gong, C.; Gillespie, M.; Griss, J.; Haw, R.; Jassal, B.; Matthews, L.; May, B.; Petryszak, R.; Ragueneau, E.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Stephan, R.; Tiwari, K.; Varusai, T.; Weiser, J.; Wright, A.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* **2024**, *52* (D1), D672–D678. https://doi.org/10.1093/nar/gkad1025.

(25) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back,

T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(26) Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L. H.; Zielinski, M.; Sargeant, T.; Schneider, R. G.; Senior, A. W.; Jumper, J.; Hassabis, D.; Kohli, P.; Avsec, Ž. Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense. *Science* **2023**, *381* (6664), eadg7492. https://doi.org/10.1126/science.adg7492.

(27) *Variant Effect Predictors*. Atlas of Variant Effects Alliance. https://www.varianteffect.org/veps (accessed 2024-05-01).

(28) Livesey, B. J.; Marsh, J. A. Updated Benchmarking of Variant Effect Predictors Using Deep Mutational Scanning. *Mol. Syst. Biol.* **2023**, *19* (8), e11474. https://doi.org/10.15252/msb.202211474.

(29) Livesey, B. J.; Marsh, J. A. Interpreting Protein Variant Effects with Computational Predictors and Deep Mutational Scanning. *Dis. Model. Mech.* **2022**, *15* (6), dmm049510. https://doi.org/10.1242/dmm.049510.

(30) Aguet, F.; Brown, A. A.; Castel, S. E.; Davis, J. R.; He, Y.; Jo, B.; Mohammadi, P.; Park, Y.; Parsana, P.; Segrè, A. V.; Strober, B. J.; Zappala, Z.; Cummings, B. B.; Gelfand, E. T.; Hadley, K.; Huang, K. H.; Lek, M.; Li, X.; Nedzel, J. L.; Nguyen, D. Y.; Noble, M. S.; Sullivan, T. J.; Tukiainen, T.; MacArthur, D. G.; Getz, G.; Addington, A.; Guan, P.; Koester, S.; Little, A. R.; Lockhart, N. C.; Moore, H. M.; Rao, A.; Struewing, J. P.; Volpi, S.; Brigham, L. E.; Hasz, R.; Hunter, M.; Johns, C.; Johnson, M.; Kopen, G.; Leinweber, W. F.; Lonsdale, J. T.; McDonald, A.; Mestichelli, B.; Myer, K.; Roe, B.; Salvatore, M.; Shad, S.; Thomas, J. A.; Walters, G.; Washington, M.; Wheeler, J.; Bridge, J.; Foster, B. A.; Gillard, B. M.; Karasik, E.; Kumar, R.; Miklos, M.; Moser, M. T.; Jewell, S. D.; Montroy, R. G.; Rohrer, D. C.; Valley, D.; Mash, D. C.; Davis, D. A.; Sobin, L.; Barcus, M. E.; Branton, P. A.; Abell, N. S.; Balliu, B.; Delaneau, O.; Frésard, L.; Gamazon, E. R.; Garrido-Martín, D.; Gewirtz, A. D. H.; Gliner, G.; Gloudemans, M. J.; Han, B.; He, A. Z.; Hormozdiari, F.; Li, X.; Liu, B.; Kang, E. Y.; McDowell, I. C.; Ongen, H.; Palowitch, J. J.; Peterson, C. B.; Quon, G.; Ripke, S.; Saha, A.; Shabalin, A. A.; Shimko, T. C.; Sul, J. H.; Teran, N. A.; Tsang, E. K.; Zhang, H.; Zhou, Y.-H.; Bustamante, C. D.; Cox, N. J.; Guigó, R.; Kellis, M.; McCarthy, M. I.; Conrad, D. F.; Eskin, E.; Li, G.; Nobel, A. B.; Sabatti, C.; Stranger, B. E.; Wen, X.; Wright, F. A.; Ardlie, K. G.; Dermitzakis, E. T.; Lappalainen, T.; Aguet, F.; Ardlie, K. G.; Cummings, B. B.; Gelfand, E. T.; Getz, G.; Hadley, K.; Handsaker, R. E.; Huang, K. H.; Kashin, S.; Karczewski, K. J.; Lek, M.; Li, X.; MacArthur, D. G.; Nedzel, J. L.; Nguyen, D. T.; Noble, M. S.; Segrè, A. V.; Trowbridge, C. A.; Tukiainen, T.; Abell, N. S.; Balliu, B.; Barshir, R.; Basha, O.; Battle, A.; Bogu, G. K.; Brown, A.; Brown, C. D.; Castel, S. E.; Chen, L. S.; Chiang, C.; Conrad, D. F.; Cox, N. J.; Damani, F. N.; Davis, J. R.; Delaneau, O.; Dermitzakis, E. T.; Engelhardt, B. E.; Eskin, E.; Ferreira, P. G.; Frésard, L.; Gamazon, E. R.; Garrido-Martín, D.; Gewirtz, A. D. H.; Gliner, G.; Gloudemans, M. J.; Guigo, R.; Hall, I. M.; Han, B.; He, Y.; Hormozdiari, F.; Howald, C.; Kyung Im, H.; Jo, B.; Yong Kang, E.; Kim, Y.; Kim-Hellmuth, S.; Lappalainen, T.; Li, G.; Li, X.; Liu, B.; Mangul, S.; McCarthy, M. I.; McDowell, I. C.; Mohammadi, P.; Monlong, J.; Montgomery, S. B.; Muñoz-Aguirre, M.; Ndungu, A. W.; Nicolae, D. L.; Nobel, A. B.; Oliva, M.; Ongen, H.; Palowitch, J. J.; Panousis, N.; Papasaikas, P.; Park, Y.; Parsana, P.; Payne, A. J.; Peterson, C. B.; Quan, J.; Reverter, F.; Sabatti, C.; Saha, A.; Sammeth, M.; Scott, A. J.; Shabalin, A. A.; Sodaei, R.; Stephens, M.; Stranger, B. E.; Strober, B. J.; Sul, J. H.; Tsang, E. K.; Urbut, S.; van de Bunt, M.; Wang, G.; Wen, X.; Wright, F. A.; Xi, H. S.; Yeger-Lotem, E.; Zappala, Z.; Zaugg, J. B.; Zhou, Y.-H.; Akey, J. M.; Bates, D.; Chan, J.; Chen, L. S.; Claussnitzer, M.; Demanelis, K.; Diegel, M.; Doherty, J. A.; Feinberg, A. P.; Fernando, M. S.; Halow, J.; Hansen, K. D.; Haugen, E.;

Hickey, P. F.; Hou, L.; Jasmine, F.; Jian, R.; Jiang, L.; Johnson, A.; Kaul, R.; Kellis, M.; Kibriya, M. G.; Lee, K.; Billy Li, J.; Li, Q.; Li, X.; Lin, J.; Lin, S.; Linder, S.; Linke, C.; Liu, Y.; Maurano, M. T.; Molinie, B.; Montgomery, S. B.; Nelson, J.; Neri, F. J.; Oliva, M.; Park, Y.; Pierce, B. L.; Rinaldi, N. J.; Rizzardi, L. F.; Sandstrom, R.; Skol, A.; Smith, K. S.; Snyder, M. P.; Stamatoyannopoulos, J.; Stranger, B. E.; Tang, H.; Tsang, E. K.; Wang, L.; Wang, M.; Van Wittenberghe, N.; Wu, F.; Zhang, R.; Nierras, C. R.; Branton, P. A.; Carithers, L. J.; Guan, P.; Moore, H. M.; Rao, A.; Vaught, J. B.; Gould, S. E.; Lockart, N. C.; Martin, C.; Struewing, J. P.; Volpi, S.; Addington, A. M.; Koester, S. E.; Little, A. R.; GTEx Consortium; Lead analysts:; Laboratory, D. A. & C. C. (LDACC):; NIH program management:; Biospecimen collection:; Pathology:; eQTL manuscript working group:; Laboratory, D. A. & C. C. (LDACC)—Analysis W. G.; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI. Genetic Effects on Gene Expression across Human Tissues. *Nature* **2017**, *550* (7675), 204–213. https://doi.org/10.1038/nature24277.

(31) Snyder, M. P.; Gingeras, T. R.; Moore, J. E.; Weng, Z.; Gerstein, M. B.; Ren, B.; Hardison, R. C.; Stamatoyannopoulos, J. A.; Graveley, B. R.; Feingold, E. A.; Pazin, M. J.; Pagan, M.; Gilchrist, D. A.; Hitz, B. C.; Cherry, J. M.; Bernstein, B. E.; Mendenhall, E. M.; Zerbino, D. R.; Frankish, A.; Flicek, P.; Myers, R. M. Perspectives on ENCODE. *Nature* **2020**, *583* (7818), 693–698. https://doi.org/10.1038/s41586-020-2449-8.

(32) Yengo, L.; Vedantam, S.; Marouli, E.; Sidorenko, J.; Bartell, E.; Sakaue, S.; Graff, M.; Eliasen, A. U.; Jiang, Y.; Raghavan, S.; Miao, J.; Arias, J. D.; Graham, S. E.; Mukamel, R. E.; Spracklen, C. N.; Yin, X.; Chen, S.-H.; Ferreira, T.; Highland, H. H.; Ji, Y.; Karaderi, T.; Lin, K.; Lüll, K.; Malden, D. E.; Medina-Gomez, C.; Machado, M.; Moore, A.; Rüeger, S.; Sim, X.; Vrieze, S.; Ahluwalia, T. S.; Akiyama, M.; Allison, M. A.; Alvarez, M.; Andersen, M. K.; Ani, A.; Appadurai, V.; Arbeeva, L.; Bhaskar, S.; Bielak, L. F.; Bollepalli, S.; Bonnycastle, L. L.; Bork-Jensen, J.; Bradfield, J. P.; Bradford, Y.; Braund, P. S.; Brody, J. A.; Burgdorf, K. S.; Cade, B. E.; Cai, H.; Cai, Q.; Campbell, A.; Cañadas-Garre, M.; Catamo, E.; Chai, J.-F.; Chai, X.; Chang, L.-C.; Chang, Y.-C.; Chen, C.-H.; Chesi, A.; Choi, S. H.; Chung, R.-H.; Cocca, M.; Concas, M. P.; Couture, C.; Cuellar-Partida, G.; Danning, R.; Daw, E. W.; Degenhard, F.; Delgado, G. E.; Delitala, A.; Demirkan, A.; Deng, X.; Devineni, P.; Dietl, A.; Dimitriou, M.; Dimitrov, L.; Dorajoo, R.; Ekici, A. B.; Engmann, J. E.; Fairhurst-Hunter, Z.; Farmaki, A.-E.; Faul, J. D.; Fernandez-Lopez, J.-C.; Forer, L.; Francescatto, M.; Freitag-Wolf, S.; Fuchsberger, C.; Galesloot, T. E.; Gao, Y.; Gao, Z.; Geller, F.; Giannakopoulou, O.; Giulianini, F.; Gjesing, A. P.; Goel, A.; Gordon, S. D.; Gorski, M.; Grove, J.; Guo, X.; Gustafsson, S.; Haessler, J.; Hansen, T. F.; Havulinna, A. S.; Haworth, S. J.; He, J.; Heard-Costa, N.; Hebbar, P.; Hindy, G.; Ho, Y.-L. A.; Hofer, E.; Holliday, E.; Horn, K.; Hornsby, W. E.; Hottenga, J.-J.; Huang, H.; Huang, J.; Huerta-Chagoya, A.; Huffman, J. E.; Hung, Y.-J.; Huo, S.; Hwang, M. Y.; Iha, H.; Ikeda, D. D.; Isono, M.; Jackson, A. U.; Jäger, S.; Jansen, I. E.; Johansson, I.; Jonas, J. B.; Jonsson, A.; Jørgensen, T.; Kalafati, I.-P.; Kanai, M.; Kanoni, S.; Kårhus, L. L.; Kasturiratne, A.; Katsuya, T.; Kawaguchi, T.; Kember, R. L.; Kentistou, K. A.; Kim, H.-N.; Kim, Y. J.; Kleber, M. E.; Knol, M. J.; Kurbasic, A.; Lauzon, M.; Le, P.; Lea, R.; Lee, J.-Y.; Leonard, H. L.; Li, S. A.; Li, X.; Li, X.; Liang, J.; Lin, H.; Lin, S.-Y.; Liu, J.; Liu, X.; Lo, K. S.; Long, J.; Lores-Motta, L.; Luan, J.; Lyssenko, V.; Lyytikäinen, L.-P.; Mahajan, A.; Mamakou, V.; Mangino, M.; Manichaikul, A.; Marten, J.; Mattheisen, M.; Mavarani, L.; McDaid, A. F.; Meidtner, K.; Melendez, T. L.; Mercader, J. M.; Milaneschi, Y.; Miller, J. E.; Millwood, I. Y.; Mishra, P. P.; Mitchell, R. E.; Møllehave, L. T.; Morgan, A.; Mucha, S.; Munz, M.; Nakatochi, M.; Nelson, C. P.; Nethander, M.; Nho, C. W.; Nielsen, A. A.; Nolte, I. M.; Nongmaithem, S. S.; Noordam, R.; Ntalla, I.; Nutile, T.; Pandit, A.; Christofidou, P.; Pärna, K.; Pauper, M.; Petersen, E. R. B.; Petersen, L. V.; Pitkänen, N.; Polašek, O.; Poveda, A.; Preuss, M. H.; Pyarajan, S.; Raffield, L. M.; Rakugi, H.; Ramirez, J.; Rasheed, A.; Raven,

D.; Rayner, N. W.; Riveros, C.; Rohde, R.; Ruggiero, D.; Ruotsalainen, S. E.; Ryan, K. A.; Sabater-Lleal, M.; Saxena, R.; Scholz, M.; Sendamarai, A.; Shen, B.; Shi, J.; Shin, J. H.; Sidore, C.; Sitlani, C. M.; Slieker, R. C.; Smit, R. A. J.; Smith, A. V.; Smith, J. A.; Smyth, L. J.; Southam, L.; Steinthorsdottir, V.; Sun, L.; Takeuchi, F.; Tallapragada, D. S. P.; Taylor, K. D.; Tayo, B. O.; Tcheandjieu, C.; Terzikhan, N.; Tesolin, P.; Teumer, A.; Theusch, E.; Thompson, D. J.; Thorleifsson, G.; Timmers, P. R. H. J.; Trompet, S.; Turman, C.; Vaccargiu, S.; van der Laan, S. W.; van der Most, P. J.; van Klinken, J. B.; van Setten, J.; Verma, S. S.; Verweij, N.; Veturi, Y.; Wang, C. A.; Wang, C.; Wang, L.; Wang, Z.; Warren, H. R.; Bin Wei, W.; Wickremasinghe, A. R.; Wielscher, M.; Wiggins, K. L.; Winsvold, B. S.; Wong, A.; Wu, Y.; Wuttke, M.; Xia, R.; Xie, T.; Yamamoto, K.; Yang, J.; Yao, J.; Young, H.; Yousri, N. A.; Yu, L.; Zeng, L.; Zhang, W.; Zhang, X.; Zhao, J.-H.; Zhao, W.; Zhou, W.; Zimmermann, M. E.; Zoledziewska, M.; Adair, L. S.; Adams, H. H. H.; Aguilar-Salinas, C. A.; Al-Mulla, F.; Arnett, D. K.; Asselbergs, F. W.; Åsvold, B. O.; Attia, J.; Banas, B.; Bandinelli, S.; Bennett, D. A.; Bergler, T.; Bharadwaj, D.; Biino, G.; Bisgaard, H.; Boerwinkle, E.; Böger, C. A.; Bønnelykke, K.; Boomsma, D. I.; Børglum, A. D.; Borja, J. B.; Bouchard, C.; Bowden, D. W.; Brandslund, I.; Brumpton, B.; Buring, J. E.; Caulfield, M. J.; Chambers, J. C.; Chandak, G. R.; Chanock, S. J.; Chaturvedi, N.; Chen, Y.-D. I.; Chen, Z.; Cheng, C.-Y.; Christophersen, I. E.; Ciullo, M.; Cole, J. W.; Collins, F. S.; Cooper, R. S.; Cruz, M.; Cucca, F.; Cupples, L. A.; Cutler, M. J.; Damrauer, S. M.; Dantoft, T. M.; de Borst, G. J.; de Groot, L. C. P. G. M.; De Jager, P. L.; de Kleijn, D. P. V.; Janaka de Silva, H.; Dedoussis, G. V.; den Hollander, A. I.; Du, S.; Easton, D. F.; Elders, P. J. M.; Eliassen, A. H.; Ellinor, P. T.; Elmståhl, S.; Erdmann, J.; Evans, M. K.; Fatkin, D.; Feenstra, B.; Feitosa, M. F.; Ferrucci, L.; Ford, I.; Fornage, M.; Franke, A.; Franks, P. W.; Freedman, B. I.; Gasparini, P.; Gieger, C.; Girotto, G.; Goddard, M. E.; Golightly, Y. M.; Gonzalez-Villalpando, C.; Gordon-Larsen, P.; Grallert, H.; Grant, S. F. A.; Grarup, N.; Griffiths, L.; Gudnason, V.; Haiman, C.; Hakonarson, H.; Hansen, T.; Hartman, C. A.; Hattersley, A. T.; Hayward, C.; Heckbert, S. R.; Heng, C.-K.; Hengstenberg, C.; Hewitt, A. W.; Hishigaki, H.; Hoyng, C. B.; Huang, P. L.; Huang, W.; Hunt, S. C.; Hveem, K.; Hyppönen, E.; Iacono, W. G.; Ichihara, S.; Ikram, M. A.; Isasi, C. R.; Jackson, R. D.; Jarvelin, M.-R.; Jin, Z.-B.; Jöckel, K.-H.; Joshi, P. K.; Jousilahti, P.; Jukema, J. W.; Kähönen, M.; Kamatani, Y.; Kang, K. D.; Kaprio, J.; Kardia, S. L. R.; Karpe, F.; Kato, N.; Kee, F.; Kessler, T.; Khera, A. V.; Khor, C. C.; Kiemeney, L. A. L. M.; Kim, B.-J.; Kim, E. K.; Kim, H.-L.; Kirchhof, P.; Kivimaki, M.; Koh, W.-P.; Koistinen, H. A.; Kolovou, G. D.; Kooner, J. S.; Kooperberg, C.; Köttgen, A.; Kovacs, P.; Kraaijeveld, A.; Kraft, P.; Krauss, R. M.; Kumari, M.; Kutalik, Z.; Laakso, M.; Lange, L. A.; Langenberg, C.; Launer, L. J.; Le Marchand, L.; Lee, H.; Lee, N. R.; Lehtimäki, T.; Li, H.; Li, L.; Lieb, W.; Lin, X.; Lind, L.; Linneberg, A.; Liu, C.-T.; Liu, J.; Loeffler, M.; London, B.; Lubitz, S. A.; Lye, S. J.; Mackey, D. A.; Mägi, R.; Magnusson, P. K. E.; Marcus, G. M.; Vidal, P. M.; Martin, N. G.; März, W.; Matsuda, F.; McGarrah, R. W.; McGue, M.; McKnight, A. J.; Medland, S. E.; Mellström, D.; Metspalu, A.; Mitchell, B. D.; Mitchell, P.; Mook-Kanamori, D. O.; Morris, A. D.; Mucci, L. A.; Munroe, P. B.; Nalls, M. A.; Nazarian, S.; Nelson, A. E.; Neville, M. J.; Newton-Cheh, C.; Nielsen, C. S.; Nöthen, M. M.; Ohlsson, C.; Oldehinkel, A. J.; Orozco, L.; Pahkala, K.; Pajukanta, P.; Palmer, C. N. A.; Parra, E. J.; Pattaro, C.; Pedersen, O.; Pennell, C. E.; Penninx, B. W. J. H.; Perusse, L.; Peters, A.; Peyser, P. A.; Porteous, D. J.; Posthuma, D.; Power, C.; Pramstaller, P. P.; Province, M. A.; Qi, Q.; Qu, J.; Rader, D. J.; Raitakari, O. T.; Ralhan, S.; Rallidis, L. S.; Rao, D. C.; Redline, S.; Reilly, D. F.; Reiner, A. P.; Rhee, S. Y.; Ridker, P. M.; Rienstra, M.; Ripatti, S.; Ritchie, M. D.; Roden, D. M.; Rosendaal, F. R.; Rotter, J. I.; Rudan, I.; Rutters, F.; Sabanayagam, C.; Saleheen, D.; Salomaa, V.; Samani, N. J.; Sanghera, D. K.; Sattar, N.; Schmidt, B.; Schmidt, H.; Schmidt, R.; Schulze, M. B.; Schunkert, H.; Scott, L. J.; Scott, R. J.; Sever, P.; Shiroma, E. J.; Shoemaker, M. B.; Shu, X.-O.; Simonsick, E. M.; Sims, M.; Singh, J. R.; Singleton, A. B.; Sinner, M. F.; Smith, J. G.; Snieder, H.; Spector, T. D.; Stampfer, M. J.; Stark, K. J.;

