HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2024 August 17.

Published in final edited form as:

Cell. 2023 August 17; 186(17): 3659–3673.e23. doi:10.1016/j.cell.2023.07.002.

Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer

Ronen E. Mukamel[1,2,3,*], **Robert E. Handsaker**[3,4,5,*], **Maxwell A. Sherman**[1,2,3,6], **Alison R. Barton**[1,2,3,7], **Margaux L. A. Hujoel**[1,2,3], **Steven A. McCarroll**[3,4,5,**], **Po-Ru Loh**[1,2,3,**,+]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.

[2]Center for Data Sciences, Brigham and Women's Hospital, Boston, Massachusetts, USA.

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

[4]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[5]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

[6]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[7]Bioinformatics and Integrative Genomics Program, Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA.

Summary

Many regions in the human genome vary in length among individuals due to variable numbers of tandem repeats (VNTRs). To assess the phenotypic impact of VNTRs genome-wide, we applied a statistical imputation approach to estimate the lengths of 9,561 autosomal VNTR loci in 418,136 unrelated UK Biobank participants and 838 GTEx participants. Association and statistical fine-mapping analyses identified 58 VNTRs that appeared to influence a complex trait in UK Biobank, 18 of which also appeared to modulate expression or splicing of a nearby gene. Non-coding VNTRs at *TMCO1* and *EIF3H* appeared to generate the largest known contributions of common

Correspondence should be addressed to R.E.M. (rmukamel@broadinstitute.org), R.E.H. (handsake@broadinstitute.org), S.A.M. (smccarro@broadinstitute.org), or P.-R.L. (poruloh@broadinstitute.org).

*These authors contributed equally to this work.

**Senior author.

+Lead contact: Po-Ru Loh

Current address for M.A.S.: Serinus Biosciences Inc, New York, New York, USA

Author contributions

R.E.M., R.E.H., S.A.M., and P.-R.L. conceived and designed the study. R.E.M., R.E.H., and P.- R.L. designed and implemented the statistical methods and performed the computational analyses. R.E.M., R.E.H, A.R.B., M.A.S., M. L. A. H., S.A.M., and P.-R.L interpreted analytical results. All authors wrote and edited the manuscript.

Declaration of interests

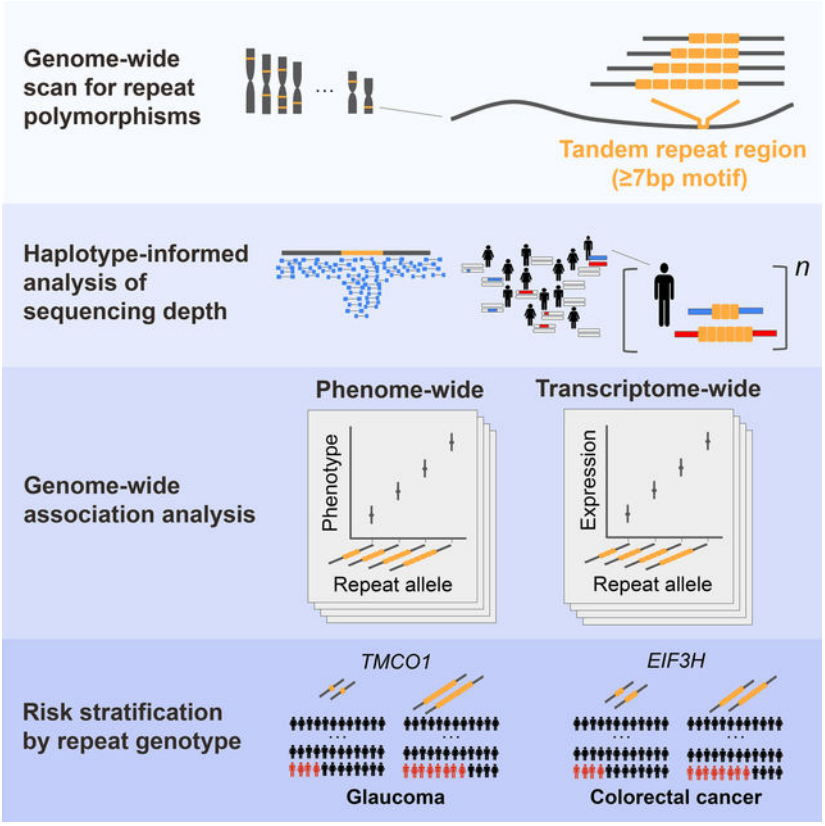
The authors declare no competing interests.

human genetic variation to risk of glaucoma and colorectal cancer, respectively. Each of these two VNTRs associated with a >2-fold range of risk across individuals. These results reveal a substantial and previously unappreciated role of non-coding VNTRs in human health and gene regulation.

In brief

Population-scale analysis of variable number tandem repeats in the human genome reveals hundreds of repeat polymorphisms that appear to influence complex traits and gene expression, including two repeats that generate the human genome's strongest associations with glaucoma and colorectal cancer.

Graphical Abstract



Introduction

Thousands of human genome segments are present in variable numbers of tandem repeats (VNTRs) in different individuals' genomes, but the effects of VNTRs on human phenotypes have been difficult to measure. At each VNTR locus, a sequence of nucleotides, from seven to thousands of base pairs long, is repeated several to hundreds of times per allele, with the number of repeats varying among individuals. Extreme VNTR alleles have been implicated in human diseases including progressive myoclonus epilepsy[1] and facioscapulohumeral muscular dystrophy[2]. However, VNTRs have not been considered in most genome-wide

association studies because such polymorphisms are not measured directly by SNP arrays and are challenging to characterize from short sequence reads.

Recent computational advances have enabled VNTR lengths to be measured or estimated from sequencing data and evaluated for association with phenotypes. Most studies to date have analyzed cohorts in which participants are both phenotyped and sequenced, measuring VNTR allele lengths either directly from spanning reads or indirectly from sequencing depth-of-coverage[3–8]. This approach has succeeded in identifying associations between VNTRs and the expression of nearby genes[4–7], but discovering associations with health and disease phenotypes[8] has proven more difficult due to the challenge of amassing phenotype and VNTR-allele information in the large number of individuals typically needed for genetic studies to discover genotype-phenotype associations, and the still-larger sample sizes required to distinguish among the effects of genomically nearby variants (such as VNTRs and nearby SNPs). An approach that has driven discovery of many SNP-phenotype associations is to impute untyped alleles based on the SNP haplotypes on which they segregate[9]; this approach has been extended to complex and multi-allelic copy number variations[10–12]. We and others recently observed that this approach can be extended to tandem repeats[13–15]. We further demonstrated that analysis of shared haplotypes can, at many loci, substantially improve the accuracy of VNTR length estimates from short-read sequencing depth by effectively combining measurements across individuals who inherited identical VNTR alleles from a recent common ancestor[14].

Our recent work applied this statistical imputation framework to analyze exome-sequencing data in UK Biobank (UKB), showing that protein-coding VNTRs underlie some of the strongest known genetic associations with diverse phenotypes including height, serum urea, and hair curl[14]. Here, we applied this approach to whole-genome sequencing data to estimate VNTR lengths genome-wide in deeply phenotyped UKB participants and donors of RNA-sequenced biosamples from the Genotype-Tissue Expression (GTEx) project to assess the role of non-coding as well as coding VNTRs in shaping human phenotypes and gene expression.

Results

Ascertainment and genotyping of 15,653 VNTR polymorphisms genome-wide

We identified VNTR loci across the human genome by analyzing the GRCh38 reference genome in conjunction with 64 haploid genome assemblies generated from long-read sequencing by the Human Genome Structural Variant Consortium (HGSVC2)[16]. At each of 100,844 autosomal repeats with repeat unit ≥7bp (identified in GRCh38 using Tandem Repeats Finder (TRF))[17], we determined the lengths of the corresponding repeat alleles in HGSVC2 assemblies by aligning flanking sequences from the human reference[18], excluding a small fraction of repeats (5.2%) for which either of the two flanking sequences failed to map uniquely in >50% of assemblies (STAR Methods). Most repeats identified by TRF were either monomorphic (51%) or biallelic (20%) in the HGSVC2 assemblies. Restricting to multiallelic repeats (≥3 distinct alleles) and removing overlapping repeats left 15,653 multiallelic VNTR loci for downstream analysis in whole-genome sequencing (WGS) data (Table S1). These VNTRs had a median repeat unit length of 34bp and a median

of 6 distinct alleles represented among HGSVC2 assemblies. VNTRs with more repeats generally exhibited greater allelic diversity (Figure 1A). VNTR allele length distributions in HGSVC2 assemblies had a median range of 199bp and median standard deviation of 46.8bp (Figure 1B).

To estimate haplotype-resolved VNTR allele lengths in UK Biobank and GTEx participants, we applied a two-stage approach in which we first generated a reference panel of 9,376 VNTR+SNP haplotypes by analyzing short-read whole-genome sequencing (WGS) data from the Simons Simplex Collection (SSC)[19,20] and subsequently imputed VNTR alleles from SSC into UKB and GTEx. To generate the reference panel, we first estimated individual-level VNTR lengths (summed across the two parental alleles) from sequencing depth-of-coverage in 8,936 SSC participants (including 4,688 unrelated individuals whose haplotypes formed the reference panel). Such read-depth-based analysis is capable of distinguishing allele length variation at the scale of hundreds of base pairs[10]; accordingly, these initial VNTR length genotypes captured allelic variation accurately (based on sibling concordance) for highly length-polymorphic VNTRs but less accurately for VNTRs with less-variable lengths (Figure 1B). To enable analysis of less-length-variable VNTRs and imputation into SNP haplotypes, we next analyzed the SNPs surrounding each VNTR to identify individuals who were likely to have inherited identical VNTR alleles from a recent common ancestor, allowing us to simultaneously reduce noise in VNTR length measurements and estimate haplotype-resolved lengths of individual VNTR alleles. We additionally determined locus-specific parameters for imputing VNTR allele lengths into SNP haplotypes, using cross-validation to assess imputation accuracy and optimize parameters[14] (STAR Methods).

For most multiallelic VNTRs, this combination of sequencing read-depth analysis with haplotype-sharing analysis enabled robust statistical imputation that correlated with actual allele-length variation more strongly than any nearby biallelic SNP did (Figure 1C, Table S1, and STAR Methods) – offering the potential for downstream discovery of genotype-phenotype associations previously invisible to or only weakly discernible from SNP-association analyses. Among 15,653 autosomal, multiallelic repeat loci, this analysis strategy typically captured a substantial proportion of allelic variation (median imputation R^2=0.48), with the most variable VNTRs (allele length s.d. >100bp; 4,462 loci) particularly well-analyzed (median imputation R^2=0.79). Excluding poorly-imputed (R^2<0.1) VNTRs and VNTR regions at which sequencing depth measurements failed quality control filters (STAR Methods) left 9,561 VNTRs for imputation into UKB and GTEx.

Limited contribution of VNTRs to common neurodevelopmental disorders

We tested each of the 9,561 VNTRs amenable to our statistical analysis for association with autism spectrum disorder in SSC and with 118 neurodevelopmental phenotypes in UKB (Figure S1A), using a study-wide significance threshold of $P<5 \times 10^{-9}$ (STAR Methods). In SSC, no VNTR reached significance in linear regression analysis adjusting for sex. In UKB, we tested imputed VNTR lengths for association with 33 mental health traits derived from participant surveys, 83 UKB-curated disease phenotypes categorized under ICD-10 codes F00-F99 and G00-G99 (Table S2), and two late-onset illnesses reported

for parents of participants (Alzheimer's disease and Parkinson's disease). We analyzed 418,136 unrelated UKB participants of European ancestry, adjusting for age, age^2, sex, UKB assessment center, genotyping array, and 20 genetic principal components (STAR Methods). These analyses identified 58 significant associations ($P<5 \times 10^{-9}$; Table S3). However, none of these associations appeared to reflect causal biological relationships: each locus contained a SNP with a stronger association to phenotype than the VNTR, and each VNTR association was assigned a low probability of causality by statistical fine-mapping (maximum FINEMAP[21] posterior inclusion probability (PIP)=0.03), suggesting linkage disequilibrium with other nearby causal variants. A previously-reported association of an intronic VNTR in *ABCA7* with Alzheimer's disease[22] replicated at nominal significance ($P=2.3 \times 10^{-4}$ for parental Alzheimer's) but also appeared unlikely to be causal (PIP=0.00).

Exploring the effects of non-coding VNTRs phenome-wide

To assess the role that VNTRs play in shaping complex traits more broadly, we expanded analysis to 668 additional phenotypes measured in UKB (including quantitative traits and binary disease phenotypes), testing each phenotype for association with the 9,561 imputed VNTRs (Table S2; Figure S1B). These analyses identified 4,910 significant VNTR-phenotype associations ($P < 5 \times 10^{-9}$), of which 107 associations (involving 58 distinct VNTRs) were assigned a high probability of causality by statistical fine-mapping (PIP>0.5; Figure 2A, Table S3, and STAR Methods). Many of the VNTRs involved in these associations overlapped regulatory or coding elements (Figure 2B): 34% (95% CI, 22–47%) of fine-mapping-supported VNTRs overlapped an annotated promoter, enhancer, or exon, a significant enrichment compared to the set of all imputed VNTRs (11% [10–12%]) or the set of all VNTRs that associated with a phenotype in UKB irrespective of fine-mapping PIP (17% [15–19%]). We additionally observed modestly greater genotyping accuracy and allelic variation among the VNTRs involved in fine-mapping-supported associations (Figure S2).

These associations included five between non-coding VNTR polymorphisms and human diseases, including a previously reported association of a VNTR upstream of the insulin gene *INS* with type 1 diabetes[23], as well as associations of VNTR length polymorphisms with risk of glaucoma, colon polyps, and hypertension. Two non-coding VNTRs (within *TMCO1* and near *EIF3H*) appeared to generate the largest known contributions of common human genetic variation to risk of glaucoma and colorectal cancer, respectively. The remaining 102 associations involved quantitative traits. Several non-coding VNTRs including a large intronic repeat in *CUL4A* associated strongly with blood cell traits ($P < 10^{-50}$), with association strengths similar to those we recently observed for coding VNTRs (Figure 2). Four of the 107 associations we identified (involving VNTRs at *INS, ACAN,* and *TCHH*[14,15,23]) have been previously reported (STAR Methods). Four other associations (near *OR5D13, SERINC2, CLCN7*, and *ITGB2*) were at loci at which no SNP reached conventional genomewide significance ($P < 5 \times 10^{-8}$; Table S4).

For three VNTRs with particularly strong and interesting phenotype associations—at *TMCO1, EIF3H,* and *CUL4A*—we performed a rigorous suite of follow-up analyses that confirmed the robustness and further elucidated the nature of their phenotype associations

(Figure S1C). First, we improved accuracy with which VNTR repeat numbers could be inferred from WGS (and directly validated this approach using HGSVC2 long-read assemblies; Figure S3) by leveraging subsequent whole-genome sequencing of 200,018 UKB participants and developing statistical models tailored to the allele distribution at each locus (STAR Methods). In each case, the absolute strength of the VNTR's association with phenotype, as well as its strength relative to nearby SNP associations, increased with this improved analysis (Table S4). We also refined definitions of disease phenotypes in UKB (STAR Methods) and analyzed data from independent cohorts to replicate associations and search for insights into potential molecular mechanisms as detailed below.

Repeat expansion at *TMCO1* associates with glaucoma risk more strongly than any SNP or indel in the genome

The strongest disease association we observed involved expansion into many repeats of an intronic 28bp sequence in *TMCO1*; this VNTR associated with glaucoma risk more strongly than any SNP or indel in the entire genome (P=1.3 × 10^{-76} vs. 2.8 × 10^{-68} for the strongest SNP association genome-wide; Figure 3A,B). All the expanded VNTR alleles (containing 5–11 repeat units vs. the one-repeat major allele; Figure 3A) segregated on a common ~70kb SNP haplotype at *TMCO1* (AF=12% in UKB, tagged by rs2790052) that was among the first-identified and strongest known influences of common genetic variation on glaucoma[24] (Figure 3B); in our analysis, excess cases among carriers of expanded alleles accounted for ~10% of primary open-angle glaucoma cases in UKB. Statistical fine-mapping pointed to the VNTR rather than the GWAS SNPs as the primary causal variant at this locus (PIP=1.00 for the VNTR), consistent with analyses of VNTR+SNP joint models (STAR Methods). Glaucoma is the leading cause of irreversible blindness worldwide[25], characterized by optic nerve damage caused in most cases by elevated intraocular pressure (IOP). Even after excluding glaucoma cases, *TMCO1* VNTR length associated with IOP more strongly than any SNP in the genome did, providing independent statistical support for the hypothesis that the VNTR rather than nearby SNPs underlies the GWAS signal at *TMCO1* (P=6.5 × 10^{-60} vs. P>2.9 × 10^{-51} for SNPs in analyses of N=94,877 UKB participants with IOP phenotypes and no reported glaucoma; Figure 3C), and that the VNTR affects glaucoma risk through its effect on IOP.

Repeat alleles of the *TMCO1* VNTR formed an allelic series with increasing effects on IOP and glaucoma risk at longer repeat lengths (Figure 3D,E). The longest VNTR alleles (top 3%) associated with larger effects on IOP and glaucoma risk than any common SNP elsewhere in the genome (glaucoma OR=1.51 [95% CI, 1.42–1.60] vs. OR≤1.34 for unlinked SNPs with MAF>0.01 and OR=1.34 [95% CI, 1.30–1.38] for rs2790052; IOP β=0.185 s.d. (SE, 0.013 s.d.) vs. β≤0.155 s.d. for unlinked common SNPs and β=0.106 s.d. (SE, 0.007 s.d.) for rs2790052). Individuals homozygous for long alleles (top 0.3% of summed allele length) exhibited >2-fold increased glaucoma risk relative to individuals with no repeat expansion (OR=2.27 [1.82–2.85]).

SNPs at *TMCO1* that tagged expanded VNTR alleles offered the opportunity to replicate these associations in independent, well-powered glaucoma and IOP genetic association data sets[26,27] (Figure 3D,E). In these replication cohorts, carriers of rs116089225:C>T, the SNP

allele associated with greatest mean VNTR allele length among carriers in UKB (AF=0.01; 11 repeats in a genotyped carrier in HGSVC2; STAR Methods), exhibited significantly elevated glaucoma risk (OR=1.70 [1.43–2.01]; Figure 3D) and IOP (β=0.201 (0.043) s.d.; Figure 3E) relative to carriers of the common risk haplotype that segregated with all expanded alleles (AF=0.12, carrier mean allele length = 7.6 repeats in UKB; glaucoma OR=1.34 [1.29–1.39], IOP β=0.084 (0.012) s.d.; Figure 3D,E). These results from studies that excluded UK Biobank provided confirmatory evidence for the series of VNTR allele effects we observed in UKB.

Though the statistical evidence points to the VNTR as the causal variant driving glaucoma associations at *TMCO1*, the molecular mechanism and causal gene underlying this association remain elusive. Consistent with previous reports[28], carriers of a rare *TMCO1* loss-of-function mutation (rs752176040:ACT>A, AF=0.00034 in exome-sequenced UKB participants[29]) did not appear to have elevated IOP (β=0.114 (0.135) s.d.) or increased glaucoma risk (OR=1.07 [0.58–1.96]). Analysis of loss-of-function mutation carriers for other nearby genes also did not provide any clues toward a candidate gene (Figure S4A). In RNA sequencing data from GTEx[30], VNTR length associated with expression at *TMCO1* in most tissues, consistent with recent SNP-based colocalization analyses[31], but these associations did not display evidence of an allelic series (Figure S4B). Additionally, joint modeling of the VNTR and nearby SNPs (STAR Methods) suggested that a variant other than the VNTR, possibly rs2790052 or rs2251768 in the promoter region of *TMCO1*, is responsible for the main eQTL at this locus and that the expression signal is unrelated to the glaucoma and IOP associations.

