HUMAN GENOMICS

Protein-coding repeat polymorphisms strongly shape diverse human phenotypes

Ronen E. Mukamel[1,2*†], Robert E. Handsaker[2,3,4*†], Maxwell A. Sherman[1,2,5], Alison R. Barton[1,2,6], Yiming Zheng[2,3], Steven A. McCarroll[2,3,4*‡], Po-Ru Loh[1,2*‡]

Many human proteins contain domains that vary in size or copy number because of variable numbers of tandem repeats (VNTRs) in protein-coding exons. However, the relationships of VNTRs to most phenotypes are unknown because of difficulties in measuring such repetitive elements. We developed methods to estimate VNTR lengths from whole-exome sequencing data and impute VNTR alleles into single-nucleotide polymorphism haplotypes. Analyzing 118 protein-altering VNTRs in 415,280 UK Biobank participants for association with 786 phenotypes identified some of the strongest associations of common variants with human phenotypes, including height, hair morphology, and biomarkers of health. Accounting for large-effect VNTRs further enabled fine-mapping of associations to many more protein-coding mutations in the same genes. These results point to cryptic effects of highly polymorphic common structural variants that have eluded molecular analyses to date.

The human genome contains thousands of variable number of tandem repeat (VNTR) polymorphisms (*1*, *2*), but the effects of these polymorphisms on human phenotypes are largely unknown. VNTRs are multiallelic variants at which a nucleotide sequence, from seven to thousands of base pairs long, is repeated several to hundreds of times, with the number of repeats varying among individuals (fig. S1). Extreme alleles of VNTRs have been implicated in diseases including progressive myoclonus epilepsy (*3*) and facioscapulohumeral muscular dystrophy (*4*). However, because most VNTRs are invisible to single-nucleotide polymorphism (SNP) arrays and difficult to measure with short-read sequencing, they have not been considered in the genotype-phenotype association studies that have been central to recent work in human genetics.

We hypothesized that exome-sequencing data might contain unknown information about VNTR lengths and that VNTR alleles might segregate on specific SNP haplotypes, enabling statistical imputation (*5*) in SNP-phenotype datasets from hundreds of thousands of people, such as participants in the UK Biobank (UKB) (*6*).

Exploring the phenotypic effects of coding VNTRs

We identified candidate VNTRs by scanning the human reference genome for tandemly repeated sequences (*7*). For each repeat, we estimated "diploid VNTR content," the sum of maternally and paternally derived allele lengths, in 49,959 exome-sequenced UKB participants (*8*) by measuring numbers of reads that aligned to the repeated sequence (*7*). We then used surrounding SNPs to identify haplotypes likely to have been coinherited from a recent common ancestor, enabling resolution of diploid measurements into allele-specific contributions and imputation of VNTR lengths into SNP-haplotypes of 437,612 additional UKB participants. We developed statistical algorithms to perform such analysis on extended SNP haplotypes for hundreds of thousands of individuals using sibling identical-by-descent information to benchmark accuracy and to optimize analysis parameters (*7*). We focused subsequent analysis on autosomal exon–overlapping repeats in 118 genes for which these measurements exhibited *cis* heritability in sibling pairs (table S1).

We applied this approach to identify relationships between coding VNTR alleles and 786 phenotypes (table S2) in up to 415,280 unrelated UKB participants (depending on phenotype) of European ancestry. This analysis found 185 statistically significant associations (table S3). To determine whether such associations were driven by VNTR length variation rather than by other variants with which the VNTRs were in linkage disequilibrium (LD), we performed fine-mapping analyses (*9*) considering nearby genotyped and imputed variants (*6*, *10*). Because variation at most VNTRs arises from three or more alleles, VNTR variation was only partially correlated with individual SNPs, enabling this analysis to distinguish VNTR from SNP effects.

Nineteen phenotype associations involving five distinct VNTRs (Table 1, table S3, and fig. S1) exhibited evidence [FINEMAP (*9*) posterior probability >0.95] that VNTR length variation, rather than nearby SNPs, drove genotype-phenotype associations. For these five VNTRs, we improved genotyping accuracy by incorporating additional information from within-repeat variation or spanning reads to confirm the associations [figs. S2 and S3 (*7*)].

These associations appeared to explain some of the largest known GWAS signals for human phenotypes, including height, serum urea, and hair phenotypes, with some associations exhibiting strength comparable to or exceeding that of any single SNP in the genome.

Three VNTRs within exons of *TENT5A*, *MUC1*, and *TCHH* had not previously been implicated at these loci; a fourth (in *ACAN*) was recently reported in parallel work (*11*). Analysis also replicated an association between the length of the KIV-2 repeat in *LPA* and lipoprotein(a) concentration (*12*) [$P = 4.4 \times 10^{-(25,121)}$, BOLT-LMM (*13*)]. All five VNTRs were genotyped and imputed accurately (root mean square error ~1 repeat unit and/or $R^2 \geq 0.7$) according to benchmarks using cross-validation (fig. S4 and table S1) and the HGSVC2 long-read sequencing dataset (figs. S5 to S9) (*7*, *14*).

Fine-mapping of *LPA* variants influencing lipoprotein(a) concentration

Complex genetics involving VNTRs and SNPs at the same locus was revealed by analyzing lipoprotein(a) concentration [Lp(a)], elevated levels of which are a major risk factor for coronary artery disease (*15*). Lp(a) is almost completely heritable, with about half of its population variance explained by a VNTR-generated size polymorphism in the second kringle-IV (KIV) domain of apo(a) (*12*). Each KIV-2 repeat unit (~5.6 kb) spans two exons of *LPA*, which together encode a 114-aa copy of this domain. Longer alleles, those with more copies of the encoded kringle repeat, are known to associate with lower Lp(a) levels (*12*, *16*), reflecting retention of longer apo(a) isoforms in the endoplasmic reticulum (*17*). In the UKB, inheritance at the *LPA* locus explained most of the variance in Lp(a) measurements [$R = 0.93$ in sibling pairs sharing both *LPA* alleles, consistent with previous work (*18*)], with KIV-2 length explaining ~61% of this variance in a nonparametric model.

To identify additional *LPA* variants that might more completely explain Lp(a) variation and to explore their interactions with KIV-2 length, we used individuals heterozygous for either of two coding variants [combined minor allele frequency (MAF) = 0.05] that create

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Program in Medical and Population Genetics, Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard University, Boston, MA, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard University, Boston, MA, USA. [4]Department of Genetics, Harvard Medical School, Boston, MA, USA. [5]Computer Science and Artificial Intelligence Laboratory, MIT, Boston, MA, USA. [6]Bioinformatics and Integrative Genomics Program, Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

*Corresponding author. Email: rmukamel@broadinstitute.org (R.E.M.); handsake@broadinstitute.org (R.E.H.); mccarroll@genetics.med.harvard.edu (S.A.M.); poruloh@broadinstitute.org (P.-R.L.)

