Science
AAAS

Supplementary Materials for

Protein-coding repeat polymorphisms strongly shape diverse human phenotypes

Ronen E. Mukamel *et al.*

Corresponding authors: Ronen E. Mukamel, rmukamel@broadinstitute.org; Robert E. Handsaker, handsake@broadinstitute.org; Steven A. McCarroll, mccarroll@genetics.med.harvard.edu; Po-Ru Loh, poruloh@broadinstitute.org

Science **373**, 1499 (2021)
DOI: 10.1126/science.abg8289

The PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S17
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S8
MDAR Reproducibility Checklist

Table of Contents

Materials and Methods

UK Biobank genetic data.

The UK Biobank resource contains extensive genetic and phenotypic data for ~500,000 participants recruited from across the UK (*45*). 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, UK10K, and 1000 Genomes Phase 3 reference panels) in a subset of 487,409 participants (*6*). We further analyzed exome-sequencing read alignments and genotype calls available from whole-exome sequencing (WES) of 49,960 participants (*8*) (which achieved >20x coverage by 76bp paired-end reads for an average of 94.6% of targeted sites). We augmented the SNP-imputation data set with 4.9 million (predominantly rare) autosomal variants from the WES genotype call set that we previously imputed into the full cohort (*10*).

Sample filters for ancestry and relatedness.

We applied strict filters to avoid confounding from population stratification and relatedness among individuals in genetic association analyses. We performed initial analyses on a stringently-filtered set of 337,466 unrelated, White British individuals identified by UKB (*6*) (based on self-report and analysis of genetic principal components (PCs)) who had not since withdrawn from the study. In follow-up analyses of VNTRs exhibiting potentially causal phenotype associations, we expanded this sample set to a larger set of 415,280 participants that we identified using less-extreme filtering on ancestry and relatedness to maximize power to fine-map associations. Specifically, starting with the set of individuals who reported White ethnicity, we (i) removed PC outliers (more than six standard deviations away from the mean in any of the first 10 PCs); and (ii) removed one individual from each ≤2nd-degree related pair (kinship coefficient > 0.0884) previously identified by UKB (*6*), prioritizing retaining individuals for whom height measurements were available. In secondary analyses of cross-population variation, we further analyzed smaller subsets of UKB participants who self-reported African, South Asian, or East Asian ancestry (comprising 1.6%, 1.9%, and 0.3% of the UKB cohort, respectively). Admixed Americans were insufficiently represented in the UKB cohort to include in analyses.

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. This set of phenotypes consisted of: (i) 636 diseases with >200 reported cases among 337,466 unrelated, White British individuals (as of Oct 10, 2019) 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 (*10*, *46*).

Protein-coding VNTR ascertainment and genotyping pipeline.

We identified and genotyped VNTR allele length variation from exome-sequencing data using an analysis pipeline consisting of three main steps (detailed in the Supplementary Text):

1. ***Identify potential VNTR loci from repeat sequences in the human reference.*** We identified approximate tandem repeats in the GRCh38 reference using two approaches: (i) Tandem Repeats Finder (*47*) v4.09 (using its suggested parameters 2 5 7 80 10 50 2000 -l 6 -h to detect repeated patterns of up to 2kb); and (ii) a separate algorithm we developed to identify large, multi-kilobase repeats such as the KIV-2 VNTR in *LPA* (Supplementary Text). We filtered to autosomal loci with a repeat unit at least 9bp long that overlapped at least one exon (of any transcript) and overlapped the set of WES targets.

2. ***Estimate VNTR lengths from exome-sequencing depth-of-coverage.*** At each potential VNTR, we estimated diploid VNTR content (i.e., the sum of VNTR allele lengths across an individual's two alleles) for exome-sequenced UKB participants by counting aligned reads overlapping the VNTR. To reduce technical noise in these measurements, we normalized depth-of-coverage estimates in each individual against corresponding estimates in the 200 other individuals in the cohort with closest-matching exome-wide sequencing profiles (Supplementary Text). To gauge the precision of these estimates, we computed the correlation coefficient between estimates in pairs of "IBD2" siblings who shared both haplotypes identical-by-descent (i.e., had inherited the same allele from their mother and inherited the same allele from their father).

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, which we then imputed from the exome-sequenced participants into the remainder of the UKB cohort. To do so, we developed a computational algorithm to efficiently phase multiallelic VNTR variants (with real-valued length estimates) using surrounding SNP-haplotype information and simultaneously compute cross-validation-based benchmarks of phasing and imputation accuracy (Supplementary Text).

The first step of this pipeline (based solely on analysis of the human reference sequence) produced a list of 8,186 exon-overlapping tandem repeat sequences in the human genome. However, we expected that the large majority of these tandem repeats did not represent true protein-coding VNTRs (either because they did not exhibit allele length variation or because they did not overlap coding sequence due to imprecise endpoint-calling) or were too short to accurately genotype from sequencing depth-of-coverage data. We therefore developed a stringent filtering pipeline to create a high-confidence subset of protein-coding VNTRs suitable for analysis, ensuring that estimated lengths of these VNTRs (based on WES depth-of-coverage) exhibited heritable variation that was not the result of being contained within a larger CNV (Supplementary Text). These filters produced a final set of 118 high-confidence protein-coding VNTRs modifying 118 distinct genes (Table S1).

We also considered applying other methods previously developed for genotyping VNTRs such as adVNTR (*48*) and ExpansionHunter (*49*). However, we were unable to use adVNTR because it requires sequencing reads to span a VNTR (which was particularly limiting here because the UK Biobank exome sequencing used 76bp reads). While ExpansionHunter is capable of genotyping repeats longer than the read length, it was designed primarily for STR genotyping and has been reported to produce incorrect estimates for VNTRs due to an assumption that different repeat units are mostly identical in sequence (*49*).

Beyond needing to overcome these specific limitations, we were also motivated to develop methods tailored to analysis of the UK Biobank exome sequencing data due to its unique property of containing an extremely large number (N≈50,000) of uniformly-processed exomes. We were able to leverage these properties to maximize VNTR genotyping accuracy by (i) normalizing read-depth by matching read-depth profiles against other samples, (ii) correcting exome capture biases, and (iii) refining genotypes via haplotype modeling (Supplementary Text). These features have not been implemented in existing software packages designed for more general-purpose VNTR genotyping.

Genotyping paralogous sequence variation within VNTRs.

We also sought to identify and genotype intra-allelic variation within repeat units—i.e., paralogous sequence variants (PSVs) —within the *LPA* and *ACAN* VNTRs, both to improve the accuracy of VNTR length estimates (Supplementary Text) and for downstream fine-mapping of the *LPA* locus for Lp(a). To do so, we catalogued within-repeat variation observed in exome-sequenced individuals and then adapted our genotyping, phasing, and imputation pipeline to analyze PSV copy number estimates derived from counting numbers of reference vs. alternate base calls at each such variant (Supplementary Text).

Initial VNTR-phenotype association and fine-mapping analyses.

We performed association tests between the 118 high-confidence coding VNTRs and 786 phenotypes in the stringently-filtered subset of 337,466 unrelated White British participants identified by UKB. We computed linear regression association statistics using BOLT-LMM (*13*) v2.3.5 including a standard set of covariates (20 PCs, assessment center, genotyping array, sex, age, and age squared), and found 185 VNTR-phenotype pairs that passed a significance threshold of $P<5 \times 10^{-8}$ (commonly used in genome-wide association studies, and slightly conservative here).

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 (*9*) 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 up to 2,000 of the most strongly associated nearby variants. The results of these analyses are summarized in Table S3.

Follow-up analyses of potentially causal VNTR-phenotype associations using refined genotypes and phenotypes.

For each of six distinct VNTRs involved in 20 VNTR-phenotype associations that were assigned a high probability of causality (>0.95), we further optimized estimates of VNTR allele lengths by refining VNTR boundaries, carefully modeling biases in exome-sequencing read-capture induced by variation within repeat units, and (for the *TENT5A* VNTR) incorporating read-level information from single sequencing reads that spanned short alleles (Supplementary Text). These genotyping optimizations either increased or maintained the evidence for causality of all associations except one, between height and a VNTR in *RRBP1*, at which the posterior probability of causality dropped to 0.33 (Supplementary Text); we therefore dropped this association from further analysis, leaving nineteen VNTR-phenotype associations involving five VNTRs (*LPA*, *ACAN*, *TENT5A*, *MUC1*, and *TCHH*).

For each of these five VNTRs, we further verified that its fine-mapping was robust to the choice of statistical fine-mapping software used for this analysis (given that modeling assumptions vary across fine-mapping software packages). Specifically, for each top VNTR-phenotype association (Table 1), we reran fine-mapping using the Sum of Single Effects model (SuSiE) (*50*) and found that in each case, the VNTR was assigned a posterior inclusion probability (PIP) of 1. We performed these analyses using susieR v0.10.1 (with default parameters, which allowed for L=10 causal variants), considering all variants (VNTR, SNPs, and indels) within 250kb of the VNTR that were nominally associated with the phenotype (P<0.01). We computed LD among these variants using LDstore (*51*).

We curated three derived phenotypes for follow-up analyses based on relevance to the associations that fine-mapped to these VNTRs. For serum lipoprotein(a), the phenotype coding provided by UKB had coded 17% of Lp(a) measurements as missing due to falling outside the reportable range (3.8-189 nmol/L). To enable analysis of individuals with such measurements, we incorporated binary information available about whether Lp(a) had been below vs. above the reportable range by creating a "cropped Lp(a)" phenotype in which we assigned Lp(a) values of 3.7 or 190 nmol/L to such individuals; we carefully modeled the effect of this cropping in subsequent association and fine-mapping analyses (Supplementary Text). We also computed an estimated glomerular filtration rate (eGFR) phenotype (relevant to *MUC1*) from serum creatinine measurements using the MDRD study equation (*52*), and we computed a male pattern baldness score phenotype (relevant to *TCHH*) following previous work (*53*).

We performed follow-up analyses using the optimized VNTR genotypes on an expanded set of 415,280 PC-filtered, unrelated European-ancestry participants (described above) and further optimized statistical power (*54*) by performing linear mixed model association analysis of top phenotype associations using BOLT-LMM (reported in Table 1 and in the European-ancestry genome-wide association plots in Fig. 2A and Fig. 3A,C; for the African-ancestry genome-wide association analysis of height in Fig. 2A, we ran linear regression using our standard covariates). To emulate mixed model association power in analyses exploring the effects of specific VNTR alleles and SNPs at the *ACAN* and *TENT5A* loci for height (Fig. 2C-F), we performed analyses on height residualized for polygenic predictions of height from array-typed SNPs (omitting those within 2Mb of each VNTR) that we generated using BOLT-LMM (--predBetasFile) in 10-fold cross-validation (*55*).

Logistic regression analyses of VNTR associations with disease outcomes.

We performed logistic regression to further investigate associations between genetically predicted Lp(a) and myocardial infarction (I21) and type-2 diabetes (E11), between *ACAN*

VNTR length and intervertebral disk disorders (M51), and between *MUC1* VNTR length and 13 kidney diseases with at least 100 cases (I12, M10 = gout, N00, N02, N03, N04, N05, N11 = chronic tubulo-interstitial nephritis, N17, N18 = chronic kidney disease, N19, N25, and Q61, using the ICD-10 coded disease phenotypes collated by UKB). For *LPA* and *MUC1*, we binned data by the variable of interest (genetically predicted Lp(a) or *MUC1* VNTR length), included an indicator variable for each bin (omitting the reference bin; Lp(a) between 10-30 nmol/L in Fig. 1D and *MUC1* VNTR copy number <55 in Fig. 3D), and reported the effect size estimated for each bin. For *ACAN*, we analyzed VNTR length as a quantitative variable. We included age, age squared, and sex as covariates in these analyses.

Estimating effects of VNTR alleles on quantitative traits.

To estimate effect sizes of VNTR alleles or groups of alleles in analyses of quantitative traits, we binned alleles by VNTR length and SNP-haplotypes (when additional likely-causal SNPs had been identified), plotting the phenotypic mean among individuals carrying an allele in the bin (counting homozygous carriers twice) against the bin-wise median VNTR length (Figs. 1A, 2D,F, 3B, and 4B,F). In Fig. 1A, we extrapolated mean Lp(a) in bins containing a large fraction (>15%) of Lp(a) measurements that exceeded the reportable range (>189 nmol/L) and had been cropped. We performed this extrapolation based on the median Lp(a) value in the and the assumption that Lp(a) within a bin was log-normally distributed with a $\sigma = 0.304$ (which appeared empirically to fit well across a broad cross-section of bins with fewer cropped values). In Fig. 2D (visualizing the effects of *ACAN* alleles on height), we rounded allele length estimates to the nearest integer and plotted integer bins with MAF>0.5% as well as extreme allele bins (for the rare 13- and 19-repeat alleles and very long alleles containing 40-42 repeats).

Computing the genetic correlation between urea and urate.

We used variance components analysis to estimate the genetic correlation between serum urea and urate in UK Biobank using the restricted maximum likelihood (REML) approach implemented in BOLT-REML (*56*). We restricted analysis to 415,280 unrelated individuals of European ancestry and included the standard set of covariates used in our association analyses.

Modeling lipoprotein(a) concentration from KIV-2 VNTR allele lengths and LPA sequence variants.

Even though Lp(a) is almost completely determined by allelic variation at *LPA*, the specific *LPA* sequence variants that influence Lp(a) and the way that they interact to determine lipoprotein(a) concentration have remained elusive (*12*). Part of the challenge of fine-mapping the *LPA* locus is the need for accurate genotyping of both KIV-2 length variation and SNP variation within the KIV-2 repeat (i.e., PSVs), which has not been available to previous large-scale studies. A further challenge is the multiple forms of nonlinearity that complicate the relationship between *LPA* sequence variation and Lp(a): (i) the nonlinear relationship of KIV-2 length with Lp(a) (even controlling for other *LPA* sequence variants), and (ii) the nonlinearities induced by the allele-specific nature of apo(a) production from individual *LPA* alleles (such that Lp(a)-modifying sequence variants on one chromosome exhibit effects that depend on the length of the KIV-2 repeat on that chromosome, while having no effect on the apo(a) production of the *LPA* allele on the other (homologous) chromosome). For these reasons, although well-powered fine-mapping studies that have applied standard stepwise conditional analyses adjusted for KIV-2 length have identified dozens of conditionally independent SNP associations at *LPA*, these lists have mostly contained noncoding variants unlikely to have causal effects (*57*, *58*).

To fully leverage our comprehensive genotyping and imputation of phased KIV-2 allele lengths, PSVs in and near KIV-2 exons, and SNPs and indels at *LPA* in ~500,000 UKB participants—and to accurately model Lp(a) measurements that had been cropped to the range 3.8-189 nmol/L—we developed novel statistical methods to fine-map the complex association pattern at *LPA* to causal SNPs, and then to perform predictive modeling of Lp(a) from genotypes of these SNPs together with KIV-2 repeat lengths. Briefly, we first performed fine-mapping within an effective-haploid model of Lp(a) created by carriers of alleles producing little or no Lp(a). This framework isolated contributions of individual *LPA* alleles, greatly elucidating the effects of variants that substantially reduced Lp(a) (Fig. S10). Stepwise conditional analyses within this framework identified an allelic series including 18 protein-altering variants and 3 variants in the 5’ UTR of *LPA* that each appeared likely to causally influence Lp(a) levels (based on achieving top or near-top association strengths in successive steps of analysis); 2 additional protein-truncating variants within KIV-2 exons had effects mostly masked by linkage disequilibrium with a canonical splice site variant (*59*) (Table S4 and Supplementary Text). Second, we created an intuitive model that accurately predicted Lp(a) as a sum of allelic contributions determined by KIV-2 length and the combination of alleles of the 23 likely-causal *LPA* SNP and indel variants carried on each haplotype. This model consisted of a low-dimensional parametrization of the “baseline curve” relating KIV-2 length to Lp(a) (in the absence of other Lp(a)-modifying variants) on top of which SNP modifiers exerted multiplicative effects (Supplementary Text).

We compared the above model of Lp(a) to two simpler models corresponding to standard analyses: a model predicting Lp(a) from KIV-2 length alone (used in Fig. 1B) and a linear model using KIV-2 length together with the 23 likely-causal *LPA* variants we identified from fine-mapping. To model Lp(a) from KIV-2 length alone, we first estimated contributions of KIV-2 alleles to Lp(a) in a SNP-unaware manner: we binned alleles solely by KIV-2 length in 2-repeat-unit windows, and we then averaged Lp(a) measurements for the alleles in each bin carried by individuals with a low-Lp(a) allele on the homologous chromosome (i.e., alleles plotted in Fig. 1A). As in the model above, we then predicted a given individual’s Lp(a) by summing the contribution of the two alleles (and cropping values outside the reportable range when performing comparisons against measured Lp(a)). In the linear model, we modeled Lp(a) as a linear combination of diploid genotypes (for the 23 *LPA* SNP and indel variants) and diploid KIV-2 content, fitting this model using linear regression against measured Lp(a).

Association analysis of medication use and liver diseases with altered Lp(a) levels.

To investigate potential effects of exposures including medication use and liver disease on lipoprotein(a) levels, we tested these exposures for association with differences between observed and genetically-predicted Lp(a) in UK Biobank. These analyses were well-powered because genetically-predicted Lp(a) explained ~80% of Lp(a) variance in participants of European ancestry (including those not in the exome-sequenced cohort, for whom we imputed KIV-2 length and *LPA* variants), thus serving as a proxy for baseline Lp(a) (prior to the exposure). We computed the log ratio of observed vs. predicted Lp(a), adjusted for age, sex, and 20 PCs, in 210,755 UK Biobank participants with predicted Lp(a) between 10-100 nmol/L (to avoid bias due to cropping of Lp(a) measurements to 3.8-189 nmol/L) in the unrelated, PC-filtered European-ancestry cohort. For the subsets of individuals who reported taking each of 1,314 medications (taken by at least 10 individuals analyzed), we computed a z-test to determine whether or not medication use associated (at Bonferroni significance, $P<4 \times 10^{-5}$) with a change in the log-ratio of observed vs. predicted Lp(a); if so, we exponentiated this change (and its 95% CI) and subtracted 1 to obtain non-log-scale changes in Lp(a). We performed analogous analyses for liver diseases (K70-K77) with at least 100 cases and for type 2 diabetes for reference. We note that these association analyses do not prove causality, which appears plausible or has been reported for many of the associations but is less clear for others; e.g., both T2D medications and T2D status associate with similar reductions in Lp(a), leaving uncertain whether the associations are driven by medication use, T2D itself, or a T2D-related comorbidity (Table S5).

Replication of the ACAN-height association in the AAAGC African-ancestry cohort.

Encouraged by the strength of the association we observed between *ACAN* VNTR length and height in UKB participants of African ancestry, we sought to replicate this finding by imputing the VNTR's association statistic into SNP association statistics from the African Ancestry Anthropometry Genetics Consortium (AAAGC; N=41,399) (*27*). We employed the approach of ImpG (*60*), 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). To better match the LD structure reflected in the AAAGC analysis (which had performed imputation using LD from AFR samples from the 1000 Genomes Project (*61*)), we estimated LD among the *ACAN* VNTR and nearby SNPs in N=661 AFR samples from 1000 Genomes Phase 3 (*62*) (after first imputing *ACAN* VNTR lengths from the UKB exome-sequencing cohort into 1000 Genomes samples; Supplementary Text). To ensure the SNPs used for imputation matched between the reference (1000 Genomes) and target (AAAGC) data sets, we restricted analysis to a subset of 4,026 SNPs within a 1Mb window centered at *ACAN* (chr15:88.9-89.9Mb) obtained by: restricting to SNPs present in both reference and target; removing rare SNPs (MAF<0.01 in AFR); removing palindromic SNPs (i.e., "A / T" and "G / C"); removing indels and multi-allelic sites; requiring allele frequencies to approximately agree in the reference and target (within 0.1); and requiring non-missing sample size (per SNP) >40,000 in the AAAGC meta-analysis. Since the published implementation of ImpG requires variants to be bi-allelic, 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 validated our implementation by verifying that it produced identical results on bi-allelic variants (e.g., the 6-repeat VNTR allele).

Consistent with our observations in UKB, the imputed association statistic for the VNTR was larger than that of any nearby SNP or indel (imputed $P=5.8 \times 10^{-40}$ for the VNTR vs. $P=3.4 \times 10^{-14}$ for to the top SNP at the *ACAN* locus; Fig. S16), and was also stronger than the strongest reported SNP association genome-wide ($P=3.3 \times 10^{-20}$). The 6-repeat allele also strongly associated with height (imputed $P=5.0 \times 10^{-15}$), and its best tag SNP analyzed in AAAGC associated with a large decrease in height (rs142149658, R^2=0.44 with the 6-repeat allele, beta=-0.34 s.d.). These results were robust to the choice of SNPs used in imputation; e.g., restricting to 137 SNPs based on LD with the VNTR (R^2>0.05) produced a similar imputed association statistic for the VNTR (imputed $P=1.6 \times 10^{-29}$).

Association analysis of TCHH VNTR length variation with hair curl in TwinsUK.
To explore the potential association of the *TCHH* 18bp (6 amino acid) repeat with hair curl (which was not phenotyped in UK Biobank), we analyzed 3,334 TwinsUK participants for whom both SNP-array genotypes and hair curl phenotypes were available and who did not report non-White ancestry. Hair curl phenotypes on a 4-point scale had previously been collected from two questionnaires (Q18_10, available for 3,015 of the 3,334 individuals we analyzed, and Q19_36, available for 1,689 of the individuals we analyzed). We normalized each hair curl phenotype separately in males and females by regressing out age, mean-centering, and dividing by the standard deviation; for individuals with both hair curl phenotypes available, we then averaged the two normalized phenotypes.

SNP-array genotyping had previously been performed using either an Illumina 610K or 317K array, which had low overlap with the Affymetrix arrays used by UK Biobank. To enable imputation of the *TCHH* VNTR from our reference panel of UK Biobank exome-sequenced participants, we first imputed SNPs in the region of chromosome 1 surrounding *TCHH* (chr1:140-165Mb) using the TOPMed imputation server (r2; 97,256 samples) (*63*, *64*). (Prior to imputation, we excluded a small fraction of A/T or C/G SNPs to avoid potential strand-flipping.) After imputation, we improved the phasing of the TOPMed-imputed TwinsUK haplotypes at UK Biobank-typed SNPs by setting imputed genotypes that were <95% confident (i.e., 0.05<HDS<0.95 for either haploid dosage) to missing and then rephasing non-missing SNP genotypes using Eagle (*65*) v2.4.1 --Kpbwt=100000, using all UK Biobank phased haplotypes (*66*) as a reference panel. We then imputed *TCHH* VNTR allele lengths into these rephased SNP-haplotypes using the exome-sequenced UKB participants as a reference panel as in our other analyses (Supplementary Text), with only the slight change of ignoring genotypes that had been set to missing when computing identity-by-state (IBS) sharing between reference and target haplotypes.

We performed linear mixed model association tests between the normalized, merged hair curl phenotype and the *TCHH* VNTR, imputed SNPs, and array-typed SNPs (with missingness <0.1) using BOLT-LMM with genotyping array as a covariate. Mixed model association analysis was necessary to account for substantial relatedness among TwinsUK participants (~1,000 monozygotic or dizygotic twin pairs among the 3,334 individuals we analyzed).

Supplementary Text

1 Identifying protein-coding variable number tandem repeats

The analysis pipeline we used to identify protein-coding VNTRs (outlined in Materials and Methods) consisted of (i) scanning the GRCh38 reference genome for exon-overlapping repeat sequences; (ii) identifying the subset of such regions that appeared to exhibit heritable length variation (as estimated by exome sequencing read-depth); and (iii) applying a series of filters to reach a final, high-confidence set of VNTRs likely to affect protein length. The primary repeat ascertainment pipeline we used for (i) (based on Tandem Repeats Finder (*47*)) is described in Materials and Methods, as is the read-depth-based length estimation procedure we used for (ii). In this note we describe an additional method we developed to identify larger VNTRs (augmenting results from Tandem Repeats Finder) in (i), and we detail the filtering procedure we used for (iii).

Scanning the genome for longer repeat regions

To find larger and more highly-diverged candidate VNTR loci than those identified by Tandem Repeats Finder, we developed an additional, more permissive method to scan the reference genome for repeat regions that were likely to be enriched for polymorphic VNTR loci. This approach was motivated by our observation that known polymorphic VNTRs often correspond to sequences in the reference genome that are in tandem repeated segments but sometimes with high divergence. The method is comprised of two components: "VNTR Scanner" and "VNTR Partitioner."

VNTR Scanner.

Step 1: Alignment scanning

Given the reference genome (GRCh38) and a fixed kmer size (k=30) and maximum genomic distance (maxDistance=1Mb), we find all kmer alignments based on the following criteria.

A kmer alignment is a list of L kmers where:

a) each of the L kmers occurs in the reference genome T times (exact matches)
b) all of the T alignments for this kmer occur within the same genomic window with maximum size maxDistance
c) the alignments of each kmer in the list correspond to consecutive positions along the genome in each of the T occurrences and
d) T >= minAlignments (=3) and T <= maxAlignments (=1000)

To make such scanning efficient, we utilized a pre-computed index of the reference genome based on the Burrows-Wheeler Transform as implemented in the bwa aligner (*67*).

Step 2: Alignment grouping

Given the set of kmer alignments from step 1, we compute a set of (often larger) genome segments that are enriched for containing such kmer alignments to evaluate as candidate VNTRs.

For each kmer alignment, we define the encompassing interval as the maximum extent of the union of the kmer alignment locations. Each base within the encompassing interval is either aligned (i.e. lies within an aligned kmer) or not. Each candidate segment S is defined by a set of kmer alignments and the encompassing interval of S is the union of the encompassing intervals of the kmer alignments. We define the (alignment) density of a segment S as the number of aligned positions in the kmer alignments of S divided by the length of the encompassing interval of S.

To group alignments into segments, we first form an initial set of candidate segments by merging together all kmer alignments whose encompassing intervals transitively overlap. Given a density threshold minDensity (=0.75), we recursively subdivide each candidate segment S until the density of S is at or above minDensity. Subdividing a segment is performed by finding the largest gap (largest contiguous run of non-aligned positions) within the encompassing interval of S, splitting the kmer alignments of S into two subsets that are (fully) on the left or right of the gap, and then regrouping the kmer alignments in each subset into smaller candidate segments (again by transitive overlap of encompassing intervals).

VNTR Partitioner.

Given a genomic segment (e.g. from VNTR Scanner), VNTR Partitioner uses heuristic methods to analyze the segment for repeating VNTR-like self-alignments. The partitioner estimates an approximate "period" of the repetitive structure of the segment (i.e. the length of the repeating unit), then estimates a set of breaks at which to partition the input segment based on the period and finally creates a multiple sequence alignment of these subsegments of the original segment, from which we derive the consensus repeat unit and final period estimate. (We used these estimates only to perform initial filtering on candidate VNTRs and later re-examined VNTR repeat units and periods manually for VNTRs of particular interest; as such, we only briefly outline the main ideas of this method below.)

The partitioner uses multiple strategies to estimate the period of the segment.

The first is an approach similar to the VNTR Scanner. Given a kmer size (k=30) a set of kmer alignments are computed within the input segment. Each base position in the segment is assigned the period of the kmer alignment containing the base and the estimated period is the mode of the period distribution across all bases in the segment.

The second strategy aligns the input segment against a linearly-shifted copy of itself scoring each position as either +1 for a match or -1 for a mismatch. The shift may be up to 75% of the length of the segment, although shifts above 50% of the length of the segment are penalized by a score of -0.5 for each base over 50%. The estimated period is the smallest shift that maximizes this alignment score.

The third strategy takes the estimated period from method 1 or method 2 (whichever is larger, but capped at 200bp) and extracts a "query" sequence of this length from the start of the input segment. This query sequence is then aligned to the input segment in sliding target windows using the Smith-Waterman-Gotoh algorithm (with parameters match=2, mismatch=1, open=2.5, extend=0.5). Each target window is padded on the left and the right by 30% of the estimated period or 20bp, whichever is less. For each target window, the best SW alignment and alignment score is obtained as well as the predicted start of this query sequence alignment within the full input segment. Consecutive target windows will often have the same start position for the alignment and often the same alignment score.

Alignment start positions are then computed that represent local maxima in the vector of the alignment scores from each target window. Several heuristics are used to reduce noise in the estimation of the local maxima: Local maxima are computed over a small window with width equal to 20bp or the estimated period, whichever is less. To be selected as a local maximum, the alignment score for this position must be within a certain fraction (50%) of the global maximum alignment score. If potential local maxima have equal alignment scores, the alignment position corresponding to the largest number of target windows is used. Local maxima must be at least 3bp apart.

After the local maxima are determined, they are used to re-estimate the period (by taking the mode) and the sliding window alignment procedure is iterated one additional time with the updated period estimate.

Finally, a multiple-sequence alignment is constructed by extracting segments bounded by the local maxima of the sliding window alignments and aligning them using MUSCLE (*68*) v3.8.425 with default parameters. From the MUSCLE alignment, a consensus repeat sequence is calculated based on the most common value in each column of the alignment and a final period estimate is determined based on the length of the consensus repeat sequence.