Common repeat polymorphism at *EIF3H* associates with a twofold range of colorectal cancer risk

Colorectal cancer is a heritable complex disease for which more than one hundred common risk alleles have been identified, each with a subtle influence on disease risk (OR<1.2)[32]. By contrast, the length of a 27bp repeat (usually ranging from 2–6 repeat units) ~20kb downstream of *EIF3H* associated strongly with risk of colorectal cancer and colon polyps ($P = 1.3 \times 10^{-24}$ and $P = 9.3 \times 10^{-34}$, respectively; Figure 4A,B), with the longest common allele (6 repeat units; AF=0.04) conferring higher colorectal cancer risk (OR=1.34 [1.24–1.45]) than any common SNP or indel in the genome (Figure 4C). The VNTR appeared to explain nearby SNP associations that were among the first associations reported for colorectal cancer[33]. Moreover, the explanatory power of this locus, which ranked first among all colorectal cancer loci genomewide ($P = 1.3 \times 10^{-24}$ for the VNTR vs. $P = 2.2 \times 10^{-19}$ for the strongest SNP association; Figure 4A), had previously been underestimated by ~50% in association studies that considered only SNPs which are in partial LD with the VNTR (maximum $R^2 = 0.27$; Figure 4A,B). Imputation of the VNTR association into summary statistics[34] (that excluded UKB) from a large colorectal cancer meta-analysis[32] replicated the VNTR association as the strongest at the locus (imputed $P = 6.7 \times 10^{-11}$ for the VNTR vs. $P \geq 7.3 \times 10^{-9}$ for nearby SNPs; Figure S5A). In UKB, the VNTR's association was driven by a series of four common alleles (3–6 repeat units) which exhibited increasing effects on risk of colorectal cancer and colon polyps. Disease risk increased linearly (on the log-odds scale) with VNTR length (Figure 4C), with each additional repeat unit associating

with a 14% (11–17%) increased risk of colorectal cancer (9% [7–10%] for colon polyps). The effects of an individual's two alleles appeared to be additive (P=0.68 for interaction term), such that common repeat length variation at *EIF3H* appeared to produce a >2-fold range of colorectal cancer risk across individuals (Figure S5B).

The length of this VNTR did not associate with expression of any nearby gene in analyses of RNA sequencing data, either from healthy tissue sequenced by GTEx[30] (P≥0.002 for each of 11 genes within 1Mb and each of up to 49 tissues) or from colorectal tumor tissues from the Cancer Genome Atlas (TCGA)[35] (P≥0.1 for each of 8 protein-coding genes in analysis of 465 tumor samples). VNTR length likewise did not significantly associate with expression of distal genes in any GTEx tissue ($P>1.7 \times 10^{-6}$ in each of ~1.2 million tests) or in TCGA ($P>4.0 \times 10^{-6}$ in each of 57,597 tests). VNTR length did associate with methylation of the nearby CpG site rs551792111 in GTEx ($P=7.1 \times 10^{-9}$ in colon samples; consistent negative effect direction in 9 of 9 tissues with available data[36]); however, this association did not replicate in TCGA colorectal samples (one-sided P=0.98).

The gene *EIF3H*, which encodes a subunit of a translation initiation factor, has been nominated as a potential causative gene at this locus[33]. However, definitive evidence linking colorectal cancer risk variants at 8q23.3 to a gene has remained elusive, and, consistent with our findings, risk alleles at this locus have not been shown to associate with *EIF3H* expression[37]. Though we have identified the VNTR as a promising candidate for the causal variant at this locus (with statistical support from analyses of two distinct phenotypes – colorectal cancer and colon polyps; between-phenotype R^2=0.02 – as well as independent replication), deciphering the molecular mechanism will require new kinds of data.

Intronic repeat expansion in *CUL4A* influences alternative splicing and erythrocyte traits

At *CUL4A*, expansion of a highly polymorphic intronic repeat (commonly consisting of ~3–100 copies of a 29–32bp repeat unit; Figure 5A) associated with decreased mean corpuscular hemoglobin ($P=6.4 \times 10^{-61}$, Figure 5B,C) and nine other erythrocyte-related traits (Figure 2, Figure 5C, and Table S4). The VNTR association was >3-fold stronger than that of nearby SNPs (none of which could effectively tag the VNTR polymorphism: maximum R^2=0.30; Figure 5B) and was driven by a series of alleles with monotonically strengthening effects on the associated phenotypes (Figure 5C). UK Biobank participants carried a multimodal allele distribution with a long tail of expanded alleles (Figure 5C), consistent with expanded alleles observed in HGSVC2 assemblies (Figure S3). The longest alleles (top 1%, >2.1kb) associated with 0.075 (0.010) s.d. reduced mean corpuscular hemoglobin (Figure 5C).

Analysis of GTEx RNA-seq data revealed that VNTR allele length strongly associated with an apparent splice defect in *CUL4A*, in which individuals carrying longer VNTR alleles were less likely to make the canonical splice over the VNTR in intron 5 (which varies in length from ~3–6kb owing to the VNTR polymorphism), instead splicing to a much more proximal sequence (122bp from the splice donor) that results in premature truncation of the *CUL4A* reading frame without the 15 downstream canonical exons ($P=1.0 \times 10^{-73}$ in cultured fibroblasts; P≤0.05 in 47 of 48 additional tissues tested; Figure 5A,D,E,F). This splice event associated with the VNTR much more strongly than with any SNP, in each of

the 30 tissues for which a variant reached Bonferroni significance (Figure 5F). In each case, longer alleles associated with greater usage of the protein-truncating isoform (Figure 5E).

The gene *CUL4A* encodes a ubiquitin ligase with an ortholog (*Cul4A*) that is required for hematopoiesis in mice[38], suggesting a molecular mechanism in which the VNTR length polymorphism might influence erythrocyte traits by interfering with *CUL4A* splicing and thereby modulating production of a truncated CUL4A isoform with reduced or lost function. Beyond the influence of the VNTR, the proportion of *CUL4A* transcripts that are mis-spliced in this way varies considerably across tissues (from an average of ~0.03 in skeletal muscle to ~0.75 in whole blood and testis; Figure 5G), suggesting that cellular context, as well as VNTR length, affects splicing outcome.

Non-coding repeat polymorphisms near *SIRPA*, *DOCK8*, and *PLEC* associate with platelet traits

The three strongest non-coding VNTR associations supported by statistical fine-mapping involved repeats near *SIRPA*, *DOCK8* and *PLEC*, with each VNTR appearing to underlie one of the top 40 associations genome-wide for a platelet phenotype, explaining >0.1% of variance (Figure 2, Figure S6, and Table S4). Approximately 50kb upstream of *SIRPA*, a 45bp repeat (with common 6-repeat and 10-repeat alleles) residing within a predicted enhancer[39] of *SIRPA* associated with mean platelet volume (MPV; P=1.6 × 10^{-167}; FINEMAP PIP=1.00). Within *DOCK8*, a gene harboring rare coding variants previously linked to MPV[40], the length of a highly length-polymorphic (~100bp–6kb) intronic 61bp repeat also associated strongly with MPV (P=6.5 × 10^{-129}; PIP=1.00). At *PLEC*, which encodes plectin, an intermediate filament binding protein with roles in cytoarchitecture and cell shape[41], the length of a 76bp intronic repeat (2–13 repeats per allele) associated with platelet distribution width (P=1.2 × 10^{-96}; PIP=0.99).

Exploring VNTR effects on gene expression and splicing

To systematically explore the role of repeat polymorphisms in gene regulation and identify potential molecular mechanisms underlying VNTR associations with complex traits, we tested the 9,561 imputed VNTRs for association with expression and splicing quantitative traits for nearby genes (<1Mb; ~6,350 single-tissue tests per VNTR) in GTEx, adjusting for technical and biological covariates (STAR Methods). These analyses identified 3,169 VNTRs that were significantly associated (P<1 × 10^{-10}) with a gene regulation trait in at least one tissue, of which 702 VNTRs were supported by statistical fine-mapping (PIP>0.5 in at least one single-tissue test; Table S5 and S6). This catalog of *cis*-regulatory associations includes a transcriptome-wide profile of VNTR-mediated regulation of splicing, which was particularly fruitful for obtaining mechanistic insights from sequence-level information about splice junctions (discussed in the next section).

VNTRs associated with *cis*-regulation exhibited enrichment near relevant genomic features: expression-associated VNTRs (eVNTRs) were enriched near transcription start sites (Figure 6A), consistent with previous eVNTR analyses[6], while splice-associated VNTRs (sVNTRs) were enriched near affected splice sites (Figure 6B). Both eVNTRs and sVNTRs were enriched in annotated regulatory regions (Figure 6C). All enrichments were accentuated

among associations supported by statistical fine-mapping, with the strongest enrichments observed among the 50 VNTRs that exhibited consistent evidence of association and causality across multiple tissues (PIP>0.5 in ≥4 tissues, and PIP<0.01 in ≤2 tissues).

Integrating these data with our phenome-wide scan of VNTR associations identified 18 VNTRs that appeared to influence both gene regulation and a complex trait in UKB. This analysis provided plausible regulatory functions for 31% (18/58) of the VNTRs that were involved in a fine-mapping-supported association in UKB (Figure 6D), including the intronic VNTR in *PLEC* (discussed above) which also associated with excision of its intron (Table S6). Other notable examples include a VNTR 1kb downstream of *GPIHBP1* (which encodes an HDL binding protein[42]) that associated with increased HDL (P=2.6 × 10^{-41}; PIP=0.99), apparently by regulating *GPIHBP1* expression (P<1 × 10^{-10} and PIP>0.5 in 13 tissues; Table S5). An intronic VNTR in *CHMP1A* associated with hypertension risk (P=1.5 × 10^{-12}; PIP=1.00), *CHMP1A* expression (P<1 × 10^{-10} and PIP>0.5 in 28 tissues), and usage of a proximal splice site 31bp upstream of the VNTR (P<1 × 10^{-10} and PIP>0.5 in 5 tissues; Figure S6,S7A). Bone mineral density was significantly associated with an eVNTR in the promoter region of *ITGB2* (P=1.8 × 10^{-12}, PIP=1.00; Figure S6) as well as an intronic sVNTR in *SBNO2* (P=3.4 × 10^{-45}, PIP=0.97; Figure S6,S7B), two genes that have previously been implicated in osteogenesis[43,44].

Hundreds of repeat polymorphisms influence splice site usage by diverse mechanisms

Repeat polymorphisms at 327 genomic loci appeared to influence splicing of a nearby gene (P<1 × 10^{-10}, FINEMAP PIP>0.5 in at least one tissue; Table S6) by a diverse set of mechanisms, with VNTRs seemingly capable of modulating usage of both proximal and distal splice sites. These included a set of 22 sVNTRs each of which displayed consistent evidence of association and causality across tissues (PIP>0.5 in ≥4 tissues, PIP<0.01 in ≤2 tissues). Among these, 21 were located within 1kb of an affected splice site and 20 associated with altered splicing of spanning introns. Further examination of the locations of these sVNTRs relative to affected splice sites, and their apparent transcriptional effects, revealed insights into the varied ways in which repeat polymorphisms can influence splicing (Figure 7 and Data S1):

- *Repeat alleles contain alternative splice sites in UPF3A and TUBGCP2.* At *UPF3A*, a 32bp intronic repeat appeared to influence usage of two types of alternative transcripts (Figure 7A). Some alternative transcripts featured splicing to an acceptor site located within the repeat region and retention of the remainder of the intron; other transcripts skipped the canonical exon at the 3' end of the spanning intron. At *TUBGCP2*, a 94bp repeat containing a canonical splice donor site influenced usage of alternative donor sites located elsewhere within the repeat (Data S1A). At both loci, alternative splice usage increased with repeat allele length.
- *Repeats containing canonical splice donor sites influence multi-exon skipping in NOC4L and RSPH1.* Long alleles of a 44bp repeat in *NOC4L* and a 41bp repeat in *RSPH1* appeared to disrupt canonical splicing and increase the frequency of splice forms lacking ≥5 skipped exons (Figure 7B, Data S1B). The

NOC4L VNTR has previously been highlighted for its association with *NOC4L* expression[6], possibly reflecting decreased expression of the exons skipped in alternatively spliced transcripts.

- *A repeat proximal to a canonical splice acceptor influences intron retention in PLIN5.* The 3' end of this 24bp repeat is located 5bp away from a canonical splice acceptor (Figure 7C). The proximity of this repeat to the affected splice acceptor, together with its pyrimidine-rich sequence composition (18 pyrimidines within the 24bp repeat), suggest the hypothesis that expanded VNTR alleles influence intron retention by disrupting branch point recognition.
- *Repeat alleles modulate retention of large intronic segments in SLC22A18, C1orf174, TARSL2, and PER1.* These VNTRs appeared to affect canonical splice sites ranging from tens to thousands of base pairs away (Data S1C–F). At *TARSL2* and *SLC22A18,* usage of alternative splicing increased with allele length, whereas *PER1* and *C1orf174* exhibited the opposite effect direction. While alternative transcripts at *SLC22A18* and *C1orf174* made use of multiple alternative splice sites, alternative transcripts at *PER1* and *TARSL2* tended to use one or two alternative sites.
- *Repeat alleles influence inclusion of a nearby exon in SLC12A7, LRRC27, PQLC1, SETX, CNN2, and PKD1.* At three loci (*SLC12A7, PKD1,* and *SETX*; Data S1G–I), the VNTR lies within an intron downstream of the variably expressed exon, while at *CNN2* and *LRRC27* (Data S1J,K), the VNTR lies within an upstream intron. Intriguingly, at *PQLC1*, the VNTR is contained within the alternatively expressed exon itself (Figure 7D).
- *Repeat alleles influence use of an alternative transcription start site for CDCA4.* The length of a 31bp repeat located 12kb upstream of the canonical transcription start site for *CDCA4* appeared to influence usage of an alternative first exon that spans the VNTR (Data S1L). Repeats at *DPH1, PRKAR1B* and *RPH3AL* (Data S1M–O) potentially reflect a similar phenomenon: at each locus, intronic transcription proximal and downstream of the VNTR increases with allele length, with no apparent alternative splice acceptor.

Contribution of VNTRs to complex trait heritability and polygenic prediction

The number and strength of the associations we observed suggest that on a per-variant basis, a VNTR tends (on average) to contribute more to human phenotypic variation than a SNP. The VNTRs we analyzed, which represent ~0.1% of common (MAF>0.01) variants, were substantially overrepresented among lead variants at GWAS loci, comprising 0.8% (90/11,858) of the lead variants identified across 31 highly heritable, polygenic blood cell traits (STAR Methods). We observed a similar pattern at expression quantitative trait loci, where 0.8% (1,459/181,627) of the lead variants were VNTRs. Comparing these numbers to an analogous estimate that common structural variants are causal at 2.66% of eQTLs[45] suggests that VNTRs contribute a sizable fraction of the heritability explained by genomic structural variation. Additionally, while VNTRs in aggregate appear to explain a modest fraction of heritability for most traits (on the order of 1% based on the above analyses),

VNTRs have a major role in shaping some phenotypes: VNTRs contribute 11% of the glaucoma heritability and 20% of the colorectal cancer heritability explained by GWAS loci in UK Biobank (STAR Methods). Incorporating VNTRs into polygenic prediction of colorectal cancer risk substantially increased accuracy: inclusion of the *EIF3H* VNTR, which is poorly tagged by nearby SNPs, boosted accuracy by ~25% (STAR Methods).

Discussion

These results identify many VNTRs that appear to have strong effects on human phenotypes and gene expression, including five VNTR length polymorphisms that are associated with risk of common diseases. Two disease associations we observed, involving VNTRs at *TMCO1* and *EIF3H*, appeared to be the strongest known genetic influences of common inherited variation on glaucoma and colorectal cancer risk, respectively. Additionally, analyses in GTEx indicated that VNTRs appear to be capable of modulating splice isoform usage by a diverse set of mechanisms and from locations both proximal and distal to affected splice sites. These discoveries were enabled by a computational approach to VNTR genotype estimation that integrated sequencing depth-of-coverage analysis with statistical phasing and imputation into SNP-array genotyping data – a framework that can similarly be applied to other genetic data sets.

While our analyses of GTEx identified plausible transcriptional mechanisms and causal genes underlying a sizable fraction (18/58) of the VNTRs linked to complex traits in UKB, most phenotype-associated VNTRs lie in non-coding regions of the genome, modifying DNA sequence length by hundreds to thousands of base pairs, yet with no obvious molecular mechanism to explain their apparent impact on phenotype. Non-coding variants linked to phenotype pose a central challenge in human genetics. Other techniques and further study will be required to elucidate the “missing regulation” and identify the mechanisms underlying the associations observed here[46].

Limitations of the study

Despite considerably expanding the set of known associations between VNTRs and human phenotypes, the results presented here likely represent an incomplete look into the landscape of repeat-mediated trait heritability, owing to genotyping challenges that we could only partially overcome as well as inherent limitations of the UK Biobank cohort we analyzed. While our read-depth and haplotype-modeling approach provided indirect VNTR length measurements that cross-validation benchmarks indicated were accurate, these estimates have yet to be validated against direct measurements from long-reads. Our strategy accurately captured larger-scale VNTR length variation (>100bp) but produced noisier genotype estimates for less-variable VNTRs, reducing power to analyze such VNTRs. Moreover, we excluded all short tandem repeat (STR) loci from analysis, for which other methods are required[13,47]. The set of VNTRs we considered was also limited by our GRCh38-based VNTR ascertainment strategy (which required multiple repeat units to be present in the human reference) and the need for mappability of short reads. Additionally, our ability to detect VNTR-phenotype associations was limited by the demographics of the UK Biobank cohort, which enrolled generally healthy participants of predominantly

European ancestry. Analyses in case-control cohorts enriched for heritable diseases will be needed to power discovery of further influences of VNTR variation on human health. Finally, our ability to identify which VNTR associations represent causal biological relationships was constrained by the limitations of statistical fine-mapping. While we chose to focus on the 107 associations assigned >50% posterior probability of causality here, some of these may not be truly causal (e.g., due to model misspecification) while other (PIP<0.5) associations may in fact be causal (e.g., 251 associations with intermediate PIP between 0.05 and 0.5; Table S3). Many of the above limitations are now beginning to be overcome as long-read sequencing data sets scale to thousands of samples[15] and short-read WGS data sets scale to hundreds of thousands of samples[48], including in diverse populations[49]. We anticipate that these recently-generated and upcoming data resources will enable many further insights into the contribution of repeat polymorphisms to heritable complex traits in the years to come.

STAR Methods

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Po-Ru Loh (poruloh@broadinstitute.org).

Materials availability—This study did not generate new unique reagents.

Data and code availability

- Individual-level VNTR genotypes imputed into UKB have been returned to the UK Biobank Resource, and VNTR+SNP haplotypes in SSC have been returned to SFARI Base (as the applicable data use conditions do not permit direct release of these data or incorporation into imputation servers). Summary statistics for VNTR-phenotype association tests have been deposited at Zenodo and are publicly available as of the date of publication (DOI listed in the key resources table).
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

UK Biobank genetic data—The UK Biobank resource contains extensive genetic and phenotypic data for ~500,000 participants recruited from across the UK[51]. We analyzed SNP and indel genotypes available from blood-derived SNP-array genotyping of 805,426 variants in 488,377 participants and subsequent imputation to 93,095,623 autosomal variants (using the Haplotype Reference Consortium and UK10K + 1000 Genomes Phase 3 reference panels) in a subset of 487,409 participants[52]. We further analyzed alignments from whole-genome sequencing (WGS) of 200,018 participants (>20x coverage by 151bp paired-end reads)[48].