†These authors contributed equally to this work.
‡These authors co-supervised this work.

Table 1. VNTRs within protein-coding sequences affect diverse human phenotypes. For each of five protein-altering VNTRs involved in phenotype associations that passed stringent fine-mapping criteria, *P* values [in linear mixed-model analyses of *N* = 415,280 unrelated UKB participants of European (EUR) ancestry] and estimated effect size ranges (across the longest and shortest alleles sufficiently common to be amenable to our computational analysis) are listed for the most strongly associated phenotype.

Gene	Cytoband	Repeat unit size	Repeat count (EUR)	Protein domain (effect)	Phenotype	Effect range ± SE	*P* value
LPA	6q25.3-q26	~5.6 kb (114 aa, two exons)	2–40	KIV (number)	Lipoprotein(a) concentration	5.1 ± 0.5 SD (= 233 ± 23 nmol/liter)	$4.4 \times 10^{-(25,121)}$
ACAN	15q26.1	57 bp (19 aa)	13–44	Chondroitin sulfate (size)	Height	0.49 ± 0.04 SD (= 3.2 ± 0.3 cm)	1.7×10^{-234}
TENT5A	6q14.1	15 bp (5 aa)	2–7	Unknown (size)	Height	0.09 ± 0.01 SD (= 0.6 ± 0.1 cm)	2.5×10^{-53}
MUC1	1q22	60 bp (20 aa)	20–125	Extracellular (size)	Serum urea	0.16 ± 0.01 SD (= 0.22 ± 0.01 mmol/liter)	2.7×10^{-163}
TCHH	1q21.3	18 bp (6 aa)	5–15	α-Helix rod (size)	Male pattern baldness score	−0.063 ± 0.006 SD	1.6×10^{-55}

null alleles that produce undetectable serum Lp(a) (*7*). This approach created an effective haploid model for Lp(a) and made it possible to systematically identify and measure the effects of Lp(a)–altering alleles (fig. S10). We performed stepwise conditional analysis to identify *LPA* sequence variants that associated with low Lp(a) despite occurring on short- or medium-length KIV-2 alleles that typically associate with higher Lp(a) levels (*7*).

These analyses identified associations with 17 protein-altering variants, each of which appeared to greatly reduce Lp(a) ($P < 1 \times 10^{-17}$ for each variant, Fisher's exact test or linear regression, table S4); 43% of European haplotypes were affected by at least one of these variants. Six variants predicted to partially or fully abolish constitutive splice sites and six missense variants achieved the strongest associations in 12 consecutive stages of stepwise analysis; five additional rare (MAF <1%) coding variants exhibited top or near-top associations in further conditional analyses (Fig. 1A, fig. S11, and table S4). The two variants with the largest impacts on Lp(a) variation in the European population (because of their high allele frequencies; MAF = 13 and 21%) were variants within the KIV-2 region that are computationally predicted to impair splicing (*19*) of KIV-2 exon 2. One of these splice variants has been experimentally validated (*20*). These variants reduced Lp(a) by 85 and 89%, respectively, when present within a single KIV-2 repeat unit; alleles carrying either variant on multiple repeat units within the VNTR produced nearly undetectable Lp(a) (fig. S12). Fine-mapping analyses identified three other common variants (MAF = 14 to 28%), two in the 5′ untranslated region (UTR) of *LPA*, which have both been observed to regulate translational activity (*21*, *22*), and one missense variant. All three variants associated with more modest effects on Lp(a) levels across a broad range of KIV-2 alleles (Fig. 1A and table S4).

The strong effects of the VNTR and SNPs at *LPA*, the large sample size of UKB, and the ability to chromosomally phase all of these variants accurately made it possible to identify nonlinear and cis-epistatic effects at *LPA*. Accounting for the effects of the 17 implicated coding variants at *LPA* showed that the inverse relationship between KIV-2 length and Lp(a) (*12*, *17*) breaks down for very short (high-protein-level) alleles (Fig. 1A). Throughout most of the KIV-2 length range (12 to 24 repeats), each one-repeat-unit decrease in KIV-2 length resulted in a 37% increase in Lp(a) (Fig. 1A). However, this effect was attenuated for alleles with fewer than 12 repeats and appeared to invert around eight repeats ($P = 9.4 \times 10^{-31}$, linear regression; Fig. 1A and fig. S13). Accounting for the nonlinear effect of KIV-2 length and for phase-resolved *LPA* sequence variants explained 90% of the heritable variance (83% of total variance) in Lp(a) [versus ~60% of total variance in earlier work (*12*, *23*)].

Serum Lp(a) levels vary across populations (*12*), with median measurements fourfold higher among Africans than among Europeans, but the reason for this cross-population variation has been unclear. We found that this variation was largely explained by population differences in the allele frequencies of *LPA* sequence variants (Fig. 1B). Elevated Lp(a) in UKB participants of African ancestry (median 80.1 nmol/liter versus 18.5 nmol/liter in Europeans) was primarily explained by the paucity of alleles carrying variants that greatly reduced Lp(a) (~13% of African alleles versus ~43% of European alleles despite sufficient discovery power in both populations) and the higher frequency of the Lp(a)–increasing 5′ UTR variant among African alleles (MAF = 46% versus 17% in European alleles for rs1800769; Fig. 1C). These allele frequency differences also explained the apparent difference in shape of the Lp(a)–KIV-2 curve in different populations (fig. S14).

The accuracy of genetically predicted Lp(a) ($R^2 = 0.83$ in Europeans) enabled insights into epidemiological associations involving Lp(a). We observed that the myocardial infarction risk–increasing effect of higher Lp(a) (*15*, *24*) extends to extreme Lp(a) levels [odds ratio (OR) = 3.1, 95% confidence interval (CI) = 1.9 to 5.2 for individuals with genetically predicted Lp(a) >400 nmol/liter; Fig. 1D]. By contrast, lower genetically predicted Lp(a) did not associate with increased type 2 diabetes (T2D) risk, suggesting that the 17% (SE 1%) lower levels of Lp(a) observed in T2D patients represents reverse causation resulting from T2D itself, T2D-related liver comorbidities, or T2D medications (Fig. 1E, fig. S15, and table S5).