Strachan, D. P.; ‘t Hart, L. M.; Tabara, Y.; Tang, H.; Tardif, J.-C.; Thanaraj, T. A.; Timpson, N. J.; Tönjes, A.; Tremblay, A.; Tuomi, T.; Tuomilehto, J.; Tusié-Luna, M.-T.; Uitterlinden, A. G.; van Dam, R. M.; van der Harst, P.; Van der Velde, N.; van Duijn, C. M.; van Schoor, N. M.; Vitart, V.; Völker, U.; Vollenweider, P.; Völzke, H.; Wacher-Rodarte, N. H.; Walker, M.; Wang, Y. X.; Wareham, N. J.; Watanabe, R. M.; Watkins, H.; Weir, D. R.; Werge, T. M.; Widen, E.; Wilkens, L. R.; Willemsen, G.; Willett, W. C.; Wilson, J. F.; Wong, T.-Y.; Woo, J.-T.; Wright, A. F.; Wu, J.-Y.; Xu, H.; Yajnik, C. S.; Yokota, M.; Yuan, J.-M.; Zeggini, E.; Zemel, B. S.; Zheng, W.; Zhu, X.; Zmuda, J. M.; Zonderman, A. B.; Zwart, J.-A.; Chasman, D. I.; Cho, Y. S.; Heid, I. M.; McCarthy, M. I.; Ng, M. C. Y.; O’Donnell, C. J.; Rivadeneira, F.; Thorsteinsdottir, U.; Sun, Y. V.; Tai, E. S.; Boehnke, M.; Deloukas, P.; Justice, A. E.; Lindgren, C. M.; Loos, R. J. F.; Mohlke, K. L.; North, K. E.; Stefansson, K.; Walters, R. G.; Winkler, T. W.; Young, K. L.; Loh, P.-R.; Yang, J.; Esko, T.; Assimes, T. L.; Auton, A.; Abecasis, G. R.; Willer, C. J.; Locke, A. E.; Berndt, S. I.; Lettre, G.; Frayling, T. M.; Okada, Y.; Wood, A. R.; Visscher, P. M.; Hirschhorn, J. N. A Saturated Map of Common Genetic Variants Associated with Human Height. *Nature* **2022**, *610* (7933), 704–712. https://doi.org/10.1038/s41586-022-05275-y.

(33) Chang, C. C.; Chow, C. C.; Tellier, L. C.; Vattikuti, S.; Purcell, S. M.; Lee, J. J. Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaScience* **2015**, *4* (1). https://doi.org/10.1186/s13742-015-0047-8.

(34) Gupta, R. M.; Hadaya, J.; Trehan, A.; Zekavat, S. M.; Roselli, C.; Klarin, D.; Emdin, C. A.; Hilvering, C. R. E.; Bianchi, V.; Mueller, C.; Khera, A. V.; Ryan, R. J. H.; Engreitz, J. M.; Issner, R.; Shoresh, N.; Epstein, C. B.; Laat, W. de; Brown, J. D.; Schnabel, R. B.; Bernstein, B. E.; Kathiresan, S. A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **2017**, *170* (3), 522-533.e15. https://doi.org/10.1016/j.cell.2017.06.049.

(35) Zhou, W.; Nielsen, J. B.; Fritsche, L. G.; Dey, R.; Gabrielsen, M. E.; Wolford, B. N.; LeFaive, J.; VandeHaar, P.; Gagliano, S. A.; Gifford, A.; Bastarache, L. A.; Wei, W.-Q.; Denny, J. C.; Lin, M.; Hveem, K.; Kang, H. M.; Abecasis, G. R.; Willer, C. J.; Lee, S. Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies. *Nat. Genet.* **2018**, *50* (9), 1335–1341. https://doi.org/10.1038/s41588-018-0184-y.

(36) Loh, P.-R.; Tucker, G.; Bulik-Sullivan, B. K.; Vilhjálmsson, B. J.; Finucane, H. K.; Salem, R. M.; Chasman, D. I.; Ridker, P. M.; Neale, B. M.; Berger, B.; Patterson, N.; Price, A. L. Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts. *Nat. Genet.* **2015**, *47* (3), 284–290. https://doi.org/10.1038/ng.3190.

(37) Mbatchou, J.; Barnard, L.; Backman, J.; Marcketta, A.; Kosmicki, J. A.; Ziyatdinov, A.; Benner, C.; O’Dushlaine, C.; Barber, M.; Boutkov, B.; Habegger, L.; Ferreira, M.; Baras, A.; Reid, J.; Abecasis, G.; Maxwell, E.; Marchini, J. Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits. *Nat. Genet.* **2021**, *53* (7), 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

(38) *Credits - PLINK 2.0*. https://www.cog-genomics.org/plink/2.0/credits (accessed 2024-04-28).

(39) Small, K. S.; Todorčević, M.; Civelek, M.; El-Sayed Moustafa, J. S.; Wang, X.; Simon, M. M.; Fernandez-Tajes, J.; Mahajan, A.; Horikoshi, M.; Hugill, A.; Glastonbury, C. A.; Quaye, L.; Neville, M. J.; Sethi, S.; Yon, M.; Pan, C.; Che, N.; Viñuela, A.; Tsai, P.-C.; Nag, A.; Buil, A.; Thorleifsson, G.; Raghavan, A.; Ding, Q.; Morris, A. P.; Bell, J. T.; Thorsteinsdottir, U.; Stefansson, K.; Laakso, M.; Dahlman, I.; Arner, P.; Gloyn, A. L.; Musunuru, K.; Lusis, A. J.; Cox, R. D.; Karpe, F.; McCarthy, M. I. Regulatory Variants at KLF14 Influence Type 2 Diabetes Risk via a Female-Specific Effect on Adipocyte Size and Body Composition. *Nat. Genet.* **2018**, *50* (4), 572–580. https://doi.org/10.1038/s41588-018-0088-x.

(40) Pan-UKB team. *Pan-ancestry genetic analysis of the UK Biobank*. https://pan.ukbb.broadinstitute.org/ (accessed 2021-08-18).

(41) Marigorta, U. M.; Rodríguez, J. A.; Gibson, G.; Navarro, A. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet.* **2018**, *34* (7), 504–517. https://doi.org/10.1016/j.tig.2018.03.005.

(42) O'Sullivan, J. W.; Ioannidis, J. P. A. Reproducibility in the UK Biobank of Genome-Wide Significant Signals Discovered in Earlier Genome-Wide Association Studies. *Sci. Rep.* **2021**, *11* (1), 18625. https://doi.org/10.1038/s41598-021-97896-y.

(43) Robertson, C. C.; Inshaw, J. R. J.; Onengut-Gumuscu, S.; Chen, W.-M.; Santa Cruz, D. F.; Yang, H.; Cutler, A. J.; Crouch, D. J. M.; Farber, E.; Bridges, S. L.; Edberg, J. C.; Kimberly, R. P.; Buckner, J. H.; Deloukas, P.; Divers, J.; Dabelea, D.; Lawrence, J. M.; Marcovina, S.; Shah, A. S.; Greenbaum, C. J.; Atkinson, M. A.; Gregersen, P. K.; Oksenberg, J. R.; Pociot, F.; Rewers, M. J.; Steck, A. K.; Dunger, D. B.; Wicker, L. S.; Concannon, P.; Todd, J. A.; Rich, S. S. Fine-Mapping, Trans-Ancestral and Genomic Analyses Identify Causal Variants, Cells, Genes and Drug Targets for Type 1 Diabetes. *Nat. Genet.* **2021**, *53* (7), 962–971. https://doi.org/10.1038/s41588-021-00880-5.

(44) Nagai, A.; Hirata, M.; Kamatani, Y.; Muto, K.; Matsuda, K.; Kiyohara, Y.; Ninomiya, T.; Tamakoshi, A.; Yamagata, Z.; Mushiroda, T.; Murakami, Y.; Yuji, K.; Furukawa, Y.; Zembutsu, H.; Tanaka, T.; Ohnishi, Y.; Nakamura, Y.; Shiono, M.; Misumi, K.; Kaieda, R.; Harada, H.; Minami, S.; Emi, M.; Emoto, N.; Daida, H.; Miyauchi, K.; Murakami, A.; Asai, S.; Moriyama, M.; Takahashi, Y.; Fujioka, T.; Obara, W.; Mori, S.; Ito, H.; Nagayama, S.; Miki, Y.; Masumoto, A.; Yamada, A.; Nishizawa, Y.; Kodama, K.; Kutsumi, H.; Sugimoto, Y.; Koretsune, Y.; Kusuoka, H.; Yanai, H.; Kubo, M. Overview of the BioBank Japan Project: Study Design and Profile. *J. Epidemiol.* **2017**, *27* (3, Supplement), S2–S8. https://doi.org/10.1016/j.je.2016.12.005.

(45) Wei, C.-Y.; Yang, J.-H.; Yeh, E.-C.; Tsai, M.-F.; Kao, H.-J.; Lo, C.-Z.; Chang, L.-P.; Lin, W.-J.; Hsieh, F.-J.; Belsare, S.; Bhaskar, A.; Su, M.-W.; Lee, T.-C.; Lin, Y.-L.; Liu, F.-T.; Shen, C.-Y.; Li, L.-H.; Chen, C.-H.; Wall, J. D.; Wu, J.-Y.; Kwok, P.-Y. Genetic Profiles of 103,106 Individuals in the Taiwan Biobank Provide Insights into the Health and History of Han Chinese. *Npj Genomic Med.* **2021**, *6* (1), 1–10. https://doi.org/10.1038/s41525-021-00178-9.

(46) Walters, R. G.; Millwood, I. Y.; Lin, K.; Schmidt Valle, D.; McDonnell, P.; Hacker, A.; Avery, D.; Edris, A.; Fry, H.; Cai, N.; Kretzschmar, W. W.; Ansari, M. A.; Lyons, P. A.; Collins, R.; Donnelly, P.; Hill, M.; Peto, R.; Shen, H.; Jin, X.; Nie, C.; Xu, X.; Guo, Y.; Yu, C.; Lv, J.; Clarke, R. J.; Li, L.; Chen, Z. Genotyping and Population Characteristics of the China Kadoorie Biobank. *Cell Genomics* **2023**, *3* (8), 100361. https://doi.org/10.1016/j.xgen.2023.100361.

(47) Lee, S.; Abecasis, G. R.; Boehnke, M.; Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **2014**, *95* (1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009.

(48) Altshuler, D.; Donnelly, P.; The International HapMap Consortium. A Haplotype Map of the Human Genome. *Nature* **2005**, *437* (7063), 1299–1320. https://doi.org/10.1038/nature04226.

(49) Howles, S. A.; Wiberg, A.; Goldsworthy, M.; Bayliss, A. L.; Gluck, A. K.; Ng, M.; Grout, E.; Tanikawa, C.; Kamatani, Y.; Terao, C.; Takahashi, A.; Kubo, M.; Matsuda, K.; Thakker, R. V.; Turney, B. W.; Furniss, D. Genetic Variants of Calcium and Vitamin D Metabolism in Kidney Stone Disease. *Nat. Commun.* **2019**, *10* (1), 5175. https://doi.org/10.1038/s41467-019-13145-x.

(50) Sanna, S.; Li, B.; Mulas, A.; Sidore, C.; Kang, H. M.; Jackson, A. U.; Piras, M. G.; Usala, G.; Maninchedda, G.; Sassu, A.; Serra, F.; Palmas, M. A.; Iii, W. H. W.; Njølstad, I.; Laakso, M.; Hveem, K.; Tuomilehto, J.; Lakka, T. A.; Rauramaa, R.; Boehnke, M.; Cucca, F.; Uda,

M.; Schlessinger, D.; Nagaraja, R.; Abecasis, G. R. Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLOS Genet.* **2011**, *7* (7), e1002198. https://doi.org/10.1371/journal.pgen.1002198.

(51) *Raymond Walters: LD Score Regression I, Heritability and Partitioning*; 2017. https://www.youtube.com/watch?v=dVrF0l9jMgE (accessed 2024-05-21).

(52) Piovesan, A.; Caracausi, M.; Antonaros, F.; Pelleri, M. C.; Vitale, L. GeneBase 1.1: A Tool to Summarize Data from NCBI Gene Datasets and Its Application to an Update of Human Gene Statistics. *Database J. Biol. Databases Curation* **2016**, *2016*, baw153. https://doi.org/10.1093/database/baw153.

(53) Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L. T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; Cortes, A.; Welsh, S.; Young, A.; Effingham, M.; McVean, G.; Leslie, S.; Allen, N.; Donnelly, P.; Marchini, J. The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **2018**, *562* (7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z.

(54) Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Altshuler, D. M.; Gibbs, R. A.; Peltonen, L.; Dermitzakis, E.; Schaffner, S. F.; Yu, F.; Peltonen, L.; Dermitzakis, E.; Bonnen, P. E.; Altshuler, D. M.; Gibbs, R. A.; de Bakker, P. I. W.; Deloukas, P.; Gabriel, S. B.; Gwilliam, R.; Hunt, S.; Inouye, M.; Jia, X.; Palotie, A.; Parkin, M.; Whittaker, P.; Yu, F.; Chang, K.; Hawes, A.; Lewis, L. R.; Ren, Y.; Wheeler, D.; Gibbs, R. A.; Marie Muzny, D.; Barnes, C.; Darvishi, K.; Hurles, M.; Korn, J. M.; Kristiansson, K.; Lee, C.; McCarroll, S. A.; Nemesh, J.; Dermitzakis, E.; Keinan, A.; Montgomery, S. B.; Pollack, S.; Price, A. L.; Soranzo, N.; Bonnen, P. E.; Gibbs, R. A.; Gonzaga-Jauregui, C.; Keinan, A.; Price, A. L.; Yu, F.; Anttila, V.; Brodeur, W.; Daly, M. J.; Leslie, S.; McVean, G.; Moutsianas, L.; Nguyen, H.; Schaffner, S. F.; Zhang, Q.; Ghori, M. J. R.; McGinnis, R.; McLaren, W.; Pollack, S.; Price, A. L.; Schaffner, S. F.; Takeuchi, F.; Grossman, S. R.; Shlyakhter, I.; Hostetter, E. B.; Sabeti, P. C.; Adebamowo, C. A.; Foster, M. W.; Gordon, D. R.; Licinio, J.; Cristina Manca, M.; Marshall, P. A.; Matsuda, I.; Ngare, D.; Ota Wang, V.; Reddy, D.; Rotimi, C. N.; Royal, C. D.; Sharp, R. R.; Zeng, C.; Brooks, L. D.; McEwen, J. E.; The International HapMap 3 Consortium; Principal investigators; Project coordination leaders; Manuscript writing group; Genotyping and QC; ENCODE 3 sequencing and SNP discovery; Copy number variation typing and analysis; Population analysis; Low frequency variation analysis; Linkage disequilibrium and haplotype sharing analysis; Imputation; Natural selection; Community engagement and sample collection groups; Scientific management. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature* **2010**, *467* (7311), 52–58. https://doi.org/10.1038/nature09298.

(55) McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A. R.; Teumer, A.; Kang, H. M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; Luo, Y.; Sidore, C.; Kwong, A.; Timpson, N.; Koskinen, S.; Vrieze, S.; Scott, L. J.; Zhang, H.; Mahajan, A.; Veldink, J.; Peters, U.; Pato, C.; van Duijn, C. M.; Gillies, C. E.; Gandin, I.; Mezzavilla, M.; Gilly, A.; Cocca, M.; Traglia, M.; Angius, A.; Barrett, J. C.; Boomsma, D.; Branham, K.; Breen, G.; Brummett, C. M.; Busonero, F.; Campbell, H.; Chan, A.; Chen, S.; Chew, E.; Collins, F. S.; Corbin, L. J.; Smith, G. D.; Dedoussis, G.; Dorr, M.; Farmaki, A.-E.; Ferrucci, L.; Forer, L.; Fraser, R. M.; Gabriel, S.; Levy, S.; Groop, L.; Harrison, T.; Hattersley, A.; Holmen, O. L.; Hveem, K.; Kretzler, M.; Lee, J. C.; McGue, M.; Meitinger, T.; Melzer, D.; Min, J. L.; Mohlke, K. L.; Vincent, J. B.; Nauck, M.; Nickerson, D.; Palotie, A.; Pato, M.; Pirastu, N.; McInnis, M.; Richards, J. B.; Sala, C.; Salomaa, V.; Schlessinger, D.; Schoenherr, S.; Slagboom, P. E.; Small, K.; Spector, T.; Stambolian, D.; Tuke, M.; Tuomilehto, J.; Van den Berg, L. H.; Van Rheenen, W.; Volker, U.; Wijmenga, C.; Toniolo, D.; Zeggini, E.; Gasparini, P.; Sampson, M. G.; Wilson, J. F.; Frayling, T.; de Bakker, P. I. W.; Swertz, M. A.; McCarroll, S.; Kooperberg, C.; Dekker, A.; Altshuler, D.; Willer, C.; Iacono, W.; Ripatti, S.; Soranzo, N.;

Walter, K.; Swaroop, A.; Cucca, F.; Anderson, C. A.; Myers, R. M.; Boehnke, M.; McCarthy, M. I.; Durbin, R.; Abecasis, G.; Marchini, J.; the Haplotype Reference Consortium. A Reference Panel of 64,976 Haplotypes for Genotype Imputation. *Nat. Genet.* **2016**, *48* (10), 1279–1283. https://doi.org/10.1038/ng.3643.

