Article

A positively selected *FBN1* missense variant reduces height in Peruvian individuals

https://doi.org/10.1038/s41586-020-2302-0

Received: 28 February 2019

Accepted: 10 March 2020

Published online: 13 May 2020

Check for updates

Samira Asgari[1,2,3,4,5], Yang Luo[1,2,3,4,5], Ali Akbari[4,6], Gillian M. Belbin[7,8,9], Xinyi Li[1,2,3,4,5], Daniel N. Harris[10,11], Martin Selig[12], Eric Bartell[4,5,13], Roger Calderon[14], Kamil Slowikowski[1,2,3,4,5], Carmen Contreras[14], Rosa Yataco[14], Jerome T. Galea[15], Judith Jimenez[14], Julia M. Coit[16], Chandel Farroñay[14], Rosalynn M. Nazarian[12], Timothy D. O'Connor[10,17], Harry C. Dietz[18,19], Joel N. Hirschhorn[4,6,13,20], Heinner Guio[21], Leonid Lecca[14], Eimear E. Kenny[7,8,9], Esther E. Freeman[22], Megan B. Murray[16] & Soumya Raychaudhuri[1,2,3,4,5,23]✉

On average, Peruvian individuals are among the shortest in the world[1]. Here we show that Native American ancestry is associated with reduced height in an ethnically diverse group of Peruvian individuals, and identify a population-specific, missense variant in the *FBN1* gene (E1297G) that is significantly associated with lower height. Each copy of the minor allele (frequency of 4.7%) reduces height by 2.2 cm (4.4 cm in homozygous individuals). To our knowledge, this is the largest effect size known for a common height-associated variant. *FBN1* encodes the extracellular matrix protein fibrillin 1, which is a major structural component of microfibrils. We observed less densely packed fibrillin-1-rich microfibrils with irregular edges in the skin of individuals who were homozygous for G1297 compared with individuals who were homozygous for E1297. Moreover, we show that the E1297G locus is under positive selection in non-African populations, and that the E1297 variant shows subtle evidence of positive selection specifically within the Peruvian population. This variant is also significantly more frequent in coastal Peruvian populations than in populations from the Andes or the Amazon, which suggests that short stature might be the result of adaptation to factors that are associated with the coastal environment in Peru.

With average heights of 165.3 cm and 152.9 cm for men and women, respectively, Peruvian individuals are among the shortest people in the world[1]. The genetic makeup of Peruvian individuals is shaped by admixture between Native American residents of Peru and the incoming Europeans, Africans and Asians who have arrived in Peru since the sixteenth century[2,3]. A previous study of height in South and Latin Americans reported that Native American ancestry is correlated with shorter height in these populations[4]; however, this association may have been the result of confounding socioeconomic or environmental factors that were not captured by socioeconomic covariates in that study (education and wealth). Even if the association between Native American ancestry and height was driven by genetic factors, the specific genes and adaptive processes remain unclear.

To define genetic factors that contribute to height in Peruvian individuals, we obtained height and genotyping data from 3,134 individuals from 1,947 households in Lima, Peru (Methods and Supplementary Information section 1). We inferred the proportion of Native American ancestry in each individual (Extended Data Fig. 1) and observed a negative correlation between height and the proportion of Native American ancestry (Pearson's correlation coefficient (r) = −0.28, 95% confidence interval = −0.31–0.25, $P = 9.3 \times 10^{-58}$) (Fig. 1a and Supplementary Information section 2). Native American ancestry remained significantly associated with decreased height after adjusting for age, sex, African and Asian ancestry proportions and a random household effect, as a proxy for unmeasured environmental factors, and a kinship matrix to account for genetic relatedness and population structure ($P = 7.2 \times 10^{-43}$, effect size = −14.75 cm for 100% Native American versus 100% European ancestry, s.e.m. = 1.06, Fig. 1b).