Filtering candidate VNTR loci to a high-confidence set of protein-coding VNTRs

Our analyses of the GRCh38 reference sequence identified 8,186 potential autosomal, exon-overlapping VNTRs (8,048 from Tandem Repeats Finder and 138 from the method above) that also overlapped the exome-capture targets used by UK Biobank. However, we expected that the large majority of these tandem repeats did not represent true protein-coding VNTRs (either because they did not vary in length in the population or because they did not overlap coding sequence due to imprecise endpoint-calling) or were too short to accurately genotype from sequencing depth-of-coverage data. We therefore developed a stringent filtering pipeline to create a high-confidence subset of protein-coding VNTRs for downstream analysis, ensuring that estimated lengths of these VNTRs (based on WES depth-of-coverage) exhibited heritable variation that was not the result of being contained within a larger CNV. This filtering pipeline proceeded in two steps.

Step 1: Initial filters.

- Removed regions with short repeat units (<9bp).
- Removed regions with fewer than 3 repeat units in the GRCh38 reference.

- Removed regions that did not appear to exhibit heritable length variation (as estimated by exome-sequencing depth-of-coverage), i.e., IBD2 sibling correlation (IBD2 R) < 0.25 or estimated imputation accuracy $\widehat{R^2} < 0.25$ (see Supplementary Text 3 for detailed definitions of these quantities).
- Removed duplicated regions, choosing randomly among pairs of regions whose endpoints were within 100bp of one another.
- Removed regions whose flanking sequences had paralogs elsewhere in the genome (indicating potential containment in a larger CNV). To identify matches, we ran BLAT (*69*) v35 on the 500bp sequence 50bp upstream and downstream of the putative VNTR, filtering it if a >97%-identity match was found for either flanking sequence.
- Removed regions that were not contained within a single gene transcript.

These initial filters produced a list of 254 exon-overlapping potential VNTR regions in 190 genes (where the number of regions exceeded the number of unique genes because some VNTR regions were still represented by multiple repeat sequence calls with different boundaries).

Step 2: Stringent filters.

- Required regions to either contain at least three entire exons, or be spanned by a single exon (of any transcript, allowing the region's end points to move by up to 5bp to satisfy this requirement).
- Eliminated pairs of non-overlapping regions whose length estimates (imputed into the full UKB cohort) were correlated (R^2>0.5) (indicating potential containment in a larger CNV).
- Eliminated regions whose depth-of-coverage length estimates in 30x-coverage WGS data (on 3,202 samples from the 1000 Genomes Project (*70*)) were correlated with depth-of-coverage on either flanking region (R>0.4), indicating potential containment in a larger CNV. Depth of coverage was measured within the putative VNTR region and in 1kb flanking segments starting 100bp outside of the VNTR region using Genome STRiP (*71*) with default parameters but removing the default genome mask that masks out non-unique portions of the reference genome.
- Eliminated regions for which an overlapping region with correlated length estimates (R^2>0.5) appeared to produce higher-quality read-depth-based length estimates (higher IBD2 sibling correlation).
- Eliminated regions entirely contained within another region.
- Eliminated regions that did not overlap canonical coding DNA sequence (CCDS from the UCSC hg38 downloads database).

These stringent filters eliminated roughly half of the 254 potential coding VNTRs that passed our initial filters. We then manually reviewed the excluded VNTRs, adding back in a few VNTRs that appeared to have been over-conservatively filtered (based on known VNTR variation at these loci from previous literature or examination of VNTR sequence together with transcript annotations on the UCSC Genome Browser (*72*)). We also manually reviewed VNTRs that passed the above filters but had estimated period lengths that were not multiples of 3, in most cases determining that the VNTR either fully spanned one or more exons or that its period had been incorrectly estimated; in other cases, we excluded the VNTR. Finally, we manually examined two remaining instances in which multiple VNTR regions were called in the same gene, selecting a representative VNTR in each case. These final filters produced our main analysis set containing 118 high-confidence protein-coding VNTRs (in 118 distinct genes).

2 Estimating diploid VNTR content from exome sequencing read depth

Depth-of-coverage is commonly used to identify copy-number variation from short-read sequencing (*73*, *74*) and has more recently been used to estimate VNTR allele lengths (combined across maternally- and paternally-inherited haplotypes) from whole-genome sequencing data (*71*, *75*, *76*). Here we applied this framework to estimate VNTR lengths from 76bp paired-end exome-sequencing reads (aligned to the GRCh38 reference using SPB pipeline used to process the 49,959-sample UK Biobank exome sequencing release (*8*)). For each repeat region ascertained in the GRCh38 reference that passed our filters for analysis (Materials and Methods), we began by counting, for each sample, the number of unique reads with a SAM flag (*77*) of 0x53, 0x63, 0x93, 0xA3, 0x51, 0x61, 0x91, or 0xA1 that aligned to the region (allowing reads with any mapping quality including zero). We considered two possible definitions of "aligning to the region": (i) having at least one mapped base fall within the region, or (ii) having all mapped bases fall within the region, choosing whichever approach produced higher IBD2 sib-pair correlation in a pilot analysis.

Counts of reads generated from polymorphic repeat regions are expected to scale linearly with total (maternal plus paternal) repeat length and with sequencing coverage, such that naively, dividing read counts by overall sequencing coverage (per-sample) produces estimates of "diploid VNTR content" (up to a suitable affine transformation). However, read counts can be strongly influenced by technical biases that vary across samples and typically contribute much more variance than Poisson noise. To mitigate this problem, computational methods that estimate depth-of-coverage usually model and correct for the effects of genomic features such as GC content that tend to influence technical biases (*73*, *74*).

Normalizing read depth against samples with similar exome-wide coverage profiles

We developed a different normalization approach that leveraged the very large number of exome sequences (49,959) available for analysis: instead of explicitly identifying features that correlated with biases in sequencing depth, we identified samples that exhibited similar exome-wide coverage profiles, and we normalized VNTR-aligned read counts from each sample against corresponding read counts from the 300 samples with most similar coverage profiles (after adjusting all samples for overall coverage). The intuition behind this approach was that in the majority of the genome that does not vary in copy number between two individuals, depth-of-coverage is influenced primarily by technical biases, such that the coverage profile of each sample provides a "barcode" that identifies which other samples were influenced by the same technical biases (e.g., due to sharing the same batch effects).

The main challenge in effectively implementing this approach is deciding which genomic regions to include in the profiles. We found that restricting to "batch-informative regions" in which read depth was especially variable among samples was helpful. Understanding the properties of outlier regions in each stage of the analysis was also important: on one hand, we needed to prevent outliers from overly influencing the results; on the other hand, outliers could contain useful information about patterns of technical artifacts.

In detail, our normalization pipeline proceeded as follows:

- Compute read depth in 1kb windows of the genome (using mosdepth (*78*) v0.2.5).
- Restrict to regions with ~15-70x mean read depth across samples (81,564 1kb regions).
- Exclude 1kb regions overlapping repeat regions identified by an initial run of our VNTR scanner (77,003 1kb regions left).
- Normalize each sample to the same mean read depth across remaining 1kb regions.
- Normalize each region:
-- subtract mean across individuals
-- divide by sqrt(mean across individuals) (proportional to Poisson standard deviation).
- Rescale all normalized read depths so that median region has std dev = 1.
- Restrict to regions with top-10% variance (the intuition being that most regions just have variation due to random noise, but a small subset are informative of batch effects).
- Crop normalized read depths to (-2, 2) to reduce the effect of outliers.
- Compute pairwise distances among 49,959 samples (where distance is defined as the sum of squared differences of final transformed read depths at remaining 1kb regions).
- For each VNTR, normalize (aligned read count / exome-wide depth) in each sample by dividing this value by the mean (aligned read count / exome-wide depth) across the 300 other samples with closest-matching read-depth profiles.

We estimated (based on benchmarks assessing correlation of VNTR length estimates across IBD2 sib-pairs) that this normalization procedure tended to reduce noise variance by slightly more than half compared to naïve estimates (aligned read count / exome-wide depth) that did not make any attempt to correct for technical biases.

We implemented a few refinements of this pipeline for our final analyses of the five VNTRs that appeared to drive strong phenotypic associations. First, instead of normalizing each sample against 300 closest-matching samples chosen from all 49,958 other samples, we normalized against 200 closets-matching samples chosen from the subset of other samples who reported European ancestry (to prevent ancestry bias among the set of "reference samples" chosen for normalization). Second, for the *LPA* and *ACAN* VNTRs, we leveraged variation within these VNTRs to separately estimate length variation of specific subtypes of repeats (by filtering to reads distinguishing the subtypes). Third, we rescaled allele length estimates (which the above pipeline normalizes to a mean of ~1 across all diploid estimates) to absolute numbers of repeats per allele (by comparison to literature and/or identification of discrete peaks in phased allele distributions). These latter two refinements are detailed for each VNTR (Supplementary Text 4).

3 Phasing and imputing VNTR lengths using surrounding SNPs

Our read-depth-based VNTR length estimates ("diploid VNTR content") for 49,959 exome-sequenced UK Biobank participants enabled haplotype-sharing analyses to phase (and simultaneously refine) VNTR length estimates within the exome-sequenced cohort and to impute allele lengths into the remainder of the UK Biobank cohort. Haplotype-based phasing and imputation methods have consistently been shown to be reliable for modeling alleles carried by at least 5-10 samples in a cohort (*63*, *79*). Consequently, in a ~50,000-sample cohort, this approach is expected to work well for alleles with frequencies as low as ~0.01% (and thus even for VNTRs with relatively high mutation rates).

We employed an iterative algorithm in which haploid allele lengths of each individual in turn were 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. This overall structure is similar in spirit to the approach applied by PHASE v2 (*80*) and many subsequent algorithms developed to phase biallelic SNPs; however, we adapted the approach in several ways to better suit the task of phasing a single multiallelic VNTR variant (with real-valued length estimates) in a very large cohort for which accurately-phased haplotypes of surrounding SNPs had previously been generated (*66*). (We performed phasing analyses on the subset of 49,796 exome-sequenced individuals with phased SNP-array haplotypes.)

We note that while the Beagle software (*81*) can perform phasing of multi-allelic variants, it requires genotypes to be modeled as discrete alleles (either in the form of hard-called genotypes (GTs) or genotype likelihoods (GLs)); here, we wished to phase and impute continuous-valued VNTR allele length estimates, which cannot be converted to GLs without making assumptions about allele frequencies. As a result, we were unable to use Beagle to phase our VNTR length estimates.

Haplotype-copying model based on identity-by-state (IBS) sharing

Most of the commonly-used statistical phasing algorithms that have been developed to date have aimed to phase many SNPs across a locus or a chromosome. In this setting, hidden Markov model (HMM)-based approaches that model haplotypes as mosaics of sub-haplotypes copied (imperfectly) from a reference panel of other haplotypes (e.g., according to the Li-Stephens model (*82*)) have been a natural and very effective approach, as HMMs facilitate efficient computation of posterior probabilities across many SNPs via dynamic programming algorithms.

Here, we faced a very different phasing task: at a typical locus, we only needed to phase a single VNTR and could assume that SNP-haplotypes had already been phased to chromosome-scale accuracy (*66*). As such, while a standard Li-Stephens HMM would produce a reasonable haplotype-copying model that could be used within a phasing routine, the HMM framework was unnecessary here, and a much simpler haplotype-copying model based on lengths of identity-by-state (IBS) sharing could produce equivalent results with less computation. Moreover, the IBS-based approach facilitated rapid tuning of the parameters of the haplotype-copying model, ultimately enabling improved phasing accuracy.

The typical length of a genomic segment shared identical-by-descent (IBD) between two haplotypes scales inversely with the time to the most recent common ancestor (TMRCA) because the TMRCA determines the number of opportunities for recombination in the coalescent tree connecting the haplotypes. In general, IBD segments are not expected to be identical-by-state (IBS)—i.e., carry exactly the same alleles for all SNPs within—due to gene conversion, recent mutation, and genotyping and phasing error. Conversely, IBS-sharing may extend beyond the ends of IBD segments due to sharing of common haplotypes. However, when considering only IBS across accurately-genotyped, accurately-phased SNPs from a modern SNP-array platform (which is unlikely to "notice" most gene conversions or recent mutations), the biological and technical complexities that break up IBS-sharing within an IBD segment tend to be quite rare (i.e., occur only once every few megabases), and portions of longer IBS segments not too near their ends tend to consistently represent true IBD.

We therefore quantified lengths of haplotype-sharing (in genetic map coordinates, i.e., centimorgans) with an IBS-like measure designed to be more robust to these complexities:

$$\ell_{\mathrm{dir}} = a \cdot \big(\text{IBS length in direction (left/right)}\big) + (1-a) \cdot (\text{IBS length in direction, allowing 1 error})$$
$$\ell = b \cdot \big(\ell_{\mathrm{left}} + \ell_{\mathrm{right}}\big) + (1-b) \cdot \min\big(\ell_{\mathrm{left}}, \ell_{\mathrm{right}}\big).$$

We set $a = 0.5$ and $b = 0.25$ based on empirical performance in phasing benchmarks (which tended to be insensitive to the choice of these parameters) and found that IBS-like lengths defined in this way served as a reasonable proxy for IBD lengths for the purpose of identifying haplotypes likely to share a recent common ancestor (and carry the same allele of a VNTR).

We then defined a haplotype-copying model in which the probability of a given haplotype having been copied from each of a set of reference haplotypes scaled roughly with:

$$\exp\left(-\ell_0/\ell\right)$$

where ℓ_0 denotes a constant factor (to be tuned from the data) and ℓ denotes the IBS-like length computed above. For computational convenience, we restricted the set of potential reference haplotypes to the top K_{top} haplotypes with longest ℓ, and to add more flexibility to the functional form of the relationship between haplotype-sharing length and copying probability, we adjusted copying probabilities by including a regularization probability p_{reg} of randomly sampling from the top K_{top} longest-matching reference haplotypes independent of ℓ. Together, these features produced copying probabilities:

$$P(\text{copy from ref. hap. } i) = \frac{\exp\left(-\frac{\ell_0}{\ell_i}\right)}{\sum_{k \text{ in } K_{\mathrm{top}}} \exp\left(-\frac{\ell_0}{\ell_k}\right)} \cdot \big(1 - p_{\mathrm{reg}}\big) + \frac{p_{\mathrm{reg}}}{K_{\mathrm{top}}}$$

We tuned the parameters ℓ_0, K_{top}, and p_{reg} independently for each VNTR as described at the end of this section.

This haplotype-copying model immediately provided a means to impute VNTR allele lengths from a phased panel of reference SNP+VNTR length haplotypes into a target SNP haplotype (as the weighted average of copied haplotypes, weighting according to the above probabilities) and provided a probabilistic framework within which to perform phasing.

Model for estimated diploid VNTR content

Assuming an individual's two haplotypes (carrying VNTR alleles of length x_1 and x_2) had been copied from reference haplotypes i and j with (phased) VNTR allele lengths x_i and x_j, we modeled the observed "diploid VNTR content" estimate obtained from read-depth as being generated under the following model:

$$x_1 = x_i + \epsilon_{\text{mut}}^{(1)}, \quad x_2 = x_j + \epsilon_{\text{mut}}^{(2)}$$

$$\text{diploid VNTR estimate} = x_1 + x_2 + \epsilon_{\text{err}} = \left(x_i + \epsilon_{\text{mut}}^{(1)}\right) + \left(x_j + \epsilon_{\text{mut}}^{(2)}\right) + \epsilon_{\text{err}}$$

where each $\epsilon_{\text{mut}} \sim N(0, \sigma_{\text{mut}}^2)$ roughly models "mutation" along the branch of the coalescent tree connecting the individual's haplotype and the copied reference haplotype (but also incorporates potential error in the reference haplotype's VNTR length) and $\epsilon_{\text{err}} \sim N(0, \sigma_{\text{err}}^2)$ models noise in our read-depth-based estimate of diploid VNTR content. We also tuned σ_{mut} and σ_{err} independently for each VNTR.

Iterative update of haploid allele length estimates

Within this probabilistic framework, we performed phasing by iteratively leaving out each individual in turn and updating the left-out-individual's (phased) VNTR allele lengths to their posterior means (assuming each of the individual's two haplotypes had been "copied with noise" under the above model from a reference panel containing all other individuals' haplotypes). Explicitly, the posterior probability of copying the individual's haplotype 1 from reference haplotype i and copying haplotype 2 from reference haplotype j satisfies:

$$P(\text{hap1 copied from ref. hap. } i, \text{hap2 copied from ref. hap. } j \mid \text{diploid VNTR estimate} = \widehat{x_{1+2}})$$
$$\propto P(\text{SNP hap1 from ref. hap. } i) \cdot P(\text{SNP hap2 from ref. hap. } j) \cdot \exp\left(-\frac{\left(\widehat{x_{1+2}} - x_i - x_j\right)^2}{2 \cdot (2\sigma_{\text{mut}}^2 + \sigma_{\text{err}}^2)}\right)$$

which provides weights from which the posterior mean can be computed as the weighted sum of posterior means assuming each copying configuration (using the formula for the conditional expectation of a bivariate Gaussian):

$$E[x_1 \mid \text{hap1 from ref. hap. } i, \text{hap2 from ref. hap. } j, \text{diploid VNTR estimate} = \widehat{x_{1+2}}]$$
$$= x_i + \frac{\sigma_{\text{mut}}^2}{(2\sigma_{\text{mut}}^2 + \sigma_{\text{err}}^2)} \cdot \left(\widehat{x_{1+2}} - x_i - x_j\right)$$

and likewise for the posterior mean of x_2.

We slightly modified the above computation to disallow copying both haplotypes from the same individual; this safeguard is commonly implemented in SNP-phasing methods to prevent circular updates in which two individuals who share both haplotypes IBD iteratively copy each other's phased allele estimates, impeding progress. We implemented this modification by zeroing out probabilities of copying from reference haplotypes i and j in the same reference individual (and renormalizing probabilities).

Continuous (real-valued) versus discrete allele length estimates

Our probabilistic framework modeled VNTR allele lengths as real-valued quantities rather than integer-valued copy numbers. This framework was sensible for modeling read-depth-based (continuous) estimates of diploid VNTR content and enabled efficient computation of posterior probabilities using Gaussian conditional expectations. In downstream analyses (e.g., association tests), we continued to analyze the real-valued VNTR allele lengths produced by our phasing and imputation pipeline, as these "dosage"-type estimates appropriately reflect uncertainty in allele lengths, maximizing power in regression analyses. However, to improve clarity of presentation, we displayed allele histograms using discrete bins in our main figures (equivalent to rounding allele lengths to the nearest integer).

Benchmarking of phasing and imputation accuracy

To benchmark the accuracy of phased VNTR length estimates, we adapted a trio-based approach commonly used for SNP-phasing benchmarks. The basic idea behind this benchmark is that if a cohort contains several complete trios (i.e., father, mother, and child), then phasing the cohort excluding all trio children will produce allele length estimates for all parental haplotypes (without using any information from within-family haplotype-sharing). Each child's diploid VNTR estimate can then be compared to the sum of the phased allele length estimates for the two transmitted haplotypes, and assuming independence of parental haplotypes, this correlation satisfies the relationship:

$$\begin{aligned}\text{"child } R^2\text{"} &= R^2(\text{diploid estimates, sum of transmitted phased estimates}) \\ &= R^2(\text{diploid estimates, truth}) \cdot R^2(\text{phased estimates, truth})\end{aligned}$$

We can thus estimate phasing accuracy by dividing the quantity on the left (obtained in trio children) by the estimated accuracy (i.e., R^2 vs. truth) of the diploid VNTR content estimates we obtain from read-depth. Assuming unbiased error in diploid VNTR estimates, this latter quantity can in turn be estimated as the correlation between diploid estimates obtained from independent sequencing of monozygotic twins, or more generally, "IBD2" sib-pairs who share both haplotypes IBD (i.e., inherited the same allele from their mother and inherited the same allele from their father):

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

Putting these two relationships together, we obtain:

$$\widehat{R^2}(\text{phased estimates, truth}) = \frac{\text{child } R^2}{\text{IBD2 } R}$$

The exome-sequenced UK Biobank cohort that we analyzed contained 613 sib-pairs, one-quarter of which are expected to be IBD2 at a given locus. We readily identified high-confidence IBD2 sib-pairs based on at most 3 mismatching SNP-array genotypes within ±1 Mb of each VNTR (computed using plink (*83*) v1.9), from which we obtained IBD2 R estimates with reasonably low noise.

The cohort contained far fewer complete trios, such that we needed to adapt the above approach to instead use “pseudo-trios”: i.e., “pseudo-children” each of whose haplotypes had a long IBD match with a “surrogate parent” in the data set. We identified pseudo-trios by starting with long IBD matches among related individuals previously identified by UK Biobank (*6*) (specifically requiring either >3 cM IBD between 1st-degree relatives or >5 cM IBD between more-distant relatives) and then identifying the best surrogate for the opposite haplotype of either member of the related pair, determining that a pseudo-trio had been completed if a surrogate with >3 cM IBD could be found. (In these computations, we approximated IBD length with the average of no-error and 1-error IBS, requiring this average to be >0.5 cM on either side of the VNTR.) This approach identified roughly 600 pseudo-trios per locus, which sufficed to obtain "child R^2" estimates with low noise.

Benchmarking imputation accuracy was much simpler using the standard cross-validation approach: we left out 5% of individuals in the cohort, phased the remainder of the cohort, and then imputed into the left-out individuals, estimating imputation accuracy as:

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

One important caveat of the above benchmarks is that they assume unbiasedness of errors in read-depth-based estimates of diploid VNTR content. In reality, 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. While these biases did not produce first-order errors in either our allele length estimates or our estimates of phasing or imputation accuracy, we also performed imputation accuracy benchmarks by comparing imputed allele lengths derived from WES-read-depth to diploid VNTR content estimated from WGS-read-depth (which was available to us for 48 UK Biobank participants included in a WGS pilot study) and subsequently to allele lengths directly measured from long-read sequencing technologies. Our analyses of these benchmarks are described in the Supplementary Text 4.

Optimization of phasing and imputation parameters

Our phasing algorithm included five constant parameters (ℓ_0, K_{top}, p_{reg}, σ_{mut}, and σ_{err}) that determined relative copying weights of reference haplotypes with different lengths of IBS-sharing and determined the relative weighting of information derived from haplotype-sharing vs. direct read-depth-based estimates. The optimal choices of these parameters are expected to differ among VNTRs depending on their mutational characteristics (e.g., mutation rate) and sequencing properties (e.g., average number of VNTR-derived fragments sequenced). We therefore incorporated parameter-tuning into our phasing algorithm to enable each VNTR to be analyzed using parameters tailored for the VNTR.

The simplest approach to tuning would be to employ a comprehensive search across the reasonable parameter space for each parameter, benchmarking phasing accuracy (as described above) for each parameter configuration and then selecting the choice of parameters that yielded the highest accuracy. However, this brute-force approach was computationally intractable given the five-dimensional parameter space and nontrivial amount of computation required to perform each phasing analysis.

To optimize parameters in a computationally tractable way, we instead employed a simulated annealing strategy (*84*) in which we explored the parameter space by iteratively: (1) proposing a random step (i.e., change to a single parameter); (2) evaluating phasing accuracy after one more full iteration (leaving out and updating each individual's allele length estimates in turn) using either the current or proposed new parameter configuration; and (3) accepting the proposed parameter change if it achieved acceptable accuracy (relative to the accuracy obtained by running the full iteration using the current parameter configuration). This stochastic search technique was intended to improve robustness to local optima.

In detail, our algorithm began by initializing the parameters to:

$\ell_0 = 2\text{ cM}$, $K_{\text{top}} = 50$, $p_{\text{reg}} = 0.05$, $\sigma_{\text{err}} = \text{std. dev. (diploid estimates)} \cdot \sqrt{\text{IBD2 } R}$, $\sigma_{\text{mut}} = \frac{\sigma_{\text{err}}}{2}$

and initializing both allele lengths of each individual to half the diploid estimate for that individual. We then ran $T = 500$ phasing iterations, each of which updated each individual's phased allele lengths once, proceeding through the individuals in random order. We designated the first 20 iterations as burn-in iterations in which no parameter modification was proposed. In each subsequent iteration, we proposed updating one parameter (cycling through the five parameters in turn) by multiplying it by a random factor drawn from the exponential of a uniform distribution: $\exp\big(U(-c, c)\big)$, where $c = 1 - \frac{t}{T}$ at iteration t of T for parameters ℓ_0, K_{top}, p_{reg} and $c = \frac{1}{4}\left(1 - \frac{t}{T}\right)$ for parameters σ_{mut} and σ_{err} (which had a smaller range of interest). We further cropped proposed values of K_{top} to stay within the interval $[25, 100]$ and proposed values of p_{reg} to at most 0.9. We then computed the "child R^2" benchmark after one full iteration using either the current parameter configuration or the new proposal. We always accepted the proposed step if it achieved a higher "child R^2" benchmark and always rejected if it reduced "child R^2" by >0.001; if it only slightly reduced accuracy, we probabilistically rejected it with increasing probability as iterations progressed:

$$P_{\text{reject}} = \min\left(1, 10^4 \cdot (\text{child } R^2 \text{ reduction}) \cdot \max\left(0.1, t/T\right)\right)$$

After 500 such iterations, we reverted to the "phasing-optimized" configuration of parameters and allele length estimates that achieved the highest phasing accuracy across all iterations. Using these allele length estimates, we computed "imputation-optimized" values of the parameters ℓ_0, K_{top}, p_{reg} by benchmarking imputation accuracy in a grid search over these three parameters. Separately, we ran 10 final iterations using the "phasing-optimized" parameters to obtain final phased allele length estimates for all individuals in the cohort (no longer leaving any individuals out for the purpose of computing accuracy benchmarks).

To minimize the effects of stochasticity, we ran the entire algorithm described above 5 times for each VNTR using 5 different random seeds (used to randomize the update order within each iteration and to randomize parameter-update proposals) and chose the run that produced the highest estimated phasing accuracy.

4 Optimizing genotyping of VNTRs with potential phenotypic effects

For each VNTR for which our association analysis and fine-mapping pipeline identified a potentially-causal phenotype association, we examined the VNTR allele length estimates produced by our default genotyping and phasing approach to identify potential improvements and to calibrate allele lengths to the absolute scale of repeat copies. (As described above, our initial analysis of normalized read-depths only produced unscaled allele length estimates relative to the mean.) For most VNTRs, we found that we could modestly improve the accuracy of our allele length estimates by optimizing the VNTR boundaries used to select aligning reads, separately genotyping and phasing repeat subtypes distinguished by paralogous sequence variants (PSVs) within the VNTR, or directly genotyping very short alleles using spanning reads. These improvements reduced both noise and bias in our estimates.

The overall changes in allele length estimates between our initial genotyping (used in the initial analyses reported in Tables S1 and S3) and optimized genotyping (used elsewhere in the paper) were modest ($R^2(\text{initial, optimized})$ = 0.97 (*LPA*), 0.88 (*ACAN*), 0.93 (*TENT5A*), 1.00 (*MUC1*), 0.82 (*TCHH*)). As such, discovery of VNTR-phenotype associations could be performed robustly without such optimizations, with further tuning performed on just the subset of VNTRs identified to be of particular interest (to ensure the accuracy of our deeper fine-mapping analyses identifying nearby SNPs likely to also exert causal effects on phenotypes). We also expect the need for such optimizations to be substantially reduced in studies that perform VNTR genotyping from whole-genome sequencing data (rather than exome-sequencing data, which creates the challenge of correcting for exome capture bias).

LPA

Sequence variation within the LPA KIV-2 VNTR.

LPA alleles typically contain 10-30 repeat units of the KIV-2 VNTR, and different repeats within an allele often harbor sequence variation. In particular, the first exon ("exon 1") of the KIV-2 repeat carries three common synonymous variants that distinguish three frequently-observed repeat types labeled A, B, and C in the literature (*12*) (Fig. S2A). Repeat type A is the most common type and accounts for 5 of the 6 KIV-2 repeats that appear in the GRCh38 reference sequence; repeat type B is also common and accounts for 1 of the 6 KIV-2 repeats in the GRCh38 reference (cf. Fig. 2 of Schmidt et al. (*12*)); while repeat type C is rare, accounting for only ~1% of all repeats (*85*).

Bias in IDT xGen v1 exome capture of KIV-2 repeat types.

This common variation within the coding region of the KIV-2 VNTR created a challenge for accurate estimation of VNTR lengths from exome sequencing read counts that was compounded by the design of the IDT xGen v1 exome capture panel used to perform exome-sequencing of UK Biobank samples. The xGen v1 panel did not contain any probes targeting any of the 6 KIV-2 repeats in the GRCh38 reference (presumably because this region had been filtered as repetitive). However, it did contain probes targeting each exon of the adjacent KIV-3 repeat, which share moderate-to-high sequence similarity with KIV-2 exons 1 and 2. Specifically, exon

1 of KIV-3 is identical to the type-B exon 1 ("1B") of KIV-2, while exon 2 of KIV-3 is similar to but somewhat diverged from exon 2 of KIV-2.