(56) Taliun, D.; Harris, D. N.; Kessler, M. D.; Carlson, J.; Szpiech, Z. A.; Torres, R.; Taliun, S. A. G.; Corvelo, A.; Gogarten, S. M.; Kang, H. M.; Pitsillides, A. N.; LeFaive, J.; Lee, S.; Tian, X.; Browning, B. L.; Das, S.; Emde, A.-K.; Clarke, W. E.; Loesch, D. P.; Shetty, A. C.; Blackwell, T. W.; Smith, A. V.; Wong, Q.; Liu, X.; Conomos, M. P.; Bobo, D. M.; Aguet, F.; Albert, C.; Alonso, A.; Ardlie, K. G.; Arking, D. E.; Aslibekyan, S.; Auer, P. L.; Barnard, J.; Barr, R. G.; Barwick, L.; Becker, L. C.; Beer, R. L.; Benjamin, E. J.; Bielak, L. F.; Blangero, J.; Boehnke, M.; Bowden, D. W.; Brody, J. A.; Burchard, E. G.; Cade, B. E.; Casella, J. F.; Chalazan, B.; Chasman, D. I.; Chen, Y.-D. I.; Cho, M. H.; Choi, S. H.; Chung, M. K.; Clish, C. B.; Correa, A.; Curran, J. E.; Custer, B.; Darbar, D.; Daya, M.; de Andrade, M.; DeMeo, D. L.; Dutcher, S. K.; Ellinor, P. T.; Emery, L. S.; Eng, C.; Fatkin, D.; Fingerlin, T.; Forer, L.; Fornage, M.; Franceschini, N.; Fuchsberger, C.; Fullerton, S. M.; Germer, S.; Gladwin, M. T.; Gottlieb, D. J.; Guo, X.; Hall, M. E.; He, J.; Heard-Costa, N. L.; Heckbert, S. R.; Irvin, M. R.; Johnsen, J. M.; Johnson, A. D.; Kaplan, R.; Kardia, S. L. R.; Kelly, T.; Kelly, S.; Kenny, E. E.; Kiel, D. P.; Klemmer, R.; Konkle, B. A.; Kooperberg, C.; Köttgen, A.; Lange, L. A.; Lasky-Su, J.; Levy, D.; Lin, X.; Lin, K.-H.; Liu, C.; Loos, R. J. F.; Garman, L.; Gerszten, R.; Lubitz, S. A.; Lunetta, K. L.; Mak, A. C. Y.; Manichaikul, A.; Manning, A. K.; Mathias, R. A.; McManus, D. D.; McGarvey, S. T.; Meigs, J. B.; Meyers, D. A.; Mikulla, J. L.; Minear, M. A.; Mitchell, B. D.; Mohanty, S.; Montasser, M. E.; Montgomery, C.; Morrison, A. C.; Murabito, J. M.; Natale, A.; Natarajan, P.; Nelson, S. C.; North, K. E.; O'Connell, J. R.; Palmer, N. D.; Pankratz, N.; Peloso, G. M.; Peyser, P. A.; Pleiness, J.; Post, W. S.; Psaty, B. M.; Rao, D. C.; Redline, S.; Reiner, A. P.; Roden, D.; Rotter, J. I.; Ruczinski, I.; Sarnowski, C.; Schoenherr, S.; Schwartz, D. A.; Seo, J.-S.; Seshadri, S.; Sheehan, V. A.; Sheu, W. H.; Shoemaker, M. B.; Smith, N. L.; Smith, J. A.; Sotoodehnia, N.; Stilp, A. M.; Tang, W.; Taylor, K. D.; Telen, M.; Thornton, T. A.; Tracy, R. P.; Van Den Berg, D. J.; Vasan, R. S.; Viaud-Martinez, K. A.; Vrieze, S.; Weeks, D. E.; Weir, B. S.; Weiss, S. T.; Weng, L.-C.; Willer, C. J.; Zhang, Y.; Zhao, X.; Arnett, D. K.; Ashley-Koch, A. E.; Barnes, K. C.; Boerwinkle, E.; Gabriel, S.; Gibbs, R.; Rice, K. M.; Rich, S. S.; Silverman, E. K.; Qasba, P.; Gan, W.; Papanicolaou, G. J.; Nickerson, D. A.; Browning, S. R.; Zody, M. C.; Zöllner, S.; Wilson, J. G.; Cupples, L. A.; Laurie, C. C.; Jaquish, C. E.; Hernandez, R. D.; O'Connor, T. D.; Abecasis, G. R. Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *Nature* **2021**, *590* (7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y.

(57) Rubinacci, S.; Delaneau, O.; Marchini, J. Genotype Imputation Using the Positional Burrows Wheeler Transform. *PLOS Genet.* **2020**, *16* (11), e1009049. https://doi.org/10.1371/journal.pgen.1009049.

(58) Browning, B. L.; Zhou, Y.; Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **2018**, *103* (3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

(59) Zheng, J.; Li, Y.; Abecasis, G. R.; Scheet, P. A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. *Genet. Epidemiol.* **2011**, *35* (2), 102–110. https://doi.org/10.1002/gepi.20552.

(60) Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature* **2007**, *447* (7145), 661–678. https://doi.org/10.1038/nature05911.

(61) Nelson, M. R.; Tipney, H.; Painter, J. L.; Shen, J.; Nicoletti, P.; Shen, Y.; Floratos, A.; Sham, P. C.; Li, M. J.; Wang, J.; Cardon, L. R.; Whittaker, J. C.; Sanseau, P. The Support of Human Genetic Evidence for Approved Drug Indications. *Nat. Genet.* **2015**, *47* (8), 856–860. https://doi.org/10.1038/ng.3314.

(62) Ochoa, D.; Karim, M.; Ghoussaini, M.; Hulcoop, D. G.; McDonagh, E. M.; Dunham, I. Human Genetics Evidence Supports Two-Thirds of the 2021 FDA-Approved Drugs. *Nat. Rev. Drug Discov.* **2022**, *21* (8), 551–551. https://doi.org/10.1038/d41573-022-00120-3.

(63) Plenge, R. M.; Scolnick, E. M.; Altshuler, D. Validating Therapeutic Targets through Human Genetics. *Nat. Rev. Drug Discov.* **2013**, *12* (8), 581–594. https://doi.org/10.1038/nrd4051.

(64) Vincent, F.; Nueda, A.; Lee, J.; Schenone, M.; Prunotto, M.; Mercola, M. Phenotypic Drug Discovery: Recent Successes, Lessons Learned and New Directions. *Nat. Rev. Drug Discov.* **2022**, *21* (12), 899–914. https://doi.org/10.1038/s41573-022-00472-w.

(65) Reay, W. R.; Cairns, M. J. Advancing the Use of Genome-Wide Association Studies for Drug Repurposing. *Nat. Rev. Genet.* **2021**, *22* (10), 658–671. https://doi.org/10.1038/s41576-021-00387-z.

(66) Schaub, M. A.; Boyle, A. P.; Kundaje, A.; Batzoglou, S.; Snyder, M. Linking Disease Associations with Regulatory Information in the Human Genome. *Genome Res.* **2012**, *22* (9), 1748–1759. https://doi.org/10.1101/gr.136127.111.

(67) van de Bunt, M.; Cortes, A.; IGAS Consortium; Brown, M. A.; Morris, A. P.; McCarthy, M. I. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS Genet.* **2015**, *11* (9), e1005535. https://doi.org/10.1371/journal.pgen.1005535.

(68) Asimit, J. L.; Rainbow, D. B.; Fortune, M. D.; Grinberg, N. F.; Wicker, L. S.; Wallace, C. Stochastic Search and Joint Fine-Mapping Increases Accuracy and Identifies Previously Unreported Associations in Immune-Mediated Diseases. *Nat. Commun.* **2019**, *10* (1), 3216. https://doi.org/10.1038/s41467-019-11271-0.

(69) Wallace, C.; Cutler, A. J.; Pontikos, N.; Pekalski, M. L.; Burren, O. S.; Cooper, J. D.; García, A. R.; Ferreira, R. C.; Guo, H.; Walker, N. M.; Smyth, D. J.; Rich, S. S.; Onengut-Gumuscu, S.; Sawcer, S. J.; Ban, M.; Richardson, S.; Todd, J. A.; Wicker, L. S. Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLOS Genet.* **2015**, *11* (6), e1005272. https://doi.org/10.1371/journal.pgen.1005272.

(70) Mahajan, A.; Taliun, D.; Thurner, M.; Robertson, N. R.; Torres, J. M.; Rayner, N. W.; Payne, A. J.; Steinthorsdottir, V.; Scott, R. A.; Grarup, N.; Cook, J. P.; Schmidt, E. M.; Wuttke, M.; Sarnowski, C.; Mägi, R.; Nano, J.; Gieger, C.; Trompet, S.; Lecoeur, C.; Preuss, M. H.; Prins, B. P.; Guo, X.; Bielak, L. F.; Below, J. E.; Bowden, D. W.; Chambers, J. C.; Kim, Y. J.; Ng, M. C. Y.; Petty, L. E.; Sim, X.; Zhang, W.; Bennett, A. J.; Bork-Jensen, J.; Brummett, C. M.; Canouil, M.; Ec kardt, K.-U.; Fischer, K.; Kardia, S. L. R.; Kronenberg, F.; Läll, K.; Liu, C.-T.; Locke, A. E.; Luan, J.; Ntalla, I.; Nylander, V.; Schönherr, S.; Schurmann, C.; Yengo, L.; Bottinger, E. P.; Brandslund, I.; Christensen, C.; Dedoussis, G.; Florez, J. C.; Ford, I.; Franco, O. H.; Frayling, T. M.; Giedraitis, V.; Hackinger, S.; Hattersley, A. T.; Herder, C.; Ikram, M. A.; Ingelsson, M.; Jørgensen, M. E.; Jørgensen, T.; Kriebel, J.; Kuusisto, J.; Ligthart, S.; Lindgren, C. M.; Linneberg, A.; Lyssenko, V.; Mamakou, V.; Meitinger, T.; Mohlke, K. L.; Morris, A. D.; Nadkarni, G.; Pankow, J. S.; Peters, A.; Sattar, N.; Stančáková, A.; Strauch, K.; Taylor, K. D.; Thorand, B.; Thorleifsson, G.; Thorsteinsdottir, U.; Tuomilehto, J.; Witte, D. R.; Dupuis, J.; Peyser, P. A.; Zeggini, E.; Loos, R. J. F.; Froguel, P.; Ingelsson, E.; Lind, L.; Groop, L.; Laakso, M.; Collins, F. S.; Jukema, J. W.; Palmer, C. N. A.; Grallert, H.; Metspalu, A.; Dehghan, A.; Köttgen, A.; Abecasis, G. R.; Meigs, J. B.; Rotter, J. I.; Marchini, J.; Pedersen, O.; Hansen, T.; Langenberg, C.; Wareham, N. J.; Stefansson, K.; Gloyn, A. L.; Morris, A. P.; Boehnke, M.; McCarthy, M. I. Fine-Mapping Type 2 Diabetes Loci to Single-Variant Resolution Using High-Density Imputation and Islet-Specific Epigenome Maps. *Nat. Genet.* **2018**, *50* (11), 1505–1513. https://doi.org/10.1038/s41588-018-0241-6.

(71) Mahajan, A.; Wessel, J.; Willems, S. M.; Zhao, W.; Robertson, N. R.; Chu, A. Y.; Gan, W.; Kitajima, H.; Taliun, D.; Rayner, N. W.; Guo, X.; Lu, Y.; Li, M.; Jensen, R. A.; Hu, Y.;

Huo, S.; Lohman, K. K.; Zhang, W.; Cook, J. P.; Prins, B. P.; Flannick, J.; Grarup, N.; Trubetskoy, V. V.; Kravic, J.; Kim, Y. J.; Rybin, D. V.; Yaghootkar, H.; Müller-Nurasyid, M.; Meidtner, K.; Li-Gao, R.; Varga, T. V.; Marten, J.; Li, J.; Smith, A. V.; An, P.; Ligthart, S.; Gustafsson, S.; Malerba, G.; Demirkan, A.; Tajes, J. F.; Steinthorsdottir, V.; Wuttke, M.; Lecoeur, C.; Preuss, M.; Bielak, L. F.; Graff, M.; Highland, H. M.; Justice, A. E.; Liu, D. J.; Marouli, E.; Peloso, G. M.; Warren, H. R.; Afaq, S.; Afzal, S.; Ahlqvist, E.; Almgren, P.; Amin, N.; Bang, L. B.; Bertoni, A. G.; Bombieri, C.; Bork-Jensen, J.; Brandslund, I.; Brody, J. A.; Burtt, N. P.; Canouil, M.; Chen, Y.-D. I.; Cho, Y. S.; Christensen, C.; Eastwood, S. V.; Eckardt, K.-U.; Fischer, K.; Gambaro, G.; Giedraitis, V.; Grove, M. L.; de Haan, H. G.; Hackinger, S.; Hai, Y.; Han, S.; Tybjærg-Hansen, A.; Hivert, M.-F.; Isomaa, B.; Jäger, S.; Jørgensen, M. E.; Jørgensen, T.; Käräjämäki, A.; Kim, B.-J.; Kim, S. S.; Koistinen, H. A.; Kovacs, P.; Kriebel, J.; Kronenberg, F.; Läll, K.; Lange, L. A.; Lee, J.-J.; Lehne, B.; Li, H.; Lin, K.-H.; Linneberg, A.; Liu, C.-T.; Liu, J.; Loh, M.; Mägi, R.; Mamakou, V.; McKean-Cowdin, R.; Nadkarni, G.; Neville, M.; Nielsen, S. F.; Ntalla, I.; Peyser, P. A.; Rathmann, W.; Rice, K.; Rich, S. S.; Rode, L.; Rolandsson, O.; Schönherr, S.; Selvin, E.; Small, K. S.; Stančáková, A.; Surendran, P.; Taylor, K. D.; Teslovich, T. M.; Thorand, B.; Thorleifsson, G.; Tin, A.; Tönjes, A.; Varbo, A.; Witte, D. R.; Wood, A. R.; Yajnik, P.; Yao, J.; Yengo, L.; Young, R.; Amouyel, P.; Boeing, H.; Boerwinkle, E.; Bottinger, E. P.; Chowdhury, R.; Collins, F. S.; Dedoussis, G.; Dehghan, A.; Deloukas, P.; Ferrario, M. M.; Ferrières, J.; Florez, J. C.; Frossard, P.; Gudnason, V.; Harris, T. B.; Heckbert, S. R.; Howson, J. M. M.; Ingelsson, M.; Kathiresan, S.; Kee, F.; Kuusisto, J.; Langenberg, C.; Launer, L. J.; Lindgren, C. M.; Männistö, S.; Meitinger, T.; Melander, O.; Mohlke, K. L.; Moitry, M.; Morris, A. D.; Murray, A. D.; de Mutsert, R.; Orho-Melander, M.; Owen, K. R.; Perola, M.; Peters, A.; Province, M. A.; Rasheed, A.; Ridker, P. M.; Rivadineira, F.; Rosendaal, F. R.; Rosengren, A. H.; Salomaa, V.; Sheu, W. H.-H.; Sladek, R.; Smith, B. H.; Strauch, K.; Uitterlinden, A. G.; Varma, R.; Willer, C. J.; Blüher, M.; Butterworth, A. S.; Chambers, J. C.; Chasman, D. I.; Danesh, J.; van Duijn, C.; Dupuis, J.; Franco, O. H.; Franks, P. W.; Froguel, P.; Grallert, H.; Groop, L.; Han, B.-G.; Hansen, T.; Hattersley, A. T.; Hayward, C.; Ingelsson, E.; Kardia, S. L. R.; Karpe, F.; Kooner, J. S.; Köttgen, A.; Kuulasmaa, K.; Laakso, M.; Lin, X.; Lind, L.; Liu, Y.; Loos, R. J. F.; Marchini, J.; Metspalu, A.; Mook-Kanamori, D.; Nordestgaard, B. G.; Palmer, C. N. A.; Pankow, J. S.; Pedersen, O.; Psaty, B. M.; Rauramaa, R.; Sattar, N.; Schulze, M. B.; Soranzo, N.; Spector, T. D.; Stefansson, K.; Stumvoll, M.; Thorsteinsdottir, U.; Tuomi, T.; Tuomilehto, J.; Wareham, N. J.; Wilson, J. G.; Zeggini, E.; Scott, R. A.; Barroso, I.; Frayling, T. M.; Goodarzi, M. O.; Meigs, J. B.; Boehnke, M.; Saleheen, D.; Morris, A. P.; Rotter, J. I.; McCarthy, M. I. Refining the Accuracy of Validated Target Identification through Coding Variant Fine-Mapping in Type 2 Diabetes. *Nat. Genet.* **2018**, *50* (4), 559–571. https://doi.org/10.1038/s41588-018-0084-1.

(72) Gallagher, M. D.; Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **2018**, *102* (5), 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002.

(73) Alsheikh, A. J.; Wollenhaupt, S.; King, E. A.; Reeb, J.; Ghosh, S.; Stolzenburg, L. R.; Tamim, S.; Lazar, J.; Davis, J. W.; Jacob, H. J. The Landscape of GWAS Validation; Systematic Review Identifying 309 Validated Non-Coding Variants across 130 Human Diseases. *BMC Med. Genomics* **2022**, *15* (1), 74. https://doi.org/10.1186/s12920-022-01216-w.

(74) *NCBI*. HapMap retired. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/ (accessed 2024-04-10).

(75) Bulik-Sullivan, B. K.; Loh, P.-R.; Finucane, H. K.; Ripke, S.; Yang, J.; Patterson, N.; Daly, M. J.; Price, A. L.; Neale, B. M. LD Score Regression Distinguishes Confounding from

Polygenicity in Genome-Wide Association Studies. *Nat. Genet.* **2015**, *47* (3), 291–295. https://doi.org/10.1038/ng.3211.

(76) Huerta-Chagoya, A.; Schroeder, P.; Mandla, R.; Deutsch, A. J.; Zhu, W.; Petty, L.; Yi, X.; Cole, J. B.; Udler, M. S.; Dornbos, P.; Porneala, B.; DiCorpo, D.; Liu, C.-T.; Li, J. H.; Szczerbiński, L.; Kaur, V.; Kim, J.; Lu, Y.; Martin, A.; Eizirik, D. L.; Marchetti, P.; Marselli, L.; Chen, L.; Srinivasan, S.; Todd, J.; Flannick, J.; Gubitosi-Klug, R.; Levitsky, L.; Shah, R.; Kelsey, M.; Burke, B.; Dabelea, D. M.; Divers, J.; Marcovina, S.; Stalbow, L.; Loos, R. J. F.; Darst, B. F.; Kooperberg, C.; Raffield, L. M.; Haiman, C.; Sun, Q.; McCormick, J. B.; Fisher-Hoch, S. P.; Ordoñez, M. L.; Meigs, J.; Baier, L. J.; González-Villalpando, C.; González-Villalpando, M. E.; Orozco, L.; García-García, L.; Moreno-Estrada, A.; Aguilar-Salinas, C. A.; Tusié, T.; Dupuis, J.; Ng, M. C. Y.; Manning, A.; Highland, H. M.; Cnop, M.; Hanson, R.; Below, J.; Florez, J. C.; Leong, A.; Mercader, J. M. The Power of TOPMed Imputation for the Discovery of Latino-Enriched Rare Variants Associated with Type 2 Diabetes. *Diabetologia* **2023**, *66* (7), 1273–1288. https://doi.org/10.1007/s00125-023-05912-9.

(77) Jun, G.; English, A. C.; Metcalf, G. A.; Yang, J.; Chaisson, M. J.; Pankratz, N.; Menon, V. K.; Salerno, W. J.; Krasheninina, O.; Smith, A. V.; Lane, J. A.; Blackwell, T.; Kang, H. M.; Salvi, S.; Meng, Q.; Shen, H.; Pasham, D.; Bhamidipati, S.; Kottapalli, K.; Arnett, D. K.; Ashley-Koch, A.; Auer, P. L.; Beutel, K. M.; Bis, J. C.; Blangero, J.; Bowden, D. W.; Brody, J. A.; Cade, B. E.; Chen, Y.-D. I.; Cho, M. H.; Curran, J. E.; Fornage, M.; Freedman, B. I.; Fingerlin, T.; Gelb, B. D.; Hou, L.; Hung, Y.-J.; Kane, J. P.; Kaplan, R.; Kim, W.; Loos, R. J. F.; Marcus, G. M.; Mathias, R. A.; McGarvey, S. T.; Montgomery, C.; Naseri, T.; Nouraie, S. M.; Preuss, M. H.; Palmer, N. D.; Peyser, P. A.; Raffield, L. M.; Ratan, A.; Redline, S.; Reupena, S.; Rotter, J. I.; Rich, S. S.; Rienstra, M.; Ruczinski, I.; Sankaran, V. G.; Schwartz, D. A.; Seidman, C. E.; Seidman, J. G.; Silverman, E. K.; Smith, J. A.; Stilp, A.; Taylor, K. D.; Telen, M. J.; Weiss, S. T.; Williams, L. K.; Wu, B.; Yanek, L. R.; Zhang, Y.; Lasky-Su, J.; Gingras, M. C.; Dutcher, S. K.; Eichler, E. E.; Gabriel, S.; Germer, S.; Kim, R.; Viaud-Martinez, K. A.; Nickerson, D. A.; Consortium, N. T.-O. for P. M. (TOPMed); Luo, J.; Reiner, A.; Gibbs, R. A.; Boerwinkle, E.; Abecasis, G.; Sedlazeck, F. J. Structural Variation across 138,134 Samples in the TOPMed Consortium. bioRxiv January 26, 2023, p 2023.01.25.525428. https://doi.org/10.1101/2023.01.25.525428.