UK Biobank phenotype data—We performed initial analyses on a set of 786 phenotypes (Table S2) that we curated from the UK Biobank "core" data set as previously described[14]. This set of phenotypes consisted of: (i) 636 diseases collated by UKB from several sources (self-report and accruing linked records from primary care, hospitalizations, and death registries) into single "first occurrence" data fields indexed by ICD-10 diagnosis codes; and (ii) 150 continuous and categorical traits selected based on high heritability or common inclusion in genome-wide association studies. Phenotypes in the latter set were derived from physical measurements and touchscreen interviews; blood count, lipid and biomarker panels of biological samples; and follow-up online questionnaires. For continuous traits, we performed quality control and normalization (outlier removal, covariate adjustment, and inverse normal transformation) as previously described[40,53]. The set of analyzed traits included 118 neurodevelopmental phenotypes tested in Stage 1 of our study (Table S2, Figure S1B). In follow-up analyses at *TMCO1* and *EIF3H*, we refined associated disease phenotypes (related to ICD-10 codes H40 and K63) and curated related phenotypes (intraocular pressure and colorectal cancer) not in our initial analysis set (STAR Methods, Phenotype refinement for disease-associated VNTRs).

Simons Simplex Collection genetic data—We analyzed aligned sequencing reads and the hg38 variant call set derived from SSC WGS 2. The SSC cohort consists of individuals from 2,600 families, each of which has one child affected with an autism spectrum disorder. Each participant was deeply whole-genome sequenced (30x mean coverage, 150bp paired-end reads). We analyzed genetic data obtained from a subset of 8,936 participants, which included 1,901 quartets (parents, proband, and unaffected sibling) and 440 trios (parents and child). In total, the analysis set contained 4,688 unrelated parents whose 9,376 haplotypes we included in our VNTR+SNP reference panel. We applied multiple rounds of quality control to generate a high-quality set of phased SNP haplotypes for SSC participants (STAR Methods, Quality-control, phasing, and IBD-calling in SNP data from SSC WGS).

Sample filters for ancestry and relatedness—For genetic association analyses in UKB, we applied strict filters to avoid confounding from population stratification and relatedness among individuals. We performed all analyses on a filtered set of 418,136 individuals that we identified by: (i) removing principal component (PC) outliers (more than six standard deviations from the mean among individuals who reported White ethnicity in any of the first 10 genetic PCs); and (ii) removing one individual from each $\leq 2^{nd}$-degree related pair (kinship coefficient > 0.0884) previously identified by UKB[52]. Each non-White ethnicity was reported by <2% of the UKB cohort, so we did not extend association analyses to other groups given low expected power.

For benchmarking accuracy of VNTR length estimation in SSC, we assigned ancestry to SSC participants using the software SNPweights v2.1[54]. Using pre-computed SNP weights for European, West African, East Asian and Native American ancestral populations (accessed from https://cdn1.sph.harvard.edu/wp-content/uploads/sites/181/2014/03/snpwt.NA_.zip on 08/20/2019), we estimated the proportion of each SSC participant's genome that derived from each ancestral population. We identified 3,904 unrelated parents whose genetic ancestry was estimated to be largely (>80%) European.

Overview of VNTR ascertainment and genotyping pipeline—We identified VNTR loci and genotyped VNTR allele length variation from whole-genome sequencing data using an analysis pipeline consisting of three main steps (detailed in the subsequent sections of STAR Methods):

1. ***Identify VNTR loci from analysis of the human reference and HGSVC2 long-read assemblies.*** We started by searching the GRCh38 reference for tandem repeats using two approaches:

   a. Tandem Repeats Finder[17] v4.09, which we ran using its suggested parameters 2 5 7 80 10 50 2000 -l 6 -h to detect patterns up to 2kb. We filtered to autosomal repeats with length ≥100bp, period ≥10bp, #repeats ≥2, and sequence identity ≥75%, and we applied a rough deduping of duplicated regions (overlap (intersection/union) >75% or both endpoints within 75bp); and

   b. VNTRScanner and VNTRPartitioner, algorithms we had previously developed to detect larger repeats with potentially greater variability within the repeat units[14].

   The combination of the two methods resulted in an initial set of 100,844 autosomal repeat loci with an estimated repeat unit ≥7bp long. We then analyzed each tandem repeat in 64 HGSVC2 long-read-based haploid genome assemblies[16] to identify which regions were multiallelic. We filtered to loci with ≥3 distinct alleles represented among the 64 HGSVC2 long-read assemblies, applied several additional quality control filters, and removed duplicated regions (using benchmarks from SSC WGS to adjudicate among substantially-overlapping regions), resulting in 15,653 VNTR loci for further analysis (STAR Methods, Identifying VNTR loci by analysis of human reference and HGSVC2 long-read assemblies).

2. ***Estimate VNTR lengths from WGS depth-of-coverage in SSC.*** At each VNTR, we estimated diploid VNTR content (i.e., the sum of VNTR lengths across an individual's two alleles) for SSC participants by analyzing the aligned WGS reads overlapping the VNTR using Genome STRiP[10], using dosage estimates from normalized read depth to estimate VNTR length (summed across parental alleles). We corrected for observed batch effects using Leiden clustering. We benchmarked the resulting (pre-refinement) VNTR genotypes using measurements from related individuals in SSC, and applied several additional variant-level quality control filters derived from these benchmarks (STAR Methods, Estimating diploid VNTR content from WGS read depth in SSC).

3. ***Phase and impute VNTR allele length estimates by modeling haplotype sharing.*** We performed statistical phasing on estimates of diploid VNTR content to estimate haploid allele lengths, and we used the resulting VNTR+SNP haplotypes for imputation of VNTR allele lengths into the UKB cohort. To do

so, we adapted the computational algorithm we had previously used[14] (STAR Methods, Phasing and imputing VNTR lengths using surrounding SNPs).

Of the 15,653 multiallelic VNTR loci identified in the first step of this pipeline, we identified a subset of 9,561 loci suitable for downstream association analyses in UKB based on genotyping quality (estimated R^2>0.1), imputation quality (estimated R^2>0.1), and other quality control filters (STAR Methods, Identifying VNTR loci by analysis of human reference and HGSVC2 long-read assemblies; Table S1).

We also considered applying other methods previously developed for genotyping repeats such as adVNTR[55] and ExpansionHunter[56]. However, we were unable to use adVNTR because it requires sequencing reads to span a VNTR, and the majority of VNTR loci (90%) had an allele longer than the read length in SSC (150bp). While ExpansionHunter is capable of genotyping repeats longer than the read length, it was designed primarily for STR genotyping and assumes that different repeat units are mostly identical in sequence, whereas many VNTRs exhibit repeat motif variability[57]. Beyond needing to overcome these specific limitations, we were also motivated to apply methods that leverage haplotype sharing among unrelated individuals in large cohorts to refine genotypes, increasing power to detect downstream associations[14].

Quality-control, phasing, and IBD-calling in SNP data from SSC WGS—We applied multiple rounds of QC to the SSC WGS 2 hg38 variant call set (9,209 samples[20]) to facilitate generating a high-quality set of phased SNP-haplotypes for SSC participants.

We first applied several variant-level filters:

- Restricted to biallelic SNPs with MAC>=5 and missingness <0.05.
- Excluded SNPs with allele frequencies (in European-ancestry SSC participants; see Methods) that differed by >0.1 compared to allele frequencies in the UK10K+1000G reference panel[58] (subsetting 1000G samples to EUR and lifted from hg19 to hg38) or to allele frequencies in the UK Biobank SNP-array data set (restricted to the British-ancestry subset curated by UK Biobank[52] and lifted from hg19 to hg38).
- Excluded SNPs with 10 or more Mendelian errors among parent-child trios (computed using the bcftools +mendelian plugin).

We then phased the filtered SNPs using Eagle2[59] and post-processed the phased haplotypes to incorporate trio relationships using the bcftools +trio-phase plugin (a component of the MoChA software package).

We subsequently observed that the set of SNPs that passed the above filters still included a small fraction of SNPs that appeared to have high error rates (often having high rates of heterozygous genotype calls and/or clustered in regions of the genome that did not lift between hg19 and hg38 and thus were not considered in the allele frequency check).

To detect remaining bad SNPs, we therefore implemented an additional round of QC consisting of a Hardy-Weinberg equilibrium check (filtering SNPs with z-score > 5 for

observed – expected heterozygotes) and a haploid Mendelian error check (filtering SNPs with >10 disagreements between phased haplotypes transmitted from parents to children in parent-child trios in the SSC data set). To facilitate the latter check, we implemented a simple hidden Markov model (HMM) to match computationally phased haplotypes of each child to computationally phased haplotypes of the child's two parents (with states corresponding to haplotype assignments, transitions modeling phase switch errors or recombinations, and emissions modeling genotype errors (treated as 10-fold less costly than state changes)). This algorithm allowed us to tabulate the number of haploid Mendelian errors observed at each SNP (based on the Viterbi decoding of each child's HMM).

After applying the additional two filters above (which together excluded ~1% of the SNPs that had passed the previous QC filters), we then reran phasing using Eagle2 followed by bcftools +trio-phase (in two rounds, first using one sibling from each quartet and then using the remaining sibling from each quartet, which appeared to improve performance). Finally, we reran the HMM above to obtain an "IBD map" matching phased haplotype segments of each child to haplotypes of the child's two parents for use in downstream analyses.

Sample exclusions for downstream analysis: Although the SSC WGS 2 variant call set contained 9,209 individuals, only 9,100 of these individuals had accessible sequence alignments (i.e., cram files) needed for downstream WGS read-depth analysis. Among these 9,100 individuals, we excluded 160 individuals whose whole-genome sequencing had been performed in a pilot WGS analysis with read-depth characteristics very different from the remainder of the data set. We further excluded 4 individuals who withdrew from SSC, leaving 8,936 individuals (including 1901 full quartets) for downstream analysis.

Estimating diploid VNTR content from WGS read depth in SSC—For each of 100,844 repeat loci we ascertained from the GRCh38 reference (STAR Methods, Overview of VNTR ascertainment and genotyping), we estimated diploid VNTR content (i.e., the sum of VNTR lengths across an individual's two alleles) for SSC participants by analyzing the aligned WGS reads overlapping the VNTR using Genome STRiP[10]. We estimated diploid VNTR content using dosage estimates from normalized read depth, without running the Genome STRiP Gaussian mixture model to determine integer copy number.

We benchmarked these VNTR content estimates by analyzing the results from siblings in SSC. We computed the following QC metrics:

1. **IBD2R**. The Pearson correlation between VNTR content measurements of SSC siblings that are identical-by-descent (IBD2) at a given VNTR locus. This quantity was used to estimate the amount of genetic signal that could be ascertained from read-depth analysis at each VNTR.
2. **PARENTR**. The correlation between VNTR content measurements among parents of IBD2 siblings (which was an indicator of batch effects, as this quantity should otherwise be close to zero).
3. **RDIFF**. The difference between IBD2R and PARENTR, i.e., RDIFF=IBD2R-PARENTR. RDIFF is meant to capture the degree to which correlation between

sibling measurements captures true genetic variation, rather than technical artifacts (e.g., batch effects) from sequencing.

4. **FLANKR**. For each VNTR locus, we measured read depth within two 1kb segments on each side of the VNTR locus, each separated by 100bp from the VNTR. We then computed the maximum of the Pearson correlation between the read-depth measurements of the segments outside of the VNTR to the read-depth measurement of the VNTR itself. This was used to control for VNTRs occurring within larger copy number variable regions.

We observed that diploid VNTR content estimates in SSC appeared to have significant batch effects, which were only partially correlated with SSC sequencing wave. To better control for these batch effects, we first excluded 160 SSC participants whose whole-genome sequencing had been performed in a pilot WGS analysis with read-depth characteristics very different from the remainder of the data set. To correct for additional batch effects, we then clustered the remaining samples as follows:

1. We selected 2,655 VNTR loci across the genome that (a) had strong evidence of batch effects (PARENTR > 0.1) and (b) were unlikely to be truly polymorphic (absolute value of RDIFF < 0.1).
2. We computed the top 100 principal components (PCs) based on read-depth estimates at these loci, and performed k-nearest-neighbor (k=25) clustering on the samples based on their coordinates in PC-space.
3. We performed Leiden clustering with resolution parameter R=1.0 on the neighbor graph, which partitioned the SSC samples into 18 clusters.

We normalized VNTR content estimates by scaling the estimates within each cluster so the median content across clusters was equal to the population median length. This clustering method and parameters described above were selected after evaluation of several strategies (hclust2 and cutree functions in R, and Louvain and Leiden clustering). We found that Leiden clustering produced the greatest decrease in PARENTR (indicating successful correction of batch effects).

Identifying VNTR loci by analysis of human reference and HGSVC2 long-read assemblies—We analyzed the 100,844 repeat regions we identified in GRCh38 (Methods) in HGSVC2 long-read haploid genome assemblies[16] to determine which were multiallelic VNTR loci. As a preliminary step, we removed duplicate loci with greater than 50% reciprocal overlap, prioritizing the loci to keep based on IBD2 sibling correlation in SSC (see Section 2 above). For each remaining repeat locus, we attempted to measure repeat length in each assembly by mapping surrounding sequence from GRCh38 to the assembly.

In detail, we extracted the flanking sequence from the reference (1kb upstream and downstream) using bedtools v2.27.1[60] and aligned the flanks to the assembly using minimap2 v2.18-r1015[18] (options --cs -x map-pb -t 7 -r 2000 -z 2000). We parsed the output, in the pairwise mapping format (PAF), to compute the length of the repeat allele in the long-read assembly. Specifically, for each flank, we selected the alignment with the largest number of matching residues (N_{res}, column 10 in PAF file), requiring:

1. N_{res}>900bp,
2. The length of the target contig containing the matched assembly sequence (column 7 in PAF file) is >10kb, and
3. No competing matches with "Number of residue matches">0.97*N_{res}

We proceeded with analysis if both flanks had an alignment satisfying (1–3). In this case, we further required that

a. The two flanks mapped to the same contig (column 6 in PAF file), and
b. The alignment directions (column 5 in PAF file) of the two flanks were consistent.

If both flanks had alignments satisfying (A-B), we measured the length of the repeat allele in the assembly by computing the distance between the flanking alignments, adjusting for any non-aligned bases at the ends of the flanks (columns 3 and 4 in the PAF file). We finally required that the computed allele length to be greater than –500bp and less than 200*(length of the repeat in GRCh38). Note that we allowed alleles to have negative lengths, possibly reflecting deletions that occurred near the ends of the VNTR or repeat loci whose boundaries were called incorrectly.

We next estimated the number of distinct alleles at each repeat locus (across the subset of the 64 HGSVC2 assemblies for which the algorithm above produced a length measurement). To do so, we counted the number of distinct allele length genotypes, requiring distinct alleles to differ by at least one-quarter of an estimated repeat unit. For repeats with short repeat units, we additionally required distinct alleles to differ by at least 7bp.

To obtain our final list of VNTRs for phasing and imputation optimization, we applied the following filters to the candidate VNTR loci:

1. >50% genotyping rate among HGSVC2 assemblies
2. ≥3 alleles represented among all N=64 assemblies
3. ≥2 alleles represented among N=12 assemblies from individuals of European descent

Finally, among all loci that satisfied (1–3), we removed regions that had substantial overlap with another VNTR. To do so, we iteratively removed each locus that had substantial overlap with another region (overlap spanning >10% of one of the two regions) where the overlapping region had higher estimated pre-refinement genotyping accuracy (estimated from IBD2 sibling correlation in all SSC participants).

This yielded a filtered set of 15,653 multi-allelic VNTR loci for further analysis.

Phasing and imputing VNTR lengths using surrounding SNPs—We performed statistical phasing on WGS read-depth-derived VNTR length estimates ("diploid VNTR content"; see Section 2 above) to estimate haploid allele lengths in SSC participants, which we then imputed from SSC into the UK Biobank cohort based on surrounding SNP-haplotypes. To do so, we adapted the computational algorithm that we previously used

to efficiently phase and impute multiallelic protein-coding VNTRs with real-valued length estimates derived from whole-exome sequencing read-depth within UKB[14]. This algorithm is described in detail in Supplementary Text 3 of ref. [14]; in brief, it employs an iterative approach (broadly similar to many algorithms that have been developed for phasing biallelic SNPs) in which haploid allele lengths of each individual in turn are updated according to a probabilistic haplotype-copying model using all other haplotypes as a reference panel, prioritizing copying from haplotypes closely matching the individual's SNP-haplotypes.

To make use of familial relatedness within the SSC cohort and to facilitate imputation from the SSC data set (containing WGS-based SNP calls in hg38 coordinates) to the UK Biobank data set (containing SNP-array genotypes in hg19 coordinates), we made the following minor modifications to our previous phasing and imputation approach. At each VNTR locus:

- We used the IBD maps we generated within SSC families (see above) to identify sib-pairs who inherited the same allele from their mother and inherited the same allele from their father ("IBD2" sibs, which we used for accuracy benchmarks; see below).
- When optimizing parameters for phasing and imputation (using the cross-validation-based procedure described in[14]), we held out diploid VNTR content estimates for:
  - 400 children (to enable optimization of phasing parameters by maximizing concordance with allele lengths estimated for transmitted parental haplotypes);
  - 400 individuals of European ancestry (to estimate European-ancestry imputation accuracy as described below, holding out full families to prevent relatedness from inflating the benchmark); and
  - 400 individuals of non-European ancestry (again holding out full families).
- For phasing (within SSC), we computed SNP-haplotype similarity based on identity-by-state (IBS) length as described in[14], which we computed using SNPs with MAF>0.01 in our QC-ed and phased version of the SSC WGS 2 (hg38) variant call set.
- For imputing from SSC into UK Biobank, we computed IBS (at the VNTR's hg19 location) using SNPs with MAF>0.001 that were present in the UKB SNP-array data set (hg19) as well as the SSC data set (lifted from hg38 to hg19).

For imputation, we restricted VNTR+SNP haplotypes to parents in SSC (N=4,688 individuals after sample exclusions; N=9,376 haplotypes) to avoid redundancy given the family structure of the SSC cohort. We post-processed the VNTR allele length assigned to each parental haplotype by taking the average of the allele length estimated for that haplotype by our phasing algorithm as well as the allele lengths estimated in any children to which the allele had been transmitted (based on our IBD maps; we restricted to confident transmissions with >2Mb of IBD-sharing).

Estimating VNTR genotyping and imputation accuracy and VNTR-SNP linkage disequilibrium—To estimate the accuracy of VNTR length estimates derived from WGS read-depth in individual genomes (before incorporating information from SNP-haplotypes, i.e., "genotype accuracy prerefinement" in Figure 1B,C), we used correlations among IBD2 sib-pairs as in our previous work[14]. Explicitly, assuming unbiased error in read-depth-based measurements of diploid VNTR content, we can estimate the accuracy (i.e., R^2 vs. truth) of these measurements as:

$$\widehat{R^2}(\text{diploid estimates, truth}) = R(\text{diploid estimate in sib 1, diploid estimate in sib 2}) = \text{"IBD2 } R\text{"}$$

To estimate imputation accuracy, we used a cross-validation-based approach as in our previous work: for 400 individuals held-out from phasing, we imputed VNTR lengths into the held-out individuals and then estimated imputation accuracy as:

$$\widehat{R^2}(\text{imputed estimates, truth}) = \frac{R^2\ (\text{imputed estimates, held out estimates})}{\text{IBD2 } R}$$

where dividing out by IBD2 R (an estimate of R^2 (held out estimates, truth)) accounts for measurement error in the held-out values. To obtain accuracy estimates indicative of imputation performance into the predominantly European-ancestry UK Biobank cohort, we restricted to SSC participants of European ancestry when selecting IBD2 sib-pairs and held-out individuals.

Two potentially counterintuitive features of these accuracy estimates are worth noting:

- Imputation accuracy can sometimes exceed pre-refinement genotype accuracy (i.e., accuracy of the diploid VNTR content measurements on which imputation is based).

  This behavior typically occurs if a VNTR has a narrow allele length distribution (such that alleles are difficult to distinguish from read-depth) but alleles are well-tagged by nearby SNPs, such that the phasing and imputation model is able to learn which SNP-haplotypes carry which alleles (and use SNPs to predict alleles more accurately than possible from read-depth).