Human height is strongly affected by VNTRs in *ACAN* and *TENT5A*

Human height associates with hundreds of common alleles (*25*), generally with small effect sizes (<0.05 SDs). By contrast, size variation of a 57-bp (19-aa) repeat in the *ACAN* gene strongly associated with height ($P = 1.7 \times 10^{-234}$, BOLT-LMM), with an effect size differential of 0.49 SDs (SE 0.04), or 3.2 cm, between the longest and shortest European alleles (Fig. 2). This association, which appears to underlie one of the first reported genetic associations with height (*26*), was also observed in a parallel study using long-read sequencing in the deCODE cohort (*11*). Here, analysis in the larger, more diverse UKB cohort, which contains double the range of allelic variation, including a very short, six-repeat African allele and European alleles with up to ~44 repeats (Fig. 2, B and D), uncovered several additional insights.

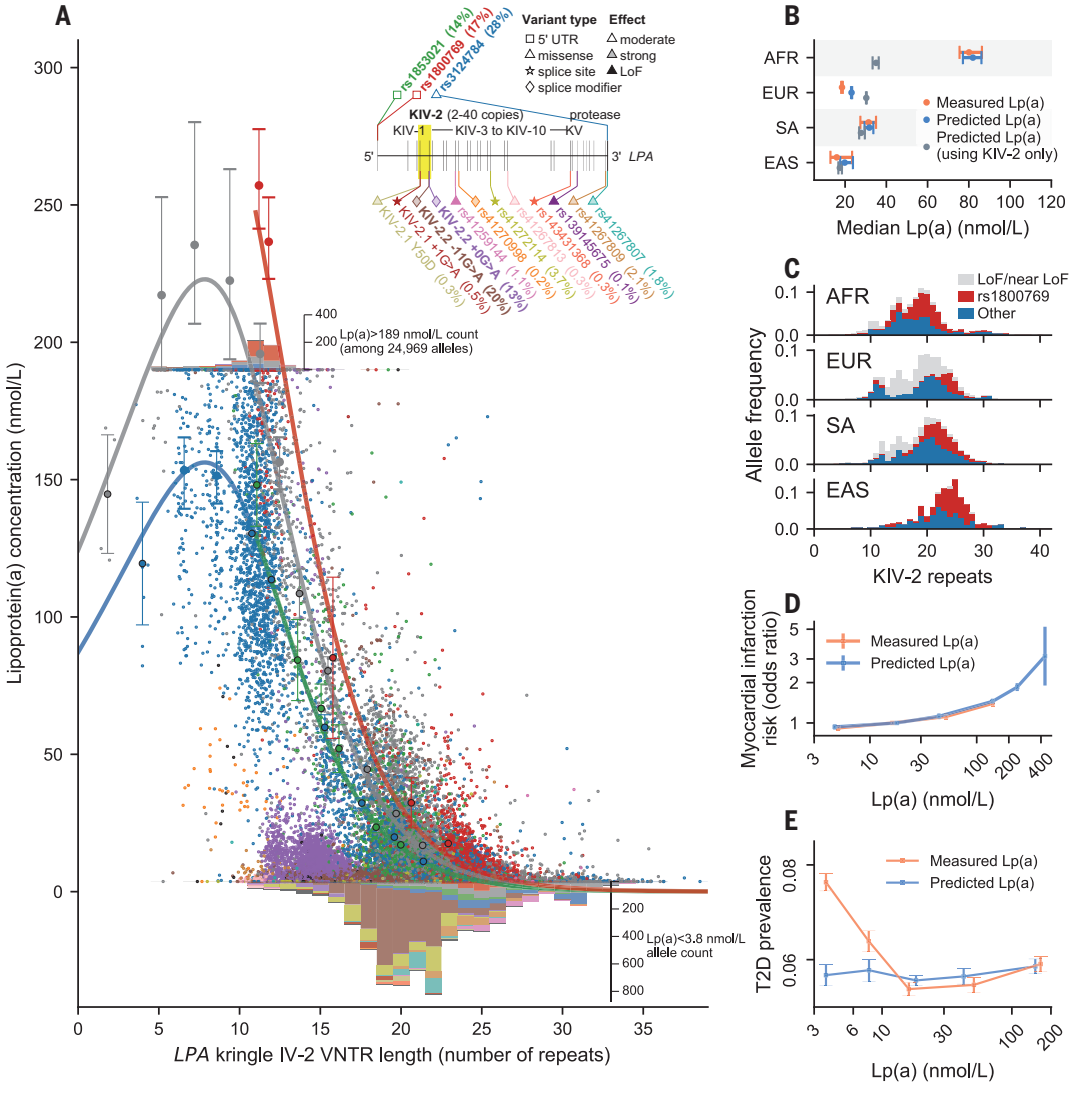

Fig. 1. Kringle IV-2 repeat length variation and 23 *LPA* SNPs together explain ~90% of lipoprotein(a) heritability. (**A**) Serum Lp(a) versus KIV-2 VNTR length in an effective-haploid model of Lp(a) involving 24,969 *LPA* alleles (in exome-sequenced UKB participants of European ancestry) for which the allele on the homologous chromosome was predicted to produce negligible Lp(a) (<4 nmol/liter). Colors indicate the 15 most common Lp(a)–modifying SNPs identified by fine-mapping analysis (full list in table S4). Curves indicate parametric fits of Lp(a) to KIV-2 length. Gray indicates alleles not carrying any Lp(a)–modifying SNPs; red, blue, and green are carriers of a single common Lp(a)–modifying SNP; large points are mean Lp(a) among such alleles in KIV-2 length bins. Error bars indicate 95% CIs. Histograms (top and bottom) show counts of Lp(a) measurements outside of the reportable range (<3.8 or >189 nmol/liter) colored by Lp(a)–modifying SNPs (*7*). (**B**) Observed and predicted median Lp(a) among individuals of African (AFR; *N* = 893), European (EUR; *N* = 42,162), South Asian (SA; *N* = 954), and East Asian (EAS; *N* = 156) ancestry. (**C**) *LPA* allele frequencies by ancestry. VNTR alleles in *cis* with a large-effect Lp(a)–reducing variant (respectively, the Lp(a)–increasing 5′ UTR variant rs1800769) are indicated in gray (respectively, red). (**D** and **E**) Myocardial infarction risk (D) and T2D prevalence (E) versus measured or genetically predicted Lp(a). Error bars indicate 95% CIs.