(78) Boettger, L. M.; Salem, R. M.; Handsaker, R. E.; Peloso, G. M.; Kathiresan, S.; Hirschhorn, J. N.; McCarroll, S. A. Recurring Exon Deletions in the HP (Haptoglobin) Gene Contribute to Lower Blood Cholesterol Levels. *Nat. Genet.* **2016**, *48* (4), 359–366. https://doi.org/10.1038/ng.3510.

(79) Sekar, A.; Bialas, A. R.; de Rivera, H.; Davis, A.; Hammond, T. R.; Kamitaki, N.; Tooley, K.; Presumey, J.; Baum, M.; Van Doren, V.; Genovese, G.; Rose, S. A.; Handsaker, R. E.; Daly, M. J.; Carroll, M. C.; Stevens, B.; McCarroll, S. A. Schizophrenia Risk from Complex Variation of Complement Component 4. *Nature* **2016**, *530* (7589), 177–183. https://doi.org/10.1038/nature16549.

(80) 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.; Perot, G.; Jonker, A. H.; Lucchesi, C.; Le Deley, M.-C.; Oberlin, O.; Marec-Bérard, P.; Véron, A. S.; Reynaud, S.; Lapouble, E.; Boeva, V.; Frio, T. R.; Alonso, J.; Bhatia, S.; Pierron, G.; Cancel-Tassin, G.; Cussenot, O.; Cox, D. G.; Morton, L. M.; Machiela, M. J.; Chanock, S. J.; Charnay, P.; Delattre, O. Chimeric EWSR1-FLI1 Regulates the Ewing Sarcoma Susceptibility Gene EGR2 via a GGAA Microsatellite. *Nat. Genet.* **2015**, *47* (9), 1073–1078. https://doi.org/10.1038/ng.3363.

(81) Mukamel, R. E.; Handsaker, R. E.; Sherman, M. A.; Barton, A. R.; Zheng, Y.; McCarroll, S. A.; Loh, P.-R. *Protein-Coding Repeat Polymorphisms Strongly Shape Diverse Human Phenotypes*; 2021; p 2021.01.19.427332. https://doi.org/10.1101/2021.01.19.427332.

(82) Hujoel, M. L. A.; Handsaker, R. E.; Sherman, M. A.; Kamitaki, N.; Barton, A. R.; Mukamel, R. E.; Terao, C.; McCarroll, S. A.; Loh, P.-R. Protein-Altering Variants at Copy Number-Variable Regions Influence Diverse Human Phenotypes. *Nat. Genet.* **2024**, 1–10. https://doi.org/10.1038/s41588-024-01684-z.

(83) Saini, S.; Mitra, I.; Mousavi, N.; Fotsing, S. F.; Gymrek, M. A Reference Haplotype Panel for Genome-Wide Imputation of Short Tandem Repeats. *Nat. Commun.* **2018**, *9* (1), 4397. https://doi.org/10.1038/s41467-018-06694-0.

(84) Ziaei Jam, H.; Li, Y.; DeVito, R.; Mousavi, N.; Ma, N.; Lujumba, I.; Adam, Y.; Maksimov, M.; Huang, B.; Dolzhenko, E.; Qiu, Y.; Kakembo, F. E.; Joseph, H.; Onyido, B.; Adeyemi, J.; Bakhtiari, M.; Park, J.; Javadzadeh, S.; Jjingo, D.; Adebiyi, E.; Bafna, V.; Gymrek, M. A Deep Population Reference Panel of Tandem Repeat Variation. *Nat. Commun.* **2023**, *14* (1), 6711. https://doi.org/10.1038/s41467-023-42278-3.

(85) Sakaue, S.; Weinand, K.; Isaac, S.; Dey, K. K.; Jagadeesh, K.; Kanai, M.; Watts, G. F. M.; Zhu, Z.; Brenner, M. B.; McDavid, A.; Donlin, L. T.; Wei, K.; Price, A. L.; Raychaudhuri, S. Tissue-Specific Enhancer–Gene Maps from Multimodal Single-Cell Data Identify Causal Disease Alleles. *Nat. Genet.* **2024**, *56* (4), 615–626. https://doi.org/10.1038/s41588-024-01682-1.

(86) Zhu, Z.; Zhang, F.; Hu, H.; Bakshi, A.; Robinson, M. R.; Powell, J. E.; Montgomery, G. W.; Goddard, M. E.; Wray, N. R.; Visscher, P. M.; Yang, J. Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets. *Nat. Genet.* **2016**, *48* (5), 481–487. https://doi.org/10.1038/ng.3538.

(87) Gusev, A.; Ko, A.; Shi, H.; Bhatia, G.; Chung, W.; Penninx, B. W. J. H.; Jansen, R.; de Geus, E. J. C.; Boomsma, D. I.; Wright, F. A.; Sullivan, P. F.; Nikkola, E.; Alvarez, M.; Civelek, M.; Lusis, A. J.; Lehtimäki, T.; Raitoharju, E.; Kähönen, M.; Seppälä, I.; Raitakari, O. T.; Kuusisto, J.; Laakso, M.; Price, A. L.; Pajukanta, P.; Pasaniuc, B. Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies. *Nat. Genet.* **2016**, *48* (3), 245–252. https://doi.org/10.1038/ng.3506.

(88) Stacey, D.; Fauman, E. B.; Ziemek, D.; Sun, B. B.; Harshfield, E. L.; Wood, A. M.; Butterworth, A. S.; Suhre, K.; Paul, D. S. ProGeM: A Framework for the Prioritization of Candidate Causal Genes at Molecular Quantitative Trait Loci. *Nucleic Acids Res.* **2019**, *47* (1), e3. https://doi.org/10.1093/nar/gky837.

(89) Nasser, J.; Bergman, D. T.; Fulco, C. P.; Guckelberger, P.; Doughty, B. R.; Patwardhan, T. A.; Jones, T. R.; Nguyen, T. H.; Ulirsch, J. C.; Lekschas, F.; Mualim, K.; Natri, H. M.; Weeks, E. M.; Munson, G.; Kane, M.; Kang, H. Y.; Cui, A.; Ray, J. P.; Eisenhaure, T. M.; Collins, R. L.; Dey, K.; Pfister, H.; Price, A. L.; Epstein, C. B.; Kundaje, A.; Xavier, R. J.; Daly, M. J.; Huang, H.; Finucane, H. K.; Hacohen, N.; Lander, E. S.; Engreitz, J. M. Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes. *Nature* **2021**, *593* (7858), 238–243. https://doi.org/10.1038/s41586-021-03446-x.

(90) Oliva, M.; Demanelis, K.; Lu, Y.; Chernoff, M.; Jasmine, F.; Ahsan, H.; Kibriya, M. G.; Chen, L. S.; Pierce, B. L. DNA Methylation QTL Mapping across Diverse Human Tissues Provides Molecular Links between Genetic Variation and Complex Traits. *Nat. Genet.* **2023**, *55* (1), 112–122. https://doi.org/10.1038/s41588-022-01248-z.

(91) THE GTEX CONSORTIUM. The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* **2020**, *369* (6509), 1318–1330. https://doi.org/10.1126/science.aaz1776.

(92) Dutta, D.; He, Y.; Saha, A.; Arvanitis, M.; Battle, A.; Chatterjee, N. Aggregative Trans-eQTL Analysis Detects Trait-Specific Target Gene Sets in Whole Blood. *Nat. Commun.* **2022**, *13* (1), 4323. https://doi.org/10.1038/s41467-022-31845-9.

(93) Sun, B. B.; Chiou, J.; Traylor, M.; Benner, C.; Hsu, Y.-H.; Richardson, T. G.; Surendran, P.; Mahajan, A.; Robins, C.; Vasquez-Grinnell, S. G.; Hou, L.; Kvikstad, E. M.; Burren, O. S.; Davitte, J.; Ferber, K. L.; Gillies, C. E.; Hedman, Å. K.; Hu, S.; Lin, T.; Mikkilineni, R.;

Pendergrass, R. K.; Pickering, C.; Prins, B.; Baird, D.; Chen, C.-Y.; Ward, L. D.; Deaton, A. M.; Welsh, S.; Willis, C. M.; Lehner, N.; Arnold, M.; Wörheide, M. A.; Suhre, K.; Kastenmüller, G.; Sethi, A.; Cule, M.; Raj, A.; Kang, H. M.; Burkitt-Gray, L.; Melamud, E.; Black, M. H.; Fauman, E. B.; Howson, J. M. M.; Kang, H. M.; McCarthy, M. I.; Nioi, P.; Petrovski, S.; Scott, R. A.; Smith, E. N.; Szalma, S.; Waterworth, D. M.; Mitnaul, L. J.; Szustakowski, J. D.; Gibson, B. W.; Miller, M. R.; Whelan, C. D. Plasma Proteomic Associations with Genetics and Health in the UK Biobank. *Nature* **2023**, *622* (7982), 329–338. https://doi.org/10.1038/s41586-023-06592-6.

(94) Yao, C.; Chen, G.; Song, C.; Keefe, J.; Mendelson, M.; Huan, T.; Sun, B. B.; Laser, A.; Maranville, J. C.; Wu, H.; Ho, J. E.; Courchesne, P.; Lyass, A.; Larson, M. G.; Gieger, C.; Graumann, J.; Johnson, A. D.; Danesh, J.; Runz, H.; Hwang, S.-J.; Liu, C.; Butterworth, A. S.; Suhre, K.; Levy, D. Genome-wide Mapping of Plasma Protein QTLs Identifies Putatively Causal Genes and Pathways for Cardiovascular Disease. *Nat. Commun.* **2018**, *9* (1), 3268. https://doi.org/10.1038/s41467-018-05512-x.

(95) Wainberg, M.; Sinnott-Armstrong, N.; Mancuso, N.; Barbeira, A. N.; Knowles, D. A.; Golan, D.; Ermel, R.; Ruusalepp, A.; Quertermous, T.; Hao, K.; Björkegren, J. L. M.; Im, H. K.; Pasaniuc, B.; Rivas, M. A.; Kundaje, A. Opportunities and Challenges for Transcriptome-Wide Association Studies. *Nat. Genet.* **2019**, *51* (4), 592–599. https://doi.org/10.1038/s41588-019-0385-z.

(96) Giambartolomei, C.; Vukcevic, D.; Schadt, E. E.; Franke, L.; Hingorani, A. D.; Wallace, C.; Plagnol, V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genet.* **2014**, *10* (5), e1004383. https://doi.org/10.1371/journal.pgen.1004383.

(97) Mai, J.; Lu, M.; Gao, Q.; Zeng, J.; Xiao, J. Transcriptome-Wide Association Studies: Recent Advances in Methods, Applications and Available Databases. *Commun. Biol.* **2023**, *6* (1), 1–10. https://doi.org/10.1038/s42003-023-05279-y.

(98) Wu, C.; Pan, W. A Powerful Fine-Mapping Method for Transcriptome-Wide Association Studies. *Hum. Genet.* **2020**, *139* (2), 199–213. https://doi.org/10.1007/s00439-019-02098-2.

(99) Mancuso, N.; Freund, M. K.; Johnson, R.; Shi, H.; Kichaev, G.; Gusev, A.; Pasaniuc, B. Probabilistic Fine-Mapping of Transcriptome-Wide Association Studies. *Nat. Genet.* **2019**, *51* (4), 675–682. https://doi.org/10.1038/s41588-019-0367-1.

(100) Zhao, S.; Crouse, W.; Qian, S.; Luo, K.; Stephens, M.; He, X. Adjusting for Genetic Confounders in Transcriptome-Wide Association Studies Improves Discovery of Risk Genes of Complex Traits. *Nat. Genet.* **2024**, *56* (2), 336–347. https://doi.org/10.1038/s41588-023-01648-9.

(101) Pers, T. H.; Karjalainen, J. M.; Chan, Y.; Westra, H.-J.; Wood, A. R.; Yang, J.; Lui, J. C.; Vedantam, S.; Gustafsson, S.; Esko, T.; Frayling, T.; Speliotes, E. K.; Boehnke, M.; Raychaudhuri, S.; Fehrmann, R. S. N.; Hirschhorn, J. N.; Franke, L. Biological Interpretation of Genome-Wide Association Studies Using Predicted Gene Functions. *Nat. Commun.* **2015**, *6* (1), 5890. https://doi.org/10.1038/ncomms6890.

(102) Weeks, E. M.; Ulirsch, J. C.; Cheng, N. Y.; Trippe, B. L.; Fine, R. S.; Miao, J.; Patwardhan, T. A.; Kanai, M.; Nasser, J.; Fulco, C. P.; Tashman, K. C.; Aguet, F.; Li, T.; Ordovas-Montanes, J.; Smillie, C. S.; Biton, M.; Shalek, A. K.; Ananthakrishnan, A. N.; Xavier, R. J.; Regev, A.; Gupta, R. M.; Lage, K.; Ardlie, K. G.; Hirschhorn, J. N.; Lander, E. S.; Engreitz, J. M.; Finucane, H. K. Leveraging Polygenic Enrichments of Gene Features to Predict Genes Underlying Complex Traits and Diseases. *Nat. Genet.* **2023**, *55* (8), 1267–1276. https://doi.org/10.1038/s41588-023-01443-6.

(103) Zhou, W.; Kanai, M.; Wu, K.-H. H.; Rasheed, H.; Tsuo, K.; Hirbo, J. B.; Wang, Y.; Bhattacharya, A.; Zhao, H.; Namba, S.; Surakka, I.; Wolford, B. N.; Lo Faro, V.; Lopera-Maya, E. A.; Läll, K.; Favé, M.-J.; Partanen, J. J.; Chapman, S. B.; Karjalainen, J.; Kurki, M.; Maasha, M.; Brumpton, B. M.; Chavan, S.; Chen, T.-T.; Daya, M.; Ding, Y.; Feng, Y.-C. A.;

Guare, L. A.; Gignoux, C. R.; Graham, S. E.; Hornsby, W. E.; Ingold, N.; Ismail, S. I.; Johnson, R.; Laisk, T.; Lin, K.; Lv, J.; Millwood, I. Y.; Moreno-Grau, S.; Nam, K.; Palta, P.; Pandit, A.; Preuss, M. H.; Saad, C.; Setia-Verma, S.; Thorsteinsdottir, U.; Uzunovic, J.; Verma, A.; Zawistowski, M.; Zhong, X.; Afifi, N.; Al-Dabhani, K. M.; Al Thani, A.; Bradford, Y.; Campbell, A.; Crooks, K.; de Bock, G. H.; Damrauer, S. M.; Douville, N. J.; Finer, S.; Fritsche, L. G.; Fthenou, E.; Gonzalez-Arroyo, G.; Griffiths, C. J.; Guo, Y.; Hunt, K. A.; Ioannidis, A.; Jansonius, N. M.; Konuma, T.; Lee, M. T. M.; Lopez-Pineda, A.; Matsuda, Y.; Marioni, R. E.; Moatamed, B.; Nava-Aguilar, M. A.; Numakura, K.; Patil, S.; Rafaels, N.; Richmond, A.; Rojas-Muñoz, A.; Shortt, J. A.; Straub, P.; Tao, R.; Vanderwerff, B.; Vernekar, M.; Veturi, Y.; Barnes, K. C.; Boezen, M.; Chen, Z.; Chen, C.-Y.; Cho, J.; Smith, G. D.; Finucane, H. K.; Franke, L.; Gamazon, E. R.; Ganna, A.; Gaunt, T. R.; Ge, T.; Huang, H.; Huffman, J.; Katsanis, N.; Koskela, J. T.; Lajonchere, C.; Law, M. H.; Li, L.; Lindgren, C. M.; Loos, R. J. F.; MacGregor, S.; Matsuda, K.; Olsen, C. M.; Porteous, D. J.; Shavit, J. A.; Snieder, H.; Takano, T.; Trembath, R. C.; Vonk, J. M.; Whiteman, D. C.; Wicks, S. J.; Wijmenga, C.; Wright, J.; Zheng, J.; Zhou, X.; Awadalla, P.; Boehnke, M.; Bustamante, C. D.; Cox, N. J.; Fatumo, S.; Geschwind, D. H.; Hayward, C.; Hveem, K.; Kenny, E. E.; Lee, S.; Lin, Y.-F.; Mbarek, H.; Mägi, R.; Martin, H. C.; Medland, S. E.; Okada, Y.; Palotie, A. V.; Pasaniuc, B.; Rader, D. J.; Ritchie, M. D.; Sanna, S.; Smoller, J. W.; Stefansson, K.; van Heel, D. A.; Walters, R. G.; Zöllner, S.; Martin, A. R.; Willer, C. J.; Daly, M. J.; Neale, B. M. Global Biobank Meta-Analysis Initiative: Powering Genetic Discovery across Human Disease. *Cell Genomics* **2022**, *2* (10), 100192. https://doi.org/10.1016/j.xgen.2022.100192.

(104) Conti, D. V.; Darst, B. F.; Moss, L. C.; Saunders, E. J.; Sheng, X.; Chou, A.; Schumacher, F. R.; Olama, A. A. A.; Benlloch, S.; Dadaev, T.; Brook, M. N.; Sahimi, A.; Hoffmann, T. J.; Takahashi, A.; Matsuda, K.; Momozawa, Y.; Fujita, M.; Muir, K.; Lophatananon, A.; Wan, P.; Le Marchand, L.; Wilkens, L. R.; Stevens, V. L.; Gapstur, S. M.; Carter, B. D.; Schleutker, J.; Tammela, T. L. J.; Sipeky, C.; Auvinen, A.; Giles, G. G.; Southey, M. C.; MacInnis, R. J.; Cybulski, C.; Wokołorczyk, D.; Lubiński, J.; Neal, D. E.; Donovan, J. L.; Hamdy, F. C.; Martin, R. M.; Nordestgaard, B. G.; Nielsen, S. F.; Weischer, M.; Bojesen, S. E.; Røder, M. A.; Iversen, P.; Batra, J.; Chambers, S.; Moya, L.; Horvath, L.; Clements, J. A.; Tilley, W.; Risbridger, G. P.; Gronberg, H.; Aly, M.; Szulkin, R.; Eklund, M.; Nordström, T.; Pashayan, N.; Dunning, A. M.; Ghoussaini, M.; Travis, R. C.; Key, T. J.; Riboli, E.; Park, J. Y.; Sellers, T. A.; Lin, H.-Y.; Albanes, D.; Weinstein, S. J.; Mucci, L. A.; Giovannucci, E.; Lindstrom, S.; Kraft, P.; Hunter, D. J.; Penney, K. L.; Turman, C.; Tangen, C. M.; Goodman, P. J.; Thompson, I. M.; Hamilton, R. J.; Fleshner, N. E.; Finelli, A.; Parent, M.-É.; Stanford, J. L.; Ostrander, E. A.; Geybels, M. S.; Koutros, S.; Freeman, L. E. B.; Stampfer, M.; Wolk, A.; Håkansson, N.; Andriole, G. L.; Hoover, R. N.; Machiela, M. J.; Sørensen, K. D.; Borre, M.; Blot, W. J.; Zheng, W.; Yeboah, E. D.; Mensah, J. E.; Lu, Y.-J.; Zhang, H.-W.; Feng, N.; Mao, X.; Wu, Y.; Zhao, S.-C.; Sun, Z.; Thibodeau, S. N.; McDonnell, S. K.; Schaid, D. J.; West, C. M. L.; Burnet, N.; Barnett, G.; Maier, C.; Schnoeller, T.; Luedeke, M.; Kibel, A. S.; Drake, B. F.; Cussenot, O.; Cancel-Tassin, G.; Menegaux, F.; Truong, T.; Koudou, Y. A.; John, E. M.; Grindedal, E. M.; Maehle, L.; Khaw, K.-T.; Ingles, S. A.; Stern, M. C.; Vega, A.; Gómez-Caamaño, A.; Fachal, L.; Rosenstein, B. S.; Kerns, S. L.; Ostrer, H.; Teixeira, M. R.; Paulo, P.; Brandão, A.; Watya, S.; Lubwama, A.; Bensen, J. T.; Fontham, E. T. H.; Mohler, J.; Taylor, J. A.; Kogevinas, M.; Llorca, J.; Castaño-Vinyals, G.; Cannon-Albright, L.; Teerlink, C. C.; Huff, C. D.; Strom, S. S.; Multigner, L.; Blanchet, P.; Brureau, L.; Kaneva, R.; Slavov, C.; Mitev, V.; Leach, R. J.; Weaver, B.; Brenner, H.; Cuk, K.; Holleczek, B.; Saum, K.-U.; Klein, E. A.; Hsing, A. W.; Kittles, R. A.; Murphy, A. B.; Logothetis, C. J.; Kim, J.; Neuhausen, S. L.; Steele, L.; Ding, Y. C.; Isaacs, W. B.; Nemesure, B.; Hennis, A. J. M.; Carpten, J.; Pandha, H.; Michael, A.; De Ruyck, K.; De Meerleer, G.; Ost, P.; Xu, J.; Razack, A.; Lim, J.; Teo, S.-H.; Newcomb, L. F.; Lin, D. W.; Fowke, J. H.; Neslund-Dudas, C.; Rybicki, B. A.; Gamulin, M.; Lessel, D.; Kulis,

T.; Usmani, N.; Singhal, S.; Parliament, M.; Claessens, F.; Joniau, S.; Van den Broeck, T.; Gago-Dominguez, M.; Castelao, J. E.; Martinez, M. E.; Larkin, S.; Townsend, P. A.; Aukim-Hastie, C.; Bush, W. S.; Aldrich, M. C.; Crawford, D. C.; Srivastava, S.; Cullen, J. C.; Petrovics, G.; Casey, G.; Roobol, M. J.; Jenster, G.; van Schaik, R. H. N.; Hu, J. J.; Sanderson, M.; Varma, R.; McKean-Cowdin, R.; Torres, M.; Mancuso, N.; Berndt, S. I.; Van Den Eeden, S. K.; Easton, D. F.; Chanock, S. J.; Cook, M. B.; Wiklund, F.; Nakagawa, H.; Witte, J. S.; Eeles, R. A.; Kote-Jarai, Z.; Haiman, C. A. Trans-Ancestry Genome-Wide Association Meta-Analysis of Prostate Cancer Identifies New Susceptibility Loci and Informs Genetic Risk Prediction. *Nat. Genet.* **2021**, *53* (1), 65–75. https://doi.org/10.1038/s41588-020-00748-0.