To identify variants that cause this effect, we performed a genome-wide association study (GWAS). We observed associations

[1]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [6]Department of Genetics, Harvard Medical School, Boston, MA, USA. [7]The Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [8]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [9]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [10]Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [11]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. [12]Pathology Service, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [13]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA. [14]Socios En Salud, Lima, Peru. [15]School of Social Work, University of South Florida, Tampa, FL, USA. [16]Department of Global Health and Social Medicine, and Division of Global Health Equity, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [17]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [18]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [19]Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [20]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [21]Instituto Nacional de Salud, Lima, Peru. [22]Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [23]Centre for Genetics and Genomics Versus Arthritis, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. ✉e-mail: soumya@broadinstitute.org

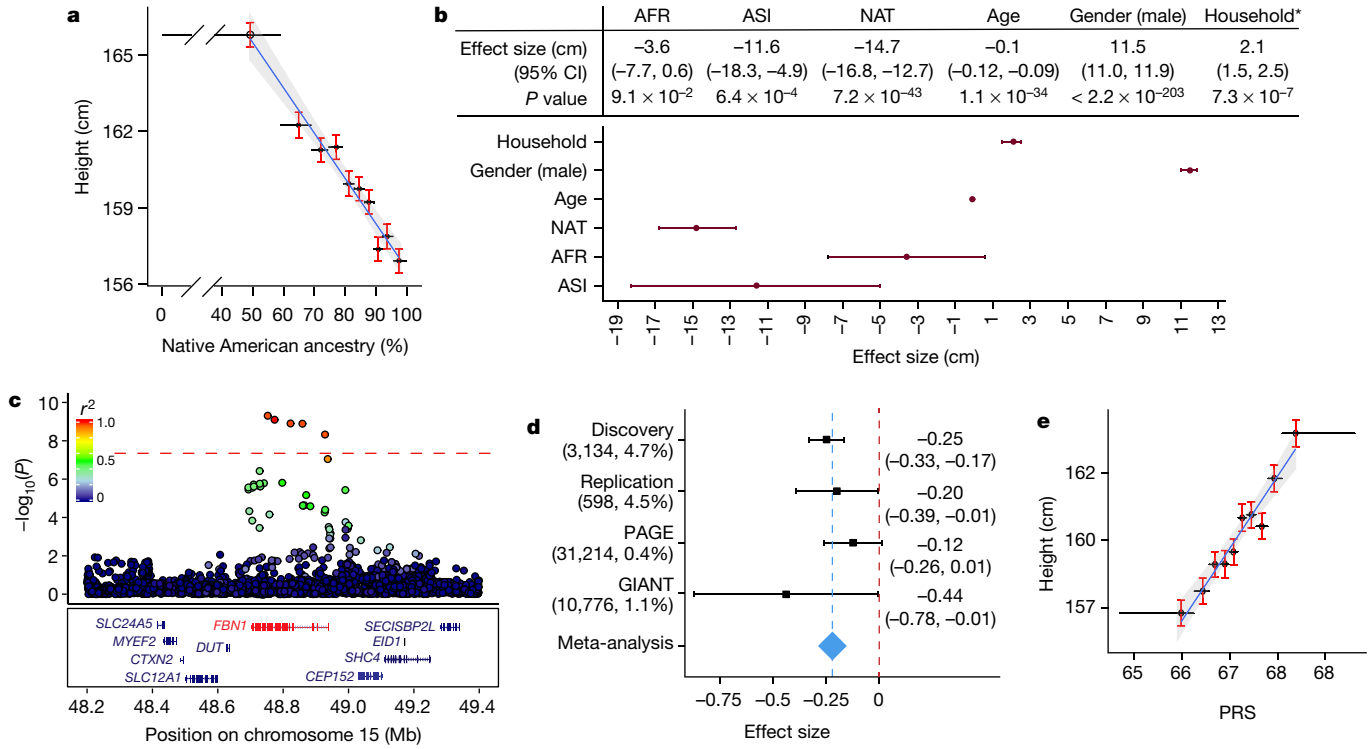


Fig. 1 | Genetic architecture of height in the Peruvian population. a, Height is negatively correlated with the proportion of Native American ancestry (n = 3,134 individuals, Pearson's r = −0.28, 95% confidence interval = −0.31–0.25, t = −16.36, d.f. = 3,132, $P = 9.3 \times 10^{-58}$, two-sided one-sample Student's t-test). Points show the median for a decile of Native American ancestry (x axis) and the average height for that decile (y axis). Error bars indicate the range (x axis) and s.e.m. (y axis). **b**, Increased Native American ancestry is associated with lower height after adjusting for age, sex, African and Asian ancestry proportions, and household as a proxy for socioeconomic factors and genetic relatedness (n = 3,134 individuals). *Household effect size is calculated as the s.d. of the intercept of the model. The effect sizes for African (AFR), Asian (ASI) and Native American (NAT) ancestry are given relative to European ancestry. P values were calculated using two-sided χ^2 difference tests. **c**, Locus-specific Manhattan plot of $-\log_{10}$-transformed GWAS P values. One locus on chromosome 15 passed the genome-wide significance threshold ($P < 5 \times 10^{-8}$, n = 3,134 individuals). P values were calculated using two-sided Wald tests. Dots show variants coloured according to their linkage disequilibrium with rs200342067 (total number of variants tested = 7,756,401, number of variants shown = 3,176). **d**, rs200342067 showed a similar MAF, direction of effect and effect size in an independent cohort of Peruvian individuals (n = 598 individuals) and two independent