Height exhibited an approximately linear relationship with length of the *ACAN* VNTR. Consistent increasing effects were observed across a series of at least nine distinct VNTR allele lengths, resulting in an association signal ($P = 1.7 \times 10^{-234}$, BOLT-LMM) stronger than that of any nearby variant, explaining 0.19% of height variance among European-ancestry UKB participants (Fig. 2, C and D). Moreover, among the 7543 UKB participants of African ancestry, the *ACAN* VNTR association was nearly 50% stronger than the association of any other variant in the genome ($P = 5.2 \times 10^{-12}$ for the VNTR versus $P = 1.4 \times 10^{-8}$ for the strongest SNP association) and explained a much larger 0.60% of height variance, primarily because of the greater VNTR length variation (SD = 3.7 repeats versus 1.5 repeats in Europeans; Fig. 2B). Imputation of the VNTR association into height association statistics from the African Ancestry Anthropometry Genetics Consortium cohort (*27*) replicated these results [with the VNTR explaining an estimated 0.42% of height variance; imputed $P = 5.8 \times 10^{-40}$ versus linear regression $P = 3.4 \times 10^{-20}$ for the strongest SNP association genome-wide; fig. S16 (*7*)].

Aggrecan, the protein encoded by *ACAN*, is a component of the extracellular matrix in growth plate cartilage and is required for normal growth plate cytoarchitecture (*28*). The VNTR generates 2.4-fold size variation in aggrecan's first chondroitin sulfate domain (CS1), in which amino acid residues are modified by long, charged polysaccharide chains that endow this extracellular matrix with key properties including the ability to hold large amounts of water (*29*).

As at *LPA*, incorporation of the *ACAN* VNTR into genetic association analysis (by stepwise conditional analysis) made it possible to identify additional genetic effects that are driven at *ACAN* by two common missense SNPs (Fig. 2C and table S6). These two missense SNPs, which affect ACAN globular domains, had two of the top three predicted deleteriousness scores (*30*) (CADD = 23.1 for rs3817428 and 27.6 for rs34949187) among common missense SNPs in *ACAN* and were corroborated by Bayesian fine-mapping (*9*) analysis (FINEMAP posterior probability of causality >0.99). A combined model including the VNTR and these SNPs explained 0.33% of height variance in Europeans.

Despite the strong effects of *ACAN* VNTR alleles on height, neither end of the allelic spectrum appeared to compromise ACAN function in any way detrimental to health. Whereas loss-of-function mutations in *ACAN* cause autosomal-dominant skeletal disorders (*31*),

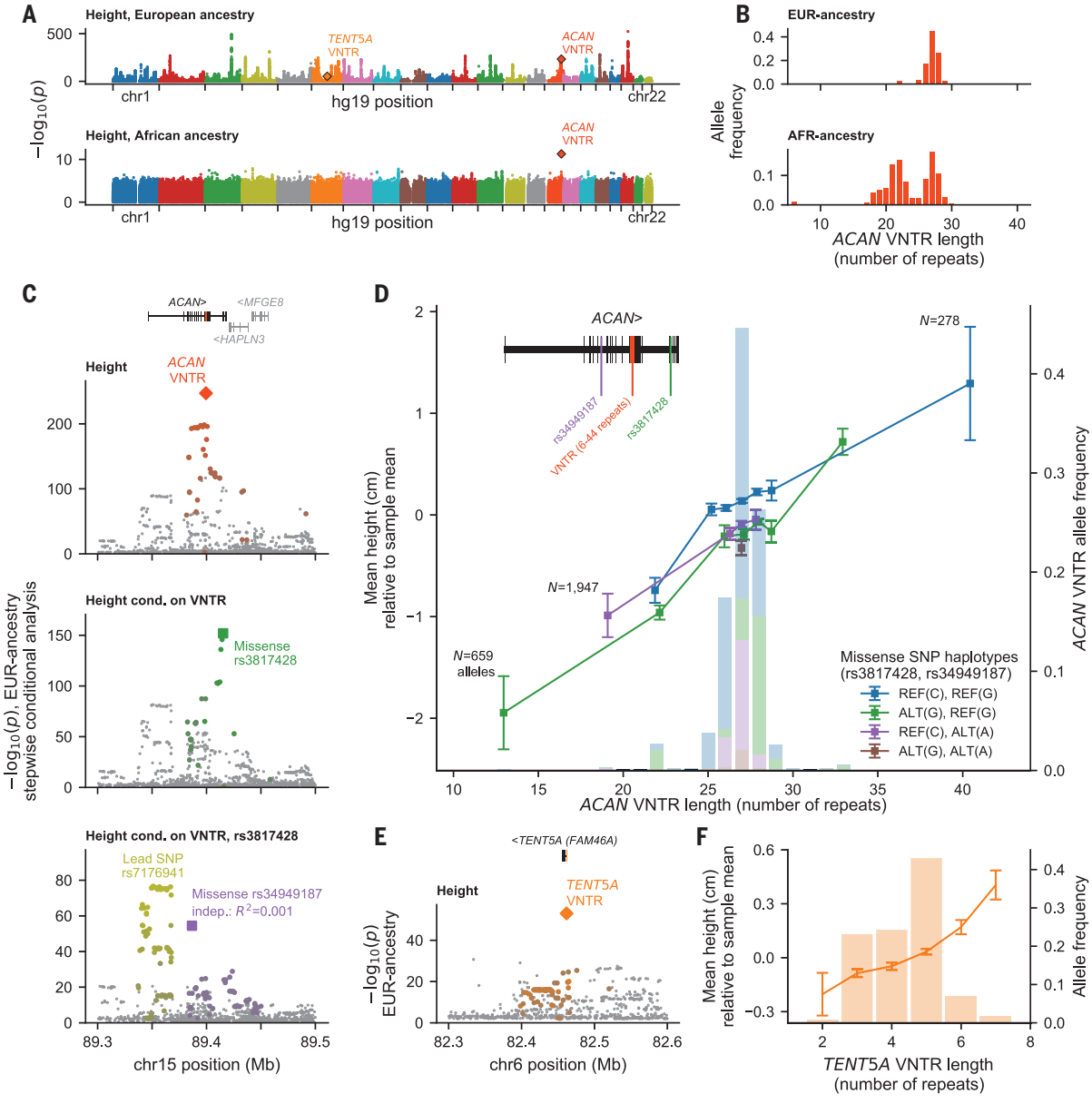

Fig. 2. Lengths of protein-coding repeat polymorphisms in *ACAN* and *TENT5A* associate with human height. (**A**) Genetic associations with height in UKB participants of EUR (top; N = 415,280) and AFR (bottom; N = 7543) ancestry. (**B**) *ACAN* VNTR allele length distributions. (**C**) Height association statistics at *ACAN* in three consecutive steps of stepwise conditional analysis (EUR N = 415,280). Large diamonds and squares indicate likely causal coding mutations; colored dots are variants in partial LD (R^2 > 0.1) with labeled variants. Height phenotypes were adjusted for genetic predictions computed using the rest of the genome (*7*). (**D**) Mean height of carriers (lines, left axis) and EUR allele frequencies (histograms, right axis) of *ACAN* alleles defined by VNTR length and missense SNP haplotype. Error bars indicate 95% CIs. Rare long alleles (40 to 42 repeats) were grouped into one bin. (**E**) Height associations at *TENT5A*. (**F**) Mean height and EUR allele frequencies for *TENT5A* VNTR alleles. Error bars indicate 95% CIs.