(105) Ni, G.; Zeng, J.; Revez, J. A.; Wang, Y.; Zheng, Z.; Ge, T.; Restuadi, R.; Kiewa, J.; Nyholt, D. R.; Coleman, J. R. I.; Smoller, J. W.; Ripke, S.; Neale, B. M.; Corvin, A.; Walters, J. T. R.; Farh, K.-H.; Holmans, P. A.; Lee, P.; Bulik-Sullivan, B.; Collier, D. A.; Huang, H.; Pers, T. H.; Agartz, I.; Agerbo, E.; Albus, M.; Alexander, M.; Amin, F.; Bacanu, S. A.; Begemann, M.; Belliveau, R. A.; Bene, J.; Bergen, S. E.; Bevilacqua, E.; Bigdeli, T. B.; Black, D. W.; Bruggeman, R.; Buccola, N. G.; Buckner, R. L.; Byerley, W.; Cahn, W.; Cai, G.; Campion, D.; Cantor, R. M.; Carr, V. J.; Carrera, N.; Catts, S. V.; Chambert, K. D.; Chan, R. C. K.; Chen, R. Y. L.; Chen, E. Y. H.; Cheng, W.; Cheung, E. F. C.; Chong, S. A.; Cloninger, C. R.; Cohen, D.; Cohen, N.; Cormican, P.; Craddock, N.; Crowley, J. J.; Davidson, M.; Davis, K. L.; Degenhardt, F.; Del Favero, J.; Demontis, D.; Dikeos, D.; Dinan, T.; Djurovic, S.; Donohoe, G.; Drapeau, E.; Duan, J.; Dudbridge, F.; Durmishi, N.; Eichhammer, P.; Eriksson, J.; Escott-Price, V.; Essioux, L.; Fanous, A. H.; Farrell, M. S.; Frank, J.; Franke, L.; Freedman, R.; Freimer, N. B.; Friedl, M.; Friedman, J. I.; Fromer, M.; Genovese, G.; Georgieva, L.; Giegling, I.; Giusti-Rodríguez, P.; Godard, S.; Goldstein, J. I.; Golimbet, V.; Gopal, S.; Gratten, J.; de Haan, L.; Hammer, C.; Hamshere, M. L.; Hansen, M.; Hansen, T.; Haroutunian, V.; Hartmann, A. M.; Henskens, F. A.; Herms, S.; Hirschhorn, J. N.; Hoffmann, P.; Hofman, A.; Hollegaard, M. V.; Hougaard, D. M.; Ikeda, M.; Joa, I.; Julià, A.; Kahn, R. S.; Kalaydjieva, L.; Karachanak-Yankova, S.; Karjalainen, J.; Kavanagh, D.; Keller, M. C.; Kennedy, J. L.; Khrunin, A.; Kim, Y.; Klovins, J.; Knowles, J. A.; Konte, B.; Kucinskas, V.; Kucinskiene, Z. A.; Kuzelova-Ptackova, H.; Kähler, A. K.; Laurent, C.; Lee, J.; Lee, S. H.; Legge, S. E.; Lerer, B.; Li, M.; Li, T.; Liang, K.-Y.; Lieberman, J.; Limborska, S.; Loughland, C. M.; Lubinski, J.; Lönnqvist, J.; Macek, M.; Magnusson, P. K. E.; Maher, B. S.; Maier, W.; Mallet, J.; Marsal, S.; Mattheisen, M.; Mattingsdal, M.; McCarley, R. W.; McDonald, C.; McIntosh, A. M.; Meier, S.; Meijer, C. J.; Melegh, B.; Melle, I.; Mesholam-Gately, R. I.; Metspalu, A.; Michie, P. T.; Milani, L.; Milanova, V.; Mokrab, Y.; Morris, D. W.; Mors, O.; Murphy, K. C.; Murray, R. M.; Myin-Germeys, I.; Müller-Myhsok, B.; Nelis, M.; Nenadic, I.; Nertney, D. A.; Nestadt, G.; Nicodemus, K. K.; Nikitina-Zake, L.; Nisenbaum, L.; Nordin, A.; O'Callaghan, E.; O'Dushlaine, C.; O'Neill, F. A.; Oh, S.-Y.; Olincy, A.; Olsen, L.; Van Os, J.; International Consortium, P. E.; Pantelis, C.; Papadimitriou, G. N.; Papiol, S.; Parkhomenko, E.; Pato, M. T.; Paunio, T.; Pejovic-Milovancevic, M.; Perkins, D. O.; Pietiläinen, O.; Pimm, J.; Pocklington, A. J.; Powell, J.; Price, A.; Pulver, A. E.; Purcell, S. M.; Quested, D.; Rasmussen, H. B.; Reichenberg, A.; Reimers, M. A.; Richards, A. L.; Roffman, J. L.; Roussos, P.; Ruderfer, D. M.; Salomaa, V.; Sanders, A. R.; Schall, U.; Schubert, C. R.; Schulze, T. G.; Schwab, S. G.; Scolnick, E. M.; Scott, R. J.; Seidman, L. J.; Shi, J.; Sigurdsson, E.; Silagadze, T.; Silverman, J. M.; Sim, K.; Slominsky, P.; Smoller, J. W.; So, H.-C.; Spencer, C. C. A.; Stahl, E. A.; Stefansson, H.; Steinberg, S.; Stogmann, E.; Straub, R. E.; Strengman, E.; Strohmaier, J.; Stroup, T. S.; Subramaniam, M.; Suvisaari, J.; Svrakic, D. M.; Szatkiewicz, J. P.; Söderman, E.; Thirumalai, S.; Toncheva, D.; Tosato, S.; Veijola, J.; Waddington, J.; Walsh, D.; Wang, D.; Wang, Q.; Webb, B. T.; Weiser, M.; Wildenauer, D. B.; Williams, N. M.; Williams, S.; Witt, S. H.; Wolen, A. R.; Wong, E. H. M.; Wormley, B. K.; Xi, H. S.; Zai, C. C.; Zheng, X.; Zimprich, F.; Wray, N. R.; Stefansson, K.;

Visscher, P. M.; Case-Control Consortium, W. T.; Adolfsson, R.; Andreassen, O. A.; Blackwood, D. H. R.; Bramon, E.; Buxbaum, J. D.; Børglum, A. D.; Cichon, S.; Darvasi, A.; Domenici, E.; Ehrenreich, H.; Esko, T.; Gejman, P. V.; Gill, M.; Gurling, H.; Hultman, C. M.; Iwata, N.; Jablensky, A. V.; Jönsson, E. G.; Kendler, K. S.; Kirov, G.; Knight, J.; Lencz, T.; Levinson, D. F.; Li, Q. S.; Liu, J.; Malhotra, A. K.; McCarroll, S. A.; McQuillin, A.; Moran, J. L.; Mortensen, P. B.; Mowry, B. J.; Nöthen, M. M.; Ophoff, R. A.; Owen, M. J.; Palotie, A.; Pato, C. N.; Petryshen, T. L.; Posthuma, D.; Rietschel, M.; Riley, B. P.; Rujescu, D.; Sham, P. C.; Sklar, P.; St Clair, D.; Weinberger, D. R.; Wendland, J. R.; Werge, T.; Daly, M. J.; Sullivan, P. F.; O'Donovan, M. C.; Wray, N. R.; Ripke, S.; Mattheisen, M.; Trzaskowski, M.; Byrne, E. M.; Abdellaoui, A.; Adams, M. J.; Agerbo, E.; Air, T. M.; Andlauer, T. F. M.; Bacanu, S.-A.; Bækvad-Hansen, M.; Beekman, A. T. F.; Bigdeli, T. B.; Binder, E. B.; Bryois, J.; Buttenschøn, H. N.; Bybjerg-Grauholm, J.; Cai, N.; Castelao, E.; Christensen, J. H.; Clarke, T.-K.; Coleman, J. R. I.; Colodro-Conde, L.; Couvy-Duchesne, B.; Craddock, N.; Crawford, G. E.; Davies, G.; Deary, I. J.; Degenhardt, F.; Derks, E. M.; Direk, N.; Dolan, C. V.; Dunn, E. C.; Eley, T. C.; Escott-Price, V.; Hassan Kiadeh, F. F.; Finucane, H. K.; Foo, J. C.; Forstner, A. J.; Frank, J.; Gaspar, H. A.; Gill, M.; Goes, F. S.; Gordon, S. D.; Grove, J.; Hall, L. S.; Hansen, C. S.; Hansen, T. F.; Herms, S.; Hickie, I. B.; Hoffmann, P.; Homuth, G.; Horn, C.; Hottenga, J.-J.; Hougaard, D. M.; Howard, D. M.; Ising, M.; Jansen, R.; Jones, I.; Jones, L. A.; Jorgenson, E.; Knowles, J. A.; Kohane, I. S.; Kraft, J.; Kretzschmar, W. W.; Kutalik, Z.; Li, Y.; Lind, P. A.; MacIntyre, D. J.; MacKinnon, D. F.; Maier, R. M.; Maier, W.; Marchini, J.; Mbarek, H.; McGrath, P.; McGuffin, P.; Medland, S. E.; Mehta, D.; Middeldorp, C. M.; Mihailov, E.; Milaneschi, Y.; Milani, L.; Mondimore, F. M.; Montgomery, G. W.; Mostafavi, S.; Mullins, N.; Nauck, M.; Ng, B.; Nivard, M. G.; Nyholt, D. R.; O'Reilly, P. F.; Oskarsson, H.; Owen, M. J.; Painter, J. N.; Pedersen, C. B.; Pedersen, M. G.; Peterson, R. E.; Peyrot, W. J.; Pistis, G.; Posthuma, D.; Quiroz, J. A.; Qvist, P.; Rice, J. P.; Riley, B. P.; Rivera, M.; Mirza, S. S.; Schoevers, R.; Schulte, E. C.; Shen, L.; Shi, J.; Shyn, S. I.; Sigurdsson, E.; Sinnamon, G. C. B.; Smit, J. H.; Smith, D. J.; Stefansson, H.; Steinberg, S.; Streit, F.; Strohmaier, J.; Tansey, K. E.; Teismann, H.; Teumer, A.; Thompson, W.; Thomson, P. A.; Thorgeirsson, T. E.; Traylor, M.; Treutlein, J.; Trubetskoy, V.; Uitterlinden, A. G.; Umbricht, D.; Van der Auwera, S.; van Hemert, A. M.; Viktorin, A.; Visscher, P. M.; Wang, Y.; Webb, B. T.; Weinsheimer, S. M.; Wellmann, J.; Willemsen, G.; Witt, S. H.; Wu, Y.; Xi, H. S.; Yang, J.; Zhang, F.; Arolt, V.; Baune, B. T.; Berger, K.; Boomsma, D. I.; Cichon, S.; Dannlowski, U.; de Geus, E. J. C.; DePaulo, J. R.; Domenici, E.; Domschke, K.; Esko, T.; Grabe, H. J.; Hamilton, S. P.; Hayward, C.; Heath, A. C.; Kendler, K. S.; Kloiber, S.; Lewis, G.; Li, Q. S.; Lucae, S.; Madden, P. A. F.; Magnusson, P. K.; Martin, N. G.; McIntosh, A. M.; Metspalu, A.; Mors, O.; Mortensen, P. B.; Müller-Myhsok, B.; Nordentoft, M.; Nöthen, M. M.; O'Donovan, M. C.; Paciga, S. A.; Pedersen, N. L.; Yang, J.; Visscher, P. M.; Wray, N. R. A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol. Psychiatry* **2021**, *90* (9), 611–620. https://doi.org/10.1016/j.biopsych.2021.04.018.

(106) Wang, Y.; Tsuo, K.; Kanai, M.; Neale, B. M.; Martin, A. R. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu. Rev. Biomed. Data Sci.* **2022**, *5* (Volume 5, 2022), 293–320. https://doi.org/10.1146/annurev-biodatasci-111721-074830.

(107) Ma, Y.; Zhou, X. Genetic Prediction of Complex Traits with Polygenic Scores: A Statistical Review. *Trends Genet.* **2021**, *37* (11), 995–1011. https://doi.org/10.1016/j.tig.2021.06.004.

(108) Fahed, A. C.; Philippakis, A. A.; Khera, A. V. The Potential of Polygenic Scores to Improve Cost and Efficiency of Clinical Trials. *Nat. Commun.* **2022**, *13* (1), 2922. https://doi.org/10.1038/s41467-022-30675-z.

(109) *Could Polygenic Risk Scores Be Useful in Psychiatry? A Review | Psychiatry and Behavioral Health | JAMA Psychiatry | JAMA Network*. https://jamanetwork.com/journals/jamapsychiatry/article-abstract/2771208 (accessed 2024-05-06).

(110) Klarin, D.; Natarajan, P. Clinical Utility of Polygenic Risk Scores for Coronary Artery Disease. *Nat. Rev. Cardiol.* **2022**, *19* (5), 291–301. https://doi.org/10.1038/s41569-021-00638-w.

(111) Shieh, Y.; Eklund, M.; Madlensky, L.; Sawyer, S. D.; Thompson, C. K.; Stover Fiscalini, A.; Ziv, E.; van't Veer, L. J.; Esserman, L. J.; Tice, J. A.; on behalf of the Athena Breast Health Network Investigators. Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *JNCI J. Natl. Cancer Inst.* **2017**, *109* (5), djw290. https://doi.org/10.1093/jnci/djw290.

(112) Roux, A.; Cholerton, R.; Sicsic, J.; Moumjid, N.; French, D. P.; Giorgi Rossi, P.; Balleyguier, C.; Guindy, M.; Gilbert, F. J.; Burrion, J.-B.; Castells, X.; Ritchie, D.; Keatley, D.; Baron, C.; Delaloge, S.; de Montgolfier, S. Study Protocol Comparing the Ethical, Psychological and Socio-Economic Impact of Personalised Breast Cancer Screening to That of Standard Screening in the "My Personal Breast Screening" (MyPeBS) Randomised Clinical Trial. *BMC Cancer* **2022**, *22* (1), 507. https://doi.org/10.1186/s12885-022-09484-6.

(113) Saya, S.; Boyd, L.; Chondros, P.; McNamara, M.; King, M.; Milton, S.; Lourenco, R. D. A.; Clark, M.; Fishman, G.; Marker, J.; Ostroff, C.; Allman, R.; Walter, F. M.; Buchanan, D.; Winship, I.; McIntosh, J.; Macrae, F.; Jenkins, M.; Emery, J. The SCRIPT Trial: Study Protocol for a Randomised Controlled Trial of a Polygenic Risk Score to Tailor Colorectal Cancer Screening in Primary Care. *Trials* **2022**, *23* (1), 810. https://doi.org/10.1186/s13063-022-06734-7.

(114) Weissbrod, O.; Kanai, M.; Shi, H.; Gazal, S.; Peyrot, W. J.; Khera, A. V.; Okada, Y.; Martin, A. R.; Finucane, H. K.; Price, A. L. Leveraging Fine-Mapping and Multipopulation Training Data to Improve Cross-Population Polygenic Risk Scores. *Nat. Genet.* **2022**, *54* (4), 450–458. https://doi.org/10.1038/s41588-022-01036-9.

(115) Darst, B. F.; Shen, J.; Madduri, R. K.; Rodriguez, A. A.; Xiao, Y.; Sheng, X.; Saunders, E. J.; Dadaev, T.; Brook, M. N.; Hoffmann, T. J.; Muir, K.; Wan, P.; Marchand, L. L.; Wilkens, L.; Wang, Y.; Schleutker, J.; MacInnis, R. J.; Cybulski, C.; Neal, D. E.; Nordestgaard, B. G.; Nielsen, S. F.; Batra, J.; Clements, J. A.; BioResource, A. P. C.; Grönberg, H.; Pashayan, N.; Travis, R. C.; Park, J. Y.; Albanes, D.; Weinstein, S.; Mucci, L. A.; Hunter, D. J.; Penney, K. L.; Tangen, C. M.; Hamilton, R. J.; Parent, M.-É.; Stanford, J. L.; Koutros, S.; Wolk, A.; Sørensen, K. D.; Blot, W. J.; Yeboah, E. D.; Mensah, J. E.; Lu, Y.-J.; Schaid, D. J.; Thibodeau, S. N.; West, C. M.; Maier, C.; Kibel, A. S.; Cancel-Tassin, G.; Menegaux, F.; John, E. M.; Grindedal, E. M.; Khaw, K.-T.; Ingles, S. A.; Vega, A.; Rosenstein, B. S.; Teixeira, M. R.; Kogevinas, M.; Cannon-Albright, L.; Huff, C.; Multigner, L.; Kaneva, R.; Leach, R. J.; Brenner, H.; Hsing, A. W.; Kittles, R. A.; Murphy, A. B.; Logothetis, C. J.; Neuhausen, S. L.; Isaacs, W. B.; Nemesure, B.; Hennis, A. J.; Carpten, J.; Pandha, H.; Ruyck, K. D.; Xu, J.; Razack, A.; Teo, S.-H.; Newcomb, L. F.; Fowke, J. H.; Neslund-Dudas, C.; Rybicki, B. A.; Gamulin, M.; Usmani, N.; Claessens, F.; Gago-Dominguez, M.; Castelao, J. E.; Townsend, P. A.; Crawford, D. C.; Petrovics, G.; Casey, G.; Roobol, M. J.; Hu, J. F.; Berndt, S. I.; Eeden, S. K. V. D.; Easton, D. F.; Chanock, S. J.; Cook, M. B.; Wiklund, F.; Witte, J. S.; Eeles, R. A.; Kote-Jarai, Z.; Watya, S.; Gaziano, J. M.; Justice, A. C.; Conti, D. V.; Haiman, C. A. Evaluating Approaches for Constructing Polygenic Risk Scores for Prostate Cancer in Men of African and European Ancestry. *Am. J. Hum. Genet.* **2023**, *110* (7), 1200–1206. https://doi.org/10.1016/j.ajhg.2023.05.010.