This capture design had two main effects on counts of read alignments in the KIV-2 region:

1. Reads aligning to exon 1B repeats in the GRCh38 reference (namely, the single KIV-2 repeat of type B which exactly matches exon 1 of KIV-3) represented "on-target" capture by the pair of 120bp probes that had been designed for exon 1 of KIV-3. In contrast, reads aligning to exon 1A repeats in the GRCh38 reference (i.e., the five KIV-2 repeats of type A) represented "off-target" capture, which resulted in generation of fewer reads per exon 1A unit within an allele than per exon 1B unit within the same allele.
2. Reads aligning to KIV-2 exon 2 repeats in the GRCh38 reference (namely, the six exon 2 repeats of KIV-2 as well as the exactly-matching exon 2 of KIV-1) all represented "off-target" capture from probes that had been intended to capture the somewhat-diverged exon 2 of KIV-3. Consequently, KIV-2 exon 2 read counts were unbiased with respect to repeat types but were greatly diminished (and thus noisier) due to reduced off-target capture efficiency.

Our initial analysis pipeline had simply counted reads mapping to the full repeat region (i.e., KIV-1 exon 2 + six repeats of KIV-2 + KIV-3 exon 1) and thus incurred downward bias of allele length estimates for alleles with larger fractions of type-A repeat units (due to the less-efficient, off-target capture of exon 1A).

Correction of capture bias by separate estimation of type-A and type-B repeat copy numbers.

To circumvent the biased capture of type-A vs. type-B repeats, we devised a modified genotyping strategy that independently estimated the copy number of each repeat type within each individual (and after phasing, within each allele). Specifically, we considered only read pairs that mapped to the seven repeats of exon 1 and its intronic flanks (i.e., the five exon 1A repeats in the KIV-2 region of the GRCh38 reference and the two exon 1B repeats, one in KIV-2 and one in KIV-3) and subdivided these read pairs into (i) those that mapped unambiguously to the 1A sequence (restricting to bases with quality ≥25); (ii) those that mapped unambiguously to the 1B sequence; and (iii) those that mapped equally well to 1A and 1B. We then performed our read-depth normalization procedure on (i) and (ii) separately (not counting reads in category (iii) or reads aligning to exon 2).

This strategy avoided the problem of capture bias because counts of reads generated from 1A vs. 1B sequence were no longer being compared to one another. As such, "1A read-depths" scaled linearly with KIV-2 type-A content and "1B read-depths" scaled linearly with KIV-2 type-B content, albeit with different constant factors due to capture bias.

Calibrating type-A and type-B read-depths to estimate absolute KIV-2 copy numbers.

Our modified genotyping procedure left one remaining task: determining the appropriate scaling factors to convert normalized 1A and 1B read-depths (that our pipeline normalized to a mean of 1 per diploid individual) to absolute estimates of repeat counts of types A and B and ultimately KIV-2. We accomplished this task by making use of the following observations:

- Absolute calibration of 1B read-depth. After phasing, the distribution of phase-resolved 1B read-depths exhibited discrete peaks with constant spacing, indicating that the appropriate scaling factor (~5.2) to use to obtain absolute type-B repeat counts could simply be selected to make the peaks become integer-valued (Fig. 2B).
- Relative calibration of 1A vs. 1B read-depth. Phased 1A read-depths were insufficiently precise to resolve integer-spaced peaks (Fig. S2B), so we instead leveraged reads aligning to exon 2 (which were equally affected by off-target capture across all KIV-2 repeat types) to calibrate 1A and 1B read-depth against one another (by regressing exon 2 read-depth on 1A and 1B read-depth). This computation yielded an estimated ratio of ~6.7 for the contribution of 1A normalized read-depth vs. 1B normalized read-depth (which corroborated with estimates of 1A vs. 1B copy number from whole-genome sequencing of 48 UK Biobank participants included in a WGS pilot study); combining this information with the 1B scaling factor of ~5.2 produced a scaling factor of ~34.9 to convert 1A read-depth to type-A repeat counts.
- Subtraction of 1 repeat corresponding to KIV-1 exon 2 + KIV-3 exon 1. Finally, we subtracted 1 from the total type-A + type-B count to adjust for our read counts including one neighboring exon on either side of KIV-2 (KIV-1 exon 2 = KIV-2 exon 2, and KIV-3 exon 1 = KIV-2 exon 1B).

Combining the above gave the calibration formula:

$$\text{KIV-2 copy number} = 34.9 \times (\text{1A phased read-depth}) + 5.2 \times (\text{1B phased read-depth}) - 1$$

for obtaining absolute KIV-2 VNTR haploid copy number estimates from phased, normalized 1A and 1B read-depths. We verified that the absolute KIV-2 copy number distribution we obtained in this way was in close agreement to literature (cf. Fig. 2 of Kraft et al. (*86*); note that this figure counts the total number of KIV repeats, which is 9 more than the number of KIV-2 repeats).

Further minor optimization of 1A read-depth measurements.

We performed one final optimization that slightly improved the accuracy of type-A repeat count estimates by making use of an additional paralogous sequence variant (119bp into the intron downstream of KIV-2 exon 1A, chr6:160616998:T>C in hg38; Fig. S2A) that marked a repeat expansion constituting ~30% of all KIV-2 repeat units (averaged across all European alleles), but present only in a minority of alleles (Fig. S2B). We labeled the type-A repeats carrying the reference allele “A1” and those carrying the alternate allele “A2”.

Subdividing reads derived from type-A repeats into A1 vs. A2 (vs. ambiguous) groups offered the opportunity to separately estimate A1 and A2 repeat content and then tailor the parameters of our phasing and imputation algorithm separately for the A2 repeat expansion vs. the A1 repeats, potentially improving accuracy. This task was complex but achievable with some care:

- Most read pairs that aligned uniquely to exon 1A do not span the 119bp-downstream variant, so directly assigning reads to A1 or A2 was not feasible.

- However, we could use reads that span the PSV to approximately partition type-A-derived reads according to the observed ratio of T:C bases at the PSV. This approach worked well because ~40% of participants were homozygous for zero copies of A2, such that the partitioning lost no accuracy for such individuals.
- We still needed to restrict reads that contain the REF allele (T) to those that arise from type-A repeats (and discard those that come from type-B repeats). We accomplished this task by separating the PSV-spanning reads into those that aligned unambiguously to 1A sequence (which we counted), those that aligned unambiguously to 1B sequence (which we discarded), and those that did not align uniquely to either (which we pro-rated by the estimated fraction of type-A repeats in the individual).
- This approach produced an approximately unbiased partition of type-A reads into A1 vs. A2 measurements that was fairly noisy (because a typical sample only had ~100 reads that spanned the PSV and could be used for partitioning, in contrast to the ~1000 reads that mapped to type-A sequence). However, both the A1 and A2 measurements phased very well, enabling an improvement in accuracy by repartitioning 1A reads according to post-phasing ("refined") A1 and A2 estimates and then running another iteration of phasing.

In our final KIV-2 length estimates, we used an average of "1A phased read-depth" obtained with and without this additional optimization (to balance the trade-off between better phasing accuracy after partitioning A1 vs. A2 but more noise in read-depth estimates).

Finally, we note that our genotyping procedure did not explicitly treat type-C repeats. As such, reads generated from type-C repeats contributed to a mixture of the other repeat type counts depending on which type-informative paralogous sequence variants they spanned (which determined whether they aligned best to A1, A2, or B). We expect that any biases introduced by this behavior are negligible given the rarity of type-C repeats (only ~1% of all KIV-2 repeat units (*85*)).

Benchmarking accuracy of KIV-2 allele length estimates.

Our cross-validation-based benchmarks of phasing accuracy (based on pseudo-trios) estimated high accuracy of phased estimates of KIV-2 repeat types A ($\widehat{R^2} = 0.95$) and B ($\widehat{R^2} = 0.98$) and subtypes A1 ($\widehat{R^2} = 0.99$) and A2 ($\widehat{R^2} = 0.99$), with similarly-accurate imputation ($\widehat{R^2} = 0.92, 0.98, 0.97, 0.99$, respectively). Assuming approximate independence of errors across different repeat types, we estimated the RMSE of total KIV-2 length created by combining A+B estimates or by combining A1+A2+B estimates (with appropriate weighting) to be ~1.1 and ~0.9, respectively (using the formula $RMSE \approx \sqrt{\sum_{types} \widehat{R^2} \cdot \text{(allele length variance)}}$), suggesting that our final KIV-2 length estimates (which averaged estimates from these two methods) likewise achieved RMSE of approximately 1 repeat unit.

We corroborated these benchmarks by separately comparing total KIV-2 allele lengths (diploid VNTR content) derived from our WES-based pipeline to WGS read-depth-based estimates we generated for 48 UK Biobank participants for whom WGS data had independently been generated by sequencing centers at the Broad Institute and BGI. Of the 48 participants, 6 had been exome-sequenced, and we imputed WES-based KIV-2 lengths into the remaining 42 participants (by separately imputing counts of each repeat type).

The WGS read-depth-based KIV-2 length estimates were highly concordant between the Broad and BGI sequencing experiments ($R^2 = 0.99$). Our WES-derived (mostly imputed) estimates also closely matched the WGS-based estimates ($R^2 = 0.95$; Fig. S4B). We further validated WES- and WGS-derived KIV-2 allele length estimates using direct measurements from Bionano optical mapping in the HGSVC2 data set (*14*), obtaining results consistent with the above benchmarks (Figs. S5-S6; see Supplementary Text 5).

ACAN

<u>*Sequence variation within the ACAN VNTR.*</u>

The 57bp repeat units of the *ACAN* VNTR harbor several common paralogous sequence variants (PSVs) that distinguish repeat units from one another. Two of the most common PSVs affect adjacent bases of the repeat unit which belong to codons for adjacent amino acids; Doege et al. previously used these "diagnostic codons" to classify *ACAN* VNTR repeats into four types (*29*) (numbered 1-4; Fig. S3A). Examination of the *ACAN* VNTR sequence in the GRCh38 reference and in other long read-based genome assemblies (*14*) showed that repeat type 4 could be further subdivided into four common or low-frequency subtypes based on three additional PSVs 17bp, 20bp, and 26bp downstream of the two "diagnostic" bases distinguishing repeat types 1-4 (Fig. S3A).

<u>*Bias in IDT xGen v1 exome capture of ACAN VNTR repeat subtypes.*</u>

The IDT xGen v1 exome capture panel tiled the *ACAN* VNTR GRCh38 reference sequence with 120bp capture probes, with 12 probes falling fully within the VNTR region (resulting in mean exome-sequencing coverage of ~400x across the VNTR). Similar to *LPA*, this capture design created biases in counts of reads generated from fragments of *ACAN* VNTR alleles containing repeats of different types. Most of these biases were more subtle than at *LPA* because (i) different repeat types usually differed by only 1-2bp out of 57bp; (ii) the GRCh38 reference contained similar numbers of type-1 repeats (6), type-2 repeats (7), and type-3 repeats (9); and (iii) most European alleles contained few type-4 repeats (e.g., 3 in GRCh38). However, some alleles (and in particular, a 22-repeat allele observed at low frequency in Europeans (*29*)) carried an expansion of a subtype of repeat 4 not contained in GRCh38 and therefore not targeted by any capture probes, which led to under-counting of this repeat4 subtype by ~20% in our initial analyses.

<u>*Correction of capture bias by separate estimation of copy numbers of repeat subtypes.*</u>

To address this bias as well as subtler biases in counts of other repeat types, we again adopted a strategy of separately genotyping each repeat subtype. In contrast to *LPA*, the *ACAN* VNTR repeat unit (57bp) was shorter than the length of an exome-sequencing read (76bp), so instead of counting reads mapping to repeat subtypes, we counted occurrences of PSV "barcodes" carried by each subtype. For repeats 1-3, these "barcodes" consisted of only the two adjacent diagnostic bases. When we observed a read alignment containing the bases (TA) indicative of repeat 4 at these positions, we further examined the bases 17bp, 20bp, and/or 26bp downstream (if present within the read) to determine whether the read could be assigned to one of the four repeat 4 subtypes. (If the read contained the REF=G base at +17bp, we assigned it to either "rep4_00x" or "rep4_01x" based on only the base at +20bp, and if it contained the ALT=T base at +17bp, we assigned it to either "rep4_1x0" or "rep4_1x1" based on only the base at +26bp (Fig. S3A).) We considered reads that either fully aligned or aligned with soft-clipping on one side (CIGAR = 76M or #M#S or #S#M) and required base quality ≥25 at all PSVs used in analysis.

For each the seven repeat subtypes (repeats 1-3 and the four subtypes of repeat 4), we applied our normalization procedure to the "barcode counts" obtained in this manner, producing an analogue of "normalized read-depth" for each repeat subtype. We then applied our phasing algorithm to these normalized barcode-depths to partition these (unscaled) estimates of copy number into their haploid components. (For rep4_00x, we slightly modified the phasing algorithm to account for extreme heteroskedasticity—as the vast majority of alleles carried 0 or 1 copies of this subtype whereas the low-frequency 22-allele carried 8 copies—by adjusting the model to assume that variance in count-based estimates increased linearly with the estimates, basing the parameters of the linear fit on observed differences between estimates in IBD2 sibs. We also applied this modification to improve phasing of repeat 2 content.) For all seven repeat subtypes, the phased estimates exhibited multimodal distributions with discrete peaks that scaled to integers upon identifying a suitable scaling transformation, enabling easy calibration of these estimates to absolute copy numbers (Fig. S3B).

In addition to genotyping these seven common and low-frequency repeat subtypes, we also genotyped an eighth VNTR variant that amounted to a deletion of the ending repeat of the VNTR (the latter half of which is moderately diverged from all of the other repeat types; Fig. S3A). This ending repeat is found in nearly all *ACAN* alleles but deleted in a low-frequency African allele (AF=1% in UKB participants of African ancestry). To genotype loss of this repeat, we simply searched for reads mapping to the *ACAN* VNTR region that contained a unique sequence (AGGACATCAGCGGGCTTCCTTCTGGAGGAGAA) generated by loss of the ending repeat. Carriers were then easily identified as individuals with >10 reads containing this unique sequence.

Combining all of these components (which we independently phased and calibrated to absolute copy-number counts) produced final VNTR allelic copy number estimates given by the formula:

$$\begin{aligned}&ACAN\ \text{VNTR copy number (haploid)}\\ &\qquad = 6.36\cdot\text{rep1} + 4.88\cdot\text{rep2} + 31.18\cdot\text{rep3} + 1.63\cdot\text{rep4}_{\text{00x}} + 1.78\cdot\text{rep4}_{\text{01x}}\\ &\qquad\quad + 0.08\cdot\text{rep4}_{\text{1x0}} + 1.89\cdot\text{rep4}_{\text{1x1}} - \text{repE}_{\text{loss}} + 2.91\end{aligned}$$

The +2.91 constant term at the end mostly arises from the literature canonically considering two additional, more-diverged repeats to be part of the VNTR (despite their sequence divergence); we therefore added 2 repeats to the count for consistency with previous work. The remaining +0.91 constant adjustment corresponds to offsets of +0.49 and +0.42 that we needed to add to counts of repeat 1 and rep4_01x, respectively, to align the modes of their allele distributions to integers. These two repeat subtypes constitute the first and last full repeats of nearly all *ACAN* VNTR alleles (based on long read assemblies (*14*)), and these first and last repeats are expected to generate fewer reads than interior repeat units due to the pre-VNTR and post-VNTR unique sequences not benefiting from capture enrichment from multiple capture probes, probably explaining these offsets.

Benchmarking accuracy of ACAN VNTR allele length estimates.

Our cross-validation-based benchmarks of phasing accuracy (based on pseudo-trios) estimated high accuracy of phased estimates of all *ACAN* VNTR repeat subtypes (estimated $\widehat{R^2}$ vs. truth of 1.00 (rep1), 1.00 (rep2), 0.97 (rep3), 1.00 (rep4_00x), 0.93 (rep4_01x), 1.00 (rep4_1x0), and 0.95 (rep4_1x1)), from which we estimated (following the same procedure we used with *LPA*) that our combined estimates of (haploid) *ACAN* VNTR copy numbers had an RMSE of ~0.9 repeat units.

Unlike with *LPA*, WGS data from the UK Biobank WGS pilot study was not useful for corroborating our allele length estimates using a different technology (because WGS read coverage at *ACAN* was much lower—and therefore noisier—than WES read coverage at ACAN, which had been greatly boosted by the tiling of the VNTR with 12 exome capture probes). However, we verified that the overall distribution of allele lengths was very consistent with literature (cf. Table 1 of Doege et al. (*29*) and Fig. 3 of Horton et al. (*87*)), with common 26-, 27-, and 28-repeat unit alleles and low-frequency alleles with 13-33 repeats. We further validated WES-derived allele length estimates against directly-measured allele lengths from the HGSVC2 long-read assemblies (*14*), observing RMSE of ~1 repeat unit as expected (Fig. S7; see Supplementary Text 5).

In addition to recovering the previously reported distribution of common alleles, we also identified rare (combined European allele frequency ~0.06%) very long alleles with ~40-45 copies of the VNTR that had not been reported in previous literature. We verified that these alleles belonged to IBD clusters (including a 16-haplotype cluster with ~42 repeats and 7-haplotype and 3-haplotype clusters with ~44 repeats), consistent with true inherited variation.

TENT5A

Short size of TENT5A VNTR alleles produces bias in read-depth measurements.

The *TENT5A* VNTR was the shortest repeat we studied in detail, consisting of a 15bp repeat unit repeated only 2-7 times per allele (*88*). Phased allele length estimates from our initial genotyping of this VNTR produced six clearly distinguishable modes, but the spacing of the modes was uneven, and in particular, the 2-allele appeared to generate much lower read-depth than expected. This behavior was probably driven by mapping bias, as reads spanning the full 30bp of the 2-allele do not map directly to the GRCh38 reference sequence (which contains a 4-allele).

Improved TENT5A VNTR genotyping by combining spanning-read and read-depth information.

The above observation suggested an alternative genotyping strategy incorporating read-level information, as smaller alleles (specifically, the 2-allele, 3-allele, and 4-allele) of the *TENT5A*

VNTR were consistently spanned by several 76bp reads indicating their presence. (We also searched for evidence of 0-alleles or 1-alleles but did not find any evidence that such alleles existed.) Additionally, 76bp reads that partially overlapped the VNTR could also inform of the presence of an allele of at least 5 repeats or an allele of at least 6 repeats (depending on the precise location of the read alignment relative to the VNTR).

We therefore implemented a hybrid genotyping strategy that combined direct read-level information (used to identify short 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 ≤5, set the individual's (unphased) genotype to be the minimum-length and maximum-length allele.
- Otherwise:
    - If the minimum-length allele has length ≤4, then estimate the length of the longer allele as 5 + 2.5 * (# reads indicating ≥6 repeats) / (# reads indicating ≥5 repeats)
    - Otherwise, estimate the sum of the lengths of the two alleles as twice the above quantity: 10 + 5 * (# reads indicating ≥6 repeats) / (# reads indicating ≥5 repeats).

The intuition behind the ratio (# reads indicating ≥6 repeats) / (# reads indicating ≥5 repeats) was that the 76bp read length almost exactly corresponds to the length of a 5-allele, such that longer alleles (specifically, 6-alleles and 7-alleles) result in generation of additional reads mapping fully within the VNTR (and indicating ≥6 repeats) in addition to reads partially overlapping the VNTR on either side (indicating ≥5 repeats). Dividing (# reads indicating ≥6 repeats) by (# reads indicating ≥5 repeats) allowed an easy adjustment for variation in sequencing depth, leaving just the need for a calibration factor which we empirically estimated to be ~2.5.

Further minor improvements to TENT5A VNTR genotyping.

The above strategy provided quite accurate, well-calibrated diploid genotype estimates that we could analyze using our standard phasing and imputation algorithm (which treated all genotype estimates as continuous, real-valued measurements). To obtain a further slight improvement in accuracy, we optimized *TENT5A* VNTR genotyping in two additional ways.

First, we adapted the above approach of estimating lengths of longer alleles to make use of our read-depth normalization pipeline (normalizing read-depths of each individual against read-depths of other individuals who exhibited the closest-matching exome-wide coverage profiles). Specifically, we ran a first round of phasing analysis on the diploid genotype estimates computed as above and identified the subset of individuals deemed to be heterozygous for a 6-allele and a shorter allele. We then calibrated counts of reads indicating ≥6 repeats (normalized for exome-wide sequencing coverage) in each individual against the mean of the corresponding quantity across a matched panel consisting of the (6-allele, shorter allele)-heterozygotes among the 500 individuals with closest-matching exome-wide coverage profiles. This strategy removed the need to calibrate counts of reads indicating ≥6 repeats against counts of reads indicating ≥5 repeats.

Second, we adapted our phasing algorithm to account for the (partially) discrete nature of genotypes we obtained from our hybrid genotyping pipeline. Specifically, to leverage the hard-called allele lengths (derived from read-level information) that our hybrid approach produced in individuals who carried shorter alleles, we modified our phasing algorithm to overwrite allele length estimates with hard-called alleles as appropriate. In detail, during each update of an individual's phasing, we ran our algorithm as usual but then post-processed the estimated allele lengths as follows: (i) if exactly one hard-called allele was available, we overwrote the shorter of the two estimated allele lengths with the hard-call; (ii) if hard-calls were available for both alleles (such that the only uncertainty was their phase), we overwrote both estimated allele lengths with the hard-calls in sorted order.

Benchmarking accuracy of TENT5A VNTR allele length estimates.

Our cross-validation-based benchmarks of phasing accuracy (based on pseudo-trios) and imputation accuracy (based on left-out samples) estimated $\widehat{R^2}$ vs. truth of 0.98 and 0.95, respectively.

We also benchmarked the accuracy of our allele length estimates in the 48 UK Biobank participants with available WGS pilot data (6 exome-sequenced and the other 42 genotyped at *TENT5A* via our imputation pipeline). *TENT5A* VNTR alleles of all lengths could be directly genotyped from 151bp reads from WGS, so we simply compared diploid genotypes from WGS to those from our WES-based pipeline, observing very high accuracy consistent with our cross-validation results ($R^2 = 0.97$; Fig. S4B).

MUC1

Sequence variation within the MUC1 VNTR.

Alleles of the *MUC1* VNTR contain abundant intra-allelic variation at multiple positions within the 60bp repeat unit (*35*). Repeat units within the GRCh38 assembly of the *MUC1* VNTR contain numerous variations of the canonical repeat sequence, partly reflecting this variation and partly reflecting the difficulty of correctly assembling this repetitive sequence: in addition to likely-true PSV variation, the *MUC1* VNTR sequence in GRCh38 also contains several erroneous frameshift-inducing insertions and deletions that prevent the full VNTR from being annotated as being contained within a large exon. For this reason, the IDT xGen v1 capture panel only contained two 120bp probes within the *MUC1* VNTR (because most of the VNTR sequence had been annotated as intronic).

Based on the effects of within-VNTR sequence variation on read-depth-based estimates of diploid VNTR content that we observed at *LPA* and *ACAN*, we expect that the *MUC1* allele lengths estimated by our genotyping, phasing, and imputation pipeline were probably also modestly biased, with repeat units more closely matching the targeted repeats producing more efficient capture (and thereby being slightly overcounted). However, in light of the complexity of intra-allelic variation at *MUC1* and the fact that our analyses of *MUC1* only explored the first-order effects of VNTR length variation (e.g., examining phenotypic effects at the resolution of short vs. long vs. very long alleles, which were clearly distinguished in our allele length estimates), we did not attempt to optimize our read-depth-based genotyping approach beyond slightly adjusting the right endpoint of the VNTR (initially called at chr1:155192051, which we revised to chr1:155192006 after examining the GRCh38 sequence). This change had a negligible effect on allele length estimates.

Calibrating MUC1 VNTR allele length estimates.

We did still need to calibrate the allele lengths produced by our genotyping and phasing algorithm (which provided only relative length estimates). To do so, we compared our estimated distribution of phased allele lengths to allele length histograms obtained by previous studies that directly measured *MUC1* allele lengths using gel electrophoresis (*34*, *89*) and to *MUC1* allele lengths derived from long reads spanning the VNTR in the HGSVC2 data set (*14*). These studies suggested a short-allele mode of ~37 repeats and a long-allele mode of ~73 repeats, so we applied a suitable affine transformation (i.e., scaling and shift) to our estimates to match these values.

Benchmarking accuracy of MUC1 VNTR allele length estimates.

Our cross-validation-based benchmarks of phasing and imputation accuracy were very high ($\widehat{R^2} \geq 0.98$), but we expected that these estimates were somewhat inflated because they were unable to assess error introduced by consistent biases from varying exome capture efficiency for alleles carrying different repeat variants.

The *MUC1* VNTR was sufficiently long and variable to be accurately measured using WGS read-depth, so we estimated diploid VNTR content in the 48 individuals in the UK Biobank WGS pilot study by analyzing read alignments using the Genome STRiP pipeline (*71*). Comparing these estimates to estimates from our WES-based pipeline for these 48 individuals (6 exome-sequenced, 42 with imputed *MUC1* alleles) showed high accuracy ($R^2 = 0.90$; Fig. S4B). This R^2 value measures both error in our WES-based estimates and error in the WGS-based estimates and thus serves as a likely lower bound on the true accuracy of allele length estimates from our WES pipeline. We subsequently also validated WES-derived allele length estimates against direct measurements from long reads spanning the VNTR in the HGSVC2 data set (*14*), obtaining results consistent with the above (Fig. S8; see Supplementary Text 5).

We note that the high R^2 values (which are somewhat counterintuitive given the poor coverage of the *MUC1* VNTR by exome capture probes) are driven in part by the bimodality of the *MUC1* VNTR allele length distribution: i.e., distinguishing alleles with ~30-40 repeats from those with ~70-80 repeats is relatively easy.

TCHH

Optimizing the boundary of the 18bp TCHH VNTR.

Our initial genotyping of the *TCHH* VNTR counted reads aligning to the repeat region defined by chr1:1521119170-152112225. However, closer inspection of this region showed that it contains two distinct repeats, as previously described (*39*): the 18bp (6 amino acid) repeat we wished to genotype, and a 39bp (13 amino acid) repeat to its right (upstream in transcription).

The latter repeat appears to be much less polymorphic (based on long-read assemblies from the HGSVC2 data set (*14*)) and had a much weaker association with male pattern baldness in UK Biobank: P=1.4 x 10^{-5} for the 39bp repeat vs. P=5 x 10^{-32} for the 18bp repeat. In our optimized genotyping, we therefore restricted read-counting to reads containing only repeats of the 18bp sequence, which corresponded to mapping within the region chr1:152111930-152112103. Doing so improved concordance of estimates in IBD2 sib-pairs to IBD2 $R = 0.91$.

Calibrating TCHH VNTR allele length estimates.

To obtain absolute *TCHH* VNTR length estimates from the relative allele length estimates produced by our genotyping and phasing pipeline, we first estimated the constant offset corresponding to the allele length that would generate zero observed reads aligning within the VNTR. We did so using the following reasoning:

- The GRCh38 reference allele for TCHH contains 8 copies of the 18bp repeat unit (followed by 1 additional copy of the same 18bp repeat that is considered to belong to the adjacent 39bp repeat).
- To align to the region chr1:152111930-152112103, a 76bp read generated from the 8-allele in the reference must start between 152111930 and 152112103 – 75 (inclusive).
- For an allele carrying a different number of repeats n, the size of the region on the allele from which such reads can be generated increases by $18(n - 8)$.
- Solving for $152111929 = 152112103 - 75 + 18(n - 8)$ gives $n = 2.5$ as the offset corresponding to the allele length that corresponds to zero observed reads.

Second, we roughly estimated the mean European *TCHH* VNTR allele length as ~8.5 repeat copies based on long read assemblies (*14*) and a corroborating analysis of insert sizes of paired-end reads from high-coverage WGS of 1000 Genomes EUR samples.

Combining these two pieces of information yielded the affine transformation we applied to obtain absolute estimates of *TCHH* VNTR allelic copy numbers.

Benchmarking accuracy of TCHH VNTR allele length estimates.

Our cross-validation-based benchmarks of phasing accuracy (based on pseudo-trios) and imputation accuracy (based on left-out samples) estimated $\widehat{R^2}$ vs. truth of 0.78 and 0.71, respectively, indicating the difficulty of imputing this VNTR, which is poorly tagged by all nearby SNPs. We expected that these cross-validation estimates of $\widehat{R^2}$ vs. truth were probably fairly accurate given that alleles of the 18bp *TCHH* VNTR harbor very little intra-allelic variation (and are thus unlikely to suffer exome capture bias). Comparison of WES-derived allele length estimates to directly-measured allele lengths from the HGSVC2 long-read assemblies (*14*) corroborated these results (Fig. S9; see Supplementary Text 5).

RRBP1

Our VNTR ascertainment pipeline identified a 30bp repeat in *RRBP1* with 4 copies in GRCh38 spanning chr20:17659123-17659247 that our initial analysis indicated exhibited heritable variation (IBD2 $R = 0.47$, phasing $\widehat{R^2} = 0.41$) that associated potentially-causally with height.

Upon closer inspection of the GRCh38 reference sequence in this region, we observed that the flanking sequence contained a total of ~40 approximate repeats (with varying levels of divergence from the 30bp repeat unit appearing in 4 consecutive perfect copies); however, expanding the region used for read-counting to include the full approximate repeat region reduced sib-pair correlation, suggesting that the repeat variation might be more local.