VNTR length variation did not associate at Bonferroni significance with any disease in the UKB ($P > 3 \times 10^{-4}$, logistic regression). A participant homozygous for the short, six-repeat allele (allele frequency = 1.2% among participants with African ancestry) had no reported musculoskeletal disease phenotypes.

A distinct coding VNTR in the *TENT5A* gene (previously named *FAM46A*) consisting of two to seven repeats of 15 bp also associated with height ($P = 2.5 \times 10^{-53}$, BOLT-LMM), with six VNTR alleles exhibiting monotonically increasing effects (Fig. 2, E and F). TENT5A, a poly(A) polymerase in which multiple coding variants have been linked to autosomal-recessive osteogenesis imperfecta (*32*), polyadenylates and increases expression in osteoblasts of the collagen genes *COL1A1* and *COL1A2* and other genes mutated in this disease (*33*).

Kidney function phenotypes shaped by a VNTR in *MUC1*

The *MUC1* gene encodes a secreted (cell surface–associated) protein (mucin 1) with cell-adhesive and anti-adhesive properties. *MUC1* harbors a VNTR that contains 20 to 125 repeats (*34*) of a 60-bp (20-aa) coding sequence that determines the length of a heavily glycosylated extracellular domain. Ultra-rare frameshift mutations within the

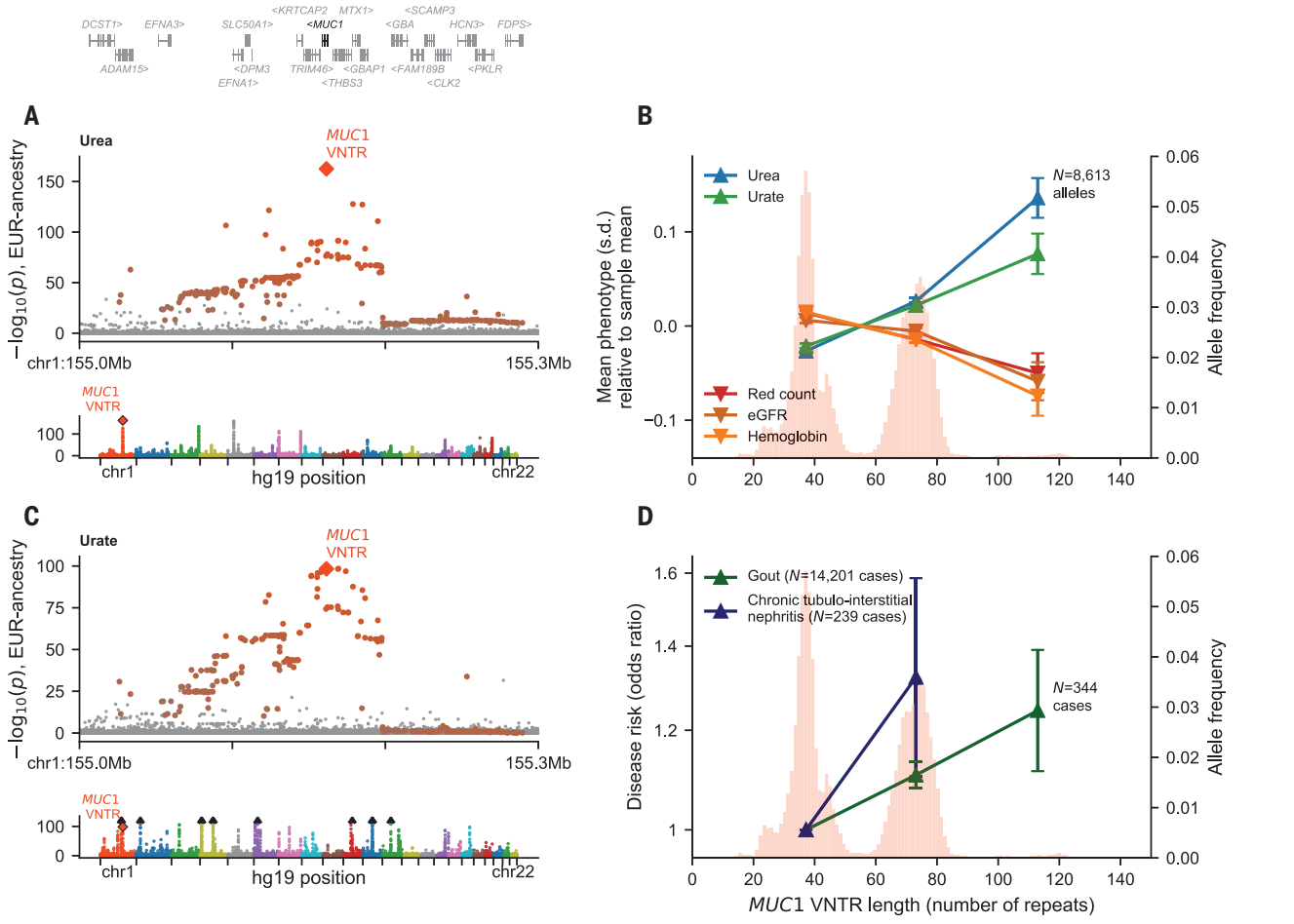

Fig. 3. *MUC1* VNTR length associates with multiple renal phenotypes. (**A** and **C**) Genetic associations with serum urea (A) and serum urate (C) at *MUC1* (top; orange dots indicate variants in LD with *MUC1* VNTR length (R^2 > 0.1) and genome-wide (bottom); *N* = 415,280 UKB EUR participants. (**B** and **D**) Mean phenotypes in carriers (B) or disease ORs (D) (lines, left axis) and allele frequencies (histograms, right axis) of *MUC1* VNTR alleles. VNTR alleles were stratified into three groups for phenotype analyses: short (<55 repeat units), long (55 to 95 repeat units), and very long (>95 repeat units). Error bars indicate 95% CIs. eGFR, estimated glomerular filtration rate.

MUC1 VNTR cause autosomal-dominant tubulointerstitial kidney disease (*35*). In our analyses, length of the *MUC1* VNTR associated with several renal phenotypes (Fig. 3), including serum urea ($P = 2.7 \times 10^{-163}$, BOLT-LMM) and serum urate ($P = 4.7 \times 10^{-99}$, BOLT-LMM). Longer VNTR alleles also associated with gout ($P = 3.6 \times 10^{-17}$, logistic regression), a disease caused by excessive uric acid crystallization in the joints.