- Imputation accuracy estimates are noisier for VNTRs with lower pre-refinement genotype accuracy (i.e., lower IBD2 R). This behavior is driven by the need to divide by IBD2 R (which can be a small quantity with sizable uncertainty) when estimating imputation accuracy using cross-validation. While our IBD2 R estimates typically used ~400 IBD2 sib-pairs at each locus, providing reasonable precision, noise in IBD2 R occasionally resulted in imputation accuracy estimates that exceeded 1 (presumably due to IBD2 R having been underestimated by chance).

One caveat of the above benchmarks is that they assume unbiasedness of errors in read-depth-based estimates of diploid VNTR content. We previously observed that exome sequencing coverage depths at VNTRs can be biased by the presence of paralogous

sequence variants (PSVs) within repeat units (that can subtly affect exome capture) or by read-mapping biases for very short alleles[14]. While the first issue has much less of an effect on whole-genome sequencing (which does not involve a capture step), to ensure robustness of our results, we performed follow-up analyses of VNTRs of particular interest in which we (i) used a variety of locus-specific techniques to optimize genotyping accuracy (see below); and (ii) validated WGS-derived genotypes against allele lengths directly measured from long-read sequencing data (Figure S3).

Computing VNTR-SNP linkage disequilibrium: To estimate VNTR-SNP linkage disequilibrium (LD) (Figure 1C, Table S1), we computed the correlation coefficient between "pre-refinement" VNTR genotypes (estimated in individual genomes from WGS depth-of-coverage) and SNP genotypes, and adjusted for the estimated accuracy of VNTR genotypes:

$$\widehat{R^2}(\text{VNTR},\ \text{SNP}) = \frac{R^2(\text{est. prerefinement VNTR genotypes, SNP genotypes})}{\text{IBD2 } R}.$$

We restricted analysis to 3,904 unrelated SSC participants of European descent. We additionally restricted to SNPs within 500kb of the VNTR, excluded variants within the VNTR, and excluded very rare (MAF<0.0005) variants. For VNTRs at *TMCO1*, *EIF3H*, and *CUL4A*, we additionally estimated VNTR-SNP LD using optimized VNTR genotypes and imp_v3 SNPs dosages in N=16,728 UKB participants, obtained by 25x-downsampling the set of 418,136 unrelated, PC-filtered individuals used in our primary analysis. (We did not adjust for VNTR genotype accuracy in UKB). We used these UKB-derived estimates for correlations reported in the main text and to color Manhattan plots (Figures 3B,C, 4A,B, and 5B,D; Figure S4).

Selection of final VNTR list for imputation into UKB: We applied the following set of QC filters to select variants suitable for taking forward for imputation into UKB:

1. IBD2R > 0.1 (in SSC)
2. RDIFF > 0.1 (in SSC)
3. FLANKR < 0.5 (in SSC)
4. Imputation R^2 > 0.1 (in SSC participants of European descent)
5. We excluded variants with the major histocompatibility comples (MHC) locus (chr6:29mb-33mb)

This resulted in the final set of 9,561 multiallelic VNTR loci for analysis in UKB.

VNTR-phenotype association and fine-mapping analyses—We performed association tests between the 9,561 imputed VNTRs and 786 phenotypes (including 118 neurodevelopmental disorders initially analyzed; Figure S1A) in our analysis set of 418,136 unrelated UK Biobank participants of genetically-determined European ancestry (see above). We computed linear regression association statistics using BOLT-LMM[61] v2.3.6 including a standard set of covariates (20 genetic PCs, assessment center, genotyping array, sex, age, and age^2), and found 4,968 VNTR-phenotype pairs that passed a Bonferroni-

corrected significance threshold of $P<5 \times 10^{-9}$ (reflecting the ~10,000 VNTRs x ~1,000 phenotypes we tested for association; Table S3). Linear regression produced well-calibrated *P*-values given that the VNTRs we analyzed exhibited common multiallelic variation and the binary phenotypes we analyzed were not ultra-rare (at least a few hundred cases in UKB[14]).

To determine which of these VNTR-phenotype associations were likely to represent causal effects of VNTR allele length variation (vs. tagging of nearby causal SNPs), we first computed linear regression association statistics for all nearby SNPs and indels imputed by UKB (within 500kb of the VNTR) as we had for the VNTR. We then applied the Bayesian fine-mapping software FINEMAP[21] v1.3.1 (options --corr-config 0.999 --sss --n-causal-snps 5) to estimate the likelihood of causality for the VNTR, accounting for linkage disequilibrium with 500 of the most strongly associated nearby variants. For fine-mapping, we excluded SNPs at multiallelic sites, rare variants (MAF<0.001), and variants called within the VNTR region. For associations for which the VNTR was assigned a high posterior probability of causality (PIP>0.5), we ran a second round of fine-mapping including 2,000 of the most strongly associated nearby variants. The results of these analyses are summarized in Table S3. In total, 107 VNTR-phenotype associations involving 58 distinct VNTRs were assigned a high posterior probability of causality by FINEMAP (PIP>0.5 in both rounds; Table S4, Figure 2).

We additionally tested each VNTR for association with autism directly in SSC. To do so, we computed linear regression association statistics, restricting to the probands and siblings in 1,901 complete quartets and including sex as a covariate. None of the VNTRs we tested reached our study-wide significance threshold ($P<5 \times 10^{-9}$).

For VNTRs at loci of particular interest, we ran additional analyses after improved genotyping and refined phenotyping of disease traits as detailed below (STAR Methods, Optimizing genotyping of VNTRs of particular interest).

Overview of optimized genotyping of VNTRs of particular interest—For each VNTR for which our association analysis and fine-mapping pipeline identified a potentially-causal phenotype association of particular interest (specifically, associations with disease traits and associations of *CUL4A* with erythrocyte traits), we performed follow-up analyses to optimize accuracy of VNTR allele length estimates and verify robustness of results. We did so by analyzing WGS data for *N*=200K UKB participants[48], which became available only after we had completed our initial imputation into UKB from SSC. We used data for sequenced UKB participants as a reference panel for imputation into the remainder of the UK Biobank cohort. Beyond the increased phasing and imputation accuracy afforded by the much larger size of this reference panel (compared to our initial analysis of *N*=8,936 SSC participants), we also obtained further improvements in VNTR genotyping accuracy by developing statistical models tailored to the allele distribution at each locus. Specifically, we incorporated information from 151bp reads that spanned short VNTR alleles (at *TMCO1* and *EIF3H*), and we optimized the selection of reads counted in read-depth-based measurements of VNTR length.

Optimized genotyping of TMCO1 VNTR

Improved TMCO1 VNTR genotyping by combining spanning-read and read-depth information.: The *TMCO1* VNTR has a bimodal allele length distribution, with the 1-repeat allele having high frequency (>0.85) in all continental populations, alleles containing 2 to 4 repeats being very rare, and expanded alleles with ≥5 repeats comprising the remainder of the allele distribution (Figure 3A,D). The repeat unit length of 28bp meant that *TMCO1* VNTR alleles with 1 to 4 repeats were consistently spanned by multiple 151bp reads indicating their presence. (We also searched for evidence of 0-alleles but did not find any evidence that such alleles existed.) While expanded alleles with ≥5 repeats could not be distinguished by single reads, the presence of such an allele could easily be detected based on observations of 151bp reads that partially overlapped the VNTR, and additionally, the lengths of expanded alleles could be estimated by counting the number of reads internal to the VNTR (similar to the read-depth-based strategy we used in initial genotyping, but greatly reducing noise by restricting to within-VNTR reads).

We therefore implemented a hybrid genotyping strategy (similar to the approach we previously used to genotype *TENT5A* alleles from WES[14]) that combined direct read-level information (used to identify short alleles and to detect the presence of expanded alleles) with read-depth information (used to estimate the lengths of longer alleles). Specifically, for each individual, we applied the following procedure:

- Identify the minimum- and maximum-length allele indicated by direct read-level evidence (which could be the same allele, indicating a homozygote).
- If the maximum-length allele indicated has length ≤4, set the individual's (unphased) genotype to be the minimum-length and maximum-length allele.
- Otherwise:
  - If the minimum-length allele has length ≤4 (i.e., the individual is heterozygous for an expanded allele), then estimate the number of repeats in the expanded allele as:

    5 + (# within-VNTR reads) / (# reads in ±5kb flanks) × (calibration factor).
  - Otherwise, estimate the total number of repeats in the two expanded alleles as: 10 + (# within-VNTR reads) / (# reads in ±5kb flanks) × (calibration factor).

Based on empirical analyses of WGS data from SSC, UKB, 1000 Genomes 30x, and GTEx, the calibration factor above that is required to convert read counts to absolute estimates of expanded allele lengths appeared to be data set-specific. We therefore estimated this calibration factor independently for each data set in which we performed analysis using the following approach:

- First, we estimated the calibration factor in the 1000 Genomes 30x data set[62] by identifying 17 heterozygous carriers of expanded alleles with lengths that could be exactly determined from a long-read assembly of either the carrier or

a related individual included in HGSVC2[16] or HPRC[63]. We set the calibration factor for 1000 Genomes 30x to the value that caused the mean estimated length of expanded alleles in these 17 individuals to equal the mean of the exact long-read-derived lengths.

- Next, we estimated mean lengths of expanded alleles in heterozygous carriers in each 1000 Genomes Project continental population (Figure 3A) by applying the calibration factor estimated above to all samples in the 1000 Genomes 30x WGS data set.
- Finally, for each other WGS data set we analyzed (all of which were predominantly EUR-ancestry), we set the calibration factor to the value that caused the mean estimated length of expanded alleles in heterozygous carriers to match the mean expanded allele length we estimated in the previous step for 1000 Genomes EUR participants.

Note that the calibration factor is relevant for genotyping VNTRs from read-depth in a sequenced cohort for the purpose of generating a reference panel, but once the reference panel has been generated, no further calibration is needed to perform imputation (as VNTR genotypes in target samples will just be estimated based on SNP-haplotypes shared with the reference panel). Additionally, calibration error in the reference panel only affects scaling of VNTR length estimates and does not impact downstream association analyses using linear models.

Optimized phasing and imputation of TMCO1 VNTR genotypes.: For each whole-genome-sequenced individual, the above strategy produced a pair of (unphased) allele length estimates with the property that calls of short alleles (≤4 repeats; usually the 1-allele) were discrete and nearly always correct, and detection of expanded alleles (≥5 repeats) was also nearly always correct, but lengths of expanded alleles were only approximately measured (by read-counting). We next needed to phase these estimates onto SNP-haplotypes in order to denoise estimated lengths of expanded alleles (by averaging estimates across individuals with long shared SNP-haplotypes) and to enable imputation into SNP-haplotypes of unsequenced UKB participants. While we could do so using our standard phasing and imputation algorithm (which treated all genotype estimates as continuous, real-valued measurements), the discrete information available here from read-level analysis allowed a simpler, more accurate approach.

To phase each individual's pair of allele length estimates onto the individual's SNP-haplotypes and refine estimates of expanded allele lengths, we did the following:

1. Determine which of the individual's two SNP-haplotypes carries the shorter allele and which SNP-haplotype carries the longer allele. We did so by counting, for each of the target individual's two SNP-haplotypes, how many of the carriers of the top 20 longest SNP-haplotype-matches had read-level support for the target individual's longer allele. We then assigned the target individual's longer allele to the SNP-haplotype with more "votes" from top haplotype matches.

2. Refine the length estimate of each detected expanded allele by taking a weighted average of the allele length estimated in the target individual together with allele lengths estimated in individuals who (i) shared a long SNP-haplotype with the target allele; and (ii) carried exactly one expanded allele (presumably on the shared haplotype). We computed this weighted average using the haplotype-copying probabilities we used in our previous work[14], which are a function of IBS-sharing length and three tunable parameters (K_{top}, ℓ_0, and p_{reg}). We tuned these parameters using a grid search that utilized cross-validation in IBD2 sib-pairs heterozygous for an expanded allele. Specifically, we held out one member of each sib-pair and chose the parameter combination that maximized correlation between held-out estimates of expanded allele lengths and refined estimates of expanded allele lengths in the non-held-out siblings.

To impute VNTR allele lengths into unsequenced individuals, we used the same haplotype-copying model but re-optimized the three parameters to maximize imputation accuracy in cross-validation (using 400 held-out samples).

Validating accuracy of TMCO1 VNTR allele length estimates.: To verify the accuracy of our genotyping strategy at *TMCO1*, we compared VNTR allele lengths we estimated in 1000 Genomes 30x WGS (after phasing together with allele lengths we estimated in UKB $N = 200$K WGS) to allele lengths derived from long-read assemblies of HGSVC2 samples (summing across each individual's two alleles). This comparison demonstrated high accuracy $\left(R^2 = 0.99\right.$; Figure S3A).

Estimating TMCO1 VNTR allele lengths in GTEx.: To estimate unphased TMCO1 allele lengths in GTEx, we used the same strategy of combining spanning-read and read-depth information that we used to analyze the UKB N=200K WGS and 1000 Genomes 30x WGS data, with just one minor difference that arose from 58 GTEx samples having been sequenced using 101bp reads instead of 151bp reads. We could still use read-level information to determine which of these individuals carried expanded (≥5-repeat) alleles, but we did not attempt to use within-VNTR read counts to estimate the lengths of these expanded alleles, instead setting their initial length estimates to the mean expanded allele length. We then phased the allele length estimates in GTEx together with allele lengths estimated in UKB N=200K WGS and 1000 Genomes 30x WGS to maximize accuracy.

Optimized genotyping of EIF3H VNTR

Improved EIF3H VNTR genotyping by modeling read-level information.: Most *EIF3H* VNTR alleles contain 2 to 6 repeats of a 27bp unit followed by a partial repeat unit (13bp). Consequently, alleles with ≤4 full repeats could usually be detected from spanning 151bp reads in the UKB N=200K WGS data set. Alleles with ≥5 repeats were too long to genotype from spanning reads, but reads that partially overlapped the VNTR could be informative of the presence of a ≥5-repeat allele, and reads internal to the VNTR indicated the presence of a ≥6-repeat allele. Altogether, read-level information was thus usually sufficient to deduce a confident (unphased) genotype call in a given individual. However, synthesizing all of this information while accounting for occasional false-positives (i.e., observations of reads

putatively supporting an allele that is not actually present) and false-negatives (i.e., absence of observations of reads supporting an allele that is present) was not straightforward, as we needed to consider how to weigh evidence from counts of reads in seven different categories:

- span1, span2, span3, span4 (i.e., reads spanning VNTR alleles with 1–4 full repeat units)
- flank4+, flank5+ (i.e., reads partially overlapping the VNTR indicating ≥4 or ≥5 repeats)
- internal (i.e., reads completely within the VNTR indicating ≥6 repeats).

We therefore developed a Bayesian genotyping strategy based on a generative model in which we assumed reads from each of the seven categories were generated independently (conditional on an individual's genotype). Letting CN1, CN2 denote the numbers of full repeat copies on the individual's two haplotypes, we assumed:

$$P(\mathrm{CN1}, \mathrm{CN2} \mid \text{ obs. reads }) \propto P(\mathrm{CN1})P(\mathrm{CN2}) \prod_{\text{category}} P\,(\#\text{ reads in category } \mid \mathrm{CN1}, \mathrm{CN2})$$

where for each of the seven categories of reads, we modeled $P\,(\#\text{ reads in category} \mid \mathrm{CN1},\ \mathrm{CN2})$ using a Poisson distribution with

$$\lambda = \lambda_{\#(\mathrm{CN1}, \mathrm{CN2}\text{ contributing to category})} \cdot (\text{local read depth in 5 kb flanks})$$

where the rate parameters λ_0, λ_1, λ_2 are defined as follows:

- λ_0 (neither CN1 nor CN2 should generate reads in the category): estimate based on empirical frequency of observing false-positive reads (in samples with strong evidence that they carry only alleles that should not produce reads in the category)
- λ_1 (exactly one of the two alleles generates reads in the category): estimate based on empirical frequency of observed reads in samples with good evidence that they carry exactly one such allele
- $\lambda_2 = 2\lambda_1$ (both alleles generate reads in the category, so twice as many reads are expected).

After estimating λ_0, λ_1, λ_2 as indicated above, we then used an expectation-maximization (EM) algorithm to estimate the frequencies of alleles with 1 to 6 repeat units to use as priors $P(\mathrm{CN1})$, $P(\mathrm{CN2})$. (We did not observe evidence of 0-repeat alleles, and while analysis of within-VNTR read counts indicated that rare 7-repeat alleles also exist, they are sufficiently rare that modeling them distinctly from 6-alleles was not necessary.)

Optimized phasing and imputation of EIF3H VNTR genotype probabilities.: For each whole-genome-sequenced individual, the above algorithm produced posterior probabilities for each possible genotype $\{\mathrm{CN1}, \mathrm{CN2}\}$ with no information about phase. For most individuals, a single genotype was by far the most likely (with only the phase of

the alleles being unknown), but for some individuals, multiple genotypes had similar posterior probabilities. We therefore leveraged information from shared SNP-haplotypes to help resolve uncertain genotypes and to phase each individual's pair of alleles onto the individual's SNP-haplotypes. We did so by running four iterations of the following algorithm, applied to each individual in turn:

- For each of the individual's two SNP-haplotypes, count how many of the five longest SNP-haplotype matches are believed to carry a 1-allele, 2-allele, …, 6-allele (adding a pseudocount of 0.5 for each allele).
- Adjust the likelihood of each $\{CN1, CN2\}$ genotype by multiplying by the relevant numbers of votes of support from SNP-haplotype-matches.
- Select the $\{CN1, CN2\}$ genotype with highest adjusted likelihood.
- Set the phase of the shorter/longer allele to match the shorter/longer of the mean allele length estimated in the five best matches for each SNP-haplotype.

We imputed VNTR allele lengths into unsequenced individuals using the same approach as at *TMCO1* (again optimizing imputation parameters via cross-validation in 400 held-out samples).

Validating accuracy of EIF3H VNTR genotypes.: To verify the accuracy of our genotyping strategy at *EIF3H*, we compared VNTR genotypes we estimated in 1000 Genomes 30x WGS (after phasing within this cohort) to allele lengths derived from long read assemblies of HGSVC2 samples (summing across each individual's two alleles). This comparison demonstrated high accuracy $\left(R^2 = 0.99\right.$; Figure S3b).

Optimized genotyping of CUL4A VNTR

Estimating CUL4A VNTR allele lengths from WGS read-depth.: To efficiently estimate diploid VNTR content at *CUL4A* in the *N*=200K UKB WGS data release, we counted reads aligning fully within the VNTR region in GRCh38 as well as in 10kb flanks on each side (restricting to reads with SAM flags 0×53, 0×63, 0×93, 0xA3, 0×51, 0×61, 0×91, or 0xA1). The count of flanking reads served as an approximate measure of local sequencing coverage for each sample, allowing us to estimate VNTR allele length (up to a constant calibration factor; see below) as the ratio of the number of within-VNTR reads to the number of flanking reads. To account for the possibility of copy-number variants influencing flanking read counts in a small fraction of samples, we excluded samples with outlier flank read counts (>2.5 s.d. from the mean on a log scale). We then phased these length estimates and imputed into the remainder of the UKB cohort using the same approach as in our previous analysis of UKB *N*=50K WES data[14].

Calibrating CUL4A VNTR allele length estimates.: The above pipeline produced unscaled allele length estimates that were not calibrated to absolute (base pair) lengths. We therefore calibrated *CUL4A* allele length estimates derived from WGS read-depth in UKB by imputing allele lengths from UKB into SNP-haplotypes for 1000 Genomes Project participants[62] and calibrating against allele lengths derived from long-read assemblies in the

HGSVC2 data set[16]. Specifically, we estimated a single scaling factor by regressing long-read-derived allele lengths on WGS-read-depth-derived (imputed) estimates (summed across each individual's two alleles), setting the intercept to 300bp (because only VNTR alleles >150bp can produce 151bp reads that align fully within the VNTR region in GRCh38). We performed this regression using the six EUR individuals included in HGSVC2 (because imputation accuracy was highest in EUR).