We attempted to optimize the boundaries of the read-counting region to improve genotyping (as measured by sib-pair correlation) and found that counting reads aligning fully within the region chr20:17659101-17659347 and containing sequence distinguishing the main repeat unit from nearby repeats (TTGCCCTGGTTCTGGGCCCCC) appeared to somewhat improve genotyping and phasing accuracy (IBD2 $R = 0.51$, phasing $\widehat{R^2} = 0.56$). However, rerunning the association and fine-mapping analysis using the updated allele length estimates reduced confidence that the association with height was causal (FINEMAP posterior probability = 0.33), so we did not perform further follow-up analyses at this locus.

Further work will be required to elucidate the nature of VNTR variation in *RRBP1* and whether it influences height. A recent study using long-read sequencing indicated that the GRCh38 assembly may be inaccurate at this locus, with *RRBP1* alleles carrying insertions (relative to GRCh38) of 15-16 repeats of a 30bp unit presumably corresponding to this VNTR (*90*).

5 Validation of VNTR genotyping using HGSVC2 long-read data

We performed additional validation of WES-derived VNTR allele lengths using the HGSVC2 data set (*14*), which included PacBio long-read sequencing data and haploid genome assemblies for *N*=64 unrelated haplotypes of 1000 Genomes participants as well as Bionano optical mapping data for most of the same individuals.

To enable validation against VNTR allele lengths directly measured by HGSVC2 data, we imputed our WES-derived VNTR length estimates from UK Biobank exomes (*N*~50K) into SNP-haplotypes of 1000 Genomes participants. For *ACAN* and *TCHH*, we imputed into the assembly-based SNP-haplotypes generated by HGSVC2 (freeze4); for *MUC1* and *LPA* (at which longer VNTRs alleles created errors in assembly), we imputed into SNP calls from 30x WGS of 1000 Genomes samples (*70*) that we rephased using Eagle2 (--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).

We then compared imputed VNTR allele lengths to allele lengths derived from the following types of data generated by HGSVC2:

- Assembled VNTR alleles. Ebert et al. (*14*) reported that the HGSVC2 assemblies accurately captured structural variants of length up to ~5kb, which was sufficient to genotype the shorter VNTRs (*ACAN* and *TCHH*; we did not analyze the very short *TENT5A* VNTR as we had already validated our *TENT5A* allele length estimates using spanning WGS reads).
- PacBio CLR reads. For the *MUC1* VNTR, which is longer (up to ~7kb) and has extreme GC content (~82%), we directly analyzed long reads that spanned the VNTR, and we genotyped haplotype-resolved allele lengths by finding optimal alignments of these spanning reads to sets of reference sequences containing each possible VNTR allele length flanked by each of the individual's two surrounding SNP haplotypes. We performed alignment using dynamic programming (the Smith-Waterman algorithm).
- Bionano optical mapping. For the *LPA* VNTR, which spans ~50-200kb (typically exceeding the length of PacBio CLR reads), we validated against structural variant calls that the HGSVC2 had generated from optical mapping of ultra-high molecular weight DNA (sufficiently long to span the full *LPA* VNTR).

The results of these analyses (Figs. S5-S9) were very concordant with our previous benchmarking described in Supplementary Text 4 above (which was based on cross-validation within UK Biobank and comparison to WGS-derived estimates).

6 Genotyping paralogous sequence variants within the *LPA* KIV-2 VNTR

The two coding exons and surrounding splice regions of the *LPA* KIV-2 VNTR contain several common variants and numerous rare variants that have been technically challenging to study because such “paralogous sequence variants” (PSVs) typically only affect 1 out of the 10-30 KIV-2 copies within an *LPA* allele, complicating variant calling from high-throughput sequencing (*85*, *91*). Previous studies have identified three such variants with clear effects of Lp(a) (a stop gain variant in exon 1 [ref. (*92*)], a canonical splice variant at the +1 donor site of exon 1 [ref. (*91*)], and a variant in the last base of exon 2 [ref. (*20*)] that impairs splicing) as well as a variant in the splice region preceding exon 2 that was observed to associate with reduced Lp(a) (*P*=0.0052 in *N*=23 carriers (*85*)).

We sought to leverage whole-exome sequencing in 49,959 UK Biobank participants (together with phasing and imputation into the remainder of the UK Biobank cohort) to comprehensively explore the landscape of variation within and surrounding the KIV-2 exons and determine the effects of such variation on Lp(a) levels. These analyses, which we describe in this note and the following note, uncovered three additional rarer variants within the region that likely result in Lp(a)-null alleles (Table S4), in addition to recovering the effects of the four previously-identified KIV-2 PSV variants (strengthening the evidence for causality of the fourth variant and quantifying the extent to which both non-LoF splice variants reduce Lp(a)).

Ascertaining and genotyping PSVs from exome-sequencing data

To ascertain and estimate (diploid) copy numbers of PSVs from exome-sequencing data, we merged read alignments across the seven paralogous copies of KIV-2 exons 1 and 2 and their intronic flanks in the GRCh38 reference (i.e., the six copies formally belonging to the KIV-2 repeat as well as KIV-1 exon 2 and KIV-3 exon 1). We then identified reads supporting non-reference variants, and for each non-reference variant identified within an individual, we estimated its diploid copy number—i.e., the number of KIV-2 repeats containing the variant, summed across the maternally- and paternally-inherited *LPA* alleles—using the formula:

$$\text{diploid PSV copy number} = \frac{\#\text{ ALT reads}}{\#\text{ ALT reads} + \#\text{ REF reads}} \cdot (\text{diploid KIV-2 copy number})$$

where diploid KIV-2 copy number had been estimated according to the genotyping and phasing algorithm described above. Intuitively, this formula allocates the estimated total number of KIV-2 repeats across those carrying the reference vs. non-reference allele at the site in question according to the ratio of the numbers of reads observed. PSV copy number estimates obtained in this way were sometimes downward-biased by capture bias or mapping bias favoring the reference allele, but such biases could be corrected in downstream analysis.

Phasing and imputing PSV genotypes

Analysis of KIV-2 PSVs using exome-sequencing data from UK Biobank was particularly technically challenging because the off-target capture of exon 2 and type-A versions of exon 1 (due to the exclusion of these exons from the IDT xGen v1 panel) resulted in generation of only

~1 read per repeat unit at some sites, such that our estimates of diploid PSV copy number tended to be very noisy. Fortunately, the large size of the UKB cohort (49,959 exome-sequenced participants) enabled us to employ haplotype-sharing analysis to simultaneously phase-resolve and refine PSV copy number estimates: intuitively, PSV measurements from individuals who shared long haplotypes (and therefore were likely to share *LPA* alleles) could be analyzed together to help inform estimates of one another's genotypes.

For 80 common and low-frequency PSVs, the same phasing algorithm that we had applied to VNTRs also accurately phased and imputed PSV genotypes (mean $\widehat{R^2}$ = 0.91 for phasing and 0.88 for imputation).

Rarer PSVs could not be phased using the same algorithm because it relied on measures of concordance of copy number estimates in IBD2 sib-pairs and in pseudo-trios to perform parameter estimation, such that if none of the 133 sib-pairs we determined to be IBD2 at *LPA* carried a rare PSV, then the parameter optimization routine broke down. We therefore implemented a separate pipeline that we applied to phase 85 rare PSVs estimated to have 2 to 400 carriers among the exome-sequenced participants. This pipeline analyzed each such PSV as follows:

- Identify any individual with at least 1 read carrying the ALT allele as a potential carrier.
- For each potential carrier:
    - For each of the individual's two haplotypes, count the number of other potential carriers among the 200 haplotypes sharing longest IBS (at *LPA*).
    - Call the individual as a homozygous carrier if both haplotypes share long IBS with several other potential carriers (total carrier count >10 across 2 x 200 "haplotype neighbors" of the two haplotypes, with reasonably balanced counts from each neighbor, i.e., <5:1 imbalance).
    - Otherwise, call the individual as a heterozygous carrier as long as one haplotype shares long IBS strictly more potential carriers than the other; if so, assign the PSV to the haplotype with more potential carriers among its IBS neighbors.

After using the above pipeline to phase the 85 rare PSVs in exome-sequenced individuals, we then imputed them into the remainder of the UK Biobank cohort using Minimac4 (*63*) (treating these rare PSVs as biallelic).

Optimizing genotyping of common and low-frequency PSVs with large effects on Lp(a)

An initial round of association and fine-mapping analysis using phased and imputed PSV genotypes from the pipelines above identified five common and low-frequency PSVs that appeared to greatly (or entirely) reduce Lp(a), including the four previously-reported variants as well as a new Tyr>Asp (Y>D) missense variant in exon 1 with MAF=0.3% in European alleles that appeared to produce an Lp(a)-null allele. Given the large effects of these variants and their influence on Lp(a) levels in a sizable fraction of the cohort, we further optimized our genotyping of these five variants to ensure accurate fine-mapping of rarer variants.

We implemented two optimizations within the copy number estimation framework described above, both aimed at increasing the fidelity of initial genotype estimates that were often based on very few reads supporting the ALT allele. First, we applied stringent filters on base qualities and within-read positions (e.g., dropping the last base of each read) that we manually optimized for each of the five PSVs by using samtools (*77*) mpileup to examine distributions of these parameters among ALT base calls from confident carriers (with multiple reads containing the ALT base) vs. potential non-carriers (with only one ALT read). Second, we identified and dropped potential duplicate reads that had not been filtered by standard duplicate read analysis (which checks for pairs of paired-end reads that map to exactly the same pair of positions) because multi-mapping reads generated from the KIV-2 region had been randomly mapped to one of the paralogous KIV-2 repeats in GRCh38.

Our cross-validation benchmarks within our phasing and imputation pipeline indicated that the optimized copy number estimates of all five of these PSVs imputed accurately ($\widehat{R^2} > 0.75$). Phased allelic copy numbers of the two most common non-LoF splice-altering mutations (MAF=20% and MAF=13% in Europeans) were sufficiently accurate to identify rarer alleles likely to carry the PSV on two KIV-2 repeats (CN=2; MAF~1%) or three KIV-2 repeats (CN=3; MAF~0.2%; Fig. S12).

7 Fine-mapping associations of *LPA* variants with lipoprotein(a)

KIV-2 size variation alone explains roughly half of population variation in Lp(a) levels, such that remaining variation in Lp(a) among individuals is largely explained by other sequence variants in *LPA* (that affect Lp(a) levels together with KIV-2 size) (*12*). Previous studies have identified several coding and splice variants in *LPA* that create null alleles (*58*, *91–95*) (producing no detectable serum Lp(a)), and a few other coding or splice region variants have been observed to associate with substantially reduced Lp(a) (with less certain causality) (*57*, *85*, *96*). Additionally, two variants in the 5' UTR of *LPA* have been observed to associate with more moderate changes in Lp(a) with support from *in vitro* (reporter assay) analyses of translational activity (*21*, *22*). However, a comprehensive list of the *LPA* variants that influence Lp(a) levels in most individuals has remained elusive (as has strong statistical evidence for causality of the variants mentioned above, aside from protein-truncating variants) because of the challenges of genotyping KIV-2 variation and accounting for nonlinear effects on Lp(a).

To comprehensively fine-map *LPA* variant associations to variants likely to causally influence Lp(a) levels, we performed a stepwise conditional analysis using a sequence of statistical tests tailored to iteratively identify likely-causal variants with different characteristics (detailed in the following paragraphs) and condition on these variants in subsequent steps. In Table S4, we list each of the 17 steps of this analysis, the likely-causal variant(s) identified, the statistical test performed, the number of alleles tested, the association statistic of the putative causal variant(s), and a description of the association statistics of other *LPA* variants. At each step of the analysis, we performed an association test for all 37,208 variants within 1Mb of *LPA* for which we could obtain or impute genotype information, including:

- SNPs and indels directly genotyped on the UK Biobank SNP-arrays;
- SNPs that we re-imputed (to obtain phased genotype estimates) from the Haplotype Reference Consortium panel r1.1 into the UKB cohort using Minimac4 (*63*) v1.0.1;
- SNPs and indels that we previously imputed from the WES cohort into the remainder of the UKB cohort (*10*); and
- SNP and indel PSVs in the KIV-2 region that we ascertained and genotyped from exome-sequencing read alignments and imputed into the rest of the UKB cohort (as described in the previous note).

We performed stepwise association analyses on subsets of the 337,466 stringently-filtered White British, unrelated UKB participants in all steps except Step 16, in which we analyzed exome-sequenced participants of self-reported African-ancestry (*N*=893).

At each step of this analysis, we tested for association of a variant on one haplotype controlling for the lipoprotein(a) contributed by the allele on the opposite haplotype (on the homologous chromosome) and controlling for the effects of variants discovered in prior steps. In the initial steps (Stage 1; Steps 1-12), we required the allele on the opposite, homologous chromosome to carry a previously identified null allele (carrying rs41272114 or rs41259144 (combined European MAF=0.049), both of which are known to create alleles that produce no detectable serum Lp(a) (*93*, *95*)), allowing us to perform association analyses within an effective haploid model of Lp(a). To increase power to detect effects of rarer variants (Stage 2; Step 13) and variants with weaker effects (Stage 3; Steps 14, 15, and 17), we then expanded our analysis to allow the allele on the opposite, homologous chromosome to carry any of the variants identified in Steps 1-12 whose carriers appeared to produce low Lp(a); this criterion was satisfied for 30% of all alleles (see legend of Table S4 for precise definition). For variant discovery in African-ancestry samples (Stage 3; Step 16), we placed no restriction on the homologous alleles, and instead modeled Lp(a) from KIV-2 length (non-parametrically) and the variants discovered in the prior steps of analysis as described below.

At each step, we removed alleles carrying large-effect variants discovered in prior steps (except at Step 17, where we specifically restricted to carriers of KIV-2.2 +0G>A (G4925A) to search for sub-haplotypes carrying causal variants). We modeled the effects of small-effect common variants identified in prior steps by including them as covariates in linear regression association tests on log(Lp(a)) (Stage 3; Steps 15-17).

Our stepwise analysis targeted variants in three stages:

Stage 1, Steps 1-12: Large-effect, common and low-frequency variants. We first sought to identify large-effect variants that drastically reduced Lp(a) despite occurring on alleles with short or medium KIV-2 lengths (that would usually be expected to produce moderate-to-high levels of Lp(a)). We performed Fisher's exact tests on variants for a binarized Lp(a) phenotype (low vs. high), restricting to alleles opposite null alleles (see above) and whose KIV-2 lengths were restricted to a particular range. The Lp(a) cutoff and maximum KIV-2 length varied between steps (Table S4).

In 12 consecutive steps of analysis, this procedure identified 12 variants each of which was predicted to be protein-altering and achieved the top association statistic in a step of analysis. The variants identified in the first two steps were noteworthy both for their high allele frequencies (European MAF=13% and 20%, respectively) and their genomic locations, which were within the KIV-2 repeat on either end of KIV-2 exon 2 (Table S4). The first variant alters the final base of KIV-2 exon 2 and was recently demonstrated to impair splicing at the adjacent splice acceptor site (*20*). The second variant is in the splice region preceding KIV-2 exon 2 and was recently identified and observed to associate with reduced Lp(a) (*85*) (P=0.0052 in N=23 carriers). Our analysis provided strong statistical evidence of the causality of this association, and we further observed that this variant was computationally predicted to impair splicing of the nearby splice donor site (SpliceAI (*19*) donor loss Δ score = 0.79).

Interestingly, the computational prediction indicated that the KIV-2 exon 2 -11G>A variant might either simply impair the splice donor site or also create a cryptic splice donor 9bp upstream (causing an inframe insertion introducing an IleSerSer amino acid sequence). To determine if this PSV created an alternative splice site, we examined RNA-seq reads from the GTEx project (*97*) (v8), across all donors and tissues, to look for any reads showing evidence of the use of a different splice site at the location of the PSV, but we found no evidence that the PSV created an alternative splice site, suggesting that it simply impairs the canonical splice donor site. (The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described above were obtained from the AnVIL GTEx portal for GTEx v8 on 01/07/21.)

Stage 2, Step 13: Rare variants producing potential null alleles. Next, we sought rare variants likely to create additional null alleles (producing undetectable Lp(a)). To do so, we performed Fisher's exact tests against Lp(a) binarized at the lower limit of the reportable range (3.8 nmol/L), restricting association analysis to alleles opposite the 30% of alleles expected to produce low Lp(a) based on Steps 1-12 (see above). We considered only variants that approximately matched the profile of known null alleles in terms of the frequency of Lp(a) measurements below the reportable range (<3.8 nmol/L) in this setup. (More precisely, 85% of Lp(a) measurements were below 3.8 nmol/L among null allele carriers for which the opposite, homologous allele was expected to produce low Lp(a), so we only considered variants for which the 95% CI of this frequency overlapped 0.85) This analysis identified 11 rare variants with top associations, of which five variants (four protein-altering and one in the 5' UTR of *LPA*) appeared likely to be causal. Two additional variants were in LD with the protein-altering variants and one variant was in LD with the 5' UTR variant. The remaining three variants were >100kb away from *LPA* and we were unable to identify plausible causal variants underlying these associations (Table S4).

Stage 3, Steps 14-17: Moderate-effect, common variants. After excluding alleles carrying large-effect variants (Steps 1-13), we sought to identify common variants with smaller effects on Lp(a). In Steps 14 and 15, we restricted to alleles with KIV-2 length in the interval 11 to 18 opposite alleles expected to produce low Lp(a). Among these alleles, the relationship between KIV-2 length and log(Lp(a)) appeared to be linear. We therefore ran linear regression analyses to test log(Lp(a)) for variant associations including KIV-2 length and KIV-2 length squared as covariates. These analyses identified two additional variants, one in the 5' UTR (with prior support from *in vitro* studies (*22*)) and a missense SNP, both of which appeared to exert a moderate and consistent influence on Lp(a) across the KIV-2 spectrum.

We next ran a similar analysis using data from African-ancestry participants, hoping to identify additional Lp(a)-modifying variants that may be less common among Europeans. To maximize power, we created an interim model for Lp(a), fit using measurements in Europeans, so we could use data from all exome-sequenced participants of African-ancestry (and place no restriction on the opposite allele). We first estimated the contribution of a "clean" allele (i.e., one without any previously identified large effect variant) to Lp(a) by averaging Lp(a) measurements for all (European and exome-sequenced) individuals with a clean allele of similar KIV-2 length (within 3 repeat units) whose opposite allele carried a null variant (rs41272114 or rs41259144). We then estimated multiplicative factors for Lp(a)-modifying variants by linear regression on log(Lp(a)) – log(Lp(a) estimated from KIV-2 length). Finally, we computed Lp(a) predictions in African-ancestry participants as the sum of the contributions of the two alleles. In Step 16, we ran linear regression analyses on log(measured Lp(a)) – log(predicted Lp(a)), identifying an additional 5' UTR SNP (also with prior support from *in vitro* studies (*21*, *22*)). Minor allele carriers of this

SNP appeared to have modestly increased Lp(a) across the KIV-2 spectrum in both Africans and Europeans (Fig. 1A).

Finally, in Step 17, we performed an analysis similar to Steps 14 and 15 among carriers of the splice modifier variant KIV-2.2 +0G>A (G4925A) and identified a missense SNP (mostly contained on alleles carrying the splice modifier variant) that appeared to reduce Lp(a) (Fig. S12).

The variants identified by this multi-stage analysis account for almost all (>97%) of the short and medium European *LPA* alleles (<19 KIV-2 repeat units) that produce small amounts of serum lipoprotein(a) (<6 nmol/L when opposite null alleles carrying rs41272114 or rs41259144). This list of variants appears to be comprehensive of all large-effect, Lp(a)-reducing variants polymorphic in Europeans except those that are very rare. We expect that additional low-frequency and rare null alleles remain to be found in other populations.

8 Predictive model of Lp(a) from KIV-2 and fine-mapped *LPA* coding and 5' UTR variants

We modeled lipoprotein(a) concentration as the sum of the contributions of maternally- and paternally-inherited *LPA* alleles consisting of phased KIV-2 lengths and SNP haplotypes for the coding and UTR SNPs identified by our fine-mapping pipeline (Table S4):

$$\mathrm{Lp(a)} = f(\mathrm{KIV2_{mat}}, \mathrm{SNPs_{mat}}) + f\big(\mathrm{KIV2_{pat}}, \mathrm{SNPs_{pat}}\big).$$

We did not include other variables in our predictive model because genetic variation at *LPA* explained the vast majority of Lp(a) variance (>80%), whereas other covariates explained very small amounts of variance (e.g., sex (0.2%), age (0.1%), and other loci in the genome (<0.1%)).

The contribution of each (haploid) *LPA* allele appeared to be well-modeled as a "baseline" function of KIV-2 length alone, on top of which coding and UTR SNPs on the same haplotype act via multiplicative factors (Fig. 1A). As such, we modeled each haplotype's contribution as:

$$f\big(\mathrm{KIV2_{hap}}, \mathrm{SNPs_{hap}}\big) = f_{\mathrm{baseline}}\big(\mathrm{KIV2_{hap}}\big) \prod_{\substack{\text{SNPs with ALT} \\ \text{allele on hap}}} c_{\mathrm{SNP}}$$

where c_{SNP} denotes a constant factor indicating the magnitude of the effect of each Lp(a)-modifying SNP that our fine-mapping pipeline identified (Table S4).

Empirically, the baseline curve $f_{\mathrm{baseline}}\big(\mathrm{KIV2_{hap}}\big)$ appeared to have two key properties: (i) an exponential decay that governs the relationship between KIV-2 length and Lp(a) for medium-length and longer alleles; and (ii) an inflection point that causes this relationship to break down for short alleles, with the shortest alleles apparently contributing less to Lp(a) than less-short alleles (Fig. 1A). Property (i) is equivalent to the logarithm of the baseline curve $\log(f_{\mathrm{baseline}})$ having a linear asymptote. To fit these behaviors while avoiding overparameterization of the baseline curve, we therefore modeled the logarithm of the baseline curve using a conic section (specifically, a hyperbola), which could produce properties (i) and (ii) using only five parameters:

$$\log\big(f_{\mathrm{baseline}}(x)\big) = c\sqrt{(x-a)^2 + b} + dx + e$$

where $x = \mathrm{KIV2_{hap}}$. We adopted this functional form only to approximately capture the empirical behavior of the relationship between KIV-2 length and Lp(a) while avoiding overfitting; we do not expect that the form itself or its parameters directly correspond to any particular feature of the underlying biology.

Model-fitting

Our main challenge in fitting the parameters of this model from UK Biobank data was the "cropping" of Lp(a) measurements provided by UK Biobank. Specifically, the Randox AU5800 instruments that UK Biobank used to perform immunoturbidimetric assays measuring Lp(a) had a reportable range of 3.8–189 nmol/L. Values of Lp(a) that fell outside this reportable range were indicated “Not reportable at assay (too low)” or “Not reportable at assay (too high)” and were not available to researchers at the time of our study.

To overcome this obstacle, we restricted our model-fitting analyses to a subset of participants whose Lp(a) measurements were largely unaffected by the cropping issue based on the combinations of *LPA* alleles they carried. Specifically, we began with the set of 40,589 exome-sequenced participants who self-reported European ancestry, passed PC-filtering (Materials and Methods) and relatedness pruning (which dropped one member of each ≤2nd-degree related pair, prioritizing retaining those with Lp(a) measurements), and had not withdrawn from UK Biobank. We then fit the coefficients of our haploid model $f\left(\mathrm{KIV2}_{\mathrm{hap}}, \mathrm{SNPs}_{\mathrm{hap}}\right)$ on the subset of haplotypes for which the “opposite haplotype” (i.e., the *LPA* allele inherited from the other parent) was predicted to produce little or no Lp(a) (<2 nmol/L). This criterion essentially isolated the effect of one *LPA* allele, which both reduced noise in Lp(a) measurements and substantially ameliorated the issue of Lp(a) measurements exceeding the reportable range (as only ~2% of European haplotypes produced >189 nmol/L of Lp(a) on their own, in contrast to ~7% of all (diploid) measurements from European-ancestry participants exceeding 189 nmol/L). This approach was a slight relaxation of the “effective-haploid” model (considering haplotypes opposite an Lp(a)-null allele) that we utilized during fine-mapping; we relaxed the criterion on the other haplotype to “predicted Lp(a) <2 nmol/L” to increase sample size (from ~5% of alleles opposite an Lp(a)-null allele to ~19% of alleles opposite an allele predicted to contribute <2 nmol/L to Lp(a)). While this approach incurred the slight inconvenience of needing a predictive model of Lp(a) to apply to opposite haplotypes during model-fitting, we overcame this difficulty by iteratively fitting the model and updating opposite-haplotype predictions until convergence.

Within each model-fitting iteration, we fit model parameters in two steps:

1. Estimate parameters of the baseline curve (5 parameters) and moderate-effect SNPs (3 multiplicative parameters for the 5’ UTR SNPs rs1800769 and rs1853021 and the missense SNP rs3124784).

   This step estimated the way in which KIV-2 length and common, moderate-effect SNPs determined Lp(a) in the majority of European haplotypes that did not carry a large-effect SNP that greatly reduced Lp(a). We fit the 8 parameters in question (5 parameters for the $f_{\mathrm{baseline}}\left(\mathrm{KIV2}_{\mathrm{hap}}\right)$ curve and 3 parameters for the SNPs) by using Matlab’s fminunc optimization routine to minimize squared error between measured Lp(a) (adjusted for the small predicted contribution of the opposite haplotype) vs. Lp(a) predicted by the model fit, with two additional subtleties.

   First, we restricted model-fitting to KIV-2 length ranges that were largely unaffected by the cropping of Lp(a) measurements to 3.8–189 nmol/L, adjusting the KIV-2 length range according to the alleles of the 3 moderate-effect SNPs. Specifically, for *LPA* alleles carrying the Lp(a)-increasing rs1800769:T minor allele (but not carrying the Lp(a)-reducing rs3124784:A minor allele), we restricted to KIV-2 in the range [14, 24]; for *LPA* alleles carrying neither of these minor alleles nor the rs1853021:A minor allele, we restricted to KIV-2 in the range [14, 23]; and for all other LPA alleles, we restricted to KIV-2 in the range [0, 23].

Second, for the rare remaining instances in which an allele satisfying the above criteria (and opposite an allele predicted to contribute <2 nmol/L to Lp(a)) still produced measured Lp(a) exceeding the reportable range, we set "measured Lp(a)" to its posterior mean assuming Lp(a) was drawn from a log-normal distribution with $\sigma = 0.304$ (which appeared empirically to approximately capture variance in Lp(a) measurements after controlling for genetic effects) and with μ set to match the fraction of Lp(a) measurements exceeding the reportable range (among alleles with the same Lp(a)-modifying SNP haplotypes and with similar KIV-2 lengths, i.e., in the same 2-percentile bin of KIV-2 length).

2. Estimate multiplicative parameters for effects of large-effect SNPs.

   Most of the large-effect coding SNPs identified by our fine-mapping procedure (including all canonical splice site variants and stop gain variants) produced no detectable Lp(a) in nearly all individuals who carried an Lp(a)-null allele on the opposite haplotype; as such, we set the multiplicative factor $c_{\mathrm{SNP}} = 0$ for these SNPs. Six large-effect SNPs appeared to each reduce Lp(a) by >70% but clearly did not completely abolish Lp(a) production: 3 cryptic splice SNPs (two near the exon-intron boundary of KIV-2 exon 2 and rs41270998) and 3 missense SNPs (rs41267807, rs41267809, and rs76144756). We estimated the multiplicative effects of these SNPs by comparing mean measured Lp(a) to mean predicted Lp(a) (predicted based on the 8 parameters fit in step 1) in carriers of these SNPs with KIV-2 length <16 (restricting to shorter KIV-2 alleles to improve signal-to-noise in these estimates.

The parameter values we obtained in this way for the 5 parameters of the baseline curve $f_{\mathrm{baseline}}(\mathrm{KIV2}_{\mathrm{hap}})$ were:

$$a = 10.11, b = 30.18, c = -0.227, d = -0.087, e = 7.437$$

The multiplicative factors c_{SNP} we estimated for the Lp(a)-modifying SNPs are provided in Table S4.

Predicting Lp(a)

Given phased KIV-2 lengths and *LPA* SNP haplotypes for an individual, our model provides a prediction of Lp(a) as the sum of predicted contributions of each *LPA* allele. These genetically-predicted values of Lp(a) were not subject to cropping to the reportable range of 3.8–189 nmol/L, offering an opportunity to examine the cardiovascular effects of extremely high Lp(a) (e.g., >400 nmol/L). However, for the purpose of assessing the accuracy of our model, we needed to compare predicted Lp(a) to measured Lp(a), so in benchmarking analyses we cropped predicted Lp(a) (after summing across each individual's two haplotypes). Stratifying by self-reported ethnicity among exome-sequenced individuals, we obtained

$$R^2\big(\text{cropped pred. Lp(a), measured Lp(a)}\big) = 0.83\ (\text{EUR}), 0.61\ (\text{SA}), 0.54\ (\text{AFR}), 0.62\ (\text{EAS}).$$

(While model-fitting was performed in a subset of the same European-ancestry participants, overfitting is not a concern given the very small number of parameters fit.) The decreased prediction accuracy in non-European ancestries likely arises from a combination of unmodeled Lp(a)-decreasing variants (with higher non-European allele frequencies) and lower accuracy of phased KIV-2 length estimates in the small fraction of non-European UK Biobank participants.