The *MUC1* VNTR length polymorphism appeared to underlie some of the strongest, earliest reported SNP associations with serum urea and serum urate, two biomarkers of renal function that otherwise have somewhat independent heritability [genetic correlation = 0.25 (SE 0.01); Fig. 3, A and C]. For urea, the VNTR exhibited the strongest association genome-wide (matching that of a SNP on chromosome 5), explaining ~1% of heritable variance (~0.2% of total variance) in Europeans and accounting for nearly all of the association signal at the *MUC1* locus [previously reported as *MTX1-GBA* (*36*); Fig. 3A]. For urate, the VNTR also appeared to be the primary causal variant at a locus previously reported as *TRIM46* (*37*) (Fig. 3C). Longer *MUC1* alleles associated with increasing levels of both serum urea and urate across the VNTR length spectrum, with an incompletely dominant effect on urea ($P = 2.3 \times 10^{-20}$ for interaction, linear regression; fig. S17) but an additive effect on urate ($P = 0.56$ for interaction).

Associations with additional renal phenotypes indicated a complex relationship between *MUC1* VNTR length and kidney function (Fig. 3, B and D). Long *MUC1* alleles (>55 repeat units) increased the risk of gout (OR = 1.10; 95% CI = 1.08 to 1.13, $P = 1.2 \times 10^{-16}$, logistic regression) and chronic tubulointerstitial nephritis (OR = 1.31, 95% CI = 1.09 to 1.57, $P = 3.4 \times 10^{-3}$, logistic regression), which remained significant after correcting for 13 kidney diseases tested. However, *MUC1* VNTR allele length did not associate with chronic kidney disease (OR = 1.01, 95% CI = 0.99 to 1.04, P = 0.33, logistic regression) reported in 14,573 cases and only weakly influenced glomerular filtration rate as estimated from serum creatinine (beta = −0.19%, 95% CI = 0.11 to 0.28, for long versus short alleles). Long *MUC1* alleles associated with modest reductions in red blood cell counts (beta = −0.029 SD, SE = 0.002, $P = 1.5 \times 10^{-39}$, linear regression) and hemoglobin levels (beta = −0.031 SD, SE = 0.002, $P = 9.9 \times 10^{-44}$, linear regression), possibly reflecting an impact of reduced kidney function on erythropoietin production.

TCHH VNTR strongly associates with hair phenotypes

Repeat length variation in a coding VNTR in *TCHH* associated strongly with male pattern baldness ($P = 1.6 \times 10^{-55}$, BOLT-LMM). *TCHH* encodes trichohyalin, a protein that associates in regular arrays with keratin intermediate filaments and confers mechanical strength to the inner root sheath (*38*). The 18-bp VNTR encodes part of a highly stabilized alpha-helix

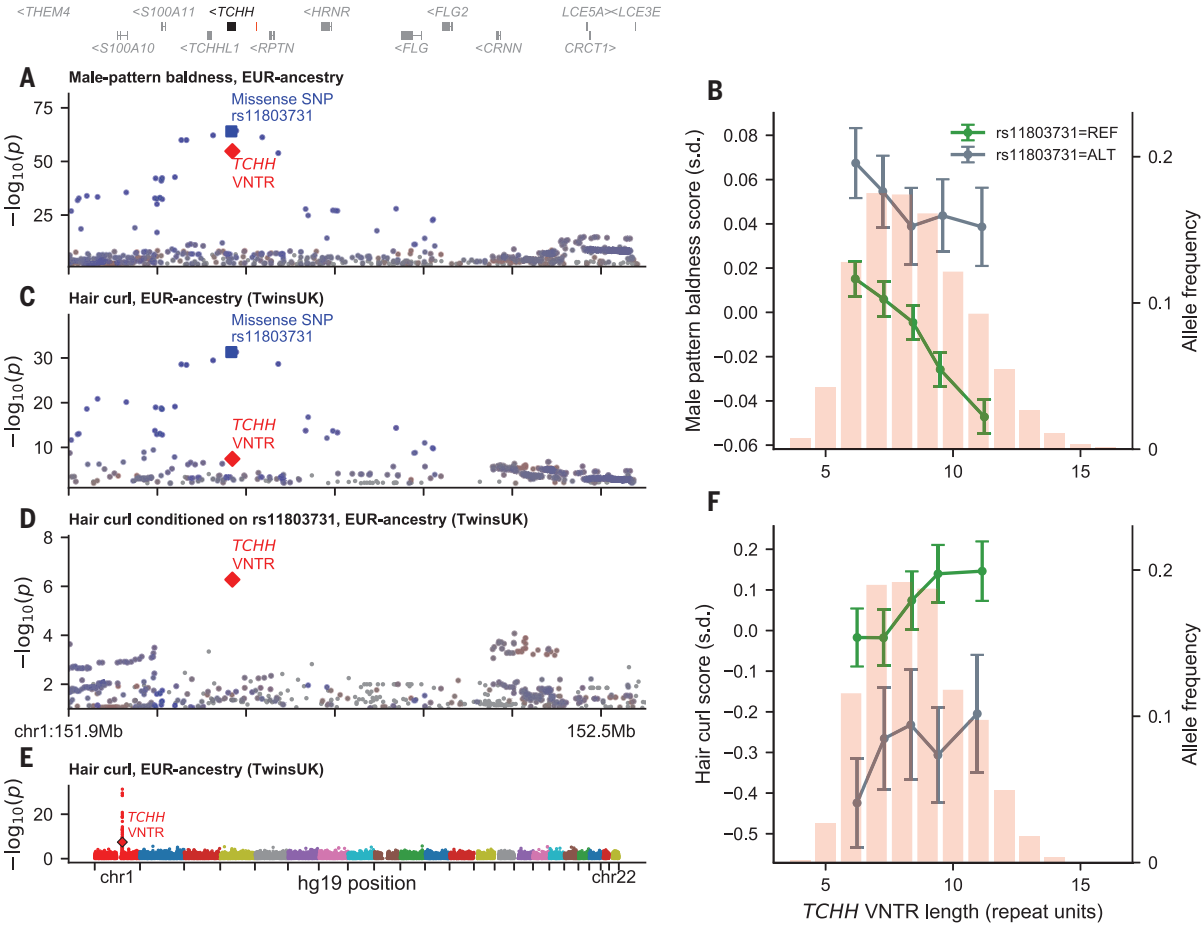

Fig. 4. *TCHH* VNTR length and missense SNP rs11803731 associate independently with hair phenotypes. (**A**) Genetic associations with male pattern baldness at *TCHH* (N = 189,537 male UKB EUR participants). Colors indicate partial LD (R > 0.1) with missense SNP rs11803731 (blue), the *TCHH* VNTR (red), or both rs11803731 and VNTR length (purple). (**B**) Mean baldness score in carriers (lines, left axis) and allele frequencies (histograms, right axis) of *TCHH* alleles. *TCHH* alleles were binned by VNTR length quintile and missense SNP rs11803731 status. (**C** and **D**) Genetic associations with hair curl at *TCHH* in N = 3334 TwinsUK participants [conditioned on rs11803731 in (D)]. (**E**) Genome-wide associations with hair curl in TwinsUK. (**F**) Relationship between *TCHH* allele length and hair curl [analogous to (B)].