Validating accuracy of CUL4A VNTR allele lengths derived from WGS read-depth.: Separately, to verify the accuracy of our WGS read-depth-based approach to measuring *CUL4A* VNTR allele lengths, we subsequently ran the same read-counting pipeline directly on WGS read alignments in the 1000 Genomes 30x data set[62]. We then compared these diploid VNTR content estimates to allele lengths derived from long read assemblies of HGSVC2 samples (summing across each individual's two alleles), observing high concordance $\left(R^2 = 0.97\right.$; Figure S3c).

Optimized genotyping of VNTRs at CHMP1A, INS, and METRNL—Similar to *CUL4A*, we estimated diploid VNTR content at *CHMP1A*, *INS*, and *METRNL* in *N*=200K UKB WGS data by counting reads aligning fully within the VNTR region and dividing by the count of reads aligning to the 10kb flanks on each side. (For each of these four VNTRs, nearly all alleles are >150bp, so counting reads aligning fully within the VNTR – i.e., excluding reads that span its left or right edges – reduces noise.) We again excluded samples with particularly low or high counts of reads aligning to the 10kb flanks, restricting to the middle 95% of the distribution (i.e., excluding samples in the top or bottom 2.5%). We then phased and imputed into the remainder of the UKB cohort as before.

Rerunning the association and fine-mapping analysis using the updated allele length estimates increased confidence in causality for the associations of the *CHMP1A* and *INS* VNTRs with hypertension and type 1 diabetes (FINEMAP posterior probability = 1.00 and 0.91, respectively) but decreased confidence in causality of the association of the *METRNL* VNTR with cataracts (FINEMAP posterior probability = 0.03). These results are reported in Table S4.

Phenotype refinement for disease-associated VNTRs—For the two strongest disease associations we observed, involving VNTRs at *TMCO1* and *EIF3H*, we sought to bolster the statistical evidence of association by: 1) refining the associated disease phenotypes via ICD-10 subcategories; and 2) curating additional, related phenotypes not included in the original set of 786 phenotypes we tested for association.

Glaucoma: We sought to increase power and statistical resolution to interrogate the relationship between the VNTR at *TMCO1* and glaucoma by refining the associated glaucoma phenotype. We initially observed a strong association between the VNTR at *TMCO1* and the glaucoma phenotype curated by UKB, categorized under the ICD-10 code H40. SNPs at *TMCO1* in LD with the VNTR had previously been associated with primary open-angle glaucoma (POAG)[24]. A substantial fraction of glaucoma cases in UKB are classified as primary angle-closure glaucoma (PACG), a disease that has little

etiological overlap with POAG[64]. Therefore, we sought to remove known PACG (ICD-10 code H40.2) from the disease phenotype. To do so, we extracted the ICD-10 codes recorded for diagnoses made during hospital inpatient admissions (UKB data field 41270, accessed via the Research Analysis Platform (RAP) on 06/03/2022). We then curated a new binary glaucoma phenotype, where we included as cases all participants with a glaucoma diagnosis (either in the original UKB-curated phenotype, or a H40 code present in data field 41270), and then removed all participants with a specific diagnosis of PACG (H40.2). Individuals with diagnoses of both POAG (H40.1) and PACG (H40.2) were considered as cases. Among the PC-filtered, unrelated set of 418,136 UKB participants in our primary analysis, we identified a total of 15,334 glaucoma cases, 1,216 of which were classified as PACG (and not POAG), leaving 14,118 cases in our final analysis. We used the resulting glaucoma phenotype for all follow-up analyses, with the exception of the estimation of the overall disease burden of expanded *TMCO1* VNTR alleles, for which we used explicit diagnoses of POAG (H40.1).

Intraocular pressure: We sought independent statistical evidence of the *TMCO1* VNTR's association with glaucoma by analysis of intraocular pressure (IOP), a major risk factor for glaucoma that was measured in ~130K UKB participants but was not in our initial analysis set. We curated a phenotype derived from IOP measurements following the practices of a recent IOP GWAS performed using UKB data[65]. We extracted UKB data fields 5254 and 5262, which recorded measurements of corneal-compensated IOP in the left and right eyes, respectively. Each participant had up to two measurements taken from each eye. We removed outlier measurements (<7 and >30 mmHg, approximately ~1% of all measurements), and averaged the remaining measurements for each participant. We used the resulting IOP phenotype for all association analyses. To assess the effects of specific *TMCO1* VNTR alleles (Figure 3E), we normalized the resulting IOP phenotype by regressing out age, age^2, and sex, and applying a linear transformation to obtain a distribution with mean 0 and standard deviation 1. To minimize confounding from IOP-lowering drugs administered to glaucoma patients, and to ensure the IOP association we observed was statistically independent of the glaucoma association, we excluded all participants with a glaucoma diagnosis from all IOP analyses.

Colon polyps: We sought to increase power and statistical resolution to interrogate the relationship between the VNTR at *EIF3H* by refining the associated phenotype categorized under ICD-10 code K63 (other diseases of the intestine). We extracted the ICD-10 codes recorded from hospital inpatient admissions (UKB data field 41270, accessed via the RAP on 06/03/2022). The majority (77%) of K63 reports were subclassified as K63.5 (colon polyps), and association analyses revealed that K63.5 was the only K63 subcategory that was significantly associated with the *EIF3H* VNTR length. In our final analyses, we analyzed a binary phenotype where cases included only individuals with specific ICD-10 reports of K63.5 in data field 41270 (22,715 cases among the PC-filtered, unrelated set of 418,136 UKB participants in our primary analysis).

Colorectal cancer: Given the strong association between the *EIF3H* VNTR and colon polyps, and previous reports that SNPs near *EIF3H* strongly associated with colorectal

cancer (CRC), we hypothesized that the *EIF3H* VNTR might also associate with CRC. We sought to test this hypothesis in UKB by direct analysis of CRC, a phenotype not included in our original list of 786 phenotypes tested for association. We extracted the ICD-10 codes obtained from UK cancer registries (UKB data field 40006 with 17 instances, accessed via RAP on 06/03/2022). We identified 6,824 participants (out of 418,136 PC-filtered unrelated individuals) with reports of colorectal cancer (ICD-10 codes C18, C19 or C20). Of these CRC cases, *N*=1,988 participants also had a K63.5 diagnosis.

SNP-based corroboration of TMCO1 VNTR length associations with glaucoma and intraocular pressure in independent cohorts—Previous genome-wide association studies of glaucoma and intraocular pressure provided the opportunity to replicate (in part) the allelic series we observed at *TMCO1* in UK Biobank (in which *TMCO1* VNTR alleles of increasing length associated with increasing glaucoma risk and IOP). To do so, we identified SNPs that tagged VNTR alleles of different lengths and then examined their effect sizes in publicly available summary association statistics from the NHGRI-EBI GWAS Catalog[66] for study GCST90011767[26] (glaucoma GWAS, downloaded on 07/15/2022) and GCST009413[27] (IOP GWAS, downloaded on 07/21/2022). These studies did not include UK Biobank data and were thus suitable for independent replication.

For this corroboratory analysis, we sought to identify a SNP that segregated with particularly long VNTR alleles and then compare its effect size to that of rs2790053, a representative of the common risk haplotype (AF=0.12) that segregates with all expanded alleles (5 or more repeats; Figure 1A) – the idea being that a SNP tagging extra-long VNTR alleles should associate with even higher glaucoma risk and IOP than the common risk haplotype that tags a mixture of all long alleles.

To find such a SNP, we examined heterozygous carriers of each SNP within 500kb of the VNTR with MAF between 0.1% and 10% and computed the mean length of the longer allele present in each carrier (which should almost always be the desired allele for SNPs that tag very long VNTR alleles). We performed this analysis using imputed genotypes available for UK Biobank participants of European ancestry (from the UKB imp_v3 data set[52]), reasoning that summary statistics available from previous GWAS rely on similar imputation. We restricted analysis to individuals in the UKB *N*=200K WGS data set and used allele lengths that we estimated via our optimized genotyping of *TMCO1* (STAR Methods, Optimizing genotyping of VNTRs of particular interest).

Upon ranking SNPs in descending order of mean estimated VNTR length in carriers, two low-frequency SNPs – rs35310077 and rs116089225, the top two SNPs on the list – were clearly the best candidates for replication, segregating with VNTR alleles with mean estimated lengths of ~9.7 repeat units. (This length estimate is probably downward-biased by imputation error resulting in regression to the mean; a single carrier of both SNPs in HGSVC2 carried an 11-repeat allele.) Closer inspection of these two SNPs showed that rs35310077 (MAF=0.007 in UKB imp_v3) tagged a sub-haplotype of rs116089225 (MAF=0.009 in UKB imp_v3), with the latter SNP exhibiting higher imputation accuracy (INFO=0.95 for rs116089225 vs. INFO=0.89 for rs35310077). We therefore proceeded with

rs116089225 (and rs2790053, the aforementioned tag SNP for the AF=0.12 common risk haplotype) for lookup in glaucoma and IOP summary statistics.

Comparison of VNTR and SNP associations at TMCO1—At *TMCO1*, the VNTR and multiple SNPs that segregated with expanded VNTR alleles all associated strongly with glaucoma. In this section, we compare the VNTR and SNP associations at this locus, obtaining several lines of evidence pointing to the VNTR as the causal variant:

1. The VNTR was the strongest associated variant at *TMCO1*, with association strength ~16% stronger than that of nearby SNPs in analysis of glaucoma (P=2.8 × 10^{-79} for the VNTR vs. P=2.4 × 10^{-62} for rs2790052) and in analysis of intraocular pressure (IOP) among participants without reported glaucoma (P=3.32 × 10^{-62} for VNTR vs. P=1.66 × 10^{-53} for rs2790052). These analyses included age, age^2, sex, and 20 PCs as covariates.

2. In a joint model including both the VNTR and rs2790052 as regressors, the VNTR remained strongly associated whereas the signal for rs2790052 was greatly diminished (P=3.8 × 10^{-15} for VNTR vs. P=0.000188 for rs2790052 in analysis of glaucoma; P=3.54 × 10^{-13} for VNTR vs. P=0.000316 for rs2790052 in analysis of IOP). The remaining signal for rs2790052 appeared to reflect non-linearity in the effects of VNTR alleles: in a model additionally including VNTR length squared, the rs2790052 association was no longer significant (P=0.09 for glaucoma; P=0.82 for IOP).

3. The VNTR was assigned a PIP of 1.00 by statistical fine-mapping, reflecting an unbiased comparison of arbitrary combinations of up to 5 causal variants at this locus. The 43 analyzed SNPs in high LD with the VNTR (R^2>0.9) were assigned PIPs of at most 0.08. Their combined posterior probability for inclusion in the causal set was more substantial (PIP sum=0.82 for glaucoma and PIP sum=0.6 for IOP); however, this appeared to reflect the ability of these variants to capture the non-linear effects of VNTR alleles. When we additionally included VNTR length squared in our analysis, the combined posterior inclusion probabilities for the 43 SNPs dropped to ~0.01 in analyses of both glaucoma and IOP.

The above results for glaucoma and IOP contrast with analyses of RNA sequencing data from GTEx[30]. VNTR length associated with expression at *TMCO1* in most tissues (e.g., in sun-exposed skin, P=2.9 × 10^{-11} for *TMCO1* expression and P=8.3 × 10^{-26} for *TMCO1-AS1* expression), consistent with recent SNP-based colocalization analyses[31]. However, these associations did not display evidence of an allelic series (Figure S5). Additionally, in joint models including both the VNTR and nearby SNPs, VNTR length did not significantly associate with expression, whereas SNPs retained significance (e.g., P=0.5 for the VNTR vs. P=0.00011 for rs2790052 for association with *TMCO1-AS1* expression in sun-exposed skin). These results suggest that a variant other than the VNTR, possibly rs2790052 or rs2251768 in the promoter region of *TMCO1*, is responsible for the main eQTL at this locus and that the expression signal is unrelated to the glaucoma and IOP associations.

Imputation-based corroboration of EIF3H VNTR length association with colorectal cancer risk—We sought to replicate the association we observed between *EIF3H* VNTR length and colorectal cancer risk by imputing the VNTR's association statistic into SNP association statistics from an independent colorectal cancer GWAS. We employed the approach of ImpG[34], which estimates a variant's association statistic based on the association statistics of variants in LD (using a multivariate normal with covariance derived from the LD matrix). Summary statistics were downloaded from the NHGRI-EBI GWAS Catalog[66] for study GCST012879[32] (accessed on 07/15/2022). We extracted statistics for all SNPs within 300kb of the VNTR that were also present in the UKB imp_v3 data set[52]. We computed LD using VNTR+SNP genotypes for 16,728 UKB participants, obtained by 25x-downsampling the set of 418,136 individuals used in our association analyses. *EIF3H* VNTR genotypes were estimated from refined genotyping using UKB *N*=200K WGS data (STAR Methods, Optimizing genotyping of VNTRs of particular interest). Since the published implementation of ImpG requires variants to be biallelic, we re-implemented the method (using the same default regularization parameter L=0.1) to apply it to continuous-valued VNTR allele length estimates; we previously validated this re-implementation of ImpG[14]. Consistent with our observations in UK Biobank, the imputed association statistic for the VNTR in these independent summary statistics (from Huyghe et al. 2019[32], excluding UKB) was larger than that of any nearby SNP or indel (imputed P=6.7 × 10^{-11} for the VNTR vs. P=7.3 × 10^{-9} for to the top SNP at the *EIF3H* locus).

Comparison to previous VNTR-phenotype association studies—In this section we compare the results of our association and fine-mapping analyses to the lists of associations reported in two recent studies cataloguing VNTR associations with complex traits:

- (Mukamel et al. 2021)[14] – our previous work, which developed the statistical approach used here and applied it to whole-exome sequencing data in UKB
- (Garg et al. 2022)[8] – which estimated VNTR lengths from whole-genome sequencing depth (without leveraging SNP-haplotype information) in TOPMed.

Our previous study[14] identified fine-mapping-supported associations involving five protein-coding VNTRs:

- Two VNTRs, in *ACAN* and *TCHH*, were involved in associations re-identified in our current WGS-based study (Table S4).
- Two more VNTRs, in *MUC1* and *LPA*, were filtered from our current analysis for technical reasons. At *MUC1*, misassembly of the VNTR in GRCh38 led to inaccurate boundary selection in our WGS pipeline, while at *LPA,* misassembly of VNTR alleles in HGSVC2 caused the VNTR to be dropped from analysis. (Had the VNTRs not been filtered, both associations would have been re-identified by the WGS-based approach in this study: P=2.1 × $10^{-14,508}$, PIP=1.00 for *LPA* VNTR association with lipoprotein(a) concentration; P=8.0 × 10^{-151}, PIP=1.00 for *MUC1* VNTR association with serum urea.)

- The final VNTR comprised a shorter repeat within an exon of *TENT5A* that was poorly genotyped from whole-genome sequencing depth-of-coverage (estimated genotyping accuracy R^2=0.019 in SSC; Table S1). Genotyping from whole-exome sequencing in UKB was much more accurate (R^2=0.66), likely reflecting high exome capture at this locus (~800 reads/sample in UKB WES). (Since *TENT5A* alleles are shorter than WGS reads, the association of *TENT5A* VNTR length with height could have been discovered in WGS using methods that consider spanning reads rather than depth-of-coverage.)

Our current study uncovered one additional coding VNTR association, involving a 39bp-repeat within an exon of *GP1BA*. The VNTR was filtered from our previous WES analysis due to low genotyping accuracy (R^2=0.085); in WGS, genotyping accuracy was slightly higher (R^2=0.127) and met our requirement for inclusion in analysis.

To summarize, while WES can occasionally provide more power than WGS for detecting coding VNTR variants (in regions at which WES coverage is drastically higher), most coding VNTRs can be analyzed from either WES or WGS data, and QC filtering choices may have a larger effect on which loci are deemed suitable for analysis. It should also be noted that WES read-depth analyses have the additional complexity of being susceptible to exome capture biases (as detailed in our previous work[14]), such that in situations where both WES and WGS are available, analyzing WGS data is technically much more straightforward.

Garg et al. 2022[8] (Table S3) reported 14 associations that reached Bonferroni significance in their analysis of ~35,000 TOPMed participants. We were initially surprised not to find any overlap with the fine-mapped associations that we reported in this study, but a closer look at the 14 associations explained this observation:

- 9 of the 14 associations involved T-cell receptor loci (*TRA* and *TRG*), where sequencing depth-of-coverage has been observed to primarily reflect T-cell fraction of blood samples due to somatic V(D)J recombination[67]. Given that our pipeline was designed to specifically capture inherited VNTR variation (by using SNP-haplotype-based imputation), we did not expect to observe these associations.
- 3 of the 14 associations involved repeats that were filtered from our analysis owing to being insufficiently polymorphic (≤2 alleles in HGSVC2).
- One association involved a phenotype not measured in UKB (factor VII).

The final association (between a VNTR at chr8:125477978–125478219 and triglycerides) replicated in our analysis, but statistical fine-mapping indicated that the VNTR was unlikely to be causal (PIP=0.00015; Table S3).

Expression and splicing quantitative trait association analyses in GTEx—We performed expression and splicing quantitative trait association analyses using data from the Genotype-Tissue Expression (GTEx) project (V8). The GTEx project analyzed 49 human tissues, measuring DNA and RNA collected from 15,201 biosamples contributed

by 838 post-mortem donors[30]. We estimated VNTR allele lengths for GTEx participants by imputation into WGS-derived SNP genotypes previously phased by GTEx using SHAPEIT2[68] (accessed on 07/22/2021 via the Terra data platform) and, at certain loci, WGS read alignments. Specifically:

- At *CHMP1A* and *SBNO2,* we imputed VNTR allele lengths using the reference VNTR+SNP haplotypes and imputation parameters we obtained from analysis of SSC.
- At *TMCO1*, we adapted the strategy for improved genotyping that we used in UKB, combining read-level information from WGS reads spanning short alleles, counts of WGS reads internal to long alleles, and nearby SNP genotypes (STAR Methods, Optimizing genotyping of VNTRs of particular interest).
- At *EIF3H* and *CUL4A*, we imputed VNTR allele lengths using reference VNTR+SNP haplotypes and imputation parameters we obtained from analysis of N=200K WGS UKB samples (STAR Methods, Optimizing genotyping of VNTRs of particular interest). We imputed into SNP haplotypes that we rephased using Eagle2[59] (--Kpbwt=100000, --pbwtIters=3) using the full UKB cohort (N~487K) as a reference panel, restricting analysis to variants typed on the UKB SNP-array with concordant EUR allele frequencies (absolute difference <0.1). For analyses at *CUL4A* that did not require haplotype-resolved estimates (Figure 5D,F), we estimated the diploid content of the highly polymorphic VNTR directly from depth-of-coverage of aligned whole-genome sequencing reads (STAR Methods, Optimizing genotyping of VNTRs of particular interest), providing estimates which sibling-derived benchmarks in SSC indicated were more accurate than imputed values.

To quantify association strengths with expression and splicing quantitative traits, we emulated the analyses performed by GTEx[30]: we obtained normalized expression and splicing quantitative trait phenotypes, as well as covariates, from the GTEx Portal (https://gtexportal.org/home/datasets (V8) accessed on 08/01/2021 and 09/08/2021), and used the software fastQTL[69] v2.0 to compute linear regression VNTR association statistics. We conducted statistical fine-mapping analysis to filter significant VNTR associations (P<1 × 10^{-10}) that likely reflect LD with nearby causal variants. At each quantitative trait locus, we analyzed SNPs (excluding those within the VNTR) with a significant association reported by GTEx (obtained from https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_eQTL.tar on 08/17/2021 and https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_sQTL.tar on 09/08/2021). We computed association statistics and LD for all variants (SNPs and the VNTR) and ran FINEMAP v1.4.1 (with option –corr-config 0.999). At most loci, we allowed up to 5 causal variants; at loci with 2–10 associated SNPs, we only allowed 2 causal variants, and at loci with only 1 associated SNP we only allowed 1 causal variant. At loci with no reported SNP associations, we assigned the VNTR a PIP of 1.00.