cohorts of Latino/Hispanic individuals (n = 31,214 and 10,776 individuals, respectively). Squares show the effect size of rs200342067 on inverse normally transformed height; the dashed blue line indicates the meta-analysis effect size; the diamond shows the meta-analysis s.e.m.; and the error bars indicate the 95% confidence intervals. The size of the cohort and the MAF of rs200342067 are shown in parentheses (left) and the effect sizes (95% confidence intervals) are shown on the right. **e**, Height is positively correlated with PRSs (n = 3,134 individuals, Pearson's r = 0.22, 95% confidence interval = 0.18–0.25, t = 12.36, d.f. = 3,132, $P = 2.7 \times 10^{-34}$, two-sided one-sample Student's t-test). Points indicate the median for a PRS decile (x axis) and the average height for that decile (y axis). Error bars show the range (x axis) and s.e.m. (y axis).

at five highly linked single-nucleotide polymorphisms (SNPs) within a single locus that overlapped with the gene *FBN1* (15q21.1, $P < 5 \times 10^{-8}$) (Extended Data Fig. 2a). One SNP, rs200342067 (minor allele frequency (MAF) = 4.72%, effect size = −2.22 cm, s.e.m. = 0.36, $P = 6.8 \times 10^{-10}$), is a missense variant (E1297G) whereas the other four SNPs are intronic (Fig. 1c and Extended Data Table 1). Accounting for additional covariates, such as population principal components or identity-by-descent[5], did not affect the association results (Supplementary Information section 3).

To replicate this association, we genotyped an independent cohort of Peruvian individuals (n = 598) (Methods and Supplementary Information section 1) and observed a similar allele frequency and effect size for rs200342067 in the replication cohort (MAF = 4.52%, effect size = −1.70 cm, s.e.m. = 0.82, P = 0.04) (Table 1). Meta-analysis of the discovery and replication cohorts increased the significance of the association (effect size = −2.14 cm, s.e.m. = 0.33, $P = 9.2 \times 10^{-11}$) (Table 1). We also tested the association of rs200342067 with inverse normally transformed height in data from the Genetic Investigation of Anthropometric Traits (GIANT)[6] and Population Architecture using Genomics and Epidemiology (PAGE)[7] studies, two publicly available datasets of Hispanic/Latino individuals. Although the allele frequencies were lower in these datasets (<1.15%), we observed similar effect sizes across cohorts (effect sizes (s.e.m.) for discovery, replication, PAGE and GIANT cohorts, respectively, were −0.25 (0.04), −0.20 (0.10), −0.12 (0.07) and −0.44 (0.22)) (Table 1). Meta-analysis of these cohorts further increased the strength of the association (effect size = −0.22 (s.e.m. = 0.03), $P = 9.8 \times 10^{-12}$) (Table 1 and Fig. 1d). These results confirm that the association between rs200342067 and height is not driven by statistical fluctuation or genotyping artefacts specific to the discovery cohort. We did not find any additional associations in the gene-based analysis of rare (MAF < 1%) or common variants (Supplementary Information section 4).