(116) Mak, T. S. H.; Porsch, R. M.; Choi, S. W.; Zhou, X.; Sham, P. C. Polygenic Scores via Penalized Regression on Summary Statistics. *Genet. Epidemiol.* **2017**, *41* (6), 469–480. https://doi.org/10.1002/gepi.22050.

(117) Vilhjálmsson, B. J.; Yang, J.; Finucane, H. K.; Gusev, A.; Lindström, S.; Ripke, S.; Genovese, G.; Loh, P.-R.; Bhatia, G.; Do, R.; Hayeck, T.; Won, H.-H.; Ripke, S.; Neale, B. M.; Corvin, A.; Walters, J. T. R.; Farh, K.-H.; Holmans, P. A.; Lee, P.; Bulik-Sullivan, B.; Collier, D. A.; Huang, H.; Pers, T. H.; Agartz, I.; Agerbo, E.; Albus, M.; Alexander, M.; Amin, F.; Bacanu, S. A.; Begemann, M.; Belliveau, R. A.; Bene, J.; Bergen, S. E.; Bevilacqua, E.; Bigdeli, T. B.; Black, D. W.; Bruggeman, R.; Buccola, N. G.; Buckner, R. L.; Byerley, W.; Cahn, W.; Cai, G.; Campion, D.; Cantor, R. M.; Carr, V. J.; Carrera, N.; Catts, S. V.; Chambert, K. D.; Chan, R. C. K.; Chen, R. Y. L.; Chen, E. Y. H.; Cheng, W.; Cheung, E. F. C.; Chong, S. A.; Cloninger, C. R.; Cohen, D.; Cohen, N.; Cormican, P.; Craddock, N.; Crowley, J. J.; Curtis, D.; Davidson, M.; Davis, K. L.; Degenhardt, F.; Del Favero, J.; DeLisi, L. E.; Demontis, D.; Dikeos, D.; Dinan, T.; Djurovic, S.; Donohoe, G.; Drapeau, E.; Duan, J.; Dudbridge, F.; Durmishi, N.; Eichhammer, P.; Eriksson, J.; Escott-Price, V.; Essioux, L.; Fanous, A. H.; Farrell, M. S.; Frank, J.; Franke, L.; Freedman, R.; Freimer, N. B.; Friedl, M.; Friedman, J. I.; Fromer, M.; Genovese, G.; Georgieva, L.; Gershon, E. S.; Giegling, I.; Giusti-Rodrguez, P.; Godard, S.; Goldstein, J. I.; Golimbet, V.; Gopal, S.; Gratten, J.; Grove, J.; de Haan, L.; Hammer, C.; Hamshere, M. L.; Hansen, M.; Hansen, T.; Haroutunian, V.; Hartmann, A. M.; Henskens, F. A.; Herms, S.; Hirschhorn, J. N.; Hoffmann, P.; Hofman, A.; Hollegaard, M. V.; Hougaard, D. M.; Ikeda, M.; Joa, I.; Julia, A.; Kahn, R. S.; Kalaydjieva, L.; Karachanak-Yankova, S.; Karjalainen, J.; Kavanagh, D.; Keller, M. C.; Kelly, B. J.; Kennedy, J. L.; Khrunin, A.; Kim, Y.; Klovins, J.; Knowles, J. A.; Konte, B.; Kucinskas, V.; Kucinskiene, Z. A.; Kuzelova-Ptackova, H.; Kahler, A. K.; Laurent, C.; Keong, J. L. C.; Lee, S. H.; Legge, S. E.; Lerer, B.; Li, M.; Li, T.; Liang, K.-Y.; Lieberman, J.; Limborska, S.; Loughland, C. M.; Lubinski, J.; Lnnqvist, J.; Macek, M.; Magnusson, P. K. E.; Maher, B. S.; Maier, W.; Mallet, J.; Marsal, S.; Mattheisen, M.; Mattingsdal, M.; McCarley, R. W.; McDonald, C.; McIntosh, A. M.; Meier, S.; Meijer, C. J.; Melegh, B.; Melle, I.; Mesholam-Gately, R. I.; Metspalu, A.; Michie, P. T.; Milani, L.; Milanova, V.; Mokrab, Y.; Morris, D. W.; Mors, O.; Mortensen, P. B.; Murphy, K. C.; Murray, R. M.; Myin-Germeys, I.; Mller-Myhsok, B.; Nelis, M.; Nenadic, I.; Nertney, D. A.; Nestadt, G.; Nicodemus, K. K.; Nikitina-Zake, L.; Nisenbaum, L.; Nordin, A.; O'Callaghan, E.; O'Dushlaine, C.; O'Neill, F. A.; Oh, S.-Y.; Olincy, A.; Olsen, L.; Van Os, J.; Pantelis, C.; Papadimitriou, G. N.; Papiol, S.; Parkhomenko, E.; Pato, M. T.; Paunio, T.; Pejovic-Milovancevic, M.; Perkins, D. O.; Pietilinen, O.; Pimm, J.; Pocklington, A. J.; Powell, J.; Price, A.; Pulver, A. E.; Purcell, S. M.; Quested, D.; Rasmussen, H. B.; Reichenberg, A.; Reimers, M. A.; Richards, A. L.; Roffman, J. L.; Roussos, P.; Ruderfer, D. M.; Salomaa, V.; Sanders, A. R.; Schall, U.; Schubert, C. R.; Schulze, T. G.; Schwab, S. G.; Scolnick, E. M.; Scott, R. J.; Seidman, L. J.; Shi, J.; Sigurdsson, E.; Silagadze, T.; Silverman, J. M.; Sim, K.; Slominsky, P.; Smoller, J. W.; So, H.-C.; Spencer, C. C. A.; Stahl, E. A.; Stefansson, H.; Steinberg, S.; Stogmann, E.; Straub, R. E.; Strengman, E.; Strohmaier, J.; Stroup, T. S.; Subramaniam, M.; Suvisaari, J.; Svrakic, D. M.; Szatkiewicz, J. P.; Sderman, E.; Thirumalai, S.; Toncheva, D.; Tooney, P. A.; Tosato, S.; Veijola, J.; Waddington, J.; Walsh, D.; Wang, D.; Wang, Q.; Webb, B. T.; Weiser, M.; Wildenauer, D. B.; Williams, N. M.; Williams, S.; Witt, S. H.; Wolen, A. R.; Wong, E. H. M.; Wormley, B. K.; Wu, J. Q.; Xi, H. S.; Zai, C. C.; Zheng, X.; Zimprich, F.; Wray, N. R.; Stefansson, K.; Visscher, P. M.; Adolfsson, R.; Andreassen, O. A.; Blackwood, D. H. R.; Bramon, E.; Buxbaum, J. D.; Børglum, A. D.; Cichon, S.; Darvasi, A.; Domenici, E.; Ehrenreich, H.; Esko, T.; Gejman, P. V.; Gill, M.; Gurling, H.; Hultman, C. M.; Iwata, N.; Jablensky, A. V.; Jonsson, E. G.; Kendler, K. S.; Kirov, G.; Knight, J.; Lencz, T.; Levinson, D. F.; Li, Q. S.; Liu, J.; Malhotra, A. K.; McCarroll, S. A.; McQuillin, A.; Moran, J. L.; Mortensen, P. B.; Mowry, B. J.; Nthen, M. M.; Ophoff, R. A.; Owen, M. J.; Palotie, A.; Pato, C. N.; Petryshen, T. L.; Posthuma, D.; Rietschel, M.; Riley, B. P.; Rujescu, D.; Sham, P. C.; Sklar, P.; St. Clair, D.; Weinberger, D. R.; Wendland, J. R.; Werge, T.; Daly, M. J.; Sullivan, P. F.; O'Donovan, M. C.; Kraft, P.; Hunter, D. J.; Adank, M.; Ahsan, H.; Aittomäki, K.; Baglietto, L.; Berndt, S.; Blomquist, C.; Canzian, F.;

Chang-Claude, J.; Chanock, S. J.; Crisponi, L.; Czene, K.; Dahmen, N.; Silva, I. dos S.; Easton, D.; Eliassen, A. H.; Figueroa, J.; Fletcher, O.; Garcia-Closas, M.; Gaudet, M. M.; Gibson, L.; Haiman, C. A.; Hall, P.; Hazra, A.; Hein, R.; Henderson, B. E.; Hofman, A.; Hopper, J. L.; Irwanto, A.; Johansson, M.; Kaaks, R.; Kibriya, M. G.; Lichtner, P.; Lindström, S.; Liu, J.; Lund, E.; Makalic, E.; Meindl, A.; Meijers-Heijboer, H.; Müller-Myhsok, B.; Muranen, T. A.; Nevanlinna, H.; Peeters, P. H.; Peto, J.; Prentice, R. L.; Rahman, N.; Sánchez, M. J.; Schmidt, D. F.; Schmutzler, R. K.; Southey, M. C.; Tamimi, R.; Travis, R.; Turnbull, C.; Uitterlinden, A. G.; van der Luijt, R. B.; Waisfisz, Q.; Wang, Z.; Whittemore, A. S.; Yang, R.; Zheng, W.; Kathiresan, S.; Pato, M.; Pato, C.; Tamimi, R.; Stahl, E.; Zaitlen, N.; Pasaniuc, B.; Belbin, G.; Kenny, E. E.; Schierup, M. H.; De Jager, P.; Patsopoulos, N. A.; McCarroll, S.; Daly, M.; Purcell, S.; Chasman, D.; Neale, B.; Goddard, M.; Visscher, P. M.; Kraft, P.; Patterson, N.; Price, A. L. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **2015**, *97* (4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001.

(118) Newcombe, P. J.; Nelson, C. P.; Samani, N. J.; Dudbridge, F. A Flexible and Parallelizable Approach to Genome-Wide Polygenic Risk Scores. *Genet. Epidemiol.* **2019**, *43* (7), 730–741. https://doi.org/10.1002/gepi.22245.

(119) Martin, A. R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B. M.; Daly, M. J. Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities. *Nat. Genet.* **2019**, *51* (4), 584–591. https://doi.org/10.1038/s41588-019-0379-x.

(120) Ding, Y.; Hou, K.; Xu, Z.; Pimplaskar, A.; Petter, E.; Boulier, K.; Privé, F.; Vilhjálmsson, B. J.; Olde Loohuis, L. M.; Pasaniuc, B. Polygenic Scoring Accuracy Varies across the Genetic Ancestry Continuum. *Nature* **2023**, *618* (7966), 774–781. https://doi.org/10.1038/s41586-023-06079-4.

(121) Duncan, L.; Shen, H.; Gelaye, B.; Meijsen, J.; Ressler, K.; Feldman, M.; Peterson, R.; Domingue, B. Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nat. Commun.* **2019**, *10* (1), 3328. https://doi.org/10.1038/s41467-019-11112-0.

(122) Wang, Y.; Guo, J.; Ni, G.; Yang, J.; Visscher, P. M.; Yengo, L. Theoretical and Empirical Quantification of the Accuracy of Polygenic Scores in Ancestry Divergent Populations. *Nat. Commun.* **2020**, *11* (1), 3865. https://doi.org/10.1038/s41467-020-17719-y.

(123) Horton, C. A.; Alexandari, A. M.; Hayes, M. G. B.; Marklund, E.; Schaepe, J. M.; Aditham, A. K.; Shah, N.; Suzuki, P. H.; Shrikumar, A.; Afek, A.; Greenleaf, W. J.; Gordân, R.; Zeitlinger, J.; Kundaje, A.; Fordyce, P. M. Short Tandem Repeats Bind Transcription Factors to Tune Eukaryotic Gene Expression. *Science* **2023**, *381* (6664), eadd1250. https://doi.org/10.1126/science.add1250.

(124) Rajan-Babu, I.-S.; Dolzhenko, E.; Eberle, M. A.; Friedman, J. M. Sequence Composition Changes in Short Tandem Repeats: Heterogeneity, Detection, Mechanisms and Clinical Implications. *Nat. Rev. Genet.* **2024**, 1–24. https://doi.org/10.1038/s41576-024-00696-z.

(125) Zane, L.; Bargelloni, L.; Patarnello, T. Strategies for Microsatellite Isolation: A Review. *Mol. Ecol.* **2002**, *11* (1), 1–16. https://doi.org/10.1046/j.0962-1083.2001.01418.x.

(126) Vieira, M. L. C.; Santini, L.; Diniz, A. L.; Munhoz, C. de F. Microsatellite Markers: What They Mean and Why They Are so Useful. *Genet. Mol. Biol.* **2016**, *39*, 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027.

(127) Depienne, C.; Mandel, J.-L. 30 Years of Repeat Expansion Disorders: What Have We Learned and What Are the Remaining Challenges? *Am. J. Hum. Genet.* **2021**, *108* (5), 764–785. https://doi.org/10.1016/j.ajhg.2021.03.011.

(128) Budowle, B.; Shea, B.; Niezgoda, S.; Chakraborty, R. CODIS STR Loci Data from 41 Sample Populations. *J. Forensic Sci.* **2001**, *46* (3), 453–489. https://doi.org/10.1520/JFS14996J.

(129) Broman, K. W.; Murray, J. C.; Sheffield, V. C.; White, R. L.; Weber, J. L. Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination. *Am. J. Hum. Genet.* **1998**, *63* (3), 861–869. https://doi.org/10.1086/302011.

(130) Malik, I.; Kelley, C. P.; Wang, E. T.; Todd, P. K. Molecular Mechanisms Underlying Nucleotide Repeat Expansion Disorders. *Nat. Rev. Mol. Cell Biol.* **2021**, *22* (9), 589–607. https://doi.org/10.1038/s41580-021-00382-6.

(131) Mirkin, S. M. Expandable DNA Repeats and Human Disease. *Nature* **2007**, *447* (7147), 932–940. https://doi.org/10.1038/nature05977.

(132) Willems, T.; Zielinski, D.; Yuan, J.; Gordon, A.; Gymrek, M.; Erlich, Y. Genome-Wide Profiling of Heritable and de Novo STR Variations. *Nat. Methods* **2017**, *14* (6), 590–592. https://doi.org/10.1038/nmeth.4267.

(133) 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.; Kristmundsdottir, S.; Sigurpalsdottir, B. D.; Stefansson, O. A.; Beyter, D.; Holley, G.; Tragante, V.; Gylfason, A.; Olason, P. I.; Zink, F.; Asgeirsdottir, M.; Sverrisson, S. T.; Sigurdsson, B.; Gudjonsson, S. A.; Sigurdsson, G. T.; Halldorsson, G. H.; Sveinbjornsson, G.; Norland, K.; Styrkarsdottir, U.; Magnusdottir, D. N.; Snorradottir, S.; Kristinsson, K.; Sobech, E.; Jonsson, H.; Geirsson, A. J.; Olafsson, I.; Jonsson, P.; Pedersen, O. B.; Erikstrup, C.; Brunak, S.; Ostrowski, S. R.; Thorleifsson, G.; Jonsson, F.; Melsted, P.; Jonsdottir, I.; Rafnar, T.; Holm, H.; Stefansson, H.; Saemundsdottir, J.; Gudbjartsson, D. F.; Magnusson, O. T.; Masson, G.; Thorsteinsdottir, U.; Helgason, A.; Jonsson, H.; Sulem, P.; Stefansson, K. The Sequences of 150,119 Genomes in the UK Biobank. *Nature* **2022**, *607* (7920), 732–740. https://doi.org/10.1038/s41586-022-04965-x.

(134) Avvaru, A. K.; Sharma, D.; Verma, A.; Mishra, R. K.; Sowpati, D. T. MSDB: A Comprehensive, Annotated Database of Microsatellites. *Nucleic Acids Res.* **2020**, *48* (D1), D155–D159. https://doi.org/10.1093/nar/gkz886.

(135) Mitra, I.; Huang, B.; Mousavi, N.; Ma, N.; Lamkin, M.; Yanicky, R.; Shleizer-Burko, S.; Lohmueller, K. E.; Gymrek, M. Patterns of de Novo Tandem Repeat Mutations and Their Role in Autism. *Nature* **2021**, *589* (7841), 246–250. https://doi.org/10.1038/s41586-020-03078-7.

(136) Ségurel, L.; Wyman, M. J.; Przeworski, M. Determinants of Mutation Rate Variation in the Human Germline. *Annu. Rev. Genomics Hum. Genet.* **2014**, *15* (Volume 15, 2014), 47–70. https://doi.org/10.1146/annurev-genom-031714-125740.

(137) Besenbacher, S.; Liu, S.; Izarzugaza, J. M. G.; Grove, J.; Belling, K.; Bork-Jensen, J.; Huang, S.; Als, T. D.; Li, S.; Yadav, R.; Rubio-García, A.; Lescai, F.; Demontis, D.; Rao, J.; Ye, W.; Mailund, T.; Friborg, R. M.; Pedersen, C. N. S.; Xu, R.; Sun, J.; Liu, H.; Wang, O.; Cheng, X.; Flores, D.; Rydza, E.; Rapacki, K.; Damm Sørensen, J.; Chmura, P.; Westergaard, D.; Dworzynski, P.; Sørensen, T. I. A.; Lund, O.; Hansen, T.; Xu, X.; Li, N.; Bolund, L.; Pedersen, O.; Eiberg, H.; Krogh, A.; Børglum, A. D.; Brunak, S.; Kristiansen, K.; Schierup, M. H.; Wang, J.; Gupta, R.; Villesen, P.; Rasmussen, S. Novel Variation and de Novo Mutation Rates in Population-Wide de Novo Assembled Danish Trios. *Nat. Commun.* **2015**, *6* (1), 5969. https://doi.org/10.1038/ncomms6969.

(138) Aponte, R. A.; Gettings, K. B.; Duewer, D. L.; Coble, M. D.; Vallone, P. M. Sequence-Based Analysis of Stutter at STR Loci: Characterization and Utility. *Forensic Sci. Int. Genet. Suppl. Ser.* **2015**, *5*, e456–e458. https://doi.org/10.1016/j.fsigss.2015.09.181.

(139) Gymrek, M. *PCR-Free Library Preparation Greatly Reduces Stutter Noise at Short Tandem Repeats*; preprint; Bioinformatics, 2016. https://doi.org/10.1101/043448.

(140) Mousavi, N.; Shleizer-Burko, S.; Yanicky, R.; Gymrek, M. Profiling the Genome-Wide Landscape of Tandem Repeat Expansions. *Nucleic Acids Res.* **2019**, *47* (15), e90. https://doi.org/10.1093/nar/gkz501.

(141) Gymrek, M.; Golan, D.; Rosset, S.; Erlich, Y. lobSTR: A Short Tandem Repeat Profiler for Personal Genomes. *Genome Res.* **2012**, *22* (6), 1154–1162. https://doi.org/10.1101/gr.135780.111.

(142) Kristmundsdóttir, S.; Sigurpálsdóttir, B. D.; Kehr, B.; Halldórsson, B. V. popSTR: Population-Scale Detection of STR Variants. *Bioinformatics* **2017**, *33* (24), 4041–4048. https://doi.org/10.1093/bioinformatics/btw568.

(143) Tankard, R. M.; Bennett, M. F.; Degorski, P.; Delatycki, M. B.; Lockhart, P. J.; Bahlo, M. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *Am. J. Hum. Genet.* **2018**, *103* (6), 858–873. https://doi.org/10.1016/j.ajhg.2018.10.015.

(144) Tang, H.; Kirkness, E. F.; Lippert, C.; Biggs, W. H.; Fabani, M.; Guzman, E.; Ramakrishnan, S.; Lavrenko, V.; Kakaradov, B.; Hou, C.; Hicks, B.; Heckerman, D.; Och, F. J.; Caskey, C. T.; Venter, J. C.; Telenti, A. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* **2017**, *101* (5), 700–715. https://doi.org/10.1016/j.ajhg.2017.09.013.