9 Effect of within-repeat variation on phenotypes

With the exception of *LPA*, within-repeat variation appeared to play at most a minor role in the VNTR-phenotype associations we identified. Specifically:

- At *LPA*, we genotyped within-repeat SNP and indel variation accessible to exome sequencing (i.e., in the coding and splice regions of the KIV-2 repeat), and we considered these variants at each step of our deep fine-mapping of Lp(a), as described above. We found several such variants that appear to strongly reduce Lp(a) (Table S4).

- At *ACAN*, we determined that copy-number genotypes for the four common repeat types explained at most a small amount of additional variance in height (0.354% of height variance explained by a joint model including VNTR length, two fine-mapped missense SNPs, and repeat-type genotypes vs. 0.338% explained by a model including VNTR length and the missense SNPs alone). We were unable to confidently determine how much of the additional 0.016% of variance explained was actually due to causal effects of the within-repeat variants (vs. SNPs in linkage disequilibrium with repeat-type composition), but this analysis indicates that the contribution of any such effects is much smaller (by ~20-fold or more) than that of the total VNTR length and the two implicated missense SNPs.

- At *MUC1*, although individual repeat units harbor extensive common variation (*35*), VNTR length appeared to explain nearly the entire association signal at the locus (linear regression P=6.2 x 10^{-153} for the VNTR, which was the variant most strongly associated with serum urea, and $P > 1 \times 10^{-8}$ for all SNPs within 500kb of *MUC1* upon conditioning on VNTR length). Thus, repeat unit variation appears to also play at most a minor role in the urea association at *MUC1*.

- At *TCHH* and *TENT5A*, we found no evidence of common variation among repeat units. For *TCHH*, long-read assemblies of 64 unrelated haplotypes generated by the Human Genome Structural Variant Consortium (HGSVC2; ref. (*14*)) contained no such variation. For *TENT5A*, the gnomAD v3.1.1 database (*98*) contained no common (MAF>1%) variants within the VNTR (which is sufficiently short to be spanned by short-read sequencing reads).

10 Cross-species comparisons of VNTR sequences

We investigated the level of conservation of each of the five VNTRs primarily studied here (in *LPA*, *ACAN*, *TENT5A*, *MUC1*, and *TCHH*) across nine primates (human, chimp, bonobo, gorilla, orangutan, gibbon, rhesus, baboon, and marmoset) and two other mammals (mouse and dog). This analysis was based solely on reference genomes for each species. We attempted to map each VNTR to each species, and for each VNTR that successfully mapped, we measured its length and consensus repeat unit in the reference genome. The results are reported in Table S7 and summarized below.

- ***LPA***: *LPA* is known to have originated through duplication of the plasminogen (*PLG*) gene and is found only in primates and old world monkeys (*12*). Analysis of the repetitive exons comprising the kringle domains (including all kringle domains, not just the KIV-2 domain that is highly polymorphic in humans) found similar assembled kringle domain lengths across primate reference sequences. However, the accuracy of reference genome assemblies across the KIV-2 region is questionable given the difficulty of assembling across such a long repeat; almost no humans have KIV-2 alleles as short as the hg38 reference (which only contains 6 KIV-2 repeats).

- ***ACAN***: The *ACAN* consensus sequence is generally well-conserved across mammals. Humans appear to have an expansion relative to other primates. Interestingly, mouse seems to have an extra codon in the homologous sequence, consistently across all of the repeat units. This extra codon is not present in humans and closely related primates, but appears in some repeat units in dog (5), marmoset (3), baboon (2), rhesus (2), orangutan (1), with the number of extra codons decreasing with decreasing divergence from human.

- ***TENT5A***: The *TENT5A* VNTR is the most diverged part of the gene. The VNTR sequence contains considerable variation not apparent in the amino acid sequence. Primates seem to have an expanded number of repeats, with human being somewhat shorter than the other great apes.

- ***MUC1***: *MUC1* is more diverged than the previous VNTRs, with apparent errors in the gene annotations for several reference genomes. We therefore restricted analysis of the *MUC1* VNTR to eight primates for which the VNTR appeared to have been assembled reasonably accurately. Human is greatly expanded compared to other primates, consistent with previous work (*2*), although gorilla and marmoset are also both longer than other primates.

- ***TCHH***: The *TCHH* protein is highly diverged among primates, especially in certain domains, including the VNTR. The gene is also quite repetitive, with multiple domains similar to the VNTR region we studied. As a result, we were unable to reliably identify clearly homologous sequences for the short (18bp, 6 amino acid) VNTR in *TCHH* across primates. This high divergence is consistent with our observation that the 18bp VNTR appears to be hypermutable in humans (exhibiting somewhat lower phasing and imputation accuracy despite being accurately genotyped from exome sequencing read-depth).

Supplementary Figures

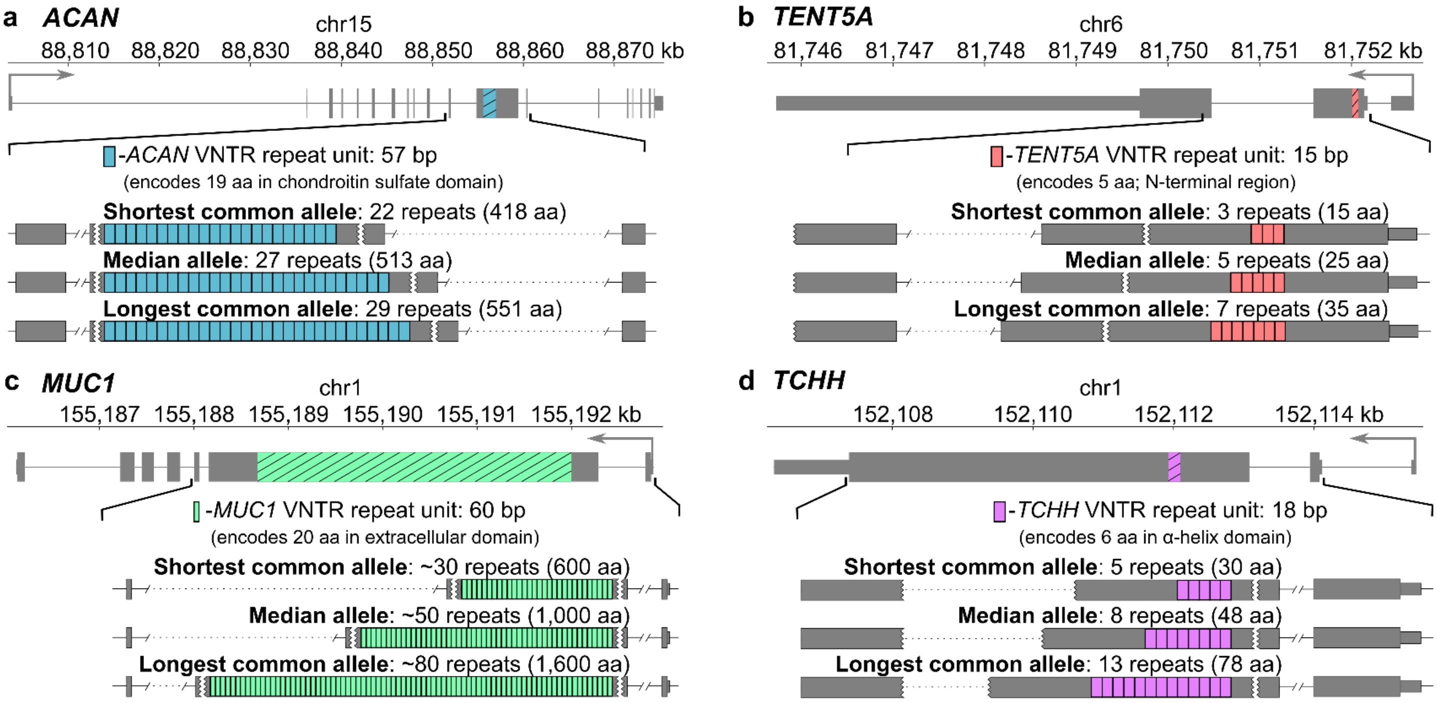
a ACAN
chr15
88,810 88,820 88,830 88,840 88,850 88,860 88,870 kb
-ACAN VNTR repeat unit: 57 bp
(encodes 19 aa in chondroitin sulfate domain)
Shortest common allele: 22 repeats (418 aa)
Median allele: 27 repeats (513 aa)
Longest common allele: 29 repeats (551 aa)
b TENT5A
chr6
81,746 81,747 81,748 81,749 81,750 81,751 81,752 kb
-TENT5A VNTR repeat unit: 15 bp
(encodes 5 aa; N-terminal region)
Shortest common allele: 3 repeats (15 aa)
Median allele: 5 repeats (25 aa)
Longest common allele: 7 repeats (35 aa)
c MUC1
chr1
155,187 155,188 155,189 155,190 155,191 155,192 kb
-MUC1 VNTR repeat unit: 60 bp
(encodes 20 aa in extracellular domain)
Shortest common allele: ~30 repeats (600 aa)
Median allele: ~50 repeats (1,000 aa)
Longest common allele: ~80 repeats (1,600 aa)
d TCHH
chr1
152,108 152,110 152,112 152,114 kb
-TCHH VNTR repeat unit: 18 bp
(encodes 6 aa in α-helix domain)
Shortest common allele: 5 repeats (30 aa)
Median allele: 8 repeats (48 aa)
Longest common allele: 13 repeats (78 aa)

Fig. S1. Exonic VNTRs create large-scale size polymorphisms of protein domains. VNTR alleles of varying lengths are depicted for four VNTRs (within *ACAN*, *TENT5A*, *MUC1*, and *TCHH*) for which we identified large-effect phenotype associations with strong implication of the VNTR by fine-mapping analysis. Gene diagrams indicate the position of the VNTR on the GRCh38 reference; callouts show examples of expanded and contracted alleles (the longest and shortest common (>1% AF) alleles identified among UKB participants of European ancestry).

a

← Direction of transcription
Human reference coordinates →

	Intron position (… +1)	**KIV-2 exon 1 position (160 … 1; minus strand)**			**Intron position (-1 …)**			
Repeat Type	**+119**	**86**	**41**	**14**	**-16**	**-17**	**-23**	**-26**
A	T (**A1**) C (**A2**)	T	A	T	G	G	C	T
B		A	G	C	A	A	-	A
C		T	G	C				
	PSV distinguishing A1/A2 subtypes	Synonymous PSVs within KIV-2 exon 1 distinguishing A/B/C repeat types			Additional intronic PSVs distinguishing A/B (C sequence not in ref)			

b

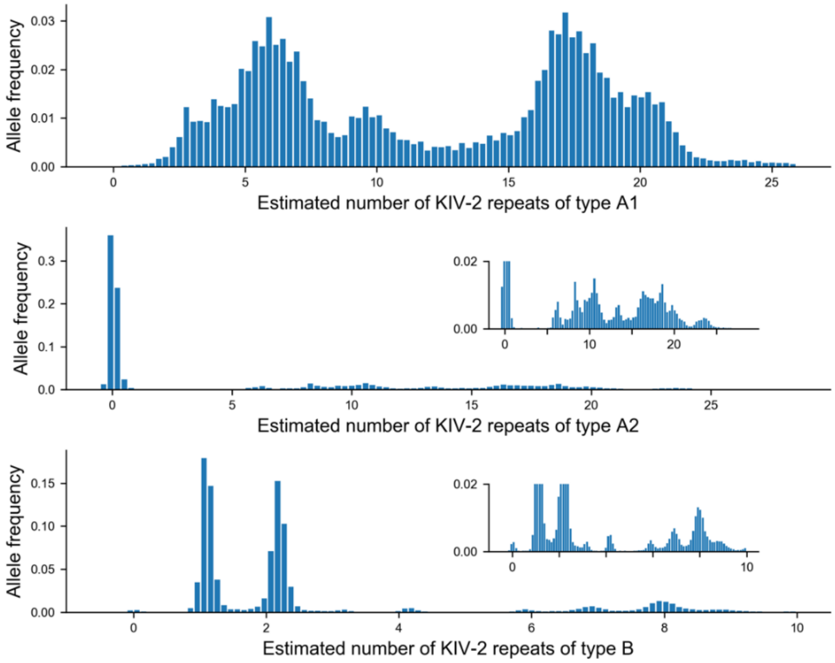

Fig. S2. Genotyping *LPA* VNTR repeat subtypes. a, Repeats within the *LPA* VNTR have previously been classified into three repeat types (A, B, C) based on sequence variation at three synonymous paralogous sequence variants (PSVs) within KIV-2 exon 1. A nearby PSV located 119bp into the downstream intron further subdivides repeat type A into two common subtypes, which we called A1 and A2. We separately estimated allelic copy numbers for repeat types A1, A2, and B by counting sequencing reads that mapped uniquely to one of these three repeat sequences. (We did not specially treat repeat type C, which only comprises ~1% of all repeats (*85*); reads generated from type-C repeats therefore contributed to a mixture of the other repeat type counts depending on which PSVs they spanned, which determined whether they aligned best to repeat A1, A2, or B.) **b**, Separately genotyping and phasing each repeat type produced multimodal copy number distributions with some haplotypes carrying large expansions of repeat types A2 and B. The distribution of type-B repeat copy number estimates exhibited clear near-integer modes. Allele distributions for *N*=34,418 exome-sequenced, unrelated British UKB participants are shown.

a

Repeat type		Repeat unit nucleotide sequence	Amino acid sequence
1		GGGCTTCCTTCTGGAGAAGTTCTAGAGACCACTGCCCCTGGAGTAGAGGACATCAGC	GLPSGEVLETTAPGVEDIS
2		GGGCTTCCTTCTGGAGAAGTTCTAGAGACCGCTGCCCCTGGAGTAGAGGACATCAGC	GLPSGEVLETAAPGVEDIS
3		GGGCTTCCTTCTGGAGAAGTTCTAGAGACTGCTGCCCCTGGAGTAGAGGACATCAGC	GLPSGEVLETAAPGVEDIS
4	00x	GGGCTTCCTTCTGGAGAAGTTCTAGAGACTACTGCCCCTGGAGTAGAGGACATCAGC	GLPSGEVLETTAPGVEDIS
4	01x	GGGCTTCCTTCTGGAGAAGTTCTAGAGACTACTGCCCCTGGAGTAGAGGAGATCAGC	GLPSGEVLETTAPGVEEIS
4	1x0	GGGCTTCCTTCTGGAGAAGTTCTAGAGACTACTGCCCCTGGAGTAGATGAGATCAGC	GLPSGEVLETTAPGVDEIS
4	1x1	GGGCTTCCTTCTGGAGAAGTTCTAGAGACTACTGCCCCTGGAGTAGATGAGATCAGT	GLPSGEVLETTAPGVDEIS
E		GGGCTTCCTTCTGGAGAAGTTCTAGAGACTTCTACCTCTGCGGTAGGGGACCTCAGT	GLPSGEVLETSTSAVGDLS
		** * * *	

2 variants in
"diagnostic codons"
-> types 1-4

3 variants
distinguish
repeat 4 types

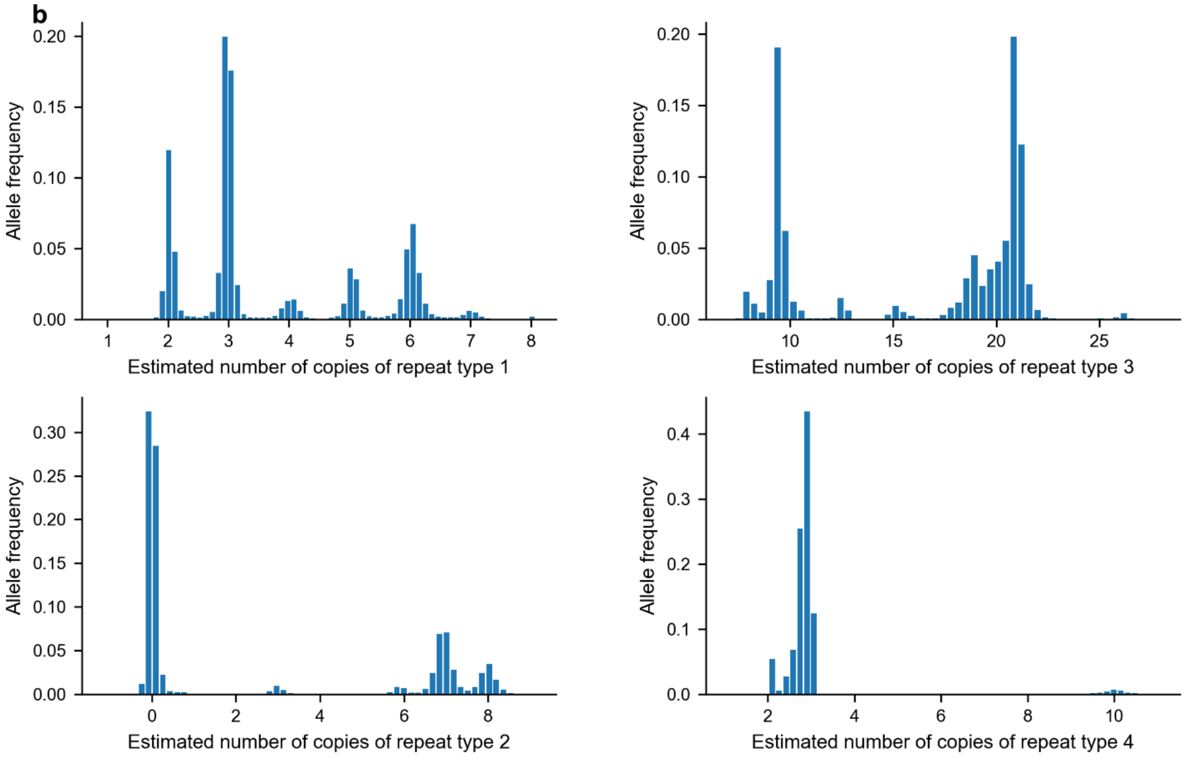

Fig. S3. Genotyping *ACAN* VNTR repeat subtypes. a, Repeats within the *ACAN* VNTR can be broadly classified into four repeat types according to sequence variation at two consecutive "diagnostic" base pairs (*29*). The fourth repeat type can be further subdivided based on three additional paralogous sequence variants (PSVs). The final, partially-diverged repeat unit ("repeat E") of the *ACAN* VNTR is lost in a few rare alleles, creating unique sequence that we identified directly from exome sequencing reads. **b**, Separately genotyping and phasing each repeat type based on counts of sequencing reads containing the corresponding PSVs enabled accurate, integer-mode estimates of repeat type counts within *ACAN* VNTR alleles. Allele distributions among *N*=46,472 exome-sequenced UKB participants of European ancestry are shown.

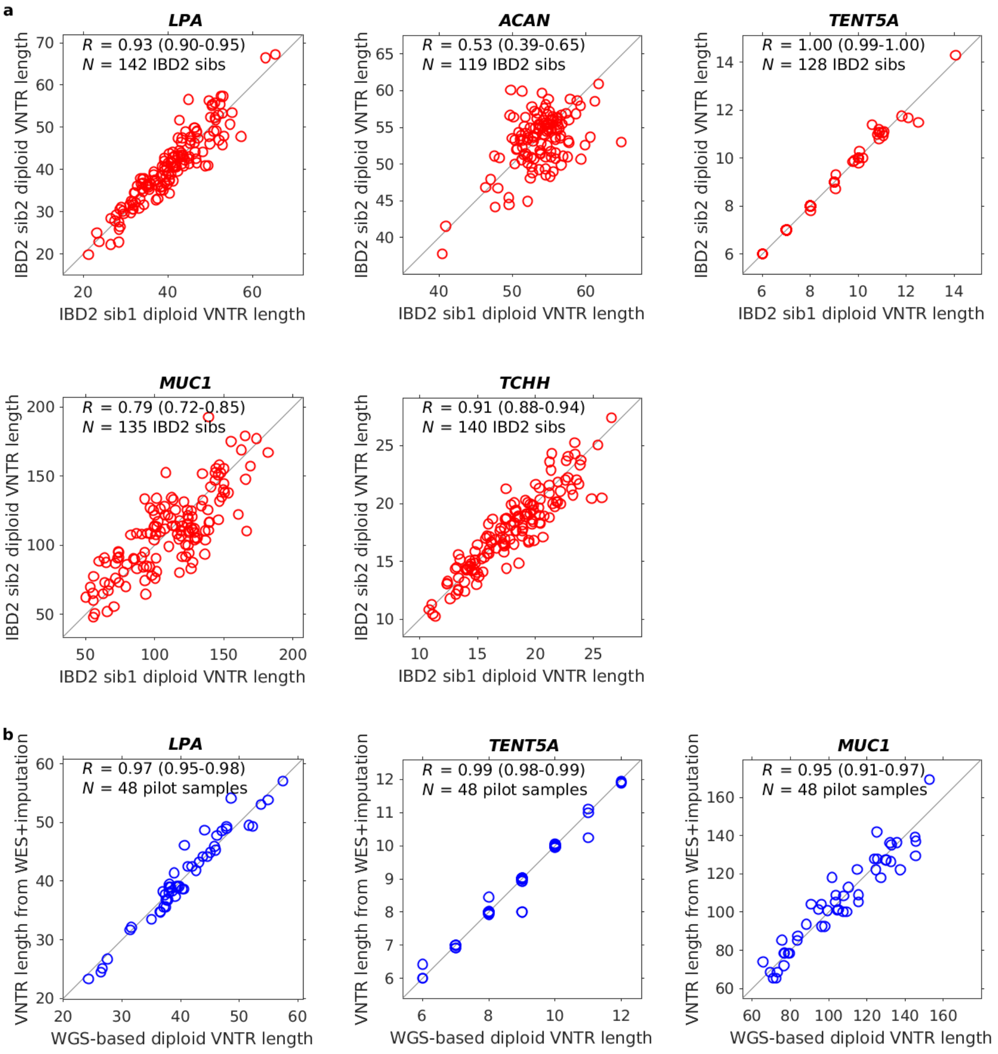

Fig. S4. Benchmarks of VNTR genotyping and imputation accuracy. **a**, Consistency of diploid VNTR length estimates (before phasing, which simultaneously refined allele length estimates) between pairs of siblings sharing both VNTR alleles (i.e., IBD2). Reported values do not exactly match Table S1 because these analyses used optimized VNTR genotypes (whereas Table S1 reports metrics from our initial genotyping pipeline). Counterintuitively, correcting for exome capture bias at *ACAN* decreased IBD2-sib correlation because the variance in the original (biased) length estimates was larger than the variance in the corrected length estimates, such that the biased estimates appeared to be more consistent across IBD2 sibs (since the contributions of particular repeat subtypes were consistently over- or underestimated). **b**, Consistency of VNTR lengths estimated from exome sequencing vs. whole-genome sequencing (available from a pilot WGS analysis of *N*=48 UK Biobank participants, of which 6 had been exome-sequenced; imputed VNTR allele length estimates are plotted for the remaining 42 participants). WGS data enabled accurate diploid VNTR length estimation at *LPA* and *MUC1* (which have high allele length variance) and at *TENT5A* (because all alleles were spanned by 151bp sequencing reads). *ACAN* and *TCHH* allele lengths could not be accurately estimated from WGS.

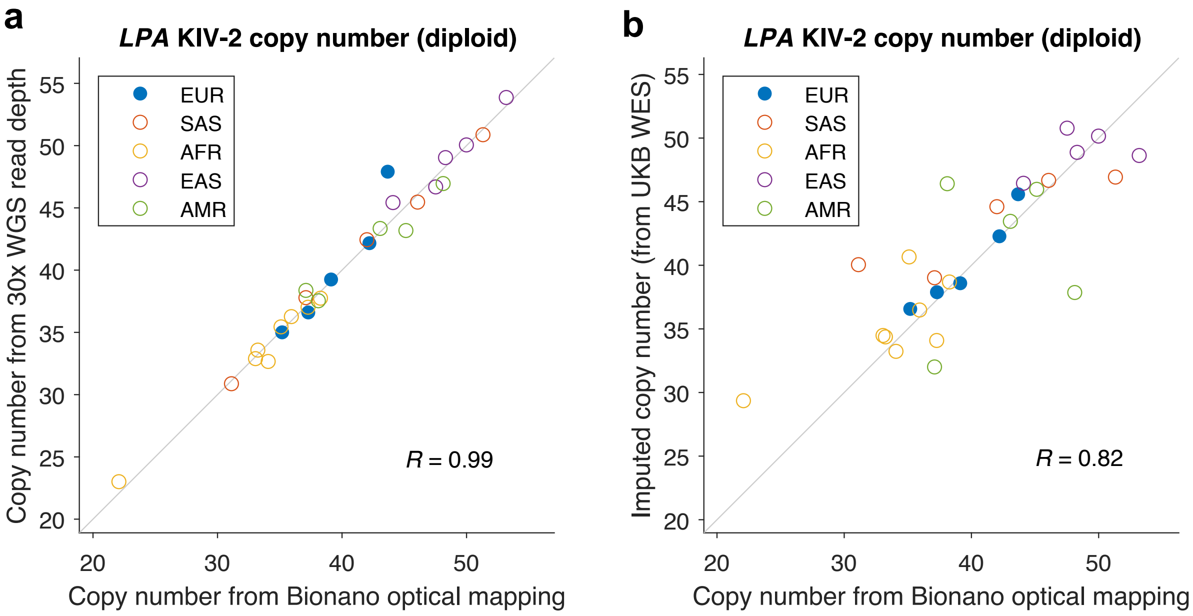

Fig. S5. Validation of WGS- and imputed WES-based *LPA* KIV-2 copy number estimates using Bionano optical mapping data in HGSVC2. **a**, KIV-2 copy number estimated from 30x WGS of 1000 Genomes samples with available Bionano data. **b**, KIV-2 copy number imputed from the *N*~50K UKB WES cohort, with KIV-2 copy numbers estimated using our optimized genotyping procedure. Because Bionano structural variant calls were unphased, both plots compare "diploid copy number" (i.e., KIV-2 copy numbers summed across each individual's two haplotypes) in *N*=28 unrelated individuals for whom both Bionano data and 30x WGS data were available. Bionano and 30x WGS-based estimates were highly concordant (*R*=0.99 considering all *N*=28 individuals), indicating high accuracy of both measurement techniques. Imputed WES-based estimates were highly concordant with Bionano across the 5 EUR samples (solid blue dots); accuracy across the full set of *N*=28 samples was somewhat lower (*R*=0.82), as expected given the difficulty of imputing from UK Biobank into non-European populations.

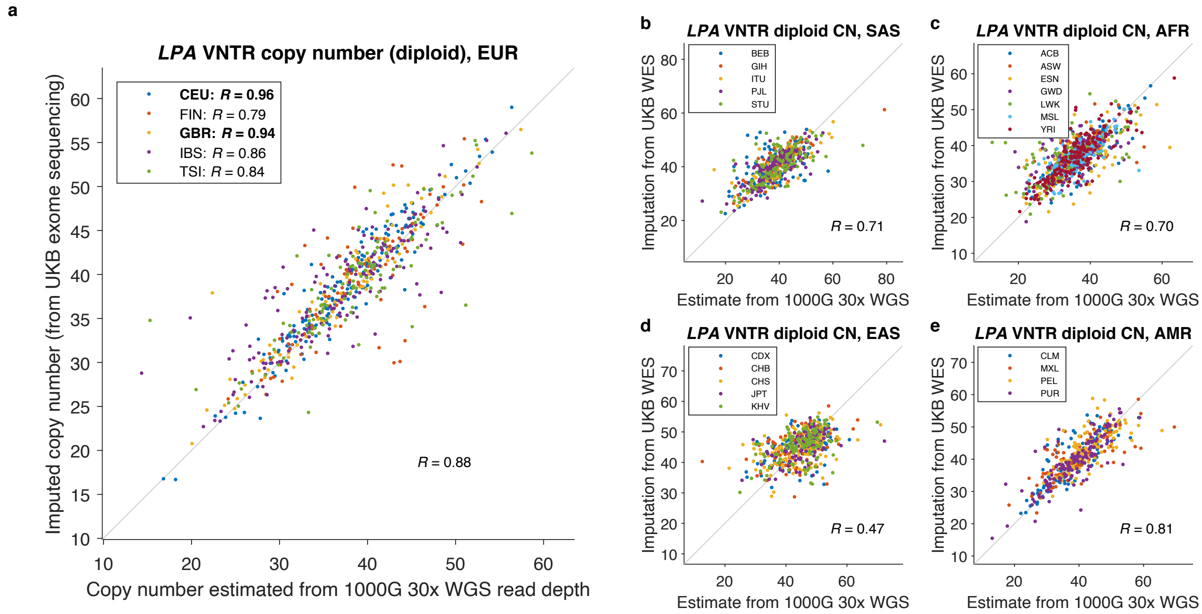

Fig. S6. Validation of WES-based *LPA* KIV-2 copy number estimates (imputed from UKB) against WGS-based estimates in the 1000 Genomes 30x WGS data set (*N*=3,202). Panels are stratified by super-population: **a**, EUR; **b**, SAS; **c**, AFR; **d**, EAS; **e**, AMR. The WGS-based estimates are expected to be nearly exact based on near-perfect concordance with Bionano optical mapping data (*R*=0.99; Fig. S5 above). These comparisons indicate high accuracy of imputed WES-based KIV-2 copy number estimates in populations of northwest European ancestry (*R*=0.96 for CEU, *R*=0.94 for GBR) well-matched to the UKB reference panel, with diminishing accuracy for less-closely-related populations.