Previous large-scale height GWAS, which were performed predominantly in Europeans, have identified 3,290 independent common height-associated variants[8]. To assess the predictive power of these European-biased variants in the Peruvian population, we generated polygenic risk scores (PRSs) using conditional effect sizes of 2,993 common height-associated variants that were present in our cohort (Methods and Supplementary Information section 5). Greater PRS values were associated with increased height (Pearson's r = 0.22, 95% confidence interval = 0.18–0.25, $P = 2.7 \times 10^{-34}$) (Fig. 1e). The estimated genetic heritability (h_g^2) of height was similar for Peruvian individuals (h_g^2 = 57.6%, s.e.m. = 9.7) and Europeans (h_g^2 = 62.5%)[9]; however, previously identified height-associated variants explained only 6.1% (95% confidence

Table 1 | Replication of rs200342067 association with height

Phenotype	Cohort	*n*	rs	Allele 1	Allele 2	MAF (%)	Effect size	S.e.m.	*z* score	Wald test *P* value
Height (cm)	Discovery	3,134	rs200342067	C	T	4.72	−2.22	0.36	−6.17	6.8×10^{-10}
	Replication	598	rs200342067	C	T	4.52	−1.70	0.82	−2.07	0.04
	Meta-analysis						−2.14	0.33	−6.48	9.2×10^{-11}
Inverse normally transformed height	Discovery	3,134	rs200342067	C	T	4.72	−0.25	0.04	−6.25	4.1×10^{-10}
	Replication	598	rs200342067	C	T	4.52	−0.20	0.10	−2.00	0.05
	PAGE	31,214	rs200342067	C	T	0.37	−0.12	0.07	−1.71	0.09
	GIANT	10,766	rs200342067	C	T	1.15	−0.44	0.22	−2.00	0.05
	Meta-analysis						−0.22	0.03	−6.81	9.8×10^{-12}

We replicated the association between rs200342067 and height in an independently collected cohort (n = 598 individuals). We also tested the association of rs200342067 with inverse normally transformed height in two publicly available datasets of Hispanic/Latino individuals (PAGE and GIANT, n = 31,214 and 10,776 individuals, respectively) and observed a similar direction of effect and effect size in these independent cohorts. P values are from two-sided Wald tests. Numbers are rounded to two decimal places.

interval = 4.6–7.8, $P = 6.7 \times 10^{-45}$) of height phenotypic variance in our cohort compared with 24.6% (95% confidence interval = 22.0–27.2) in the original European cohort.

The lower predictive power of the PRS calculated based on a European GWAS in a non-European population could be the result of differences in factors related to population demography (such as linkage disequilibrium, allele frequency, sex and age composition)[10–12], non-transmitted genetic factors (such as the genetic makeup of the parent[13] and peers[14]), non-genetic factors (such as environmental exposure[15]) or genetic interactions with non-genetic factors[16]. In line with previous reports[11,12], we observed that the European-biased PRS explains a larger proportion of height phenotypic variance in individuals with a high proportion of European ancestry compared with individuals with a low proportion of European ancestry, suggesting that the reduced effect of PRS in Peruvian individuals may—at least in part—be related to genetic differences (Supplementary Information section 5).

Of previously identified common height-associated variants, 99% have effect sizes of less than 0.5 cm per allele[6] (Fig. 2a). By contrast, rs200342067 reduces height by 2.2 cm per allele and explains 0.9% of height phenotypic variance in our cohort (Extended Data Fig. 2b). This effect size is comparable to a few other extremely rare (MAF < 0.5%) height-associated variants that are believed to be under purifying selection[6,8]. In the 1000 Genomes Project[17], rs200342067 is specific to Mexican (MAF = 0.78%) and Peruvian (MAF = 4.12%) populations. However, the genomic region that overlaps with rs200342067 is under a hard selective sweep in some European, south Asian, east Asian and South American populations[18,19] (Supplementary Information section 6). This observation led us to the hypothesis that rs200342067 might have risen in frequency in the Peruvian population as a result of positive selection.