We quantified (unnormalized) alternative splice usage at *CUL4A* (Figure 5E,G) using the intron excision ratio

$$\frac{N_{alt}}{N_{alt} + N_{can}},$$

where N_{alt} and N_{can} are counts of reads from LeafCutter[50] (accessed on 12/12/2021 from Terra) for reads supporting excision of the alternative intron (chr13:113229519–113229642) and the canonical intron (chr13:113229519–113233177), respectively (Figure 5A). We similarly computed alternative splice usage (Figure 7 and Data S1) using LeafCutter counts for alternative introns (indicated in red in the figures) and canonical introns (indicated in blue). We obtained tissue-level estimates of median *CUL4A* expression (Figure 5G) from the GTEx Portal (accessed from https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz on 07/08/2022).

Expression quantitative trait association analyses in TCGA—We performed expression quantitative trait association analysis in colorectal cancer tissue using data from the Cancer Genome Atlas (TCGA; Cancer Genome Atlas Network 2012). We obtained phased SNP data from the National Cancer Institute's Genomic Data Commons (GDC, http://gdc.cancer.gov/, accessed on 03/31/2022), previously generated[70] using the TOPMed imputation server. For VNTR imputation, we used VNTR+SNP reference haplotypes and parameters obtained from analysis of *EIF3H* in UKB N=200K WGS samples (STAR Methods, Optimizing genotyping of VNTRs of particular interest). We obtained gene expression quantification, measured in fragments per kilobase per million mapped reads (FPKM), in colorectal cancer tissue derived from RNA-sequencing via the GDC portal (http://portal.gdc.cancer.gov/, accessed on 08/23/2022; files matching search terms primary site=colon or rectum, program=TCGA, sample type=primary tumor, workflow=STAR – Counts, data category=transcriptome profiling, data format=tsv, and data type=gene expression quantification). We performed linear regression association tests to quantify the strength of association between imputed VNTR lengths and expression of each of 57,597 transcripts expressed in colorectal cancer cells (465 samples with imputed VNTR lengths and expression data available). For each of 8 genes that reside within 1Mb of the VNTR, we ran an additional analysis including somatic copy number as a covariate.

Methylation association analyses in GTEx and TCGA—We performed methylation quantitative trait analyses using data from the Enhancing GTEx (eGTEx) project. We obtained DNAm values that were measured and background-adjusted using the single sample normal-exponential out-of-band (ssnoob) method with dye bias correction (gs://fc-secure-ba7a45c3-e08a-4a09–834b-d6707eef6b96/GTEx_v9_EPIC_data/noob_final_BMIQ_all_tissues_987.txt.gz, accessed on 3/17/23)[36]. For each of 754,119 measured sites with available methylation data, we tested inverse-normal transformed DNAm values for association with imputed *EIF3H* VNTR lengths using linear regression, restricting analysis to data from 189 transverse colon samples and adjusting for technical and biological covariates curated by eGTEx (https://gtexportal.org/home/datasets, accessed on 3/17/23)[36]. The only association that reached Bonferroni significance ($P<5 \times 10^{-8}$) involved measurements at 8:117635401 (hg19; 252bp away from the VNTR which spans

8:117635054–117635149; P=7.05 × 10^{-9}). We additionally tested measurements at this site for association with *EIF3H* VNTR length in the 8 other tissues with available DNAm measurements.

Assessing the impact of LOF variants near TMCO1 on glaucoma and IOP—We identified variants predicted to cause loss-of-function (pLoF) of genes near *TMCO1* using data derived from whole-exome sequencing of 454,787 UK Biobank participants[29]. We first extracted a set of pLoF variants for each gene, selecting variants previously annotated as "LoF" by SnpEff. We then used plink2[71] to extract carriers of each pLoF variant from the 450k interim release of population level exome OQFE variants derived from WES. We computed effects on glaucoma risk via logistic regression (including age, age^2, sex and 20 PCs as covariates) and effects on IOP by taking the phenotypic mean among carriers (adjusted for the same covariates) (Figure S2).

Estimating contributions of VNTRs to trait heritability—To quantify the relative importance of VNTRs vs. SNPs in shaping complex traits, we determined the fraction of GWAS loci for blood cell traits at which a VNTR was the lead variant. This approach has previously been used to provide a rough estimate of the contribution of structural variants to heritability[45,72], the idea being that even though lead variants are not always causal, misattribution to one class or the other should approximately wash out. (We also considered using LD score regression but determined that this approach would not provide sufficient statistical resolution.)

We applied this approach to blood cell traits in UK Biobank, which were the most suitable traits for this assessment given their high heritability, polygenicity, and phenotyping rate. For each of the 31 blood cell traits in our analysis set (phenotypes beginning with "blood" in Table S2), we computed linear regression association statistics for all MAF>0.01 imputed SNPs, indels, and VNTRs genome-wide as described above. VNTRs comprised ~0.1% of all variants tested (9,561/9,825,142) and were involved in ~0.1% of significant associations (2,318/2,247,970).

We then iteratively generated a list of lead variants for each trait by, at each stage, adding to the list the strongest associated variant (SNP or VNTR) that was >500kb away from other variants already on the list (using the same P<5 × 10^{-9} threshold we used in our primary VNTR-phenotype association analyses). VNTRs comprised 0.75% (90/11,858) of all lead variants.

We similarly determined the fraction of *cis*-eQTL/sQTL loci for which a VNTR was the lead variant. At each locus with either a VNTR association discovered in our analysis or a SNP association reported by GTEx reaching the significance threshold for our analysis of gene regulation traits (P<1 × 10^{-10} in a single-tissue analysis), we compared the strength of the strongest VNTR association with the strongest SNP association. The lead variant was a VNTR at 0.80% (1,459/181,627) of eQTLs and 0.54% (1,096/199,866) of sQTLs.

For glaucoma and colorectal cancer, we additionally estimated the contributions of the VNTRs at *TMCO1* and *EIF3H* to heritability attributable to GWAS loci identifiable in UK

Biobank. To do so, we computed linear regression association statistics for all MAF>0.01 SNPs, excluding multiallelic sites. We then used plink (v1.9, option –clump) to clump SNPs that reached canonical genomewide significance ($P<5 \times 10^{-8}$). For glaucoma, we compared the explanatory power (R^2) of linear regression models fit with and without the *TMCO1* VNTR, including all clump representatives except SNPs at *TMCO1* as additional regressors. The model that included the VNTR explained, on the observed scale, h^2_{GWAS}=0.815% compared to 0.725% for the model without the VNTR, such that the VNTR explained ~11% of h^2_{GWAS}. In a similar analysis for colorectal cancer and the *EIF3H* VNTR, the model including the VNTR explained h^2_{GWAS}=0.142% compared to 0.114% for the model excluding the VNTR, such that the VNTR explained ~20% of h^2_{GWAS}.

Improvement in colorectal cancer risk prediction by incorporating EIF3H VNTR—To evaluate the potential benefit of including the *EIF3H* VNTR in a polygenic risk score (PRS) for colorectal cancer (CRC), we compared scores computed using all CRC GWAS SNPs with scores computed using the VNTR and GWAS SNPs, excluding SNPs near *EIF3H* (117.5–117.7Mb in hg19). To ensure that estimates of model accuracy are not biased by overfitting, we used 5-fold cross validation. Namely, we partitioned the analysis set of 418,136 unrelated individuals of European descent into 5 non-overlapping test sets. For each test set:

1. We computed genome-wide association statistics using BOLT-LMM[61] v2.3.6, including our standard set of covariates (20 genetic PCs, assessment center, genotyping array, sex, age, and age^2) and restricting analysis to individuals in the training set (i.e., the complement of the test set).
2. We used plink (v1.9, option –clump) to clump all MAF>0.01 SNPs at biallelic sites that reached genome-wide significance ($P<5 \times 10^{-8}$).
3. We used linear regression to estimate parameters for two models of CRC (measured on the observed scale and residualized for age, age^2, sex, and 20 PCs): i) a model that included all clump representatives included as regressors; and ii) a model that included the VNTR and all clump representatives, excluding clump representatives near *EIF3H*. The covariates (age, age^2, sex, and 20 PCs) were included as covariates in model fitting, and parameters were estimated using data from the training set.
4. We estimated CRC in the test set from genotypes only (SNPs and SNPs+VNTR) using parameters estimated in (3).

We compared predictions merged across all folds with CRC measurements (on the observed scale and residualized for covariates). Predictions of the SNPs-only model achieved an R^2=0.00049, compared to R^2=0.00062 for the model that incorporated VNTR genotypes, such that inclusion of the VNTR increased risk prediction accuracy by ~25%.

QUANTIFICATION AND STATISTICAL ANALYSIS

Details of exact analyses, statistical tests, and tools can be found in the main text and STAR Methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank A. Segre, A. Lee, N. Kamitaki, and R. Gupta for helpful discussions related to this work. This research was conducted using the UK Biobank Resource under application #40709. Computational analyses were performed on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School (http://rc.hms.harvard.edu) and the UK Biobank Research Analysis Platform (RAP). We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to genetic data on SFARI Base. The results presented here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. 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. R.E.M. was supported by US National Institutes of Health (NIH) grant K25 HL150334. R.E.H. and S.A.M. were supported by NIH grant R01 HG006855. M.A.S. was supported by the MIT John W. Jarve (1978) Seed Fund for Science Innovation and NIH fellowship F31 MH124393. A.R.B. was supported by NIH fellowship F31 HL154537 and training grant T32 HG 2295-16. M.L.A.H. was supported by US NIH Fellowship F32 HL160061. P.-R.L. was supported by NIH grant DP2 ES030554, a Burroughs Wellcome Fund Career Award at the Scientific Interfaces, the Next Generation Fund at the Broad Institute of MIT and Harvard, and a Sloan Research Fellowship.

References

1. Lalioti MD, Scott HS, Buresi C, Rossier C, Bottani A, Morris MA, Malafosse A, and Antonarakis SE (1997). Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. Nature 386, 847–851. 10.1038/386847a0. [PubMed: 9126745]
2. Wijmenga C, Hewitt JE, Sandkuijl LA, Clark LN, Wright TJ, Dauwerse HG, Gruter A-M, Hofker MH, Moerer P, Williamson R, et al. (1992). Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. Nat. Genet. 2, 26–30. 10.1038/ng0992-26. [PubMed: 1363881]
3. Course MM, Gudsnuk K, Smukowski SN, Winston K, Desai N, Ross JP, Sulovari A, Bourassa CV, Spiegelman D, Couthouis J, et al. (2020). Evolution of a Human-Specific Tandem Repeat Associated with ALS. Am. J. Hum. Genet. 107, 445–460. 10.1016/j.ajhg.2020.07.004. [PubMed: 32750315]
4. Bakhtiari M, Park J, Ding Y-C, Shleizer-Burko S, Neuhausen SL, Halldórsson BV, Stefánsson K, Gymrek M, and Bafna V (2021). Variable number tandem repeats mediate the expression of proximal genes. Nat. Commun. 12, 2075. 10.1038/s41467-021-22206-z. [PubMed: 33824302]
5. Eslami Rasekh M, Hernández Y, Drinan SD, Fuxman Bass JI, and Benson G (2021). Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. Nucleic Acids Res. 49, 4308–4324. 10.1093/nar/gkab224. [PubMed: 33849068]
6. Garg P, Martin-Trujillo A, Rodriguez OL, Gies SJ, Hadelia E, Jadhav B, Jain M, Paten B, and Sharp AJ (2021). Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. Am. J. Hum. Genet. 108, 809–824. 10.1016/j.ajhg.2021.03.016. [PubMed: 33794196]
7. Lu T-Y, Human Genome Structural Variation Consortium T, and Chaisson MJP (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. Nat. Commun. 12, 4250. 10.1038/s41467-021-24378-0. [PubMed: 34253730]
8. Garg P, Jadhav B, Lee W, Rodriguez OL, Martin-Trujillo A, and Sharp AJ (2022). A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. Am. J. Hum. Genet. 109, 1065–1076. 10.1016/j.ajhg.2022.04.016. [PubMed: 35609568]

9. Marchini J, Howie B, Myers S, McVean G, and Donnelly P (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913. 10.1038/ng2088. [PubMed: 17572673]
10. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, and McCarroll SA (2015). Large multiallelic copy number variations in humans. Nat. Genet. 47, 296–303. 10.1038/ng.3200. [PubMed: 25621458]
11. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, 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. 10.1038/nature16549. [PubMed: 26814963]
12. Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, Hirschhorn JN, and McCarroll SA (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat. Genet. 48, 359–366. 10.1038/ng.3510. [PubMed: 26901066]
13. Saini S, Mitra I, Mousavi N, Fotsing SF, 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. [PubMed: 30353011]
14. Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, and Loh P-R (2021). Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science 373, 1499–1505. 10.1126/science.abg8289. [PubMed: 34554798]
15. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet, 1–8. 10.1038/s41588-021-00865-4. [PubMed: 33414547]
16. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372. 10.1126/science.abf7117.
17. Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580. 10.1093/nar/27.2.573. [PubMed: 9862982]
18. Li H (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. 10.1093/bioinformatics/bty191. [PubMed: 29750242]
19. Fischbach GD, and Lord C (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron 68, 192–195. 10.1016/j.neuron.2010.10.006. [PubMed: 20955926]
20. An J-Y, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science 362, eaat6576. 10.1126/science.aat6576. [PubMed: 30545852]
21. Benner C, Spencer CCA, Havulinna AS, 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. [PubMed: 26773131]
22. De Roeck A, Duchateau L, Van Dongen J, Cacace R, Bjerke M, Van den Bossche T, Cras P, Vandenberghe R, De Deyn PP, Engelborghs S, et al. (2018). An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. Acta Neuropathol. (Berl.) 135, 827–837. 10.1007/s00401-018-1841-z. [PubMed: 29589097]
23. Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, et al. (1995). Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat. Genet. 9, 284–292. 10.1038/ng0395-284. [PubMed: 7773291]
24. Burdon KP, Macgregor S, Hewitt AW, Sharma S, Chidlow G, Mills RA, Danoy P, Casson R, Viswanathan AC, Liu JZ, et al. (2011). Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. Nat. Genet. 43, 574–578. 10.1038/ng.824. [PubMed: 21532571]
25. Steinmetz JD, Bourne RRA, Briant PS, Flaxman SR, Taylor HRB, Jonas JB, Abdoli AA, Abrha WA, Abualhasan A, Abu-Gharbieh EG, et al. (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020:

the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob. Health 9, e144–e160. 10.1016/S2214-109X(20)30489-7. [PubMed: 33275949]

26. Gharahkhani P, Jorgenson E, Hysi P, Khawaja AP, Pendergrass S, Han X, Ong JS, Hewitt AW, Segrè AV, Rouhana JM, et al. (2021). Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. Nat. Commun. 12, 1258. 10.1038/s41467-020-20851-4. [PubMed: 33627673]
27. Bonnemaijer PWM, Leeuwen E.M. van, Iglesias AI, Gharahkhani P, Vitart V, Khawaja AP, Simcoe M, Höhn R, Cree AJ, Igo RP, et al. (2019). Multi-trait genomewide association study identifies new loci associated with optic disc parameters. Commun. Biol. 2, 1–12. 10.1038/s42003-019-0634-9. [PubMed: 30740537]
28. Sharma S, Burdon KP, Chidlow G, Klebe S, Crawford A, Dimasi DP, Dave A, Martin S, Javadiyan S, Wood JPM, et al. (2012). Association of Genetic Variants in the TMCO1 Gene with Clinical Parameters Related to Glaucoma and Characterization of the Protein in the Eye. Invest. Ophthalmol. Vis. Sci. 53, 4917–4925. 10.1167/iovs.11-9047. [PubMed: 22714896]
29. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature 599, 628–634. 10.1038/s41586-021-04103-z. [PubMed: 34662886]
30. Aguet F, Barbeira AN, Bonazzola Rodrigo, Brown A, and Castel SE (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330. 10.1126/science.aaz1776. [PubMed: 32913098]
31. Hamel AR, Rouhana JM, Yan W, Monovarfeshani A, Jiang X, Liang Q, Mehta PA, Wang J, Shrivastava A, Duchinski K, et al. (2022). Integrating genetic regulation and single-cell expression with GWAS prioritizes causal genes and cell types for glaucoma. medRxiv. 10.1101/2022.05.14.22275022.
32. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, Conti DV, Qu C, Jeon J, Edlund CK, et al. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nat. Genet. 51, 76–87. 10.1038/s41588-018-0286-6. [PubMed: 30510241]
33. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, et al. (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat. Genet. 40, 623–630. 10.1038/ng.111. [PubMed: 18372905]
34. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, and Price AL (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics 30, 2906–2914. 10.1093/bioinformatics/btu416. [PubMed: 24990607]
35. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, Kovar CL, Lewis LR, Morgan MB, Newsham IF, et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337. 10.1038/nature11252. [PubMed: 22810696]
36. Oliva M, Demanelis K, Lu Y, Chernoff M, Jasmine F, Ahsan H, Kibriya MG, Chen LS, and Pierce BL (2023). DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. Nat. Genet. 55, 112–122. 10.1038/s41588-022-01248-z. [PubMed: 36510025]
37. Carvajal-Carmona LG, Cazier J-B, Jones AM, Howarth K, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J, et al. (2011). Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum. Mol. Genet. 20, 2879–2888. 10.1093/hmg/ddr190. [PubMed: 21531788]
38. Waning DL, Li B, Jia N, Naaldijk Y, Goebel WS, HogenEsch H, and Chun KT (2008). Cul4A is required for hematopoietic cell viability and its deficiency leads to apoptosis. Blood 112, 320–329. 10.1182/blood-2007-11-126300. [PubMed: 18339895]
39. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017, bax028. 10.1093/database/bax028. [PubMed: 28605766]

40. Barton AR, Sherman MA, Mukamel RE, and Loh P-R (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. Nat. Genet, 1–10. 10.1038/s41588-021-00892-1. [PubMed: 33414547]
41. Svitkina TM, Verkhovsky AB, and Borisy GG (1996). Plectin sidearms mediate interaction of intermediate filaments with microtubules and other components of the cytoskeleton. J. Cell Biol. 135, 991–1007. 10.1083/jcb.135.4.991. [PubMed: 8922382]
42. Ioka RX, Kang M-J, Kamiyama S, Kim D-H, Magoori K, Kamataki A, Ito Y, Takei YA, Sasaki M, Suzuki T, et al. (2003). Expression Cloning and Characterization of a Novel Glycosylphosphatidylinositol-anchored High Density Lipoprotein-binding Protein, GPI-HBP1 *. J. Biol. Chem. 278, 7344–7349. 10.1074/jbc.M211932200. [PubMed: 12496272]
43. Miura Y, Miura M, Gronthos S, Allen MR, Cao C, Uveges TE, Bi Y, Ehirchiou D, Kortesidis A, Shi S, et al. (2005). Defective osteogenesis of the stromal stem cells predisposes CD18-null mice to osteoporosis. Proc. Natl. Acad. Sci. 102, 14022–14027. 10.1073/pnas.0409397102. [PubMed: 16172402]
44. Maruyama K, Uematsu S, Kondo T, Takeuchi O, Martino MM, Kawasaki T, and Akira S (2013). Strawberry notch homologue 2 regulates osteoclast fusion by enhancing the expression of DC-STAMP. J. Exp. Med. 210, 1947–1960. 10.1084/jem.20130512. [PubMed: 23980096]
45. Scott AJ, Chiang C, and Hall IM (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res. 31, 2249–2257. 10.1101/gr.275488.121. [PubMed: 34544830]
46. Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, Cotsapas C, Cassa CA, and Sunyaev SR (2022). The missing link between genetic association and regulatory function. eLife 11, e74970. 10.7554/eLife.74970. [PubMed: 36515579]
47. Margoliash J, Fuchs S, Li Y, Massarat A, Goren A, and Gymrek M (2022). Polymorphic short tandem repeats make widespread contributions to blood and serum traits. bioRxiv. 10.1101/2022.08.01.502370.
48. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, Palsson G, Hardarson MT, Oddsson A, Jensson BO, et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature 607, 732–740. 10.1038/s41586-022-04965-x. [PubMed: 35859178]
49. "All of Us" Research Program Investigators (2019). The "All of Us" Research Program. N. Engl. J. Med. 381, 668–676. [PubMed: 31412182]
50. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, and Pritchard JK (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. 50, 151–158. 10.1038/s41588-017-0004-9. [PubMed: 29229983]
51. 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. 10.1371/journal.pmed.1001779. [PubMed: 25826379]
52. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, 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. 10.1038/s41586-018-0579-z. [PubMed: 30305743]
53. Loh P-R, Kichaev G, Gazal S, Schoech AP, and Price AL (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. 50, 906–908. 10.1038/s41588-018-0144-6. [PubMed: 29892013]
54. Chen C-Y, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, and Price AL (2013). Improved ancestry inference using weights from external reference panels. Bioinformatics 29, 1399–1406. 10.1093/bioinformatics/btt144. [PubMed: 23539302]
55. 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, 1709–1719. 10.1101/gr.235119.118. [PubMed: 30352806]
56. Dolzhenko E, Vugt J.J.F.A. van, Shaw RJ, Bekritsky MA, Blitterswijk M. van, Narzisi G, Ajay SS, Rajan V, Lajoie BR, Johnson NH, et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 10.1101/gr.225672.117.