that forms an elongated rod structure (*39*). A rare nonsense mutation in *TCHH* has been implicated in uncombable hair syndrome (*40*), and a common haplotype containing the *TCHH* missense SNP rs11803731 (encoding a leucine to methionine substitution in TCHH) is by far the strongest genetic determinant of hair curl in individuals of European ancestry (*41*, *42*). In the UKB, the *TCHH* VNTR and rs11803731 exhibited independent associations with male pattern baldness (Fig. 4, A and B).

The *TCHH* VNTR appeared to be hypermutable and was poorly tagged by all nearby individual SNPs (R^2 < 0.1), leading us to wonder whether it might also contribute to hair curl in a way invisible to genome-wide association studies of this phenotype. Imputing *TCHH* VNTR alleles into the TwinsUK cohort (*43*) (N = 3334 genotyped individuals with hair curl phenotypes) revealed that the *TCHH* VNTR appeared to be the human genome's second-largest contributor to hair curl variation genome-wide (explaining ~1% of variance; $P = 3.6 \times 10^{-8}$, BOLT-LMM) after the missense SNP rs11803731 in *TCHH* (which explained ~4% of variance; Fig. 4, C to F). LD between the VNTR and rs11803731 further explained an association reported near *LCE3E* (450 kb upstream of *TCHH*) previously thought to be independent of *TCHH* (*42*) (Fig. 4, C and D).

Discussion

These results identify many strong effects of protein-coding VNTRs on human phenotypes. Most were among the strongest effects of all common variants identified for these phenotypes to date and resolved previously mysterious genetic associations for multiple traits. Incorporation of multiallelic VNTRs into fine-mapping analyses also helped to identify many more functional variants at the same loci, revealing the importance of incorporating allelic series of SNP and VNTR alleles into functional studies and epidemiological research.

These results are likely just the leading edge of a far larger set of VNTR-phenotype associations that future studies will reveal. In this work with exome-sequencing data, we were unable to analyze VNTRs that exist in noncoding sequences, are too short for depth-of-coverage to accurately measure length variation, or are too mutable to segregate well with SNP haplotypes. We anticipate that newer sequencing technologies applied to large, diverse cohorts will yield further insights into the mutational and evolutionary processes of VNTRs and their contribution to the "missing heritability" of human phenotypes.

A frustration in the study of human genetics has been that most reported genetic associations involve haplotypes of noncoding and missense SNPs with potential phenotypic

contributions that are challenging to disentangle from one another and have first-order molecular effects that are opaque. VNTRs have several attributes that help to overcome these challenges. First, multiallelic VNTRs usually share only partial LD with nearby diallelic SNP and indel variants. Second, associations with protein-coding VNTRs implicate the size and copy number of specific protein domains, leading to specific, testable hypotheses about the effects of protein domains in biological systems. Third, the directions of coding VNTR associations have clear meaning, revealing whether risk is generated by having more or less of a domain. Finally, VNTRs generate natural allelic series of functionally distinct alleles that can be used for dose–response studies in human tissues and cellular models. We anticipate that these attributes will lead to new insights about the mechanisms by which gene and protein variation affect human biology.

REFERENCES AND NOTES

1. P. H. Sudmant *et al*., *Nature* **526**, 75–81 (2015).
2. A. Sulovari *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23243–23253 (2019).
3. M. D. Lalioti *et al*., *Nature* **386**, 847–851 (1997).
4. C. Wijmenga *et al*., *Nat. Genet.* **2**, 26–30 (1992).
5. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, *Nat. Genet.* **39**, 906–913 (2007).
6. C. Bycroft *et al*., *Nature* **562**, 203–209 (2018).
7. Materials and methods are available as supplementary materials.
8. C. V. Van Hout *et al*., *Nature* **586**, 749–756 (2020).
9. C. Benner *et al*., *Bioinformatics* **32**, 1493–1501 (2016).
10. A. R. Barton, M. A. Sherman, R. E. Mukamel, P.-R. Loh, *Nat. Genet.* **53**, 1260–1269 (2021).
11. D. Beyter *et al*., *Nat. Genet.* **53**, 779–786 (2021).
12. K. Schmidt, A. Noureen, F. Kronenberg, G. Utermann, *J. Lipid Res.* **57**, 1339–1359 (2016).
13. P.-R. Loh *et al*., *Nat. Genet.* **47**, 284–290 (2015).
14. P. Ebert *et al*., *Science* **372**, eabf7117 (2021).
15. R. Clarke *et al*., *N. Engl. J. Med.* **361**, 2518–2528 (2009).
16. G. Utermann *et al*., *J. Clin. Invest.* **80**, 458–465 (1987).
17. A. L. White, J. E. Hixson, D. L. Rainwater, R. E. Lanford, *J. Biol. Chem.* **269**, 9060–9066 (1994).
18. E. Boerwinkle *et al*., *J. Clin. Invest.* **90**, 52–60 (1992).
19. K. Jaganathan *et al*., *Cell* **176**, 535–548.e24 (2019).
20. S. Coassin *et al*., *Eur. Heart J.* **38**, 1823–1831 (2017).
21. B. R. Zysow, G. E. Lindahl, D. P. Wade, B. L. Knight, R. M. Lawn, *Arterioscler. Thromb. Vasc. Biol.* **15**, 58–64 (1995).
22. K. Suzuki, M. Kuriyama, T. Saito, A. Ichinose, *J. Clin. Invest.* **99**, 1361–1366 (1997).
23. M. Trinder, M. M. Uddin, P. Finneran, K. G. Aragam, P. Natarajan, *JAMA Cardiol.* (2020).
24. D. F. Gudbjartsson *et al*., *J. Am. Coll. Cardiol.* **74**, 2982–2994 (2019).
25. L. Yengo *et al*., *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
26. M. N. Weedon *et al*., *Nat. Genet.* **40**, 575–583 (2008).
27. M. Graff *et al*., *Am. J. Hum. Genet.* **108**, 564–582 (2021).
28. K. L. Lauing *et al*., *Dev. Biol.* **396**, 224–236 (2014).
29. K. J. Doege, S. N. Coulter, L. M. Meek, K. Maslen, J. G. Wood, *J. Biol. Chem.* **272**, 13974–13979 (1997).
30. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, *Nucleic Acids Res.* **47** (D1), D886–D894 (2019).
31. L. Gleghorn, R. Ramesar, P. Beighton, G. Wallis, *Am. J. Hum. Genet.* **77**, 484–490 (2005).
32. M. Doyard *et al*., *J. Med. Genet.* **55**, 278–284 (2018).
33. O. Gewartowska *et al*., *Cell Rep.* **35**, 109015 (2021).
34. J. C. Fowler, A. S. Teixeira, L. E. Vinall, D. M. Swallow, *Hum. Genet.* **113**, 473–479 (2003).
35. A. Kirby *et al*., *Nat. Genet.* **45**, 299–303 (2013).
36. Y. Okada *et al*., *Nat. Genet.* **44**, 904–909 (2012).
37. A. Köttgen *et al*., *Nat. Genet.* **45**, 145–154 (2013).
38. P. M. Steinert, D. A. D. Parry, L. N. Marekov, *J. Biol. Chem.* **278**, 41409–41419 (2003).
39. S. C. Lee *et al*., *J. Biol. Chem.* **268**, 12164–12176 (1993).
40. F. B. Ü. Basmanav *et al*., *Am. J. Hum. Genet.* **99**, 1292–1304 (2016).
41. S. E. Medland *et al*., *Am. J. Hum. Genet.* **85**, 750–755 (2009).
42. F. Liu *et al*., *Hum. Mol. Genet.* **27**, 559–575 (2018).
43. A. Moayyeri, C. J. Hammond, D. J. Hart, T. D. Spector, *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
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.