To test this hypothesis, we used integrated Selection of Allele Favoured by Evolution (iSAFE) analysis[20] to search for variants under positive selection in a 1.2-megabase (Mb) region around rs200342067. The top positive selection signal was from rs12441775 (Fig. 2b), an intronic variant in *FBN1* with unknown function. The derived rs12441775*G allele has a much higher frequency in all non-African populations than African populations (derived allele frequency (DAF) = 58% (interquartile range (IQR) = 51–64) in non-African populations versus 4% (IQR = 1–5) in African populations)[17] (Extended Data Fig. 3). This allele shows evidence of positive selection (measured using integrated haplotype score[18,19] (iHS) and extended haplotype homozygosity[21] (EHH) statistics) in European, south Asian and South American populations including the Peruvian population (DAF = 61%, iHS = −2.16) (Fig. 2c, Extended Data Fig. 3 and Supplementary Information section 6) suggesting an out-of-Africa positive selection on rs12441775.

As rs12441775 is located 77 kilobases (kb) upstream of rs200342067, we considered that the increased frequency of rs200342067 in the Peruvian population may be the result of positive selection at rs12441775. Notably, rs12441775*G (derived/major) and rs200342067*C (derived/minor) alleles are out of phase with each other and rarely co-occur on the same extended haplotypes. In our cohort, only 3% (9 out of 297) of the haplotypes that carried the rs200342067*C allele (allele frequency = 4.7%) also carried rs12441775*G (allele frequency = 64.8%) (Fig. 2d and Supplementary Information section 6). Therefore, positive selection at rs12441775 cannot explain the increased frequency of rs200342067*C in Peruvians.

The presence of strong positive selection at haplotypes that carry the rs200342067*T allele prevents the detection of potentially weaker selection signals in haplotypes carrying the rs200342067*C allele using methods such as iHS[18] or pairwise nucleotide diversity (π)[22]. However, if rs200342067*C is under independent positive selection, the length of the haplotype sequence carrying this allele is expected to be longer than the sequence of haplotypes carrying other derived alleles with similar allele frequencies in neutral regions of the genome[23]. Indeed, we observed that haplotypes carrying rs200342067*C are longer than 99.2% of haplotypes with similar alleles in the neutral genomic regions (n = 2,380 variants, n = 3,134 individuals and 6,268 haplotypes) (Fig. 2e and Methods). Excluding the nine haplotypes that carry both rs200342067*C and rs12441775*G alleles does not change this result (Extended Data Fig. 4). Similarly, haplotypes that carry the rs200342067*C allele are longer than 100% of haplotypes simulated under a neutral demographic model that matches the population history of Peru (Methods and Extended Data Fig. 5). Taken together, these results suggest that the rs200342067*C allele is under positive selection independent of rs12441775*G. Almost all other missense variants in *FBN1* are under purifying selection, causing this gene to have a significantly lower burden of missense variants than expected (z score = 5.53, $P = 3.2 \times 10^{-8}$, Exome Aggregation Consortium (ExAC), n = 60,706 individuals)[24].

The selection signal at rs200342067*C is weaker than rs12441775*G. This difference may be due to the difference in the age of the alleles (484 (95% confidence interval = 373–605) versus 2,382 (95% confidence interval = 2,286–2,479) generations old[25] for rs200342067*C and rs12441775*G, respectively). It is also not clear whether the same evolutionary pressures are driving selection at both alleles. We also note that the positive selection signal at rs200342067 is weaker than known examples of recent hard selective sweeps (such as *SLC24A5* or *LCT*)[19]. Whereas alleles under strong positive selection have |iHS| values of >2[19], the iHS value for rs200342067 is −1.5. This value is more extreme than previously reported[19] iHS values of 95.3% variants with a similar DAF and local recombination rate in the Peruvian population (Extended Data Fig. 6a). Similarly, the EHH for rs200342067 is more extreme than the EHH of 97.5% variants with a similar DAF and recombination rates in our cohort (Extended Data Fig. 6b, c).

FBN1 is 266 kb downstream of *SLC24A5* (Fig. 1c), a well-known example of positive selection due to its role in skin pigmentation[26,27].