(145) Dashnow, H.; Lek, M.; Phipson, B.; Halman, A.; Sadedin, S.; Lonsdale, A.; Davis, M.; Lamont, P.; Clayton, J. S.; Laing, N. G.; MacArthur, D. G.; Oshlack, A. STRetch: Detecting and Discovering Pathogenic Short Tandem Repeat Expansions. *Genome Biol.* **2018**, *19* (1), 1–13. https://doi.org/10.1186/s13059-018-1505-2.

(146) Dolzhenko, E.; Deshpande, V.; Schlesinger, F.; Krusche, P.; Petrovski, R.; Chen, S.; Emig-Agius, D.; Gross, A.; Narzisi, G.; Bowman, B.; Scheffler, K.; van Vugt, J. J. F. A.; French, C.; Sanchis-Juan, A.; Ibáñez, K.; Tucci, A.; Lajoie, B. R.; Veldink, J. H.; Raymond, F. L.; Taft, R. J.; Bentley, D. R.; Eberle, M. A. ExpansionHunter: A Sequence-Graph-Based Tool to Analyze Variation in Short Tandem Repeat Regions. *Bioinformatics* **2019**, *35* (22), 4754–4756. https://doi.org/10.1093/bioinformatics/btz431.

(147) Bakhtiari, M.; Shleizer-Burko, S.; Gymrek, M.; Bansal, V.; Bafna, V. Targeted Genotyping of Variable Number Tandem Repeats with adVNTR. *Genome Res.* **2018**, *28* (11), 1709–1719. https://doi.org/10.1101/gr.235119.118.

(148) Mukamel, R. E.; Handsaker, R. E.; Sherman, M. A.; Barton, A. R.; Zheng, Y.; McCarroll, S. A.; Loh, P.-R. Protein-Coding Repeat Polymorphisms Strongly Shape Diverse Human Phenotypes. *Science* **2021**, *373* (6562), 1499–1505. https://doi.org/10.1126/science.abg8289.

(149) Tw, H.; Jd, G.; Ce, Y.; Gr, C. A Variable Dinucleotide Repeat in the CFTR Gene Contributes to Phenotype Diversity by Forming RNA Secondary Structures That Alter Splicing. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (10). https://doi.org/10.1073/pnas.0400182101.

(150) Hui, J.; Stangl, K.; Lane, W. S.; Bindereif, A. HnRNP L Stimulates Splicing of the eNOS Gene by Binding to Variable-Length CA Repeats. *Nat. Struct. Biol.* **2003**, *10* (1), 33–37. https://doi.org/10.1038/nsb875.

(151) Afek, A.; Schipper, J. L.; Horton, J.; Gordân, R.; Lukatsky, D. B. Protein−DNA Binding in the Absence of Specific Base-Pair Recognition. *Proc. Natl. Acad. Sci.* **2014**, *111* (48), 17140–17145. https://doi.org/10.1073/pnas.1410569111.

(152) Tsuge, M.; Hamamoto, R.; Silva, F. P.; Ohnishi, Y.; Chayama, K.; Kamatani, N.; Furukawa, Y.; Nakamura, Y. A Variable Number of Tandem Repeats Polymorphism in an E2F-1 Binding Element in the 5′ Flanking Region of SMYD3 Is a Risk Factor for Human Cancers. *Nat. Genet.* **2005**, *37* (10), 1104–1107. https://doi.org/10.1038/ng1638.

(153) Raveh-Sadka, T.; Levo, M.; Shabi, U.; Shany, B.; Keren, L.; Lotan-Pompan, M.; Zeevi, D.; Sharon, E.; Weinberger, A.; Segal, E. Manipulating Nucleosome Disfavoring Sequences Allows Fine-Tune Regulation of Gene Expression in Yeast. *Nat. Genet.* **2012**, *44* (7), 743–750. https://doi.org/10.1038/ng.2305.

(154) Quilez, J.; Guilmatre, A.; Garg, P.; Highnam, G.; Gymrek, M.; Erlich, Y.; Joshi, R. S.; Mittelman, D.; Sharp, A. J. Polymorphic Tandem Repeats within Gene Promoters Act as Modifiers of Gene Expression and DNA Methylation in Humans. *Nucleic Acids Res.* **2016**, *44* (8), 3750–3762. https://doi.org/10.1093/nar/gkw219.

(155) Rothenburg, S.; Koch-Nolte, F.; Rich, A.; Haag, F. A Polymorphic Dinucleotide Repeat in the Rat Nucleolin Gene Forms Z-DNA and Inhibits Promoter Activity. *Proc. Natl. Acad. Sci.* **2001**, *98* (16), 8985–8990. https://doi.org/10.1073/pnas.121176998.

(156) Murat, P.; Guilbaud, G.; Sale, J. E. DNA Polymerase Stalling at Structured DNA Constrains the Expansion of Short Tandem Repeats. *Genome Biol.* **2020**, *21*, 209. https://doi.org/10.1186/s13059-020-02124-x.

(157) Reddy, K.; Schmidt, M. H. M.; Geist, J. M.; Thakkar, N. P.; Panigrahi, G. B.; Wang, Y.-H.; Pearson, C. E. Processing of Double-R-Loops in (CAG)·(CTG) and C9orf72 (GGGGCC)·(GGCCCC) Repeats Causes Instability. *Nucleic Acids Res.* **2014**, *42* (16), 10473–10487. https://doi.org/10.1093/nar/gku658.

(158) Martin-Trujillo, A.; Garg, P.; Patel, N.; Jadhav, B.; Sharp, A. J. Genome-Wide Evaluation of the Effect of Short Tandem Repeat Variation on Local DNA Methylation. *Genome Res.* **2023**, *33* (2), 184–196. https://doi.org/10.1101/gr.277057.122.

(159) Liu, X. S.; Wu, H.; Krzisch, M.; Wu, X.; Graef, J.; Muffat, J.; Hnisz, D.; Li, C. H.; Yuan, B.; Xu, C.; Li, Y.; Vershkov, D.; Cacace, A.; Young, R. A.; Jaenisch, R. Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. *Cell* **2018**, *172* (5), 979-992.e6. https://doi.org/10.1016/j.cell.2018.01.012.

(160) *association tests with multi-allelic variants from VCF*. https://groups.google.com/g/plink2-users/c/3N2JZihdmgI (accessed 2024-05-01).

(161) *Genetic association tests using SAIGE*. GitHub. https://github.com/weizhouUMICH/SAIGE/wiki/Genetic-association-tests-using-SAIGE (accessed 2024-05-01).

(162) *file format · Issue #15 · rgcgithub/regenie*. GitHub. https://github.com/rgcgithub/regenie/issues/15 (accessed 2024-05-01).

(163) Gymrek, M.; Willems, T.; Guilmatre, A.; Zeng, H.; Markus, B.; Georgiev, S.; Daly, M. J.; Price, A. L.; Pritchard, J. K.; Sharp, A. J.; Erlich, Y. Abundant Contribution of Short Tandem Repeats to Gene Expression Variation in Humans. *Nat. Genet.* **2016**, *48* (1), 22–29. https://doi.org/10.1038/ng.3461.

(164) Wang, Q. S.; Huang, H. Methods for Statistical Fine-Mapping and Their Applications to Auto-Immune Diseases. *Semin. Immunopathol.* **2022**, *44* (1), 101–113. https://doi.org/10.1007/s00281-021-00902-8.

(165) Hormozdiari, F.; Kostem, E.; Kang, E. Y.; Pasaniuc, B.; Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **2014**, *198* (2), 497–508. https://doi.org/10.1534/genetics.114.167908.

(166) Wakefield, J. Bayes Factors for Genome-Wide Association Studies: Comparison with P-Values. *Genet. Epidemiol.* **2009**, *33* (1), 79–86. https://doi.org/10.1002/gepi.20359.

(167) Maller, J. B.; McVean, G.; Byrnes, J.; Vukcevic, D.; Palin, K.; Su, Z.; Howson, J. M. M.; Auton, A.; Myers, S.; Morris, A.; Pirinen, M.; Brown, M. A.; Burton, P. R.; Caulfield, M. J.; Compston, A.; Farrall, M.; Hall, A. S.; Hattersley, A. T.; Hill, A. V. S.; Mathew, C. G.; Pembrey, M.; Satsangi, J.; Stratton, M. R.; Worthington, J.; Craddock, N.; Hurles, M.; Ouwehand, W.; Parkes, M.; Rahman, N.; Duncanson, A.; Todd, J. A.; Kwiatkowski, D. P.; Samani, N. J.; Gough, S. C. L.; McCarthy, M. I.; Deloukas, P.; Donnelly, P. Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases. *Nat. Genet.* **2012**, *44* (12), 1294–1301. https://doi.org/10.1038/ng.2435.

(168) Cui, R.; Elzur, R. A.; Kanai, M.; Ulirsch, J. C.; Weissbrod, O.; Daly, M. J.; Neale, B. M.; Fan, Z.; Finucane, H. K. Improving Fine-Mapping by Modeling Infinitesimal Effects. *Nat. Genet.* **2024**, *56* (1), 162–169. https://doi.org/10.1038/s41588-023-01597-3.

(169) Trynka, G.; Hunt, K. A.; Bockett, N. A.; Romanos, J.; Mistry, V.; Szperl, A.; Bakker, S. F.; Bardella, M. T.; Bhaw-Rosun, L.; Castillejo, G.; de la Concha, E. G.; de Almeida, R. C.; Dias, K.-R. M.; van Diemen, C. C.; Dubois, P. C. A.; Duerr, R. H.; Edkins, S.; Franke, L.; Fransen, K.; Gutierrez, J.; Heap, G. A. R.; Hrdlickova, B.; Hunt, S.; Izurieta, L. P.; Izzo, V.; Joosten, L. A. B.; Langford, C.; Mazzilli, M. C.; Mein, C. A.; Midah, V.; Mitrovic, M.; Mora, B.; Morelli, M.; Nutland, S.; Núñez, C.; Onengut-Gumuscu, S.; Pearce, K.; Platteel, M.; Polanco, I.; Potter, S.; Ribes-Koninckx, C.; Ricaño-Ponce, I.; Rich, S. S.; Rybak, A.; Santiago, J. L.; Senapati, S.; Sood, A.; Szajewska, H.; Troncone, R.; Varadé, J.; Wallace, C.; Wolters, V. M.; Zhernakova, A.; Thelma, B. K.; Cukrowska, B.; Urcelay, E.; Bilbao, J. R.; Mearin, M. L.; Barisani, D.; Barrett, J. C.; Plagnol, V.; Deloukas, P.; Wijmenga, C.; van Heel, D. A. Dense Genotyping Identifies and Localizes Multiple Common and Rare Variant Association Signals in Celiac Disease. *Nat. Genet.* **2011**, *43* (12), 1193–1201. https://doi.org/10.1038/ng.998.

(170) Ripke, S.; Sanders, A. R.; Kendler, K. S.; Levinson, D. F.; Sklar, P.; Holmans, P. A.; Lin, D.-Y.; Duan, J.; Ophoff, R. A.; Andreassen, O. A.; Scolnick, E.; Cichon, S.; St. Clair, D.; Corvin, A.; Gurling, H.; Werge, T.; Rujescu, D.; Blackwood, D. H. R.; Pato, C. N.; Malhotra, A. K.; Purcell, S.; Dudbridge, F.; Neale, B. M.; Rossin, L.; Visscher, P. M.; Posthuma, D.; Ruderfer, D. M.; Fanous, A.; Stefansson, H.; Steinberg, S.; Mowry, B. J.; Golimbet, V.; De Hert, M.; Jönsson, E. G.; Bitter, I.; Pietiläinen, O. P. H.; Collier, D. A.; Tosato, S.; Agartz, I.; Albus, M.; Alexander, M.; Amdur, R. L.; Amin, F.; Bass, N.; Bergen, S. E.; Black, D. W.; Børglum, A. D.; Brown, M. A.; Bruggeman, R.; Buccola, N. G.; Byerley, W. F.; Cahn, W.; Cantor, R. M.; Carr, V. J.; Catts, S. V.; Choudhury, K.; Cloninger, C. R.; Cormican, P.; Craddock, N.; Danoy, P. A.; Datta, S.; de Haan, L.; Demontis, D.; Dikeos, D.; Djurovic, S.; Donnelly, P.; Donohoe, G.; Duong, L.; Dwyer, S.; Fink-Jensen, A.; Freedman, R.; Freimer, N. B.; Friedl, M.; Georgieva, L.; Giegling, I.; Gill, M.; Glenthøj, B.; Godard, S.; Hamshere, M.; Hansen, M.; Hansen, T.; Hartmann, A. M.; Henskens, F. A.; Hougaard, D. M.; Hultman, C. M.; Ingason, A.; Jablensky, A. V.; Jakobsen, K. D.; Jay, M.; Jürgens, G.; Kahn, R. S.; Keller, M. C.; Kenis, G.; Kenny, E.; Kim, Y.; Kirov, G. K.; Konnerth, H.; Konte, B.; Krabbendam, L.; Krasucki, R.; Lasseter, V. K.; Laurent, C.; Lawrence, J.; Lencz, T.; Lerer, F. B.; Liang, K.-Y.; Lichtenstein, P.; Lieberman, J. A.; Linszen, D. H.; Lönnqvist, J.; Loughland, C. M.; Maclean, A. W.; Maher, B. S.; Maier, W.; Mallet, J.; Malloy, P.; Mattheisen, M.; Mattingsdal, M.; McGhee, K. A.; McGrath, J. J.; McIntosh, A.; McLean, D. E.; McQuillin, A.; Melle, I.; Michie, P. T.; Milanova, V.; Morris, D. W.; Mors, O.; Mortensen, P. B.; Moskvina, V.; Muglia, P.; Myin-Germeys, I.; Nertney, D. A.; Nestadt, G.; Nielsen, J.; Nikolov, I.; Nordentoft, M.; Norton, N.; Nöthen, M. M.; O'Dushlaine, C. T.; Olincy, A.; Olsen, L.; O'Neill, F. A.; Ørntoft, T. F.; Owen, M. J.; Pantelis, C.; Papadimitriou, G.; Pato, M. T.; Peltonen, L.; Petursson, H.; Pickard, B.; Pimm, J.; Pulver, A. E.; Puri, V.; Quested, D.; Quinn, E. M.; Rasmussen, H. B.; Réthelyi, J. M.; Ribble, R.; Rietschel, M.; Riley, B. P.; Ruggeri, M.; Schall, U.; Schulze, T. G.; Schwab, S. G.; Scott, R. J.; Shi, J.; Sigurdsson, E.; Silverman, J. M.; Spencer, C. C. A.; Stefansson, K.; Strange, A.; Strengman, E.; Stroup, T. S.; Suvisaari, J.; Terenius, L.; Thirumalai, S.; Thygesen, J. H.; Timm, S.; Toncheva, D.; van den Oord, E.; van Os, J.; van Winkel, R.; Veldink, J.; Walsh, D.; Wang, A. G.; Wiersma, D.; Wildenauer, D. B.; Williams, H. J.; Williams, N. M.; Wormley, B.; Zammit, S.; Sullivan, P. F.; O'Donovan, M. C.; Daly, M. J.; Gejman, P. V.; The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-Wide Association Study Identifies Five New Schizophrenia Loci. *Nat. Genet.* **2011**, *43* (10), 969–976. https://doi.org/10.1038/ng.940.

(171) Sklar, P.; Ripke, S.; Scott, L. J.; Andreassen, O. A.; Cichon, S.; Craddock, N.; Edenberg, H. J.; Nurnberger, J. I.; Rietschel, M.; Blackwood, D.; Corvin, A.; Flickinger, M.; Guan, W.; Mattingsdal, M.; McQuillin, A.; Kwan, P.; Wienker, T. F.; Daly, M.; Dudbridge, F.; Holmans, P. A.; Lin, D.; Burmeister, M.; Greenwood, T. A.; Hamshere, M. L.; Muglia, P.; Smith, E. N.;

Zandi, P. P.; Nievergelt, C. M.; McKinney, R.; Shilling, P. D.; Schork, N. J.; Bloss, C. S.; Foroud, T.; Koller, D. L.; Gershon, E. S.; Liu, C.; Badner, J. A.; Scheftner, W. A.; Lawson, W. B.; Nwulia, E. A.; Hipolito, M.; Coryell, W.; Rice, J.; Byerley, W.; McMahon, F. J.; Schulze, T. G.; Berrettini, W.; Lohoff, F. W.; Potash, J. B.; Mahon, P. B.; McInnis, M. G.; Zöllner, S.; Zhang, P.; Craig, D. W.; Szelinger, S.; Barrett, T. B.; Breuer, R.; Meier, S.; Strohmaier, J.; Witt, S. H.; Tozzi, F.; Farmer, A.; McGuffin, P.; Strauss, J.; Xu, W.; Kennedy, J. L.; Vincent, J. B.; Matthews, K.; Day, R.; Ferreira, M. A.; O'Dushlaine, C.; Perlis, R.; Raychaudhuri, S.; Ruderfer, D.; Lee, P. H.; Smoller, J. W.; Li, J.; Absher, D.; Bunney, W. E.; Barchas, J. D.; Schatzberg, A. F.; Jones, E. G.; Meng, F.; Thompson, R. C.; Watson, S. J.; Myers, R. M.; Akil, H.; Boehnke, M.; Chambert, K.; Moran, J.; Scolnick, E.; Djurovic, S.; Melle, I.; Morken, G.; Gill, M.; Morris, D.; Quinn, E.; Mühleisen, T. W.; Degenhardt, F. A.; Mattheisen, M.; Schumacher, J.; Maier, W.; Steffens, M.; Propping, P.; Nöthen, M. M.; Anjorin, A.; Bass, N.; Gurling, H.; Kandaswamy, R.; Lawrence, J.; McGhee, K.; McIntosh, A.; McLean, A. W.; Muir, W. J.; Pickard, B. S.; Breen, G.; St. Clair, D.; Caesar, S.; Gordon-Smith, K.; Jones, L.; Fraser, C.; Green, E. K.; Grozeva, D.; Jones, I. R.; Kirov, G.; Moskvina, V.; Nikolov, I.; O'Donovan, M. C.; Owen, M. J.; Collier, D. A.; Elkin, A.; Williamson, R.; Young, A. H.; Ferrier, I. N.; Stefansson, K.; Stefansson, H.; Þorgeirsson, Þ.; Steinberg, S.; Gustafsson, Ó.; Bergen, S. E.; Nimgaonkar, V.; Hultman, C.; Landén, M.; Lichtenstein, P.; Sullivan, P.; Schalling, M.; Osby, U.; Backlund, L.; Frisén, L.; Langstrom, N.; Jamain, S.; Leboyer, M.; Etain, B.; Bellivier, F.; Petursson, H.; Sigur∂sson, E.; Müller-Mysok, B.; Lucae, S.; Schwarz, M.; Fullerton, J. M.; Schofield, P. R.; Martin, N.; Montgomery, G. W.; Lathrop, M.; Óskarsson, H.; Bauer, M.; Wright, A.; Mitchell, P. B.; Hautzinger, M.; Reif, A.; Kelsoe, J. R.; Purcell, S. M.; Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-Scale Genome-Wide Association Analysis of Bipolar Disorder Identifies a New Susceptibility Locus near ODZ4. *Nat. Genet.* **2011**, *43* (10), 977–983. https://doi.org/10.1038/ng.943.

(172) Yang, J.; Ferreira, T.; Morris, A. P.; Medland, S. E.; Madden, P. A. F.; Heath, A. C.; Martin, N. G.; Montgomery, G. W.; Weedon, M. N.; Loos, R. J.; Frayling, T. M.; McCarthy, M. I.; Hirschhorn, J. N.; Goddard, M. E.; Visscher, P. M. Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits. *Nat. Genet.* **2012**, *44* (4), 369–375. https://doi.org/10.1038/ng.2213.

(173) Newcombe, P. J.; Conti, D. V.; Richardson, S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet. Epidemiol.* **2016**, *40* (3), 188–201. https://doi.org/10.1002/gepi.21953.

(174) Wang, G.; Sarkar, A.; Carbonetto, P.; Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2020**, *82* (5), 1273–1300. https://doi.org/10.1111/rssb.12388.