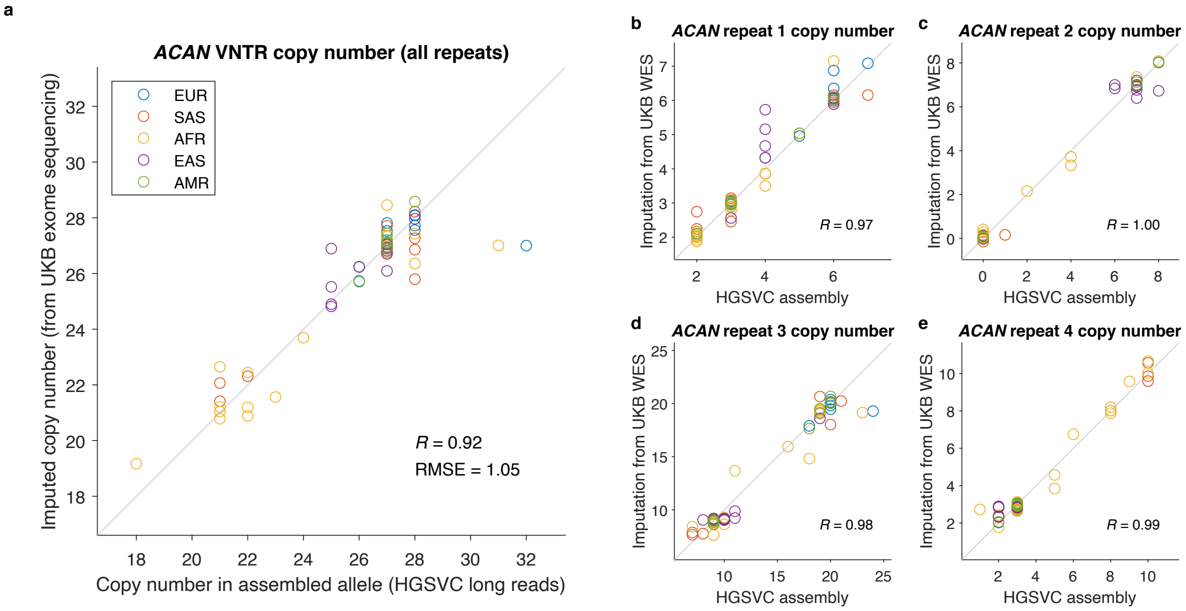

Fig. S7. Validation of WES-based *ACAN* VNTR copy number estimates (imputed from UKB) using haploid genome assemblies from HGSVC2 (*N*=64 haplotypes). a, Total *ACAN* VNTR copy number across all repeat types; **b-e**, copy numbers of repeat types 1-4. *ACAN* VNTR allele lengths were accurately imputed from UKB WES in all 1000 Genomes populations. One EUR haplotype is an outlier (blue circle at right of panel a): the paternally-inherited haplotype of NA12329 (CEU) inherited from NA06984 was imputed as a ~27-allele but appears to actually be a 32-allele based on long-read data and confirmatory evidence from 30x WGS. This haplotype appears to be poorly represented in UKB; the closest-matching haplotype in the *N*=50K WES cohort shared IBS for only ~0.1 cM on either side of the VNTR.

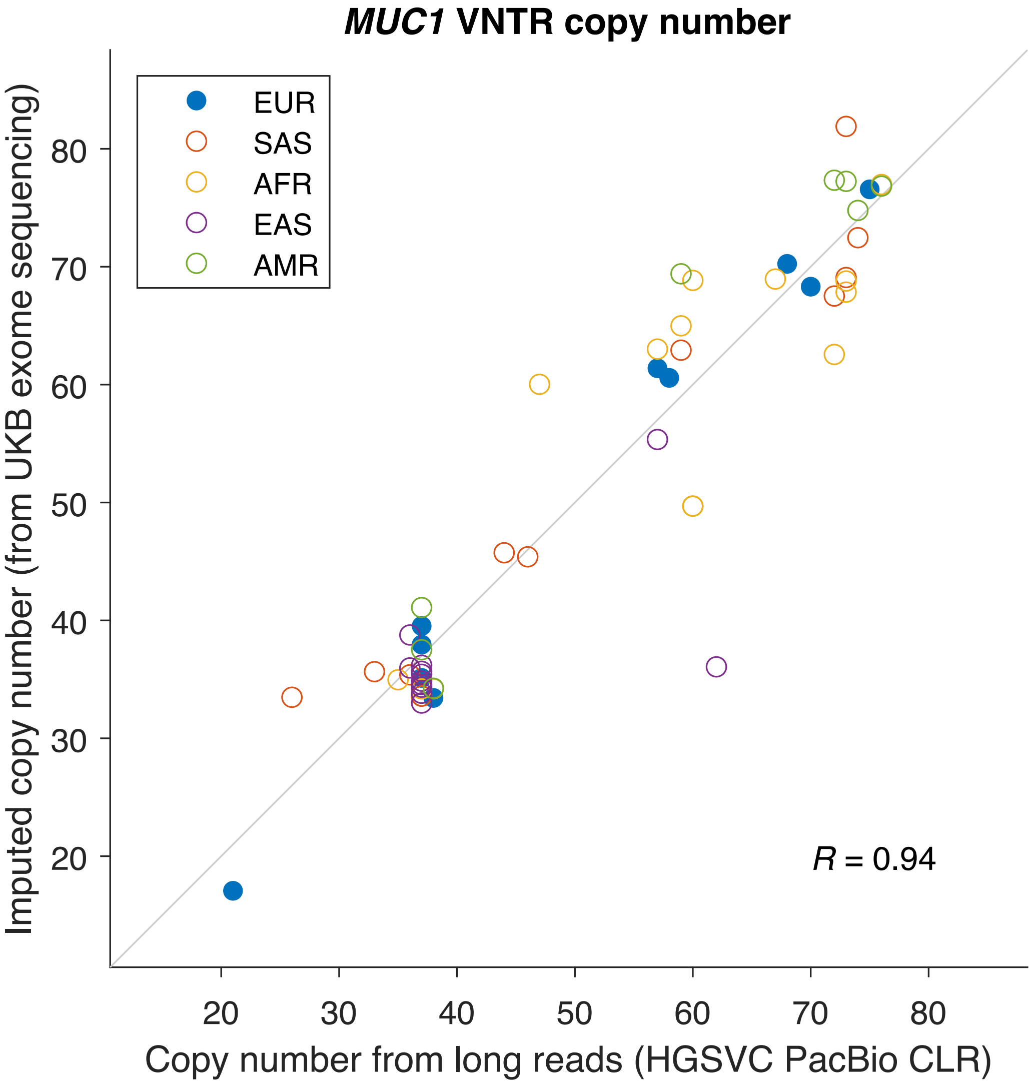

Fig. S8. Validation of WES-based *MUC1* VNTR copy number estimates (imputed from UKB) using PacBio CLR spanning reads from HGSVC2. Haplotype-resolved *MUC1* VNTR allele lengths were determined by reanalysis of PacBio CLR long-read data (Supplementary Text 5). Plotted data points are for *N*=54 haplotypes from *N*=27 unrelated individuals for whom both CLR and 30x WGS data were available). *MUC1* VNTR allele lengths were accurately imputed from UKB WES in all 1000 Genomes populations (*R*=0.94 considering all *N*=54 haplotypes).

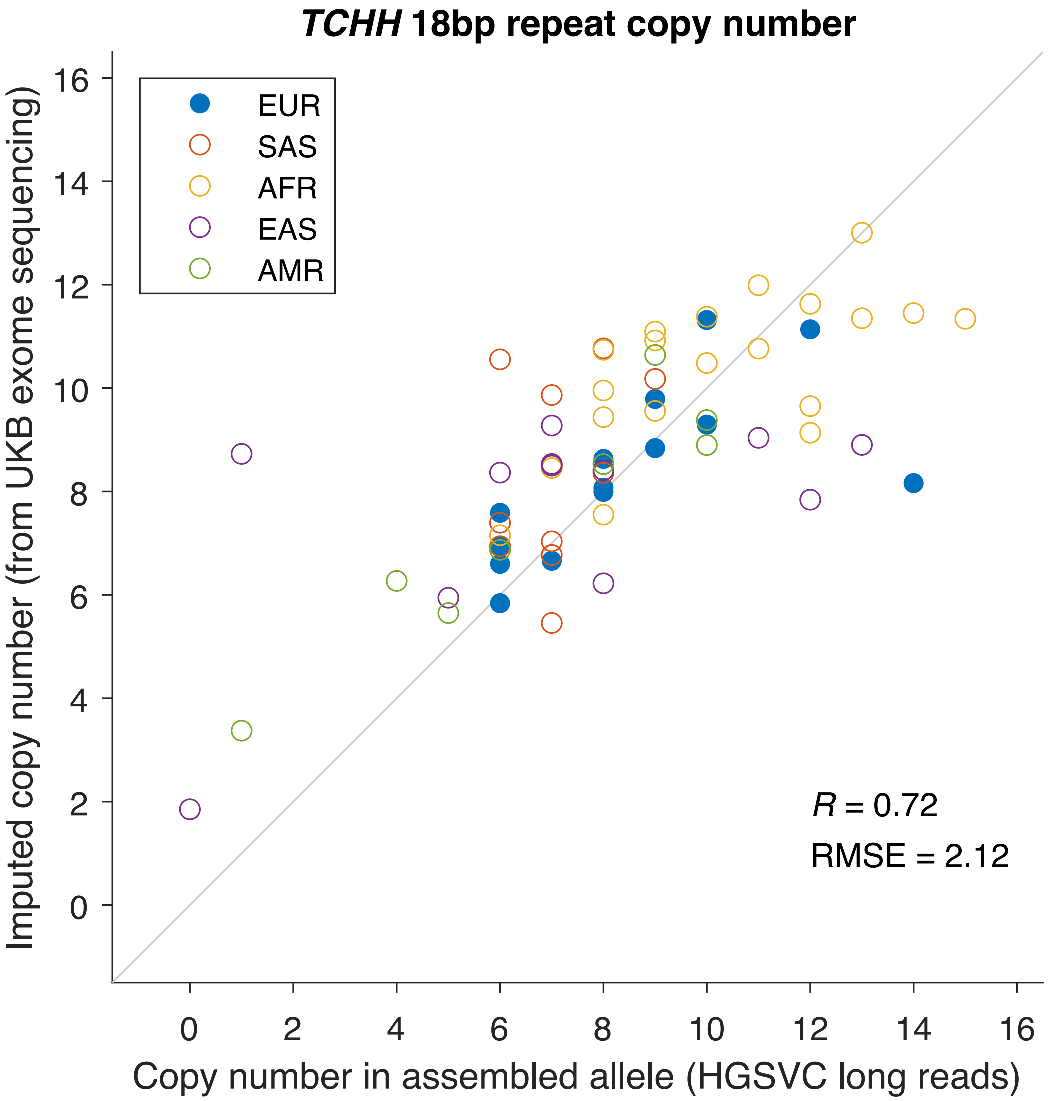

Fig. S9. Validation of WES-based *TCHH* VNTR copy number estimates (imputed from UKB) using haploid genome assemblies from HGSVC2 (*N*=64 haplotypes). The length of the 18bp repeat in *TCHH* was imputed quite accurately into 1000 Genomes EUR samples, whereas accuracy was lower in other populations (*R*=0.72 considering all *N*=64 haplotypes), as expected based on the apparent hypermutability of this repeat. The same EUR haplotype that was an outlier for *ACAN* is again an outlier here (blue dot at right): the paternally-inherited haplotype of NA12329 (CEU) inherited from NA06984 was imputed as an ~8-allele but appears to actually be a 14-allele based on the long-read assembly. This haplotype may have ancestry not well-represented in UKB.

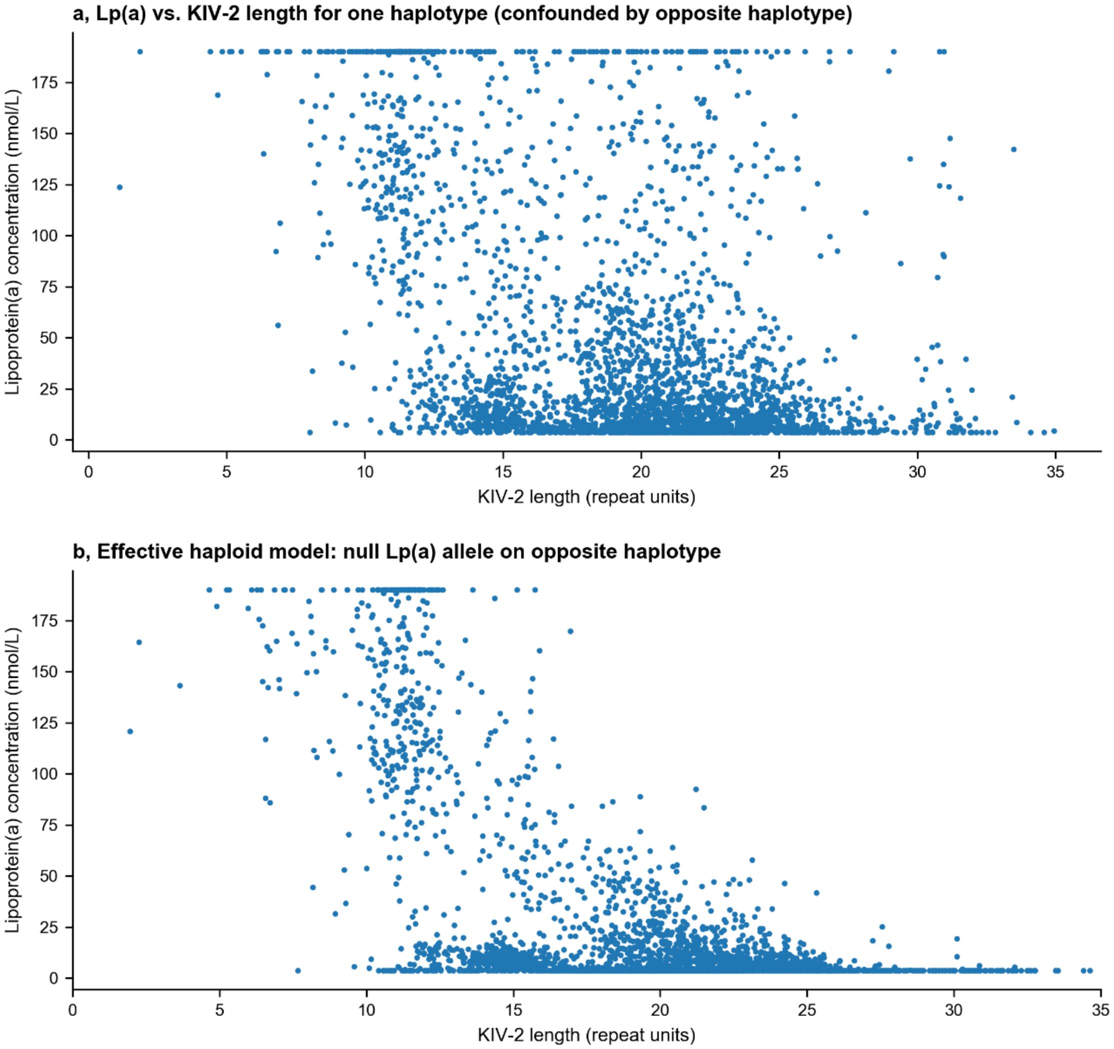

Fig. S10. An effective-haploid model of Lp(a) created by carriers of null-Lp(a) alleles isolates contributions of individual *LPA* alleles. a, Scatter plot of Lp(a) vs. estimated kringle IV-2 length (of one allele) in exome-sequenced UK Biobank participants. The relationship between KIV-2 length and Lp(a) is blurred by the contribution of the *LPA* allele on the "opposite" chromosome (i.e., the homologous chromosome inherited from the other parent). **b**, Scatter plot of Lp(a) vs. KIV-2 length restricted to *N*=3,343 alleles for which the opposite haplotype carries a low-frequency variant known to produce a null-Lp(a) allele (rs41272114 **(*93*)** or rs41259144 **(*95*)**). In both panels, alleles are drawn from exome-sequenced, unrelated, British UKB participants; in panel **a**, alleles are down-sampled to match the count in panel **b**.

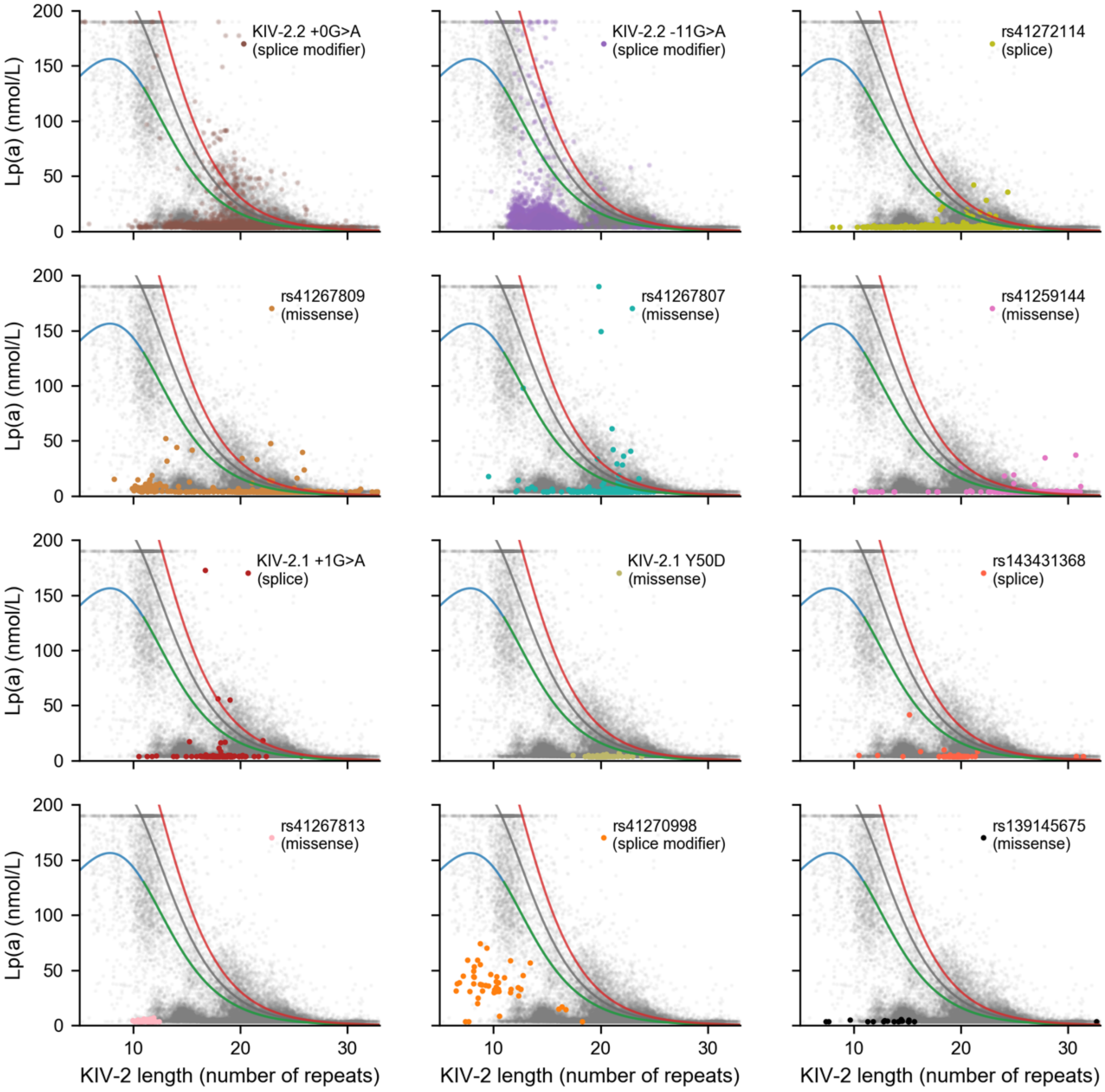

Fig. S11. Lp(a)-reducing effects of 12 *LPA* coding or splice variants. Scatter plots of Lp(a) vs. KIV-2 length, with carriers of a single Lp(a)-modifying variant identified in one of the first 12 steps of conditional fine-mapping analyses (Table S4) highlighted in each panel. Points in each panel correspond to *LPA* alleles in *N*=24,969 exome-sequenced UKB participants of European ancestry for which the allele on the homologous chromosome was predicted to produce little or no Lp(a) (i.e., <4 nmol/L).

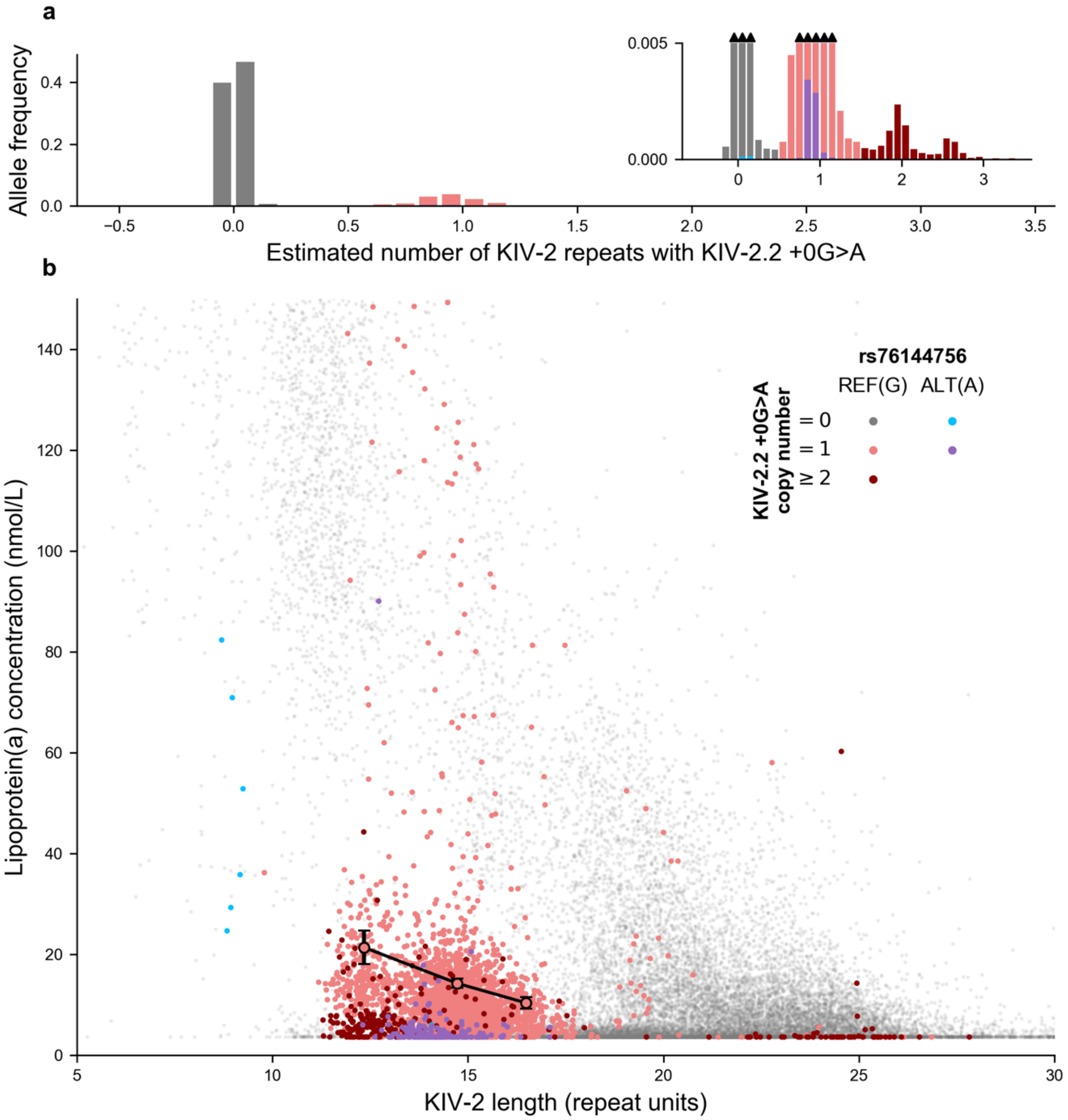

Fig. S12. Epistatic effects of splice modifier variant KIV-2.2 +0G>A (G4925A) dosage, KIV-2 length, and *LPA* missense SNP rs76144756 on Lp(a). a, Distribution (across *N*=68,836 European *LPA* alleles) of the estimated number of KIV-2 repeats carrying a common splice variant in the last base pair of KIV-2 exon 2 that impairs splicing at the adjacent splice junction (G4925A in the nomenclature of ref. (*20*)). **b,** Lp(a) vs. KIV-2 length for alleles in a near-haploid model as in Fig. 2a, with color indicating KIV-2.2 +0G>A dosage and rs76144756 status. Alleles carrying a single copy of the +0G>A splice variant exhibit substantially reduced Lp(a); alleles that additionally carry either another copy of the splice variant (on an additional KIV-2 repeat within the same allele) or the missense SNP rs76144756 exhibit further reduction in Lp(a). The inverse relationship between KIV-2 length and Lp(a) is still evident among alleles with impaired splicing due to carrying one copy of the KIV-2.2 +0G>A splice variant: large markers indicate mean Lp(a) among such alleles carrying the reference allele of rs76144756, binned by KIV-2 length (error bars, 95% CIs).

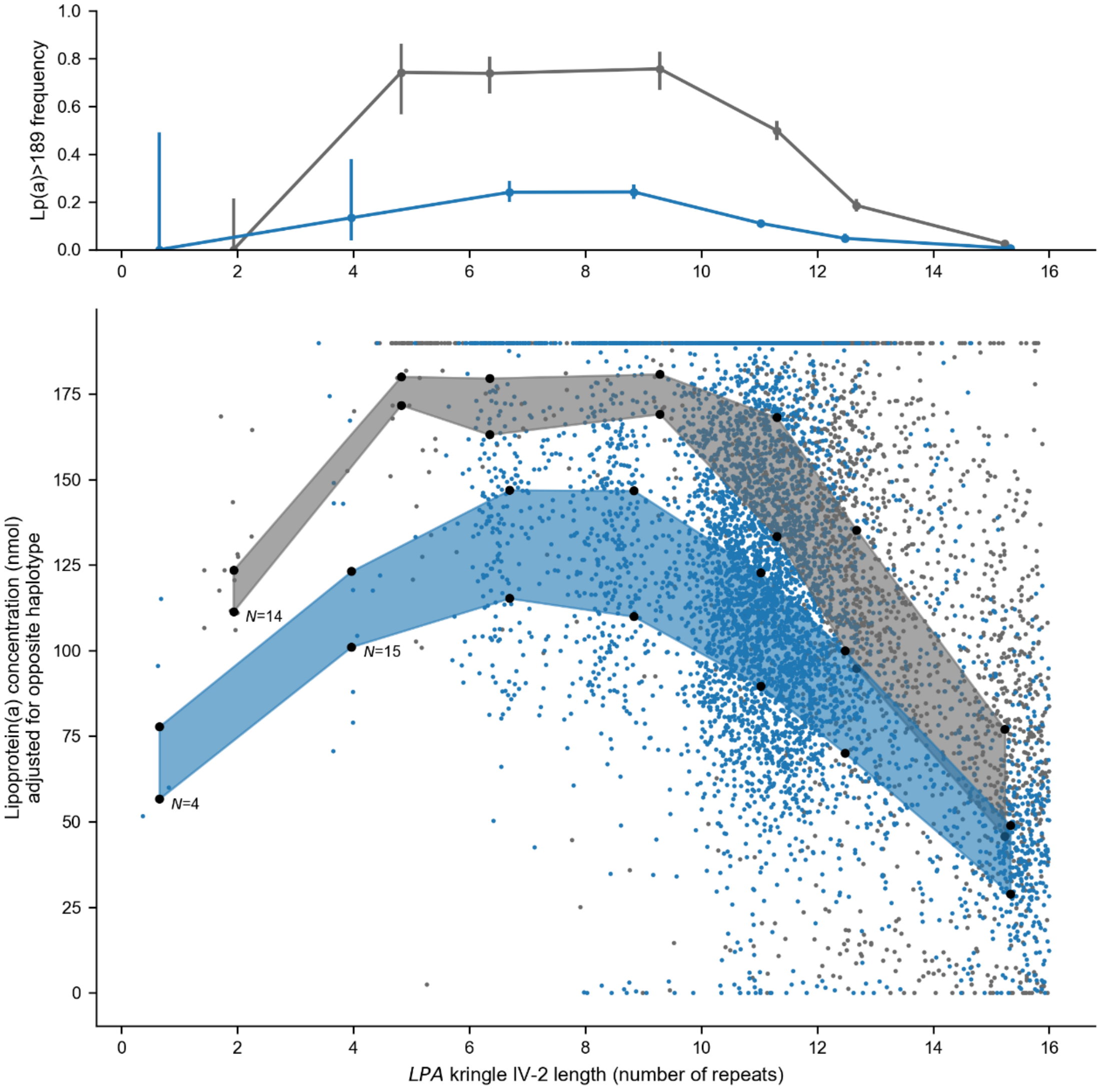

Fig. S13. The relationship between Lp(a) and KIV-2 copy number appears to invert at the short end of the KIV-2 allele length spectrum. Top, Frequency of Lp(a) measurements exceeding the reportable range (>189 nmol/L) as a function of KIV-2 length for short (≤16 repeat units) alleles in exome-sequenced, unrelated European individuals. This analysis was restricted to *N*=8,939 alleles for which (i) the haplotype carrying the allele did not carry any large-effect Lp(a)-reducing variants, and (ii) the allele on the homologous chromosome was predicted to produce relatively low levels of Lp(a) (<40 nmol/L). Error bars, 95% CIs (Wilson score interval for binomial proportion). **Bottom,** Lp(a) vs. KIV-2 length scatter plot for alleles considered in the top panel, with the shaded region indicating the 20^{th}-50^{th} percentiles of Lp(a) (in bins of KIV-2 length). Lp(a) measurements that did not exceed the reportable range (<189 nmol/L) were adjusted for predicted Lp(a) produced by the opposite haplotype.