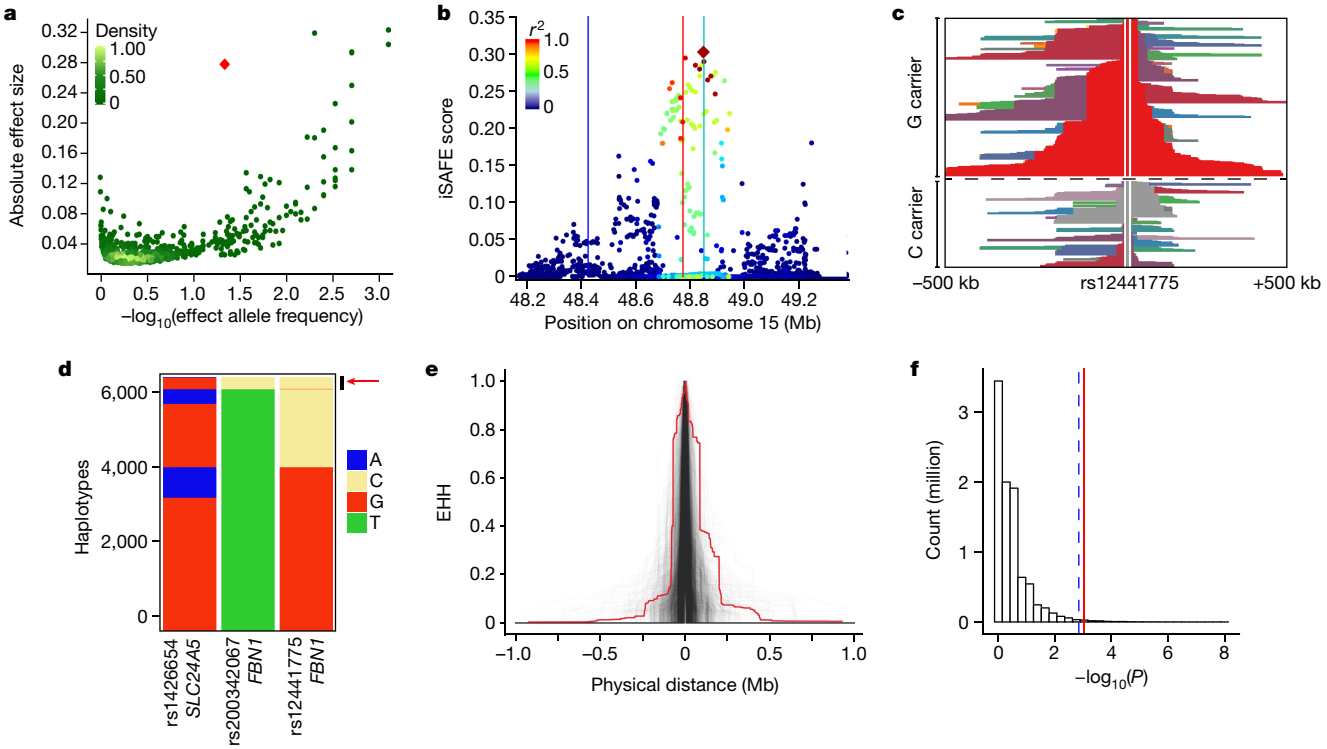


Fig. 2 | rs200342067 is positively selected in the Peruvian population. **a**, Conditional effect sizes and allele frequencies of 3,290 previously identified height-associated variants in the European population ($n \approx 700{,}000$ individuals; green dots) compared with the effect size and allele frequency of rs200342067 (red diamond) from this study ($n = 3{,}134$ individuals, MAF = 4.7%). Effect sizes are shown as the absolute effect size on inverse normally transformed height. **b**, iSAFE plot for a 1.2-Mb region around rs200342067 in our cohort ($n = 3{,}134$ individuals). Dots indicate variants coloured according to their linkage disequilibrium with rs12441775 (red diamond); red, cyan and blue vertical lines show the physical position of rs200342067, rs12441775 and rs1426654, respectively. **c**, Haplotype decay around rs12441775 in our cohort ($n = 3{,}134$ individuals). The position of rs12441775 is marked below the haplotype, rs12441775*G haplotypes are shown above the dashed line (derived/major, $n = 4{,}063$ haplotypes) and rs12441775*C haplotypes are shown below the dashed line (ancestral/minor, $n = 2{,}205$ haplotypes). **d**, Stacked bar plot of the rs200342067, rs12441775 and rs1426654 haplotypes in our cohort ($n = 6{,}268$ haplotypes). Only 3% of the haplotypes that carry the rs200342067*C allele (red arrow) also carry the rs12441775*G allele (allele frequency = 64.8%) and only 4% carry rs1426654*A (allele frequency = 17.9%). The *x* axis shows the indicated SNPs, and the *y* axis shows the number of haplotypes with the derived or alternate allele of rs200342067, rs12441775, and rs1426654.The red arrow and the black line indicate the haplotypes carrying rs200342067*C allele. **e**, EHH plots for haplotypes carrying the rs200342067*C allele (red line, $n = 297$ haplotypes) compared with haplotypes carrying 2,380 variants that overlap the neutral regions of the genome and have a similar DAF to rs200342067*C (4.7 ± 1%; grey lines) in our cohort. Haplotypes carrying the rs200342067*C allele are longer than 99.2% of the haplotypes carrying similar alleles in the neutral genomic regions. **f**, Histogram of Fisher's exact test results comparing the extent of allele frequency differences between coastal ($n = 46$ individuals) and non-coastal ($n = 104$ individuals) regions in Peru. The *x* axis shows the $-\log_{10}$-transformed *P* value of the two-sided Fisher's exact test ($n = 9{,}381{,}550$ variants); the *y* axis shows the variant count in millions; the dashed blue line shows the 99th percentile; the solid red line shows the $-\log_{10}$-transformed *P* value of rs200342067 (0.7th percentile, $P = 0.0005$, two-sided Fisher's exact test).