57. Course MM, Sulovari A, Gudsnuk K, Eichler EE, and Valdmanis PN (2021). Characterizing nucleotide variation and expansion dynamics in human-specific variable number tandem repeats. Genome Res. 31, 1313–1324. 10.1101/gr.275560.121. [PubMed: 34244228]
58. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, 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. 10.1038/ncomms9111. [PubMed: 26368830]
59. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48, 1443–1448. 10.1038/ng.3679. [PubMed: 27694958]
60. Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. 10.1093/bioinformatics/btq033. [PubMed: 20110278]
61. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 47, 284–290. 10.1038/ng.3190. [PubMed: 25642633]
62. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 185, 3426–3440.e19. 10.1016/j.cell.2022.08.004. [PubMed: 36055201]
63. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. (2023). A draft human pangenome reference. Nature 617, 312–324. 10.1038/s41586-023-05896-x. [PubMed: 37165242]
64. Wiggs JL, and Pasquale LR (2017). Genetics of glaucoma. Hum. Mol. Genet. 26, R21–R27. 10.1093/hmg/ddx184. [PubMed: 28505344]
65. Khawaja AP, Cooke Bailey JN, Wareham NJ, Scott RA, Simcoe M, Igo RP, Song YE, Wojciechowski R, Cheng C-Y, Khaw PT, et al. (2018). Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. Nat. Genet. 50, 778–782. 10.1038/s41588-018-0126-8. [PubMed: 29785010]
66. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012. 10.1093/nar/gky1120. [PubMed: 30445434]
67. Bentham R, Litchfield K, Watkins TBK, Lim EL, Rosenthal R, Martínez-Ruiz C, Hiley CT, Bakir MA, Salgado R, Moore DA, et al. (2021). Using DNA sequencing data to quantify T cell fraction and therapy response. Nature 597, 555–560. 10.1038/s41586-021-03894-5. [PubMed: 34497419]
68. Delaneau O, Howie B, Cox AJ, Zagury J-F, and Marchini J (2013). Haplotype Estimation Using Sequencing Reads. Am. J. Hum. Genet. 93, 687–696. 10.1016/j.ajhg.2013.09.002. [PubMed: 24094745]
69. Ongen H, Buil A, Brown AA, Dermitzakis ET, and Delaneau O (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 32, 1479–1485. 10.1093/bioinformatics/btv722. [PubMed: 26708335]
70. Sayaman RW, Saad M, Thorsson V, Hu D, Hendrickx W, Roelands J, Porta-Pardo E, Mokrab Y, Farshidfar F, Kirchhoff T, et al. (2021). Germline genetic contribution to the immune landscape of cancer. Immunity 54, 367–386.e8. 10.1016/j.immuni.2021.01.011. [PubMed: 33567262]
71. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4. 10.1186/s13742-015-0047-8.
72. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, et al. (2017). The impact of structural variation on human gene expression. Nat. Genet. 49, 692–699. 10.1038/ng.3834. [PubMed: 28369037]

Highlights

- Haplotype-informed analysis accurately genotypes many tandem repeat polymorphisms
- Hundreds of repeat polymorphisms influence complex human traits and gene expression
- Repeat expansion at *TMCO1* generates the genome's strongest association with glaucoma
- Repeat polymorphism at *EIF3H* associates with twofold range of colorectal cancer risk

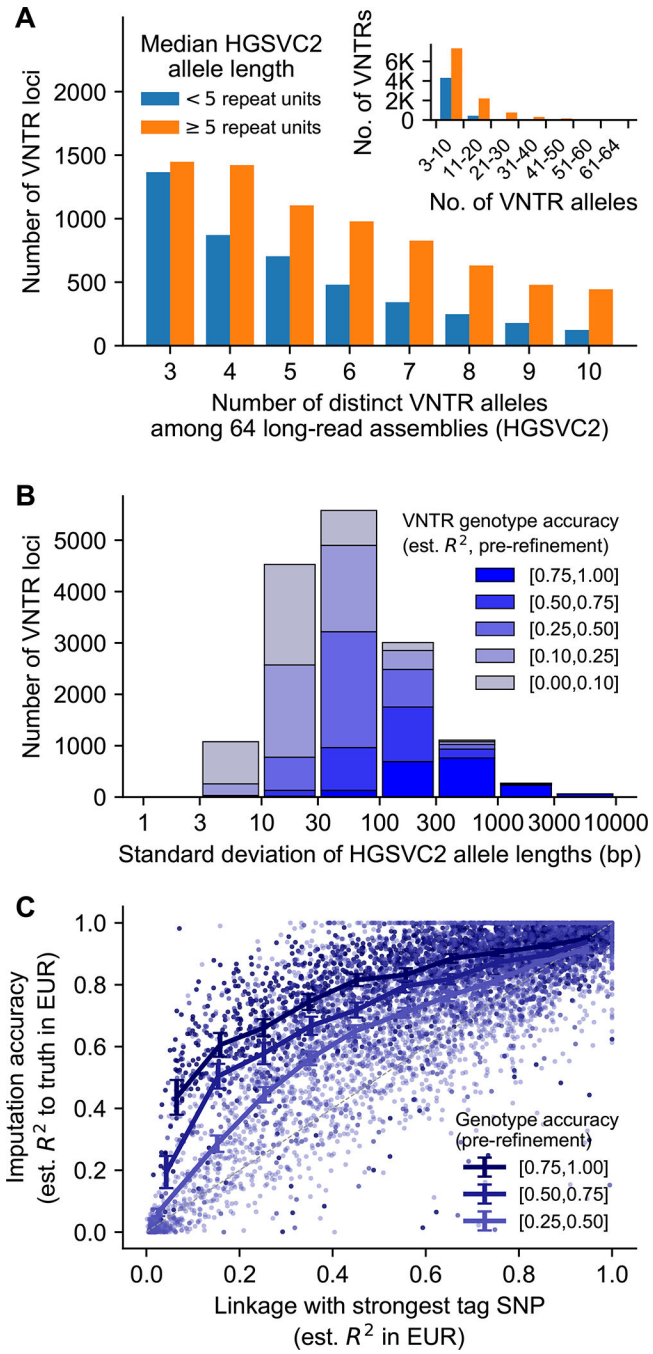


Figure 1. Ascertainment, genotyping and imputation of 15,653 multiallelic VNTR loci.

A) Counts of VNTR loci stratified by number of distinct alleles observed among *N*=64 long-read haploid genome assemblies from HGSVC2 (x-axis) and the median number of repeats per allele (blue/orange bars). Inset, same counts binned at coarser scale. **B)** Counts of VNTR loci stratified by HGSVC2 allele length distribution width (standard deviation) and estimated accuracy of VNTR genotypes pre-refinement (i.e., measured from WGS depth-of-coverage in individual genomes; STAR Methods). **C)** Scatter of imputation accuracy vs. level of linkage disequilibrium with the best tag SNP for each VNTR. Color indicates

pre-refinement genotype accuracy as in b); VNTRs with noisy estimates of imputation accuracy due to low pre-refinement genotype accuracy (R^2<0.25) were omitted, leaving N=7,145 VNTRs for plotting. Lines represent mean imputation accuracy at loci binned by level of linkage with SNPs. Error bars, 95% CIs; EUR, European-ancestry; est., estimated.

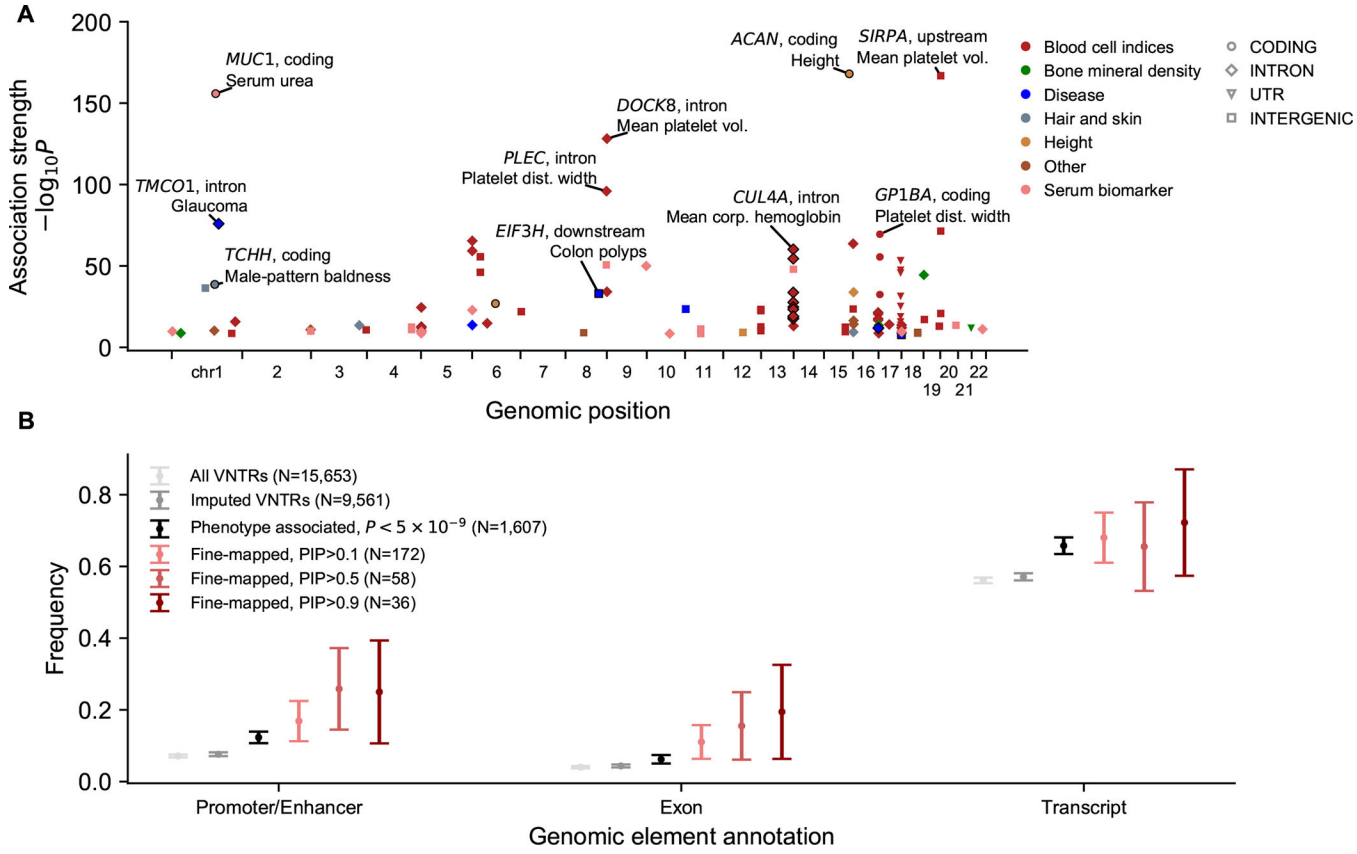


Figure 2. Phenome-wide association and statistical fine-mapping analyses identify 58 VNTRs linked to complex traits.

A) Manhattan plot displaying 107 VNTR-phenotype associations (involving 58 distinct VNTRs) that reached Bonferroni significance ($P<5 \times 10^{-9}$) and for which the VNTR was assigned a high posterior probability of causality by FINEMAP (PIP>0.5). Marker color indicates phenotype category, and marker shape indicates genic context. Outlined markers indicate associations for which we improved VNTR genotyping or refined the associated phenotype (Table S4). For context, the plot also includes two associations to protein-coding VNTRs (at *MUC1* and *TENT5A*[14]) that we previously identified in analysis of whole-exome sequencing data. **B)** Frequency of VNTR overlap with GeneHancer[39] annotated promoters and enhancers (left), GENCODE (v26) exons (middle), and GENCODE (v26) transcripts (right) for VNTRs grouped by association and fine-mapping status. Error bars, 95%CIs.

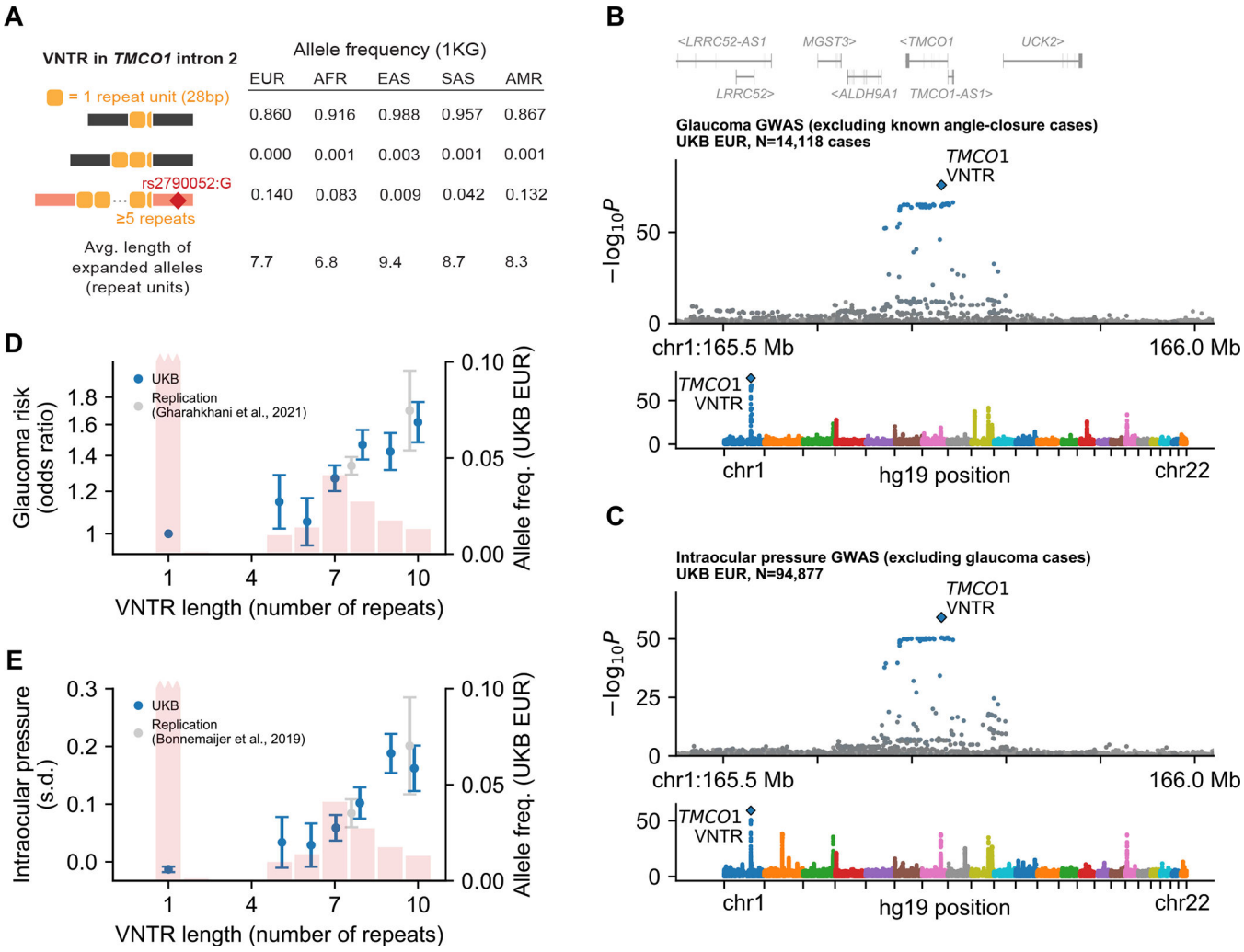


Figure 3. An intronic repeat expansion within *TMCO1* associates with glaucoma risk and intraocular pressure.

A) Frequencies of the 1-, 2-, and ≥5 repeat unit alleles in each of the continental populations represented in the 1000 Genomes Project. Expanded alleles (5–11 repeat units) segregated with a ~70kb SNP haplotype (red) represented by rs2790052:G. Each allele in HGSVC2 also contains a partial repeat (7bp of the 28bp unit) depicted in the haplotype diagrams. **B,C)** Associations of SNPs and VNTR with glaucoma (B) and intraocular pressure (C). SNP and VNTR associations are shown at the *TMCO1* locus (top) and genome-wide (bottom). Colored markers in locus plots, variants in partial LD with the VNTR (R^2>0.01). **D,E)** Effect sizes of VNTR alleles for glaucoma risk (D, left axis) and mean intraocular pressure in carriers of each allele (E, left axis). Values in UK Biobank are shown in blue; values inferred based on SNP associations in independent replication cohorts are shown in gray (STAR Methods). Histograms (right axis), frequencies of VNTR allele lengths estimated in European-ancestry UKB participants. Error bars, 95% CIs.