ACKNOWLEDGMENTS

We thank R. Gupta, J. Hirschhorn, M. Hujoel, S. Raychaudhuri, and M. Warman for helpful discussions. This research was conducted using the UKB resource under application no. 40709. Computational analyses were performed on the O2 High Performance Compute Cluster, which is supported by the Research Computing Group, at Harvard Medical School (http://rc.hms.harvard.edu). **Funding:** R.E.M. was supported by National Science Foundation (NSF) grant DMS-1939015 and National Institutes of Health (NIH) grant K25 HL150334. R.E.H. and S.A.M. were supported by NIH grant R01 HG006855. M.A.S. was supported by the MIT John W. Jarve (1978) Seed Fund for Science Innovation and by NIH fellowship F31 MH124393. A.R.B. was supported by NIH fellowship F31 HL154537 and training grant T32 HG 2295-16. P.-R.L. was supported by NIH grant DP2 ES030554, a Burroughs Wellcome Fund Career Award at the Scientific Interfaces, the Next Generation Fund at the Broad Institute of MIT and Harvard, and a Sloan Research Fellowship. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (CDRF), Zoe Global Ltd., and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and Biomedical Research Centre based at Guy's and St. Thomas' NHS Foundation Trust in partnership with King's College London. **Author contributions:** R.E.M., R.E.H., S.A.M., and P.-R.L. conceived and designed the study. R.E.M., R.E.H., and P.-R.L. designed and implemented the statistical methods and performed the computational analyses. Y.Z. helped to design and implement the VNTR partitioner algorithm. R.E.M., R.E.H, A.R.B., M.A.S., S.A.M., and P.-R.L. interpreted analytical results. All authors wrote and edited the manuscript. **Competing interests:** The authors declare no competing interests. **Data availability:** Access to the following data resources is available to all bona fide researchers by application: UKB (https://www.ukbiobank.ac.uk/), Twins UK (https://twinsuk.ac.uk/), the Haplotype Reference Consortium imputation panel (http://www.haplotype-reference-consortium.org/), and AAAGC height summary statistics (https://www.ebi.ac.uk/gwas/). Individual-level VNTR allele length estimates (resolved to phased SNP haplotypes) and genetically predicted Lp(a) values are available from UKB as a return from application no. 40709. **Code availability:** The following publicly available software resources were used to perform analyses in this work: Eagle2 (v2.3.5), https://data.broadinstitute.org/alkesgroup/Eagle/; Minimac4 (v1.0.1), https://genome.sph.umich.edu/wiki/Minimac4; BOLT-LMM (v2.3.5), https://data.broadinstitute.org/alkesgroup/BOLT-LMM/; FINEMAP (v1.3.1), http://www.christianbenner.com/; plink (v1.9 and v2.0), https://www.cog-genomics.org/plink2/; Tandem Repeats Finder (v4.09.1), tandem.bu.edu/trf/trf.html; the TOPMed Imputation Server, https://imputation.biodatacatalyst.nhlbi.nih.gov/; BLAT (v35), http://hgdownload.soe.ucsc.edu/admin/exe/; susieR (v0.10.1), https://stephenslab.github.io/susieR/; LDstore (v2.0), http://www.finemap.me; and ImpG (v1.0.1), https://github.com/huwenboshi/ImpG. Code and scripts used to perform analyses are available at Zenodo (*44*).

SUPPLEMENTARY MATERIALS

https://science.org/doi/10.1126/science.abg8289
Materials and Methods
Supplementary Text
Figs. S1 to S17
Tables S1 to S8
References (*45*–*98*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

29 January 2021; accepted 20 August 2021
10.1126/science.abg8289

Protein-coding repeat polymorphisms strongly shape diverse human phenotypes

Ronen E. MukamelRobert E. HandsakerMaxwell A. ShermanAlison R. BartonYiming ZhengSteven A. McCarrollPo-Ru Loh

Science, 373 (6562), • DOI: 10.1126/science.abg8289

Repeats associated with phenotype

The degree to which repeated sequences within a genome affect human phenotypes has been difficult to establish. Mukamel *et al*. examined thousands of genomes in the UK Biobank and found that some of the largest effects of common genetic variants on human phenotypes, including those with clinical relevance, arise from protein-coding repeat polymorphisms (see the Perspective by Gymrek and Goren). Mapping the effects of the size and copy number of these repeated protein domains links genetic variation to human phenotypes, including lipoprotein(a) concentration, height, and male pattern balding. Furthermore, the alleles and frequencies of these repeated sequences differ between individuals of African and European descent, resulting in differences between the populations with clinical relevance for traits including lipoprotein(a) levels, a risk factor for coronary artery disease. —LMZ

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.