However, positive selection at rs200342067 is unlikely to be the result of selection at extended haplotypes that contain positively selected alleles in *SLC24A5*, as there is no linkage between variants that overlap *FBN1* and variants that overlap *SLC24A5* ($r^2 < 0.05$). We also investigated the structure of the haplotypes with rs1426654*A, a *SLC24A5* allele associated with light skin pigmentation[26,28] that is known to be under positive selection[29] specifically; we observed that rs200342067*C and rs1426654*A are out of phase with each other and almost never co-occur on the same extended haplotypes. Only 4% (12 out of 297) of haplotypes that carried the rs200342067*C allele (allele frequency = 4.7%) also carried the rs12441775*G allele (allele frequency = 17.9%) (Fig. 2d and Supplementary Information section 6). Moreover, *FBN1* and *SLC24A5* are in different topologically associating domains, suggesting that rs200342067 (or other *FBN1* variants) are unlikely to have been selected owing to long-range regulatory effects on *SLC24A5*.

As adaptation to the local environment can drive considerable allele frequency shifts, we compared the frequency of rs200342067 among 150 individuals who were recruited separately from our cohort through the Peruvian Genome Project[3] (PGP) from three different geographical regions in Peru: the coast ($n = 46$), Amazon ($n = 28$) and Andes ($n = 76$). The rs200342067 variant is more frequent in the individuals from the coast compared to individuals from the Andes or Amazon (MAF = 9.7%, 1.7% and 0%, respectively; coastal versus non-coastal Fisher's exact test $P = 0.0005$) (Fig. 2f and Supplementary Information section 7). Allele frequency differences as extreme as this are observed in less than 0.7% of all variants ($n = 9{,}381{,}550$ variants) (Fig. 2f) and in less than 1.1% of variants that were matched on DAF and local recombination rate to rs200342067 ($n = 2{,}062$ variants) (Extended Data Fig. 6d). We also used Bayenv2[30] to check the deviation of rs200342067 from a neutral population structure model after correction for population stratification. The deviation of rs200342067 from the neutral population structure was more extreme than 91.7% of variants in the same DAF and recombination bin ($n = 2{,}062$ variants) (Methods and Extended Data Fig. 6e). Among coastal populations, the Moches population—who are from the north coast of Peru—had an especially high frequency of rs200342067 ($n = 21$, minor allele count = 4, MAF = 9.5%). Notably, the average height of the Moches people is far below the average height in Peru (158 cm and 147 cm for Moches men and women[31] versus 164 cm and 152 cm for Peruvian men and women measured in the same year[1]), suggesting

that rs200342067 may have been selected as a result of adaptation to factors associated with the coastal environment.