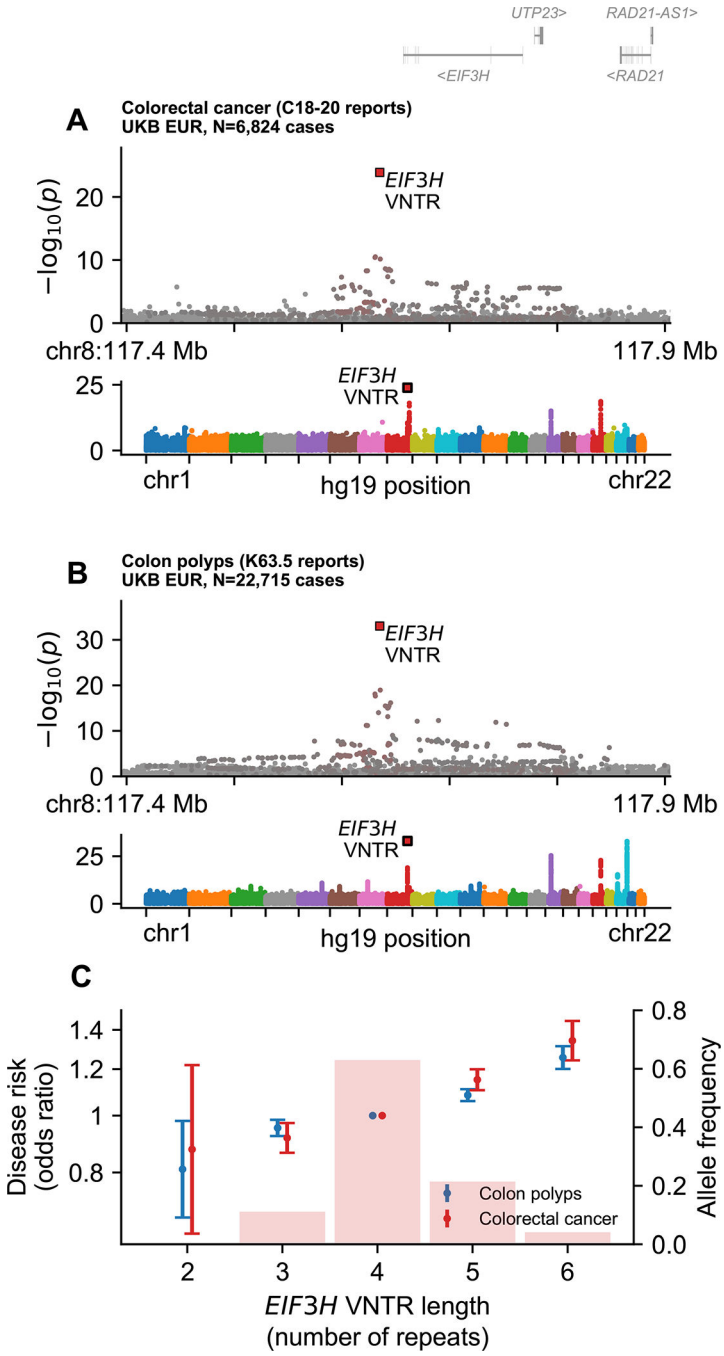


Figure 4. A repeat expansion downstream of *EIF3H* associates with colorectal cancer risk and colon polyps.

A,B) Associations of inherited variants with colorectal cancer (A) and colon polyps (B) at the *EIF3H* locus (top) and genomewide (bottom). Colored markers in locus plots, variants in partial LD with the VNTR (R^2>0.01). **C)** Frequencies of VNTR alleles observed in European-ancestry UKB participants (histogram, right axis) and their effect sizes (markers, left axis) for colorectal cancer (red) and colon polyps (blue). Error bars, 95% CIs.

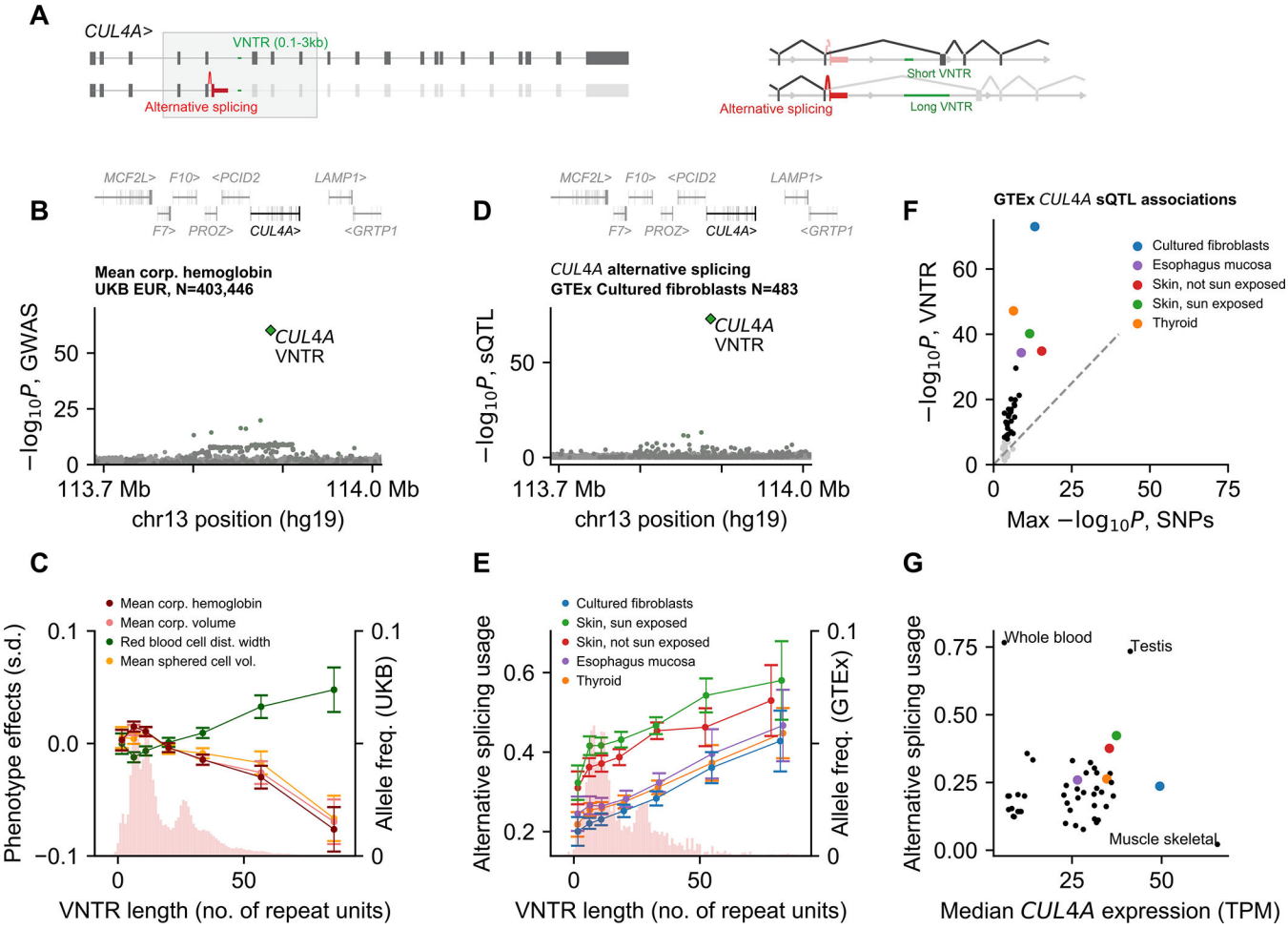


Figure 5. An intronic repeat expansion at *CUL4A* associates with erythrocyte traits and splice isoform usage.

A) Alternative splicing of two commonly expressed *CUL4A* isoforms. The fifth intron of the canonical transcript contains a highly length-polymorphic VNTR (0.1–3kb, green). The image on the right is zoomed in on the region of *CUL4A* containing the alternative splice. **B)** VNTR and SNP associations with mean corpuscular hemoglobin at the *CUL4A* locus. Colored markers, variants in partial LD with the VNTR (R^2>0.01). **C)** VNTR allele length distribution in European-ancestry UKB participants (histogram, right axis) and mean phenotype in carriers of VNTR alleles (binned by length) for the four most strongly associated blood cell traits (lines, left axis). **D)** VNTR and SNP associations with *CUL4A* alternative splicing usage in cultured fibroblasts. Colored markers, variants in partial LD with the VNTR (R^2>0.01). **E)** VNTR allele distribution in GTEx (histogram, right axis) and mean alternative splice usage in carriers of VNTR alleles (binned by length) for the five tissues with the strongest VNTR association (lines, left axis). Alternative splice usage is the proportion of *CUL4A* transcripts that are alternatively spliced as indicated in panel (a) (as quantified by LeafCutter[50]; STAR Methods). **F)** Scatter plot of VNTR association strength vs. strength of the strongest SNP association with alternative splicing in each of the N=49 tissues analyzed by GTEx. Gray dots, tissues for which no variant significantly associated

with splicing. **G)** Scatter of median alternative splice usage vs. median *CUL4A* expression for each of *N*=49 tissues. Error bars, 95% CIs.

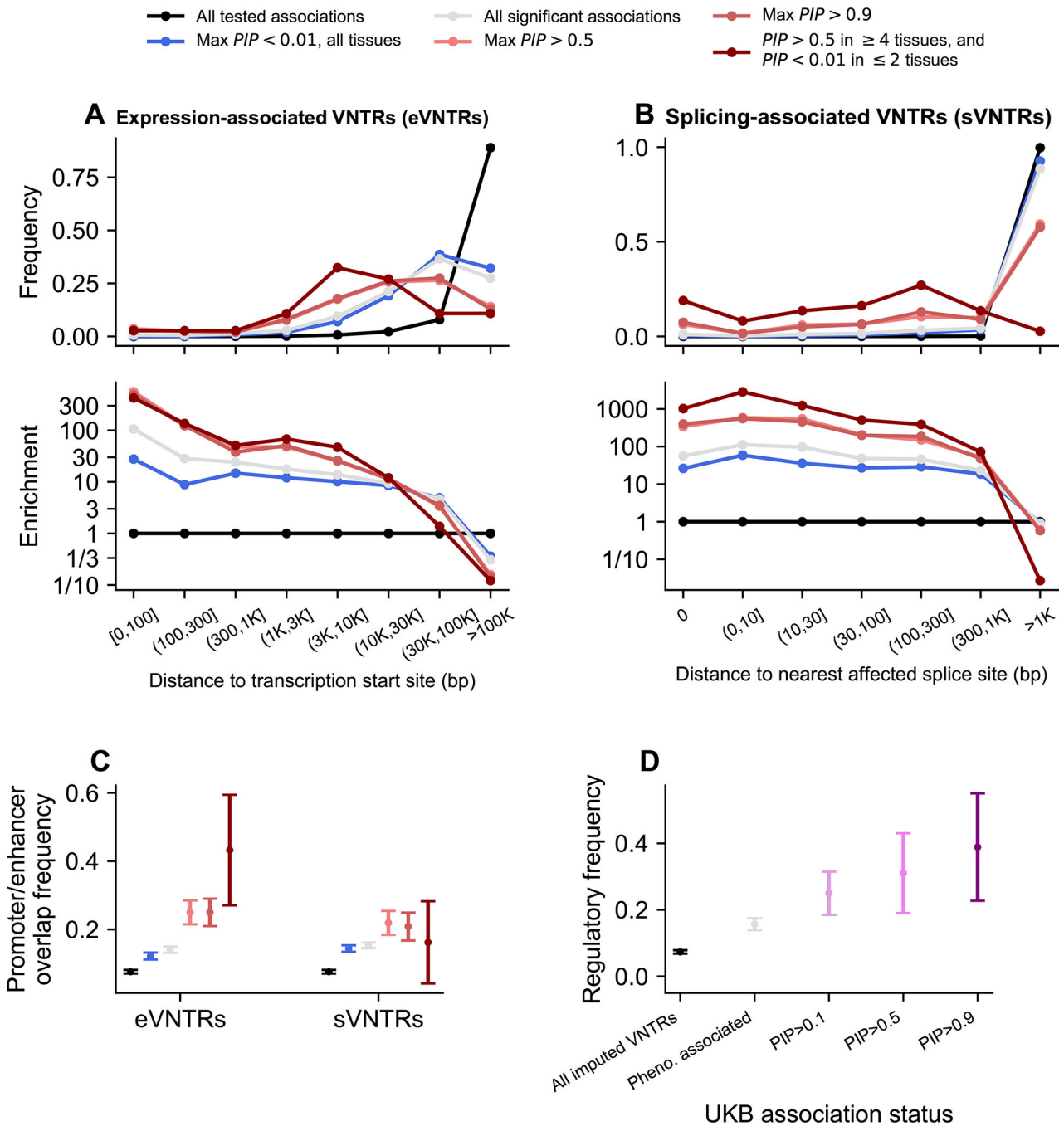


Figure 6. VNTRs associated with gene regulation are enriched near relevant genomic elements and implicate genes mediating complex trait associations.

A) Frequency distribution of distances between eVNTRs and transcription start sites of associated genes (top). VNTRs are stratified by association status and fine-mapping PIP. Fold-change in frequency (i.e., enrichment) relative to baseline distribution (bottom) derived from all tested VNTR-gene pairs (black). **B)** Similar to panel (A) for distribution of distances between sVNTRs and affected splice sites. Affected splice sites are endpoints of introns whose excision counts are tabulated in the denominator of the associated splicing

quantitative trait. **C)** Frequency of overlap with a GeneHancer[39] annotated promoter or enhancer, stratifying VNTRs as in panels (A,B). **D)** Proportion of VNTRs involved in a fine-mapping-supported (PIP>0.5) association with a splicing or expression quantitative trait, stratifying VNTRs by association status and fine-mapping PIP in analyses of complex traits in UKB. Error bars, 95% CIs.

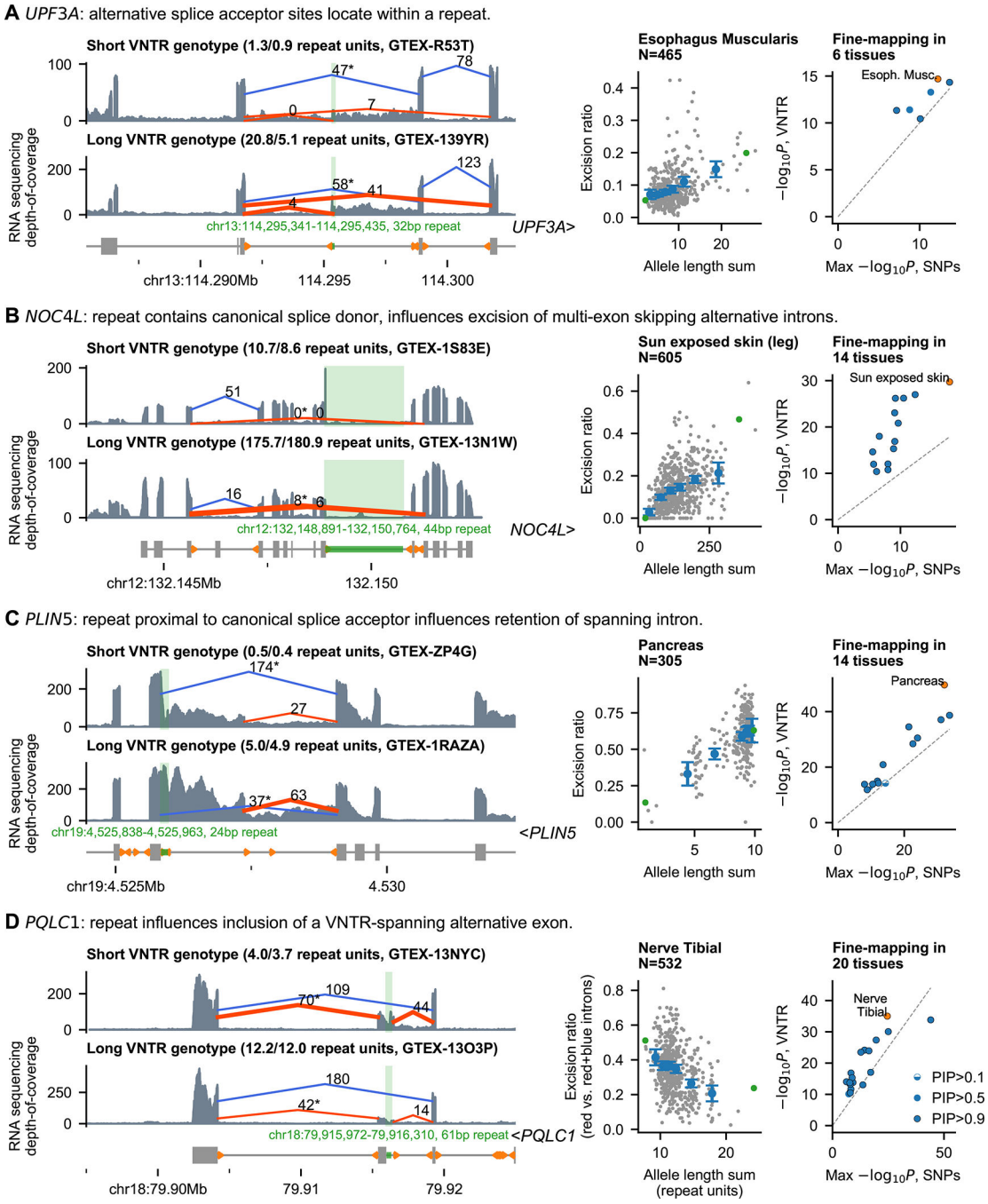


Figure 7. Repeat polymorphisms influence splicing by diverse mechanisms.

VNTRs at *UPF3A* (**A**), *NOC4L* (**B**), *PLIN5* (**C**), and *PLQC1* (**D**) exhibit consistent evidence of regulating splicing across multiple tissues. See Data S1 for additional examples of splice-regulating VNTRs. At each locus: Sashimi plot (left) displaying RNA sequencing depth-of-coverage and LeafCutter intron excision counts for GTEx samples from individuals with short (top) or long (bottom) VNTR genotypes. Coverage within VNTR (green) is normalized to account for VNTR allele length. Orange arrows, splice sites identified by LeafCutter. Scatter plot of excision ratio vs. VNTR allele length sum (middle); dots

correspond to samples from a single representative tissue. Excision ratios are computed from excision counts for the red vs. red plus blue introns (STAR Methods). Green markers, samples displayed in sashimi plots; large blue markers, means across samples binned by VNTR genotype; error bars, 95% CIs. Scatter of VNTR vs. SNP association statistics (right) for the splicing quantitative trait derived from the intron with starred excision count (left panel). Statistics are displayed for all tissues for which the VNTR reached study-wide significance ($P<1 \times 10^{-10}$). Marker fill, posterior probability of the VNTR's inclusion in the causal set (PIP).

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Individual-level VNTR genotypes imputed into UKB	This paper	http://www.ukbiobank.ac.uk/
VNTR+SNP haplotypes in SSC	This paper	https://base.sfari.org
Summary VNTR-phenotype association statistics	This paper	10.5281/zenodo.8087857
UK Biobank genetic and phenotype data	Bycroft et al. 2018; Backman et al. 2021; Halldorsson et al. 2022	http://www.ukbiobank.ac.uk/
Simons Simplex Collection whole-genome sequencing data	An et al. 2018	https://base.sfari.org
Genotype-Tissue Expression (GTEx) whole-genome sequencing, RNA sequencing, and methylation data (v8)	Aguet et al. 2020; Oliva et al. 2023	http://gtexportal.org/ and https://www.ncbi.nlmmih.gov/gap/; dbGaP accession phs000424.v8.p2
1000 Genomes Project high-coverage WGS data	Byrska-Bishop et al. 2022	https://www.internationalgenome.org/data-portal/datacollection/30xgrch38
HGSVC2 long-read assemblies	Ebert et al. 2021	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/assemblies/
GWAS summary statistics from independent studies of glaucoma, IOP, and colorectal cancer	Gharahkhani et al. 2021; Bonnemaijer et al. 2019; Huyghe et al. 2019	http://ebi.ac.uk/gwas/; accessions GCST009413, GCST90011767, and GCST012879
The Cancer Genome Atlas (TCGA) genotype, RNA sequencing, and methylation data (colorectal cancer cohort)	Muzny et al. 2012	https://www.ncbi.nlm.nih.gov/gap/; dbGaP accession phs000178.v11.p8
Software and algorithms		
Code and scripts for estimating VNTR allele lengths in WGS cohorts and imputing into SNP haplotypes	This paper	10.5281/zenodo.8087857
FINEMAP	Benner et al. 2016	http://christianbenner.com/
TandemRepeatsFinder	Benson 1999	https://tandem.bu.edu/trf/trf.html
Minimap2	Li 2018	https://github.com/lh3/minimap2
fastQTL	Ongen et al. 2016	https://github.com/francois-a/fastqtl
SNPweights	Chen et al. 2013	https://www.hsph.harvard.edu/alkes-price/software/
BOLT-LMM	Loh et al. 2015; Loh et al. 2018	https://data.broadinstitute.org/alkesgroup/BOLT-LMM/
plink	Chang et al. 2015	https://www.cog-genomics.org/plink/
Genome STRiP	Handsaker et al. 2015	https://software.broadinstitute.org/software/genomestrip/
bedtools	Quinlan and Hall 2010	https://github.com/arq5x/bedtools2