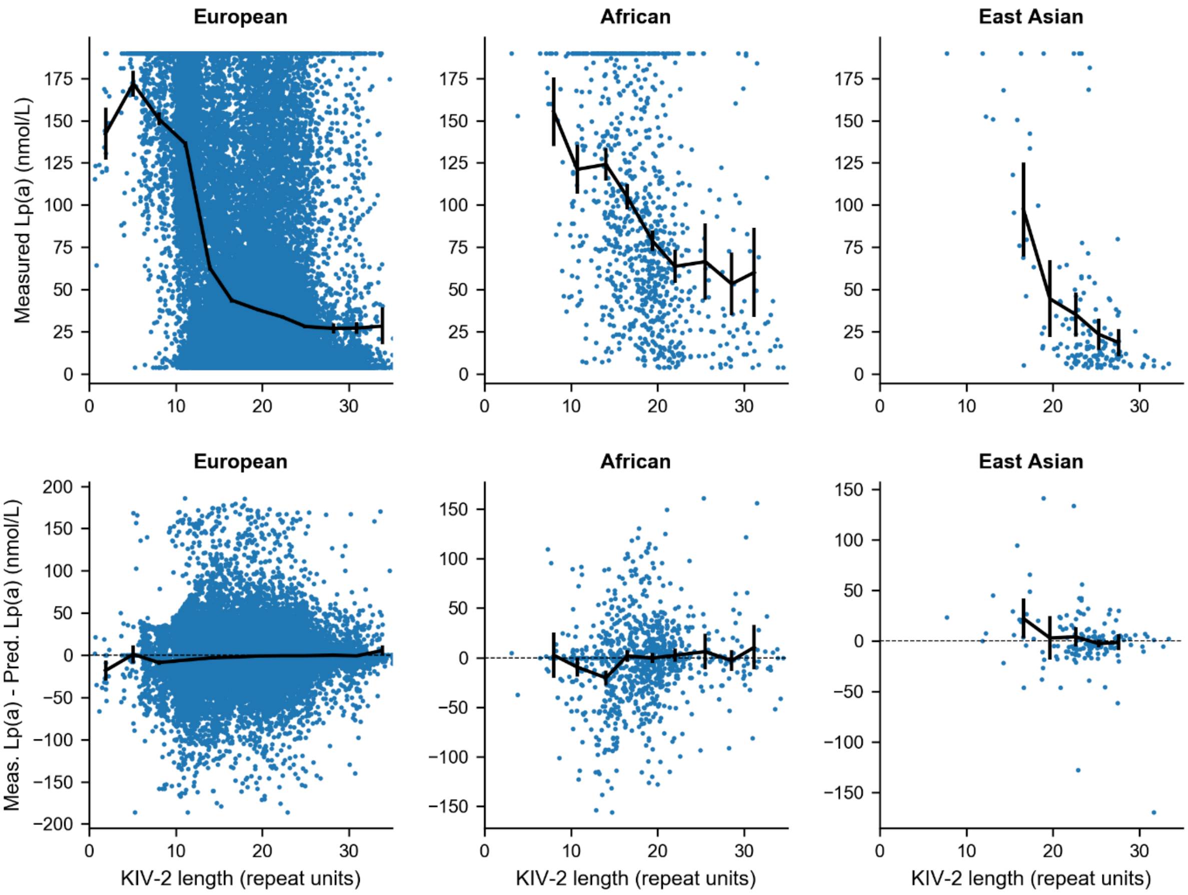

Fig. S14. Allele frequency differences in KIV-2 VNTR length and *LPA* coding and 5' UTR SNPs explain cross-population differences in the relationship between KIV-2 length and Lp(a). Top, Measured Lp(a) vs. KIV-2 allele length (of one haplotype) in exome-sequenced UKB participants of self-reported European ancestry (left; *N*=46,472), African ancestry (middle; *N*=1,008), and East Asian ancestry (right; *N*=173). Mean Lp(a) across KIV-2 length bins is shown in black (error bars, 95% CIs). **Bottom**, Same as in top panel but with measured Lp(a) residualized by genetically predicted Lp(a). The genetic architecture of Lp(a) appears to be consistent across populations, with predictions (based on the model we fit using European-ancestry samples) accounting for the KIV-2 length association in all populations.

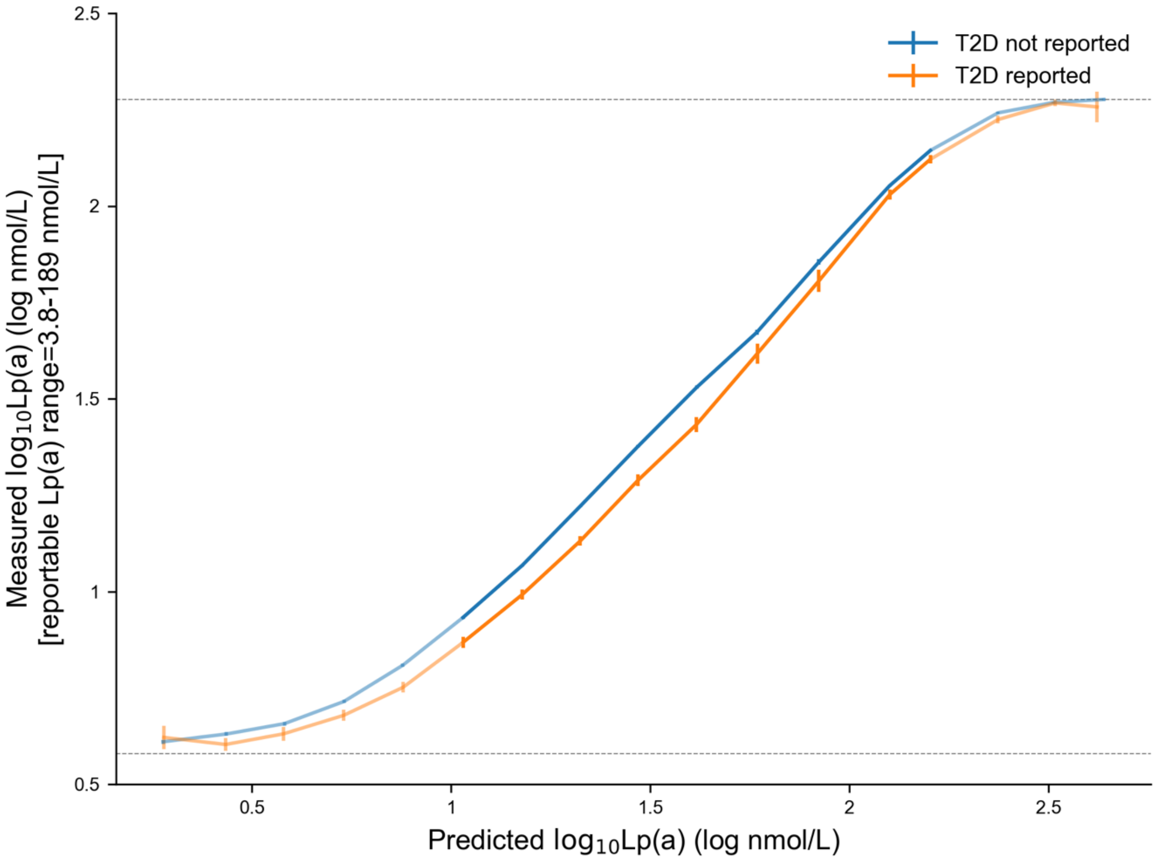

Fig. S15. Individuals with type 2 diabetes exhibit reduced Lp(a) across the range of genetically predicted Lp(a). Log-log plot of measured Lp(a) vs. genetically predicted Lp(a), stratified by type 2 diabetes (T2D) status, in *N*= 311,854 unrelated, British UKB participants. Means are plotted for bins of predicted Lp(a); error bars, 95% CIs. The reportable range of Lp(a) measurements was 3.8-189 nmol/L (horizontal lines), such that the cropping of reported measurements to this range tended to caused bias toward the mean in Lp(a) measurements of individuals with very low or very high Lp(a).

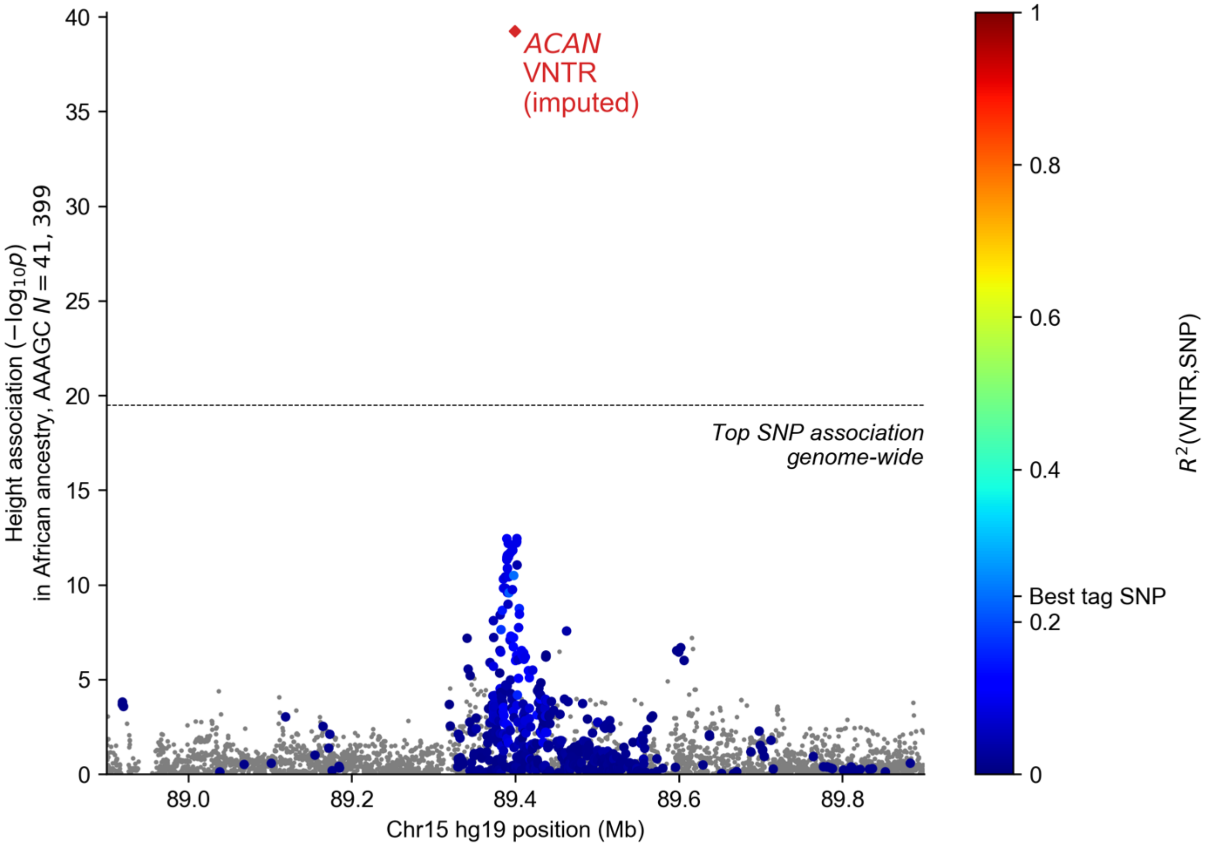

Fig. S16. Replication of association between *ACAN* VNTR length and height in African-ancestry individuals by summary-statistic imputation. Height association *P*-values at the *ACAN* locus are displayed for SNPs analyzed by AAAGC (*N*=41,399) (*27*) and for the *ACAN* VNTR (for which strength of association with height was imputed into AAAGC via modeling of linkage disequilibrium with nearby SNPs). Variants exhibiting non-negligible LD with the VNTR (R^2>0.01) are colored by the amount of linkage, and the strength of the top SNP association genome-wide is indicated by the dashed line.

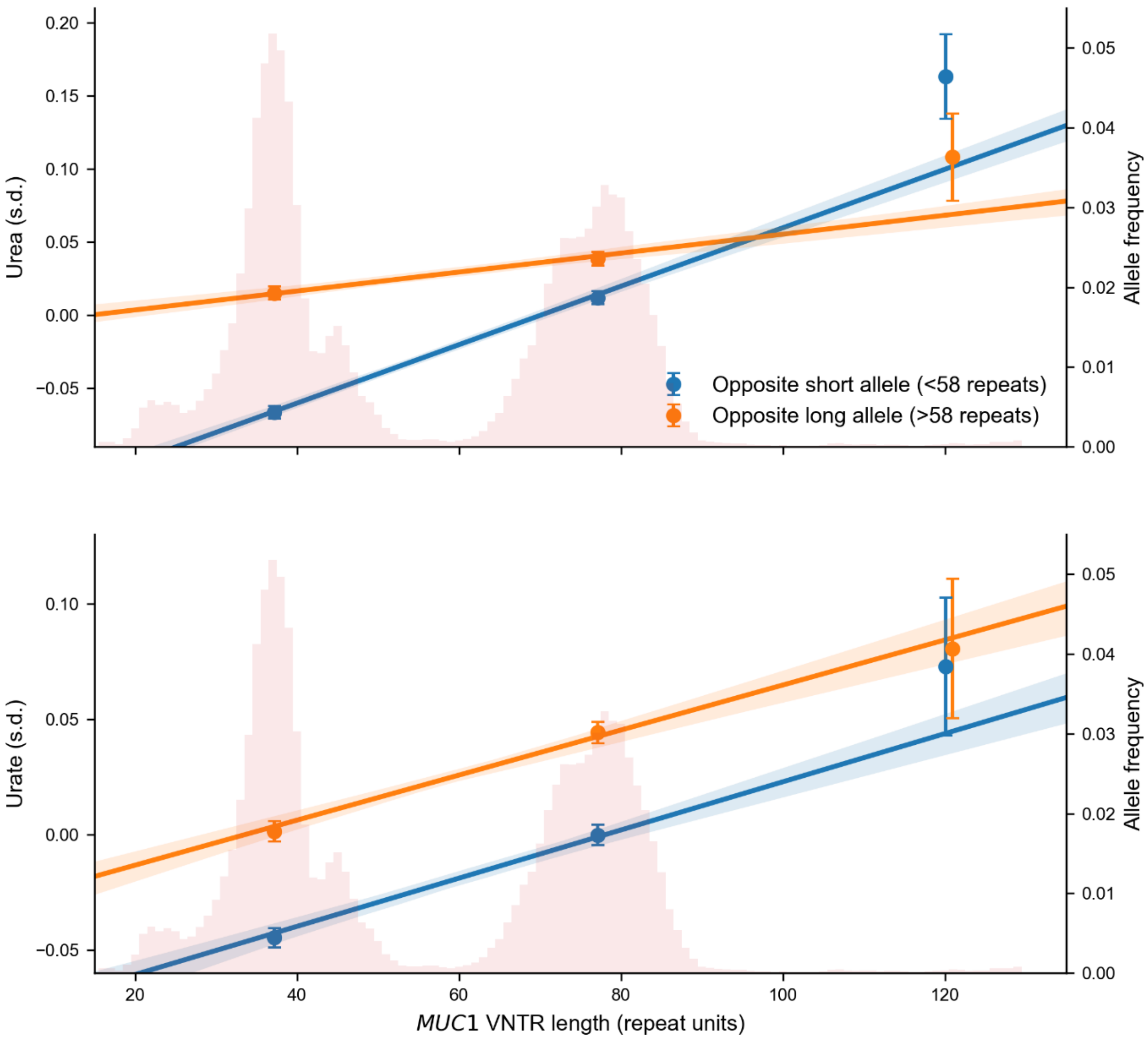

Fig. S17. Incomplete dominance in the association of *MUC1* VNTR length with serum urea. a, Serum urea vs. *MUC1* VNTR length, stratified by length of the VNTR allele on the homologous chromosome, in *N*=415,280 unrelated UKB participants of European ancestry. Plot markers indicate mean urea for alleles binned by *MUC1* VNTR length; least-squares linear fit (solid) also shown. **b,** Serum urate vs. *MUC1* VNTR length, stratified and plotted as in panel **a**. Error bars, 95% CIs. We formally tested for a dominance effect using linear regression in which we included an interaction term (the product of maternally- and paternally-derived allele lengths) in addition to summed allele lengths, age, age squared, sex, and 20 PCs, obtaining P=2.3 x 10^{-20} for urea and P=0.56 for urate.

References and Notes

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

2. A. Sulovari, R. Li, P. A. Audano, D. Porubsky, M. R. Vollger, G. A. Logsdon, W. C. Warren, A. A. Pollen, M. J. P. Chaisson, E. E. Eichler; Human Genome Structural Variation Consortium, Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23243–23253 (2019). doi:10.1073/pnas.1912175116 Medline

3. M. D. Lalioti, H. S. Scott, C. Buresi, C. Rossier, A. Bottani, M. A. Morris, A. Malafosse, S. E. Antonarakis, Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386**, 847–851 (1997). doi:10.1038/386847a0 Medline

4. C. Wijmenga, J. E. Hewitt, L. A. Sandkuijl, L. N. Clark, T. J. Wright, H. G. Dauwerse, A.-M. Gruter, M. H. Hofker, P. Moerer, R. Williamson, G.-J. B. van Ommen, G. W. Padberg, R. R. Frants, Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet.* **2**, 26–30 (1992). doi:10.1038/ng0992-26 Medline

5. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007). doi:10.1038/ng2088 Medline

6. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). doi:10.1038/s41586-018-0579-z Medline

7. Materials and methods are available as supplementary materials.

8. C. V. Van Hout, I. Tachmazidou, J. D. Backman, J. D. Hoffman, D. Liu, A. K. Pandey, C. Gonzaga-Jauregui, S. Khalid, B. Ye, N. Banerjee, A. H. Li, C. O'Dushlaine, A. Marcketta, J. Staples, C. Schurmann, A. Hawes, E. Maxwell, L. Barnard, A. Lopez, J. Penn, L. Habegger, A. L. Blumenfeld, X. Bai, S. O'Keeffe, A. Yadav, K. Praveen, M. Jones, W. J. Salerno, W. K. Chung, I. Surakka, C. J. Willer, K. Hveem, J. B. Leader, D. J. Carey, D. H. Ledbetter, L. Cardon, G. D. Yancopoulos, A. Economides, G. Coppola, A. R. Shuldiner, S. Balasubramanian, M. Cantor, M. R. Nelson, J. Whittaker, J. G. Reid, J. Marchini, J. D. Overton, R. A. Scott, G. R. Abecasis, L. Yerges-Armstrong, A. Baras; Geisinger-Regeneron DiscovEHR Collaboration; Regeneron Genetics Center, Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020). doi:10.1038/s41586-020-2853-0 Medline

9. C. Benner, C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, M. Pirinen, FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016). doi:10.1093/bioinformatics/btw018 Medline

10. A. R. Barton, M. A. Sherman, R. E. Mukamel, P.-R. Loh, Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021). Medline

11. D. Beyter, H. Ingimundardottir, A. Oddsson, H. P. Eggertsson, E. Bjornsson, H. Jonsson, B. A. Atlason, S. Kristmundsdottir, S. Mehringer, M. T. Hardarson, S. A. Gudjonsson, D. N. Magnusdottir, A. Jonasdottir, A. Jonasdottir, R. P. Kristjansson, S. T. Sverrisson, G. Holley, G. Palsson, O. A. Stefansson, G. Eyjolfsson, I. Olafsson, O. Sigurdardottir, B. Torfason, G. Masson, A. Helgason, U. Thorsteinsdottir, H. Holm, D. F. Gudbjartsson, P. Sulem, O. T. Magnusson, B. V. Halldorsson, K. Stefansson, Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021). Medline

12. K. Schmidt, A. Noureen, F. Kronenberg, G. Utermann, Structure, function, and genetics of lipoprotein (a). *J. Lipid Res.* **57**, 1339–1359 (2016). doi:10.1194/jlr.R067314 Medline

13. P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, A. L. Price, Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015). doi:10.1038/ng.3190 Medline

14. P. Ebert, P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. H. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, H. Ashraf, N. T. Chuang, X. Yang, K. M. Munson, A. P. Lewis, S. Fairley, L. J. Tallon, W. E. Clarke, A. O. Basile, M. Byrska-Bishop, A. Corvelo, U. S. Evani, T.-Y. Lu, M. J. P. Chaisson, J. Chen, C. Li, H. Brand, A. M. Wenger, M. Ghareghani, W. T. Harvey, B. Raeder, P. Hasenfeld, A. A. Regier, H. J. Abel, I. M. Hall, P. Flicek, O. Stegle, M. B. Gerstein, J. M. C. Tubio, Z. Mu, Y. I. Li, X. Shi, A. R. Hastie, K. Ye, Z. Chong, A. D. Sanders, M. C. Zody, M. E. Talkowski, R. E. Mills, S. E. Devine, C. Lee, J. O. Korbel, T. Marschall, E. E. Eichler, Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021). doi:10.1126/science.abf7117 Medline

15. R. Clarke, J. F. Peden, J. C. Hopewell, T. Kyriakou, A. Goel, S. C. Heath, S. Parish, S. Barlera, M. G. Franzosi, S. Rust, D. Bennett, A. Silveira, A. Malarstig, F. R. Green, M. Lathrop, B. Gigante, K. Leander, U. de Faire, U. Seedorf, A. Hamsten, R. Collins, H. Watkins, M. Farrall; PROCARDIS Consortium, Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* **361**, 2518–2528 (2009). doi:10.1056/NEJMoa0902604 Medline

16. G. Utermann, H. J. Menzel, H. G. Kraft, H. C. Duba, H. G. Kemmler, C. Seitz, Lp(a) glycoprotein phenotypes. Inheritance and relation to Lp(a)-lipoprotein concentrations in plasma. *J. Clin. Invest.* **80**, 458–465 (1987). doi:10.1172/JCI113093 Medline

17. A. L. White, J. E. Hixson, D. L. Rainwater, R. E. Lanford, Molecular basis for “null” lipoprotein(a) phenotypes and the influence of apolipoprotein(a) size on plasma lipoprotein(a) level in the baboon. *J. Biol. Chem.* **269**, 9060–9066 (1994). doi:10.1016/S0021-9258(17)37076-X Medline

18. E. Boerwinkle, C. C. Leffert, J. Lin, C. Lackner, G. Chiesa, H. H. Hobbs, Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. *J. Clin. Invest.* **90**, 52–60 (1992). doi:10.1172/JCI115855 Medline

19. K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, K. K.-H. Farh, Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019). doi:10.1016/j.cell.2018.12.015 Medline

20. S. Coassin, G. Erhart, H. Weissensteiner, M. Eca Guimarães de Araújo, C. Lamina, S. Schönherr, L. Forer, M. Haun, J. L. Losso, A. Köttgen, K. Schmidt, G. Utermann, A. Peters, C. Gieger, K. Strauch, A. Finkenstedt, R. Bale, H. Zoller, B. Paulweber, K.-U. Eckardt, A. Hüttenhofer, L. A. Huber, F. Kronenberg, A novel but frequent variant in *LPA* KIV-2 is associated with a pronounced Lp(a) and cardiovascular risk reduction. *Eur. Heart J.* **38**, 1823–1831 (2017). doi:10.1093/eurheartj/ehx174 Medline

21. B. R. Zysow, G. E. Lindahl, D. P. Wade, B. L. Knight, R. M. Lawn, C/T polymorphism in the 5′ untranslated region of the apolipoprotein(a) gene introduces an upstream ATG and reduces in vitro translation. *Arterioscler. Thromb. Vasc. Biol.* **15**, 58–64 (1995). doi:10.1161/01.ATV.15.1.58 Medline

22. K. Suzuki, M. Kuriyama, T. Saito, A. Ichinose, Plasma lipoprotein(a) levels and expression of the apolipoprotein(a) gene are dependent on the nucleotide polymorphisms in its 5′-flanking region. *J. Clin. Invest.* **99**, 1361–1366 (1997). doi:10.1172/JCI119295 Medline

23. M. Trinder, M. M. Uddin, P. Finneran, K. G. Aragam, P. Natarajan, Clinical utility of lipoprotein(a) and LPA genetic risk score in risk prediction of incident atherosclerotic cardiovascular disease. *JAMA Cardiol.* (2020). doi:10.1001/jamacardio.2020.5398 Medline

24. D. F. Gudbjartsson, G. Thorgeirsson, P. Sulem, A. Helgadottir, A. Gylfason, J. Saemundsdottir, E. Bjornsson, G. L. Norddahl, A. Jonasdottir, A. Jonasdottir, H. P. Eggertsson, S. Gretarsdottir, G. Thorleifsson, O. S. Indridason, R. Palsson, F. Jonasson, I. Jonsdottir, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, R. Danielsen, S. E. Matthiasson, S. Kristmundsdottir, B. V. Halldorsson, A. B. Hreidarsson, E. M. Valdimarsson, T. Gudnason, R. Benediktsson, V. Steinthorsdottir, U. Thorsteinsdottir, H. Holm, K. Stefansson, Lipoprotein(a) Concentration and Risks of Cardiovascular Disease and Diabetes. *J. Am. Coll. Cardiol.* **74**, 2982–2994 (2019). doi:10.1016/j.jacc.2019.10.019 Medline

25. L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher; GIANT Consortium, Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018). doi:10.1093/hmg/ddy271 Medline

26. M. N. Weedon, H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans, M. Mangino, R. M. Freathy, J. R. B. Perry, S. Stevens, A. S. Hall, N. J. Samani, B. Shields, I. Prokopenko, M. Farrall, A. Dominiczak, T. Johnson, S. Bergmann, J. S. Beckmann, P. Vollenweider, D. M. Waterworth, V. Mooser, C. N. A. Palmer, A. D. Morris, W. H. Ouwehand, J. H. Zhao, S. Li, R. J. Loos, I. Barroso, P. Deloukas, M. S. Sandhu, E. Wheeler, N. Soranzo, M. Inouye, N. J. Wareham, M. Caulfield, P. B. Munroe, A. T. Hattersley, M. I. McCarthy, T. M. Frayling; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium; Cambridge GEM Consortium, Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **40**, 575–583 (2008). doi:10.1038/ng.121 Medline

27. M. Graff, A. E. Justice, K. L. Young, E. Marouli, X. Zhang, R. S. Fine, E. Lim, V. Buchanan, K. Rand, M. F. Feitosa, M. K. Wojczynski, L. R. Yanek, Y. Shao, R. Rohde, A. A. Adeyemo, M. C. Aldrich, M. A. Allison, C. B. Ambrosone, S. Ambs, C. Amos, D. K. Arnett, L. Atwood, E. V. Bandera, T. Bartz, D. M. Becker, S. I. Berndt, L. Bernstein, L. F. Bielak, W. J. Blot, E. P. Bottinger, D. W. Bowden, J. P. Bradfield, J. A. Brody, U. Broeckel, G. Burke, B. E. Cade, Q.