To ensure that the association between rs200342067 and height in the Peruvian population is not driven by population structure and stratification between individuals from different geographical regions, we performed a principal component analysis in the PGP cohort[3] (n = 150) using a set of common variants (MAF ≥ 5%) and used SNP loadings from the principal component analysis in the PGP cohort to infer population principal components in our cohort (n = 3,134) (Methods and Supplementary Information section 7). Correction for these principal components did not change the effect size or the strength of the observed association between rs200342067 and height (n = 3,134, MAF = 4.72%, effect size = −2.30 cm, s.e.m. = 0.36, $P = 3.0 \times 10^{-10}$), confirming that the observed association between rs200342067 and height is not a result of confounding population structure.

The rs200342067 variant changes the conserved T (major/ancestral) allele to a C (minor/derived) allele in exon 31 of *FBN1* (g.48773926T>C) (Extended Data Fig. 7). This change substitutes a large, negatively charged glutamic acid for a glycine, the smallest amino acid in fibrillin 1, encoded by *FBN1* (FBN1(E1297G)). Fibrillin 1 is an extracellular matrix glycoprotein that serves as the structural backbone of force-bearing microfibrils in elastic and non-elastic tissues[32] and is also involved in tissue development, homeostasis and repair by interacting with transforming growth factor (TGFβ) and other growth factors[32]. Although the clinical importance of FBN1(E1297G) is not known, other fibrillin 1 mutations cause nine dominantly inherited Mendelian diseases[33]. Most of these diseases include skeletal anomalies and changes in skin elasticity[33]. To investigate the possible clinical consequences of FBN1(E1297G), we performed dermatological and rheumatological clinical exams on 11 individuals from our cohort: 2 homozygous (C/C) individuals, 2 heterozygous (C/T) individuals and 7 matched controls with the reference (T/T) genotype (Methods). Although the musculoskeletal examination revealed no differences between individuals, one individual with the C/C genotype had a notably thicker skin as assessed in a total body skin examination and appeared older than the stated age. The other individual with the C/C genotype had no clinically abnormal cutaneous findings and none of the C/T or T/T individuals had an abnormal skin exam (Supplementary Information section 8).

We also obtained skin biopsies from two individuals with the rs200342067 C/C genotype (alternate homozygous) and two with rs200342067 T/T genotypes (reference homozygous, Methods). We matched each individual with the C/C genotype with individuals with the T/T genotype based on age, sex and ancestry proportions. Immunohistochemical staining showed that the individuals with C/C genotype had shorter microfibrillar projections from the dermal–epidermal junction into the superficial (papillary) dermis as well as less fibrillin 1 deposition in the deeper dermis (Methods and Extended Data Fig. 8). Scanning electron microscopy analyses showed that individuals with the C/C genotype have less densely packed microfibrils with irregular edges and with microfibrils embedded in less dense collagen bundles, confirming the abnormal appearance of fibrillin 1 observed in immunohistochemical analysis of the skin biopsies (Fig. 3 and Extended Data Fig. 9). Together, these experiments suggest that rs200342067 alters the amount and architecture of microfibrillar deposits in the skin.

Whereas all of the reported mutations in FBN1 causing short stature phenotypes occur in the TGFβ-binding domains, mutations in the calcium-binding epidermal growth factor (cbEGF) domains of FBN1 predominantly lead to Marfan or Marfan-like syndromes[34]. Notably, missense mutations in the cbEGF domains 11–18 of fibrillin 1 encoded by exons 24–32 of *FBN1* (also known as the neonatal region) (Extended Data Fig. 7) have previously been associated with severe neonatal forms of Marfan syndrome, mortality within the first two years of postnatal life and poor disease prognosis in adults[33,35]. To our knowledge, the E1297G mutation is the first mutation in the neonatal region of fibrillin 1 that leads to short stature, in contrast to the tall stature common in individuals with Marfan syndrome.