(175) Chen, W.; Larrabee, B. R.; Ovsyannikova, I. G.; Kennedy, R. B.; Haralambieva, I. H.; Poland, G. A.; Schaid, D. J. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **2015**, *200* (3), 719–736. https://doi.org/10.1534/genetics.115.176107.

(176) Benner, C.; Spencer, C. C. A.; Havulinna, A. S.; Salomaa, V.; Ripatti, S.; Pirinen, M. FINEMAP: Efficient Variable Selection Using Summary Data from Genome-Wide Association Studies. *Bioinformatics* **2016**, *32* (10), 1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

(177) Zou, Y.; Carbonetto, P.; Wang, G.; Stephens, M. Fine-Mapping from Summary Data with the "Sum of Single Effects" Model. *PLOS Genet.* **2022**, *18* (7), e1010299. https://doi.org/10.1371/journal.pgen.1010299.

(178) Kichaev, G.; Roytman, M.; Johnson, R.; Eskin, E.; Lindström, S.; Kraft, P.; Pasaniuc, B. Improved Methods for Multi-Trait Fine Mapping of Pleiotropic Risk Loci. *Bioinformatics* **2017**, *33* (2), 248–255. https://doi.org/10.1093/bioinformatics/btw615.

(179) Hernández, N.; Soenksen, J.; Newcombe, P.; Sandhu, M.; Barroso, I.; Wallace, C.; Asimit, J. L. The Flashfm Approach for Fine-Mapping Multiple Quantitative Traits. *Nat. Commun.* **2021**, *12* (1), 6147. https://doi.org/10.1038/s41467-021-26364-y.

(180) Kanai, M.; Elzur, R.; Zhou, W.; Kanai, M.; Wu, K.-H. H.; Rasheed, H.; Tsuo, K.; Hirbo, J. B.; Wang, Y.; Bhattacharya, A.; Zhao, H.; Namba, S.; Surakka, I.; Wolford, B. N.; Faro, V. L.; Lopera-Maya, E. A.; Läll, K.; Favé, M.-J.; Partanen, J. J.; Chapman, S. B.; Karjalainen, J.; Kurki, M.; Maasha, M.; Brumpton, B. M.; Chavan, S.; Chen, T.-T.; Daya, M.; Ding, Y.; Feng, Y.-C. A.; Guare, L. A.; Gignoux, C. R.; Graham, S. E.; Hornsby, W. E.; Ingold, N.; Ismail, S. I.; Johnson, R.; Laisk, T.; Lin, K.; Lv, J.; Millwood, I. Y.; Moreno-Grau, S.; Nam, K.; Palta, P.; Pandit, A.; Preuss, M. H.; Saad, C.; Setia-Verma, S.; Thorsteinsdottir, U.; Uzunovic, J.; Verma, A.; Zawistowski, M.; Zhong, X.; Afifi, N.; Al-Dabhani, K. M.; Thani, A. A.; Bradford, Y.; Campbell, A.; Crooks, K.; Bock, G. H. de; Damrauer, S. M.; Douville, N. J.; Finer, S.; Fritsche, L. G.; Fthenou, E.; Gonzalez-Arroyo, G.; Griffiths, C. J.; Guo, Y.; Hunt, K. A.; Ioannidis, A.; Jansonius, N. M.; Konuma, T.; Lee, M. T. M.; Lopez-Pineda, A.; Matsuda, Y.; Marioni, R. E.; Moatamed, B.; Nava-Aguilar, M. A.; Numakura, K.; Patil, S.; Rafaels, N.; Richmond, A.; Rojas-Muñoz, A.; Shortt, J. A.; Straub, P.; Tao, R.; Vanderwerff, B.; Vernekar, M.; Veturi, Y.; Barnes, K. C.; Boezen, M.; Chen, Z.; Chen, C.-Y.; Cho, J.; Smith, G. D.; Finucane, H. K.; Franke, L.; Gamazon, E. R.; Ganna, A.; Gaunt, T. R.; Ge, T.; Huang, H.; Huffman, J.; Katsanis, N.; Koskela, J. T.; Lajonchere, C.; Law, M. H.; Li, L.; Lindgren, C. M.; Loos, R. J. F.; MacGregor, S.; Matsuda, K.; Olsen, C. M.; Porteous, D. J.; Shavit, J. A.; Snieder, H.; Takano, T.; Trembath, R. C.; Vonk, J. M.; Whiteman, D. C.; Wicks, S. J.; Wijmenga, C.; Wright, J.; Zheng, J.; Zhou, X.; Awadalla, P.; Boehnke, M.; Bustamante, C. D.; Cox, N. J.; Fatumo, S.; Geschwind, D. H.; Hayward, C.; Hveem, K.; Kenny, E. E.; Lee, S.; Lin, Y.-F.; Mbarek, H.; Mägi, R.; Martin, H. C.; Medland, S. E.; Okada, Y.; Palotie, A. V.; Pasaniuc, B.; Rader, D. J.; Ritchie, M. D.; Sanna, S.; Smoller, J. W.; Stefansson, K.; Heel, D. A. van; Walters, R. G.; Zöllner, S.; Americas, B. of the; Project, B. J.; BioMe; BioVU; Study, C.-O. H.; Group, C. K. B. C.; Medicine, C. C. for P.; Genetics, deCODE; Estonian Biobank, F.; Scotland, G.; Team, G. & H. R.; LifeLines; Biobank, M. G. B.; Initiative, M. G.; Korea, N. B. of; BioBank, P. M.; Biobank, Q.; Study, T. Q. S. and H.; Biobank, T.; Study, T. H.; Initiative, U. A. C. H.; Resource, U. G.; Biobank, U.; Martin, A. R.; Willer, C. J.; Daly, M. J.; Neale, B. M.; Daly, M. J.; Finucane, H. K. Meta-Analysis Fine-Mapping Is Often Miscalibrated at Single-Variant Resolution. *Cell Genomics* **2022**, *2* (12). https://doi.org/10.1016/j.xgen.2022.100210.

(181) Zhang, W.; Najafabadi, H.; Li, Y. SparsePro: An Efficient Fine-Mapping Method Integrating Summary Statistics and Functional Annotations. *PLOS Genet.* **2023**, *19* (12), e1011104. https://doi.org/10.1371/journal.pgen.1011104.

(182) Yang, Z.; Wang, C.; Liu, L.; Khan, A.; Lee, A.; Vardarajan, B.; Mayeux, R.; Kiryluk, K.; Ionita-Laza, I. CARMA Is a New Bayesian Model for Fine-Mapping in Genome-Wide Association Meta-Analyses. *Nat. Genet.* **2023**, *55* (6), 1057–1065. https://doi.org/10.1038/s41588-023-01392-0.

(183) Wang, Q. S.; Kelley, D. R.; Ulirsch, J.; Kanai, M.; Sadhuka, S.; Cui, R.; Albors, C.; Cheng, N.; Okada, Y.; Aguet, F.; Ardlie, K. G.; MacArthur, D. G.; Finucane, H. K. Leveraging Supervised Learning for Functionally Informed Fine-Mapping of Cis-eQTLs Identifies an Additional 20,913 Putative Causal eQTLs. *Nat. Commun.* **2021**, *12* (1), 3394. https://doi.org/10.1038/s41467-021-23134-8.

(184) Weissbrod, O.; Hormozdiari, F.; Benner, C.; Cui, R.; Ulirsch, J.; Gazal, S.; Schoech, A. P.; van de Geijn, B.; Reshef, Y.; Márquez-Luna, C.; O'Connor, L.; Pirinen, M.; Finucane, H. K.; Price, A. L. Functionally Informed Fine-Mapping and Polygenic Localization of Complex Trait Heritability. *Nat. Genet.* **2020**, *52* (12), 1355–1363. https://doi.org/10.1038/s41588-020-00735-5.

(185) Jiang, J.; Cole, J. B.; Freebern, E.; Da, Y.; VanRaden, P. M.; Ma, L. Functional Annotation and Bayesian Fine-Mapping Reveals Candidate Genes for Important Agronomic Traits in Holstein Bulls. *Commun. Biol.* **2019**, *2* (1), 1–12. https://doi.org/10.1038/s42003-019-0454-y.

(186) Kichaev, G.; Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-Ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **2015**, *97* (2), 260–271. https://doi.org/10.1016/j.ajhg.2015.06.007.

(187) Kichaev, G.; Yang, W.-Y.; Lindstrom, S.; Hormozdiari, F.; Eskin, E.; Price, A. L.; Kraft, P.; Pasaniuc, B. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genet.* **2014**, *10* (10), e1004722. https://doi.org/10.1371/journal.pgen.1004722.

(188) Wallace, C. A More Accurate Method for Colocalisation Analysis Allowing for Multiple Causal Variants. *PLOS Genet.* **2021**, *17* (9), e1009440. https://doi.org/10.1371/journal.pgen.1009440.

(189) Zhang, X.; Jiang, W.; Zhao, H. Integration of Expression QTLs with Fine Mapping via SuSiE. *PLOS Genet.* **2024**, *20* (1), e1010929. https://doi.org/10.1371/journal.pgen.1010929.

(190) Zou, Y.; Carbonetto, P.; Xie, D.; Wang, G.; Stephens, M. Fast and Flexible Joint Fine-Mapping of Multiple Traits via the Sum of Single Effects Model. bioRxiv April 18, 2024, p 2023.04.14.536893. https://doi.org/10.1101/2023.04.14.536893.

(191) Arvanitis, M.; Tayeb, K.; Strober, B. J.; Battle, A. Redefining Tissue Specificity of Genetic Regulation of Gene Expression in the Presence of Allelic Heterogeneity. *Am. J. Hum. Genet.* **2022**, *109* (2), 223–239. https://doi.org/10.1016/j.ajhg.2022.01.002.

(192) *About The Developmental Genotype-Tissue Expression (dGTEx) Project*. https://www.gtexportal.org/home/aboutdGTEx (accessed 2024-05-22).

(193) Urbut, S. M.; Wang, G.; Carbonetto, P.; Stephens, M. Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions. *Nat. Genet.* **2019**, *51* (1), 187–195. https://doi.org/10.1038/s41588-018-0268-8.

(194) Zou, Y.; Carbonetto, P.; Xie, D.; Wang, G.; Stephens, M. Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions. bioRxiv February 11, 2024, p 2023.04.14.536893. https://doi.org/10.1101/2023.04.14.536893.

(195) Shi, Y.; Niu, Y.; Zhang, P.; Luo, H.; Liu, S.; Zhang, S.; Wang, J.; Li, Y.; Liu, X.; Song, T.; Xu, T.; He, S. Characterization of Genome-Wide STR Variation in 6487 Human Genomes. *Nat. Commun.* **2023**, *14* (1), 2092. https://doi.org/10.1038/s41467-023-37690-8.

(196) Cui, Y.; Ye, W.; Li, J. S.; Li, J. J.; Vilain, E.; Sallam, T.; Li, W. A Genome-Wide Spectrum of Tandem Repeat Expansions in 338,963 Humans. *Cell* **2024**, *0* (0). https://doi.org/10.1016/j.cell.2024.03.004.

(197) Bustos, B. I.; Billingsley, K.; Blauwendraat, C.; Gibbs, J. R.; Gan-Or, Z.; Krainc, D.; Singleton, A. B.; Lubbe, S. J.; International Parkinson's Disease Genomics Consortium (IPDGC). Genome-Wide Contribution of Common Short-Tandem Repeats to Parkinson's Disease Genetic Risk. *Brain* **2023**, *146* (1), 65–74. https://doi.org/10.1093/brain/awac301.

(198) Verbiest, M. A.; Lundström, O.; Xia, F.; Baudis, M.; Bilgin Sonay, T.; Anisimova, M. Short Tandem Repeat Mutations Regulate Gene Expression in Colorectal Cancer. *Sci. Rep.* **2024**, *14* (1), 3331. https://doi.org/10.1038/s41598-024-53739-0.

(199) Dolzhenko, E.; English, A.; Dashnow, H.; De Sena Brandine, G.; Mokveld, T.; Rowell, W. J.; Karniski, C.; Kronenberg, Z.; Danzi, M. C.; Cheung, W. A.; Bi, C.; Farrow, E.; Wenger, A.; Chua, K. P.; Martínez-Cerdeño, V.; Bartley, T. D.; Jin, P.; Nelson, D. L.; Zuchner, S.; Pastinen, T.; Quinlan, A. R.; Sedlazeck, F. J.; Eberle, M. A. Characterization and Visualization of Tandem Repeats at Genome Scale. *Nat. Biotechnol.* **2024**, 1–9. https://doi.org/10.1038/s41587-023-02057-3.

(200) Manigbas, C. A.; Jadhav, B.; Garg, P.; Shadrina, M.; Lee, W.; Martin-Trujillo, A.; Sharp, A. J. A Phenome-Wide Association Study of Tandem Repeat Variation in 168,554

Individuals from the UK Biobank. medRxiv January 23, 2024, p 2024.01.22.24301630. https://doi.org/10.1101/2024.01.22.24301630.

(201) Loh, P.-R. Uncovering Complex Trait Heritability Hidden in the Repeatome. *Cell Genomics* **2023**, *3* (12). https://doi.org/10.1016/j.xgen.2023.100461.

(202) 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.; Barann, M.; Wieland, T.; Greger, L.; van Iterson, M.; Almlöf, J.; Ribeca, P.; Pulyakhina, I.; Esser, D.; Giger, T.; Tikhonov, A.; Sultan, M.; Bertier, G.; MacArthur, D. G.; Lek, M.; Lizano, E.; Buermans, H. P. J.; Padioleau, I.; Schwarzmayr, T.; Karlberg, O.; Ongen, H.; Kilpinen, H.; Beltran, S.; Gut, M.; Kahlem, K.; Amstislavskiy, V.; Stegle, O.; Pirinen, M.; Montgomery, S. B.; Donnelly, P.; McCarthy, M. I.; Flicek, P.; Strom, T. M.; Lehrach, H.; Schreiber, S.; Sudbrak, R.; Carracedo, Á.; Antonarakis, S. E.; Häsler, R.; Syvänen, A.-C.; van Ommen, G.-J.; Brazma, A.; Meitinger, T.; Rosenstiel, P.; Guigó, R.; Gut, I. G.; Estivill, X.; Dermitzakis, E. T. Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans. *Nature* **2013**, *501* (7468), 506–511. https://doi.org/10.1038/nature12531.

(203) Bick, A. G.; Metcalf, G. A.; Mayo, K. R.; Lichtenstein, L.; Rura, S.; Carroll, R. J.; Musick, A.; Linder, J. E.; Jordan, I. K.; Nagar, S. D.; Sharma, S.; Meller, R.; Basford, M.; Boerwinkle, E.; Cicek, M. S.; Doheny, K. F.; Eichler, E. E.; Gabriel, S.; Gibbs, R. A.; Glazer, D.; Harris, P. A.; Jarvik, G. P.; Philippakis, A.; Rehm, H. L.; Roden, D. M.; Thibodeau, S. N.; Topper, S.; Blegen, A. L.; Wirkus, S. J.; Wagner, V. A.; Meyer, J. G.; Cicek, M. S.; Muzny, D. M.; Venner, E.; Mawhinney, M. Z.; Griffith, S. M. L.; Hsu, E.; Ling, H.; Adams, M. K.; Walker, K.; Hu, J.; Doddapaneni, H.; Kovar, C. L.; Murugan, M.; Dugan, S.; Khan, Z.; Boerwinkle, E.; Lennon, N. J.; Austin-Tse, C.; Banks, E.; Gatzen, M.; Gupta, N.; Henricks, E.; Larsson, K.; McDonough, S.; Harrison, S. M.; Kachulis, C.; Lebo, M. S.; Neben, C. L.; Steeves, M.; Zhou, A. Y.; Smith, J. D.; Frazar, C. D.; Davis, C. P.; Patterson, K. E.; Wheeler, M. M.; McGee, S.; Lockwood, C. M.; Shirts, B. H.; Pritchard, C. C.; Murray, M. L.; Vasta, V.; Leistritz, D.; Richardson, M. A.; Buchan, J. G.; Radhakrishnan, A.; Krumm, N.; Ehmen, B. W.; Schwartz, S.; Aster, M. M. T.; Cibulskis, K.; Haessly, A.; Asch, R.; Cremer, A.; Degatano, K.; Shergill, A.; Gauthier, L. D.; Lee, S. K.; Hatcher, A.; Grant, G. B.; Brandt, G. R.; Covarrubias, M.; Banks, E.; Able, A.; Green, A. E.; Carroll, R. J.; Zhang, J.; Condon, H. R.; Wang, Y.; Dillon, M. K.; Albach, C. H.; Baalawi, W.; Choi, S. H.; Wang, X.; Rosenthal, E. A.; Ramirez, A. H.; Lim, S.; Nambiar, S.; Ozenberger, B.; Wise, A. L.; Lunt, C.; Ginsburg, G. S.; Denny, J. C.; The All of Us Research Program Genomics Investigators; Manuscript Writing Group; All of Us Research Program Genomics Principal Investigators; Biobank, M.; Genome Center: Baylor-Hopkins Clinical Genome Center; Genome Center: Broad, C., and Mass General Brigham Laboratory for Molecular Medicine; Genome Center: University of Washington; Data and Research Center; All of Us Research Demonstration Project Teams; NIH All of Us Research Program Staff. Genomic Data in the All of Us Research Program. *Nature* **2024**, *627* (8003), 340–346. https://doi.org/10.1038/s41586-023-06957-x.

(204) Hunter-Zinck, H.; Shi, Y.; Li, M.; Gorman, B. R.; Ji, S.-G.; Sun, N.; Webster, T.; Liem, A.; Hsieh, P.; Devineni, P.; Karnam, P.; Gong, X.; Radhakrishnan, L.; Schmidt, J.; Assimes, T. L.; Huang, J.; Pan, C.; Humphries, D.; Brophy, M.; Moser, J.; Muralidhar, S.; Huang, G. D.; Przygodzki, R.; Concato, J.; Gaziano, J. M.; Gelernter, J.; O’Donnell, C. J.; Hauser, E. R.; Zhao, H.; O’Leary, T. J.; Tsao, P. S.; Pyarajan, S. Genotyping Array Design and Data Quality Control in the Million Veteran Program. *Am. J. Hum. Genet.* **2020**, *106* (4), 535–548. https://doi.org/10.1016/j.ajhg.2020.03.004.

(205) Kristmundsdottir, S.; Eggertsson, H. P.; Arnadottir, G. A.; Halldorsson, B. V. popSTR2 Enables Clinical and Population-Scale Genotyping of Microsatellites. *Bioinformatics* **2020**, *36* (7), 2269–2271. https://doi.org/10.1093/bioinformatics/btz913.

(206) Jam, H. Z.; Zook, J. M.; Javadzadeh, S.; Park, J.; Sehgal, A.; Gymrek, M. Genome-Wide Profiling of Genetic Variation at Tandem Repeat from Long Reads. bioRxiv January 23, 2024, p 2024.01.20.576266. https://doi.org/10.1101/2024.01.20.576266.

(207) Vinces, M. D.; Legendre, M.; Caldara, M.; Hagihara, M.; Verstrepen, K. J. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science* **2009**, *324* (5931), 1213–1216. https://doi.org/10.1126/science.1170097.

(208) Mölder, F.; Jablonski, K. P.; Letcher, B.; Hall, M. B.; Tomkins-Tinch, C. H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S. O.; Kanitz, A.; Wilm, A.; Holtgrewe, M.; Rahmann, S.; Nahnsen, S.; Köster, J. Sustainable Data Analysis with Snakemake. F1000Research April 19, 2021. https://doi.org/10.12688/f1000research.29032.2.

(209) *wdl-docs*. https://docs.openwdl.org/en/stable/ (accessed 2024-05-13).

(210) Voss, K.; Auwera, G. V. der; Gentry, J. Full-Stack Genomics Pipelining with GATK4 + WDL + Cromwell. *F1000Research* **2017**, *6*. https://doi.org/10.7490/f1000research.1114634.1.

(211) *UK Biobank Research Analysis Platform*. https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform (accessed 2024-05-19).

(212) *Snakemake plugin catalog*. https://snakemake.github.io/snakemake-plugin-catalog/ (accessed 2024-05-13).