Cai, N. Caporaso, C. Carlson, J. Carpten, G. Casey, S. J. Chanock, G. Chen, M. Chen, Y. I. Chen, W.-M. Chen, A. Chesi, C. W. K. Chiang, L. Chu, G. A. Coetzee, D. V. Conti, R. S. Cooper, M. Cushman, E. Demerath, S. L. Deming, L. Dimitrov, J. Ding, W. R. Diver, Q. Duan, M. K. Evans, A. G. Falusi, J. D. Faul, M. Fornage, C. Fox, B. I. Freedman, M. Garcia, E. M. Gillanders, P. Goodman, O. Gottesman, S. F. A. Grant, X. Guo, H. Hakonarson, T. Haritunians, T. B. Harris, C. C. Harris, B. E. Henderson, A. Hennis, D. G. Hernandez, J. N. Hirschhorn, L. H. McNeill, T. D. Howard, B. Howard, A. W. Hsing, Y.-H. H. Hsu, J. J. Hu, C. D. Huff, D. Huo, S. A. Ingles, M. R. Irvin, E. M. John, K. C. Johnson, J. M. Jordan, E. K. Kabagambe, S. J. Kang, S. L. Kardia, B. J. Keating, R. A. Kittles, E. A. Klein, S. Kolb, L. N. Kolonel, C. Kooperberg, L. Kuller, A. Kutlar, L. Lange, C. D. Langefeld, L. Le Marchand, H. Leonard, G. Lettre, A. M. Levin, Y. Li, J. Li, Y. Liu, Y. Liu, S. Liu, K. Lohman, V. Lotay, Y. Lu, W. Maixner, J. E. Manson, B. McKnight, Y. Meng, K. L. Monda, K. Monroe, J. H. Moore, T. H. Mosley, P. Mudgal, A. B. Murphy, R. Nadukuru, M. A. Nalls, K. L. Nathanson, U. Nayak, A. N'Diaye, B. Nemesure, C. Neslund-Dudas, M. L. Neuhouser, S. Nyante, H. Ochs-Balcom, T. O. Ogundiran, A. Ogunniyi, O. Ojengbede, H. Okut, O. I. Olopade, A. Olshan, B. Padhukasahasram, J. Palmer, C. D. Palmer, N. D. Palmer, G. Papanicolaou, S. R. Patel, C. A. Pettaway, P. A. Peyser, M. F. Press, D. C. Rao, L. J. Rasmussen-Torvik, S. Redline, A. P. Reiner, S. K. Rhie, J. L. Rodriguez-Gil, C. N. Rotimi, J. I. Rotter, E. A. Ruiz-Narvaez, B. A. Rybicki, B. Salako, M. M. Sale, M. Sanderson, E. Schadt, P. J. Schreiner, C. Schurmann, A. G. Schwartz, D. A. Shriner, L. B. Signorello, A. B. Singleton, D. S. Siscovick, J. A. Smith, S. Smith, E. Speliotes, M. Spitz, J. L. Stanford, V. L. Stevens, A. Stram, S. S. Strom, L. Sucheston, Y. V. Sun, S. M. Tajuddin, H. Taylor, K. Taylor, B. O. Tayo, M. J. Thun, M. A. Tucker, D. Vaidya, D. J. Van Den Berg, S. Vedantam, M. Vitolins, Z. Wang, E. B. Ware, S. Wassertheil-Smoller, D. R. Weir, J. K. Wiencke, S. M. Williams, L. K. Williams, J. G. Wilson, J. S. Witte, M. Wrensch, X. Wu, J. Yao, N. Zakai, K. Zanetti, B. S. Zemel, W. Zhao, J. H. Zhao, W. Zheng, D. Zhi, J. Zhou, X. Zhu, R. G. Ziegler, J. Zmuda, A. B. Zonderman, B. M. Psaty, I. B. Borecki, L. A. Cupples, C.-T. Liu, C. A. Haiman, R. Loos, M. C. Y. Ng, K. E. North, Discovery and fine-mapping of height loci via high-density imputation of GWASs in individuals of African ancestry. *Am. J. Hum. Genet.* **108**, 564–582 (2021). doi:10.1016/j.ajhg.2021.02.011 Medline

28. K. L. Lauing, M. Cortes, M. S. Domowicz, J. G. Henry, A. T. Baria, N. B. Schwartz, Aggrecan is required for growth plate cytoarchitecture and differentiation. *Dev. Biol.* **396**, 224–236 (2014). doi:10.1016/j.ydbio.2014.10.005 Medline

29. K. J. Doege, S. N. Coulter, L. M. Meek, K. Maslen, J. G. Wood, A human-specific polymorphism in the coding region of the aggrecan gene. Variable number of tandem repeats produce a range of core protein sizes in the general population. *J. Biol. Chem.* **272**, 13974–13979 (1997). doi:10.1074/jbc.272.21.13974 Medline

30. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47** (D1), D886–D894 (2019). doi:10.1093/nar/gky1016 Medline

31. L. Gleghorn, R. Ramesar, P. Beighton, G. Wallis, A mutation in the variable repeat region of the aggrecan gene (*AGC1*) causes a form of spondyloepiphyseal dysplasia associated with severe, premature osteoarthritis. *Am. J. Hum. Genet.* **77**, 484–490 (2005). doi:10.1086/444401 Medline

32. M. Doyard, S. Bacrot, C. Huber, M. Di Rocco, A. Goldenberg, M. S. Aglan, P. Brunelle, S. Temtamy, C. Michot, G. A. Otaify, C. Haudry, M. Castanet, J. Leroux, J.-P. Bonnefont, A. Munnich, G. Baujat, P. Lapunzina, S. Monnot, V. L. Ruiz-Perez, V. Cormier-Daire, *FAM46A* mutations are responsible for autosomal recessive osteogenesis imperfecta. *J. Med. Genet.* **55**, 278–284 (2018). doi:10.1136/jmedgenet-2017-104999 Medline

33. O. Gewartowska, G. Aranaz-Novaliches, P. S. Krawczyk, S. Mroczek, M. Kusio-Kobiałka, B. Tarkowski, F. Spoutil, O. Benada, O. Kofroňová, P. Szwedziak, D. Cysewski, J. Gruchota, M. Szpila, A. Chlebowski, R. Sedlacek, J. Prochazka, A. Dziembowski, Cytoplasmic polyadenylation by TENT5A is required for proper bone formation. *Cell Rep.* **35**, 109015 (2021). doi:10.1016/j.celrep.2021.109015 Medline

34. J. C. Fowler, A. S. Teixeira, L. E. Vinall, D. M. Swallow, Hypervariability of the membrane-associated mucin and cancer marker MUC1. *Hum. Genet.* **113**, 473–479 (2003). doi:10.1007/s00439-003-1011-8 Medline

35. A. Kirby, A. Gnirke, D. B. Jaffe, V. Barešová, N. Pochet, B. Blumenstiel, C. Ye, D. Aird, C. Stevens, J. T. Robinson, M. N. Cabili, I. Gat-Viks, E. Kelliher, R. Daza, M. DeFelice, H. Hůlková, J. Sovová, P. Vylet'al, C. Antignac, M. Guttman, R. E. Handsaker, D. Perrin, S. Steelman, S. Sigurdsson, S. J. Scheinman, C. Sougnez, K. Cibulskis, M. Parkin, T. Green, E. Rossin, M. C. Zody, R. J. Xavier, M. R. Pollak, S. L. Alper, K. Lindblad-Toh, S. Gabriel, P. S. Hart, A. Regev, C. Nusbaum, S. Kmoch, A. J. Bleyer, E. S. Lander, M. J. Daly, Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in *MUC1* missed by massively parallel sequencing. *Nat. Genet.* **45**, 299–303 (2013). doi:10.1038/ng.2543 Medline

36. Y. Okada, X. Sim, M. J. Go, J.-Y. Wu, D. Gu, F. Takeuchi, A. Takahashi, S. Maeda, T. Tsunoda, P. Chen, S.-C. Lim, T.-Y. Wong, J. Liu, T. L. Young, T. Aung, M. Seielstad, Y.-Y. Teo, Y. J. Kim, J.-Y. Lee, B.-G. Han, D. Kang, C.-H. Chen, F.-J. Tsai, L.-C. Chang, S.-J. C. Fann, H. Mei, D. C. Rao, J. E. Hixson, S. Chen, T. Katsuya, M. Isono, T. Ogihara, J. C. Chambers, W. Zhang, J. S. Kooner, E. Albrecht, K. Yamamoto, M. Kubo, Y. Nakamura, N. Kamatani, N. Kato, J. He, Y.-T. Chen, Y. S. Cho, E.-S. Tai, T. Tanaka; KidneyGen Consortium; CKDGen Consortium; GUGC consortium, Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* **44**, 904–909 (2012). doi:10.1038/ng.2352 Medline

37. A. Köttgen, E. Albrecht, A. Teumer, V. Vitart, J. Krumsiek, C. Hundertmark, G. Pistis, D. Ruggiero, C. M. O'Seaghdha, T. Haller, Q. Yang, T. Tanaka, A. D. Johnson, Z. Kutalik, A. V. Smith, J. Shi, M. Struchalin, R. P. S. Middelberg, M. J. Brown, A. L. Gaffo, N. Pirastu, G. Li, C. Hayward, T. Zemunik, J. Huffman, L. Yengo, J. H. Zhao, A. Demirkan, M. F. Feitosa, X. Liu, G. Malerba, L. M. Lopez, P. van der Harst, X. Li, M. E. Kleber, A. A. Hicks, I. M. Nolte, A. Johansson, F. Murgia, S. H. Wild, S. J. L. Bakker, J. F. Peden, A. Dehghan, M. Steri, A. Tenesa, V. Lagou, P. Salo, M. Mangino, L. M. Rose, T. Lehtimäki, O. M. Woodward, Y. Okada, A. Tin, C. Müller, C. Oldmeadow, M. Putku, D. Czamara, P. Kraft, L. Frogheri, G. A. Thun, A. Grotevendt, G. K. Gislason, T. B. Harris, L. J. Launer, P. McArdle, A. R. Shuldiner, E. Boerwinkle, J. Coresh, H. Schmidt, M. Schallert, N. G. Martin, G. W. Montgomery, M. Kubo, Y. Nakamura, T. Tanaka, P. B. Munroe, N. J. Samani, D. R. Jacobs Jr., K. Liu, P. D'Adamo, S. Ulivi, J. I. Rotter, B. M. Psaty, P. Vollenweider, G. Waeber, S. Campbell, O. Devuyst, P. Navarro, I. Kolcic, N. Hastie, B. Balkau, P. Froguel, T. Esko, A. Salumets, K. T. Khaw, C. Langenberg, N. J. Wareham, A. Isaacs, A. Kraja, Q. Zhang, P. S. Wild, R. J. Scott, E. G. Holliday, E. Org, M. Viigimaa, S. Bandinelli, J. E. Metter, A. Lupo, E. Trabetti, R. Sorice, A. Döring, E. Lattka, K. Strauch, F. Theis, M. Waldenberger, H.-E. Wichmann, G. Davies, A. J. Gow, M. Bruinenberg, R. P. Stolk, J. S. Kooner, W. Zhang, B. R. Winkelmann, B. O. Boehm, S. Lucae, B. W. Penninx, J. H. Smit, G. Curhan, P. Mudgal, R. M. Plenge, L. Portas, I. Persico, M. Kirin, J. F. Wilson, I. Mateo Leach, W. H. van Gilst, A. Goel, H. Ongen, A. Hofman, F. Rivadeneira, A. G. Uitterlinden, M. Imboden, A. von Eckardstein, F. Cucca, R. Nagaraja, M. G. Piras, M. Nauck, C. Schurmann, K. Budde, F. Ernst, S. M. Farrington, E. Theodoratou, I. Prokopenko, M. Stumvoll, A. Jula, M. Perola, V. Salomaa, S.-Y. Shin, T. D. Spector, C. Sala, P. M. Ridker, M. Kähönen, J. Viikari, C. Hengstenberg, C. P. Nelson, J. F. Meschia, M. A. Nalls, P.

Sharma, A. B. Singleton, N. Kamatani, T. Zeller, M. Burnier, J. Attia, M. Laan, N. Klopp, H. L. Hillege, S. Kloiber, H. Choi, M. Pirastu, S. Tore, N. M. Probst-Hensch, H. Völzke, V. Gudnason, A. Parsa, R. Schmidt, J. B. Whitfield, M. Fornage, P. Gasparini, D. S. Siscovick, O. Polašek, H. Campbell, I. Rudan, N. Bouatia-Naji, A. Metspalu, R. J. F. Loos, C. M. van Duijn, I. B. Borecki, L. Ferrucci, G. Gambaro, I. J. Deary, B. H. R. Wolffenbuttel, J. C. Chambers, W. März, P. P. Pramstaller, H. Snieder, U. Gyllensten, A. F. Wright, G. Navis, H. Watkins, J. C. M. Witteman, S. Sanna, S. Schipf, M. G. Dunlop, A. Tönjes, S. Ripatti, N. Soranzo, D. Toniolo, D. I. Chasman, O. Raitakari, W. H. L. Kao, M. Ciullo, C. S. Fox, M. Caulfield, M. Bochud, C. Gieger; LifeLines Cohort Study; CARDIoGRAM Consortium; DIAGRAM Consortium; ICBP Consortium; MAGIC Consortium, Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–154 (2013). doi:10.1038/ng.2500 Medline

38. P. M. Steinert, D. A. D. Parry, L. N. Marekov, Trichohyalin mechanically strengthens the hair follicle: Multiple cross-bridging roles in the inner root shealth. *J. Biol. Chem.* **278**, 41409–41419 (2003). doi:10.1074/jbc.M302037200 Medline

39. S. C. Lee, I. G. Kim, L. N. Marekov, E. J. O'Keefe, D. A. Parry, P. M. Steinert, The structure of human trichohyalin. Potential multiple roles as a functional EF-hand-like calcium-binding protein, a cornified cell envelope precursor, and an intermediate filament-associated (cross-linking) protein. *J. Biol. Chem.* **268**, 12164–12176 (1993). doi:10.1016/S0021-9258(19)50322-2 Medline

40. F. B. Ü. Basmanav, L. Cau, A. Tafazzoli, M.-C. Méchin, S. Wolf, M. T. Romano, F. Valentin, H. Wiegmann, A. Huchenq, R. Kandil, N. Garcia Bartels, A. Kilic, S. George, D. J. Ralser, S. Bergner, D. J. P. Ferguson, A.-M. Oprisoreanu, M. Wehner, H. Thiele, J. Altmüller, P. Nürnberg, D. Swan, D. Houniet, A. Büchner, L. Weibel, N. Wagner, R. Grimalt, A. Bygum, G. Serre, U. Blume-Peytavi, E. Sprecher, S. Schoch, V. Oji, H. Hamm, P. Farrant, M. Simon, R. C. Betz, Mutations in three genes encoding proteins involved in hair shaft formation cause uncombable hair syndrome. *Am. J. Hum. Genet.* **99**, 1292–1304 (2016). doi:10.1016/j.ajhg.2016.10.004 Medline

41. S. E. Medland, D. R. Nyholt, J. N. Painter, B. P. McEvoy, A. F. McRae, G. Zhu, S. D. Gordon, M. A. R. Ferreira, M. J. Wright, A. K. Henders, M. J. Campbell, D. L. Duffy, N. K. Hansell, S. Macgregor, W. S. Slutske, A. C. Heath, G. W. Montgomery, N. G. Martin, Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.* **85**, 750–755 (2009). doi:10.1016/j.ajhg.2009.10.009 Medline

42. F. Liu, Y. Chen, G. Zhu, P. G. Hysi, S. Wu, K. Adhikari, K. Breslin, E. Pośpiech, M. A. Hamer, F. Peng, C. Muralidharan, V. Acuna-Alonzo, S. Canizales-Quinteros, G. Bedoya, C. Gallo, G. Poletti, F. Rothhammer, M. C. Bortolini, R. Gonzalez-Jose, C. Zeng, S. Xu, L. Jin, A. G. Uitterlinden, M. A. Ikram, C. M. van Duijn, T. Nijsten, S. Walsh, W. Branicki, S. Wang, A. Ruiz-Linares, T. D. Spector, N. G. Martin, S. E. Medland, M. Kayser, Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. *Hum. Mol. Genet.* **27**, 559–575 (2018). doi:10.1093/hmg/ddx416 Medline

43. A. Moayyeri, C. J. Hammond, D. J. Hart, T. D. Spector, The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013). doi:10.1017/thg.2012.89 Medline

44. R. E. Mukamel, R. E. Handsaker, M. A. Sherman, A. R. Barton, Y. Zheng, S. A. McCarroll, P.-R. Loh, Codes and scripts for: Protein-coding repeat polymorphisms strongly shape diverse human phenotypes, Zenodo (2021); https://doi.org/10.5281/zenodo.4776804.

45. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T.

Peakman, R. Collins, 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 (2015). doi:10.1371/journal.pmed.1001779 Medline

46. P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, A. L. Price, Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018). doi:10.1038/s41588-018-0144-6 Medline

47. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). doi:10.1093/nar/27.2.573 Medline

48. M. Bakhtiari, S. Shleizer-Burko, M. Gymrek, V. Bansal, V. Bafna, Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018). doi:10.1101/gr.235119.118 Medline

49. E. Dolzhenko, J. J. F. A. van Vugt, R. J. Shaw, M. A. Bekritsky, M. van Blitterswijk, G. Narzisi, S. S. Ajay, V. Rajan, B. R. Lajoie, N. H. Johnson, Z. Kingsbury, S. J. Humphray, R. D. Schellevis, W. J. Brands, M. Baker, R. Rademakers, M. Kooyman, G. H. P. Tazelaar, M. A. van Es, R. McLaughlin, W. Sproviero, A. Shatunov, A. Jones, A. Al Khleifat, A. Pittman, S. Morgan, O. Hardiman, A. Al-Chalabi, C. Shaw, B. Smith, E. J. Neo, K. Morrison, P. J. Shaw, C. Reeves, L. Winterkorn, N. S. Wexler, D. E. Housman, C. W. Ng, A. L. Li, R. J. Taft, L. H. van den Berg, D. R. Bentley, J. H. Veldink, M. A. Eberle, Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017). doi:10.1101/gr.225672.117 Medline

50. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020). doi:10.1111/rssb.12388

51. C. Benner, A. S. Havulinna, M.-R. Järvelin, V. Salomaa, S. Ripatti, M. Pirinen, Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017). doi:10.1016/j.ajhg.2017.08.012 Medline

52. A. S. Levey, J. P. Bosch, J. B. Lewis, T. Greene, N. Rogers, D. Roth; Modification of Diet in Renal Disease Study Group, A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. *Ann. Intern. Med.* **130**, 461–470 (1999). doi:10.7326/0003-4819-130-6-199903160-00002 Medline

53. C. X. Yap, J. Sidorenko, Y. Wu, K. E. Kemper, J. Yang, N. R. Wray, M. R. Robinson, P. M. Visscher, Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nat. Commun.* **9**, 5407 (2018). doi:10.1038/s41467-018-07862-y Medline

54. J. Listgarten, C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, D. Heckerman, Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012). doi:10.1038/nmeth.2037 Medline

55. J. Mefford, D. Park, Z. Zheng, A. Ko, M. Ala-Korpela, M. Laakso, P. Pajukanta, J. Yang, J. Witte, N. Zaitlen, Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models. *J. Comput. Biol.* **27**, 599–612 (2020). doi:10.1089/cmb.2019.0325 Medline

56. P.-R. Loh, G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, Schizophrenia Working Group of Psychiatric Genomics Consortium, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler, M. C. O’Donovan, B. M. Neale, N. Patterson, A. L. Price; , Contrasting genetic

architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015). doi:10.1038/ng.3431 Medline

57. S. Mack, S. Coassin, R. Rueedi, N. A. Yousri, I. Seppälä, C. Gieger, S. Schönherr, L. Forer, G. Erhart, P. Marques-Vidal, J. S. Ried, G. Waeber, S. Bergmann, D. Dähnhardt, A. Stöckl, O. T. Raitakari, M. Kähönen, A. Peters, T. Meitinger, K. Strauch, L. Kedenko, B. Paulweber, T. Lehtimäki, S. C. Hunt, P. Vollenweider, C. Lamina, F. Kronenberg, A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *J. Lipid Res.* **58**, 1834–1844 (2017). doi:10.1194/jlr.M076232 Medline

58. S. M. Zekavat, S. Ruotsalainen, R. E. Handsaker, M. Alver, J. Bloom, T. Poterba, C. Seed, J. Ernst, M. Chaffin, J. Engreitz, G. M. Peloso, A. Manichaikul, C. Yang, K. A. Ryan, M. Fu, W. C. Johnson, M. Tsai, M. Budoff, R. S. Vasan, L. A. Cupples, J. I. Rotter, S. S. Rich, W. Post, B. D. Mitchell, A. Correa, A. Metspalu, J. G. Wilson, V. Salomaa, M. Kellis, M. J. Daly, B. M. Neale, S. McCarroll, I. Surakka, T. Esko, A. Ganna, S. Ripatti, S. Kathiresan, P. Natarajan; NHLBI TOPMed Lipids Working Group, Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. *Nat. Commun.* **9**, 2606 (2018). doi:10.1038/s41467-018-04668-w Medline

59. S. Di Maio, R. Grüneis, G. Streiter, C. Lamina, M. Maglione, S. Schoenherr, D. Öfner, B. Thorand, A. Peters, K.-U. Eckardt, A. Köttgen, F. Kronenberg, S. Coassin, Investigation of a nonsense mutation located in the complex KIV-2 copy number variation region of apolipoprotein(a) in 10,910 individuals. *Genome Med.* **12**, 74 (2020). doi:10.1186/s13073-020-00771-0 Medline

60. B. Pasaniuc, N. Zaitlen, H. Shi, G. Bhatia, A. Gusev, J. Pickrell, J. Hirschhorn, D. P. Strachan, N. Patterson, A. L. Price, Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014). doi:10.1093/bioinformatics/btu416 Medline

61. 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). doi:10.1038/nature11632 Medline

62. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi:10.1038/nature15393 Medline

63. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016). doi:10.1038/ng.3656 Medline

64. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S. B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A.

Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi:10.1038/s41586-021-03205-y Medline

65. P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, A. L Price, Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016). doi:10.1038/ng.3679 Medline

66. P.-R. Loh, G. Genovese, S. A. McCarroll, Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020). doi:10.1038/s41586-020-2430-6 Medline

67. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi:10.1093/bioinformatics/btp324 Medline

68. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). doi:10.1093/nar/gkh340 Medline

69. W. J. Kent, BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). doi:10.1101/gr.229202 Medline

70. M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy-Gallego, P. Flicek, S. Germer, H. Brand, I. M. Hall, M. E. Talkowski, G. Narzisi, M. C. Zody, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021), doi:10.1101/2021.02.06.430068.

71. R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, S. A. McCarroll, Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015). doi:10.1038/ng.3200 Medline

72. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi:10.1101/gr.229102 Medline

73. R. E. Handsaker, J. M. Korn, J. Nemesh, S. A. McCarroll, Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011). doi:10.1038/ng.768 Medline

74. A. Abyzov, A. E. Urban, M. Snyder, M. Gerstein, CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011). doi:10.1101/gr.114876.110 Medline

75. T.-Y. Lu, Human Genome Structural Variation Consortium, M. J. P. Chaisson, Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nat. Commun.* **12**, 4250 (2021). doi:10.1038/s41467-021-24378-0 Medline

76. P. Garg, A. Martin-Trujillo, O. L. Rodriguez, S. J. Gies, E. Hadelia, B. Jadhav, M. Jain, B. Paten, A. J. Sharp, 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 (2021). doi:10.1016/j.ajhg.2021.03.016 Medline

77. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

78. B. S. Pedersen, A. R. Quinlan, Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018). doi:10.1093/bioinformatics/btx699 Medline

79. B. L. Browning, S. R. Browning, Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016). doi:10.1016/j.ajhg.2015.11.020 Medline

80. M. Stephens, P. Scheet, Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005). doi:10.1086/428594 Medline

81. B. L. Browning, S. R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009). doi:10.1016/j.ajhg.2009.01.005 Medline

82. N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003). doi:10.1093/genetics/165.4.2213 Medline

83. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). doi:10.1186/s13742-015-0047-8 Medline

84. S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, Optimization by simulated annealing. *Science* **220**, 671–680 (1983). doi:10.1126/science.220.4598.671 Medline

85. S. Coassin, S. Schönherr, H. Weissensteiner, G. Erhart, L. Forer, J. L. Losso, C. Lamina, M. Haun, G. Utermann, B. Paulweber, G. Specht, F. Kronenberg, A comprehensive map of single-base polymorphisms in the hypervariable *LPA* kringle IV type 2 copy number variation region. *J. Lipid Res.* **60**, 186–199 (2019). doi:10.1194/jlr.M090381 Medline

86. H. G. Kraft, A. Lingenhel, R. W. C. Pang, R. Delport, M. Trommsdorff, H. Vermaak, E. D. Janus, G. Utermann, Frequency distributions of apolipoprotein(a) kringle IV repeat alleles and their effects on lipoprotein(a) levels in Caucasian, Asian, and African populations: The distribution of null alleles is non-random. *Eur. J. Hum. Genet.* **4**, 74–87 (1996). doi:10.1159/000472175 Medline

87. W. E. Horton Jr., M. Lethbridge-Çejku, M. C. Hochberg, R. Balakir, P. Precht, C. C. Plato, J. D. Tobin, L. Meek, K. Doege, An association between an aggrecan polymorphic allele and bilateral hand osteoarthritis in elderly white men: Data from the Baltimore Longitudinal Study of Aging (BLSA). *Osteoarthritis Cartilage* **6**, 245–251 (1998). doi:10.1053/joca.1998.0117 Medline

88. I. Barragán, S. Borrego, M. M. Abd El-Aziz, M. F. El-Ashry, L. Abu-Safieh, S. S. Bhattacharya, G. Antiñolo, Genetic analysis of FAM46A in Spanish families with autosomal recessive retinitis pigmentosa: Characterisation of novel VNTRs. *Ann. Hum. Genet.* **72**, 26–34 (2008). Medline

89. L. E. Vinall, M. King, M. Novelli, C. A. Green, G. Daniels, J. Hilkens, M. Sarner, D. M. Swallow, Altered expression and allelic association of the hypervariable membrane mucin MUC1 in Helicobacter pylori gastritis. *Gastroenterology* **123**, 41–49 (2002). doi:10.1053/gast.2002.34157 Medline

90. P. A. Audano, A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, W. C. Warren, V. Magrini, S. D. McGrath, Y. I. Li, R. K. Wilson, E. E. Eichler, Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019). doi:10.1016/j.cell.2018.12.019 Medline

91. A. Noureen, F. Fresser, G. Utermann, K. Schmidt, Sequence variation within the KIV-2 copy number polymorphism of the human LPA gene in African, Asian, and European populations. *PLOS ONE* **10**, e0121582 (2015). doi:10.1371/journal.pone.0121582 Medline

92. W. Parson, H. G. Kraft, H. Niederstätter, A. W. Lingenhel, S. Köchl, F. Fresser, G. Utermann, A common nonsense mutation in the repetitive Kringle IV-2 domain of human apolipoprotein(a) results in a truncated protein and low plasma Lp(a). *Hum. Mutat.* **24**, 474–480 (2004). doi:10.1002/humu.20101 Medline

93. M. Ogorelkova, A. Gruber, G. Utermann, Molecular basis of congenital lp(a) deficiency: A frequent apo(a) 'null' mutation in caucasians. *Hum. Mol. Genet.* **8**, 2087–2096 (1999). doi:10.1093/hmg/8.11.2087 Medline

94. E. T. Lim, P. Würtz, A. S. Havulinna, P. Palta, T. Tukiainen, K. Rehnström, T. Esko, R. Mägi, M. Inouye, T. Lappalainen, Y. Chan, R. M. Salem, M. Lek, J. Flannick, X. Sim, A. Manning, C. Ladenvall, S. Bumpstead, E. Hämäläinen, K. Aalto, M. Maksimow, M. Salmi, S. Blankenberg, D. Ardissino, S. Shah, B. Horne, R. McPherson, G. K. Hovingh, M. P. Reilly, H. Watkins, A. Goel, M. Farrall, D. Girelli, A. P. Reiner, N. O. Stitziel, S. Kathiresan, S. Gabriel, J. C. Barrett, T. Lehtimäki, M. Laakso, L. Groop, J. Kaprio, M. Perola, M. I. McCarthy, M. Boehnke, D. M. Altshuler, C. M. Lindgren, J. N. Hirschhorn, A. Metspalu, N. B. Freimer, T. Zeller, S. Jalkanen, S. Koskinen, O. Raitakari, R. Durbin, D. G. MacArthur, V. Salomaa, S. Ripatti, M. J. Daly, A. Palotie; Sequencing Initiative Suomi (SISu) Project, Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLOS Genet.* **10**, e1004494 (2014). doi:10.1371/journal.pgen.1004494 Medline

95. B. M. Morgan, A. N. Brown, N. Deo, T. W. R. Harrop, G. Taiaroa, P. D. Mace, S. M. Wilbanks, T. R. Merriman, M. J. A. Williams, S. P. A. McCormick, Nonsynonymous SNPs in *LPA* homologous to plasminogen deficiency mutants represent novel null apo(a) alleles. *J. Lipid Res.* **61**, 432–444 (2020). doi:10.1194/jlr.M094540 Medline

96. M. A. Said, M. W. Yeung, Y. J. van de Vegte, J. W. Benjamins, R. P. F. Dullaart, S. Ruotsalainen, S. Ripatti, P. Natarajan, L. E. Juarez-Orozco, N. Verweij, P. van der Harst, Genome-wide association study and identification of a protective missense variant on lipoprotein(a) concentration: Protective missense variant on lipoprotein(a) concentration-brief report. *Arterioscler. Thromb. Vasc. Biol.* **41**, 1792–1800 (2021). doi:10.1161/ATVBAHA.120.315300 Medline

97. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi:10.1126/science.aaz1776 Medline

98. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E.

Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O’Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi:10.1038/s41586-020-2308-7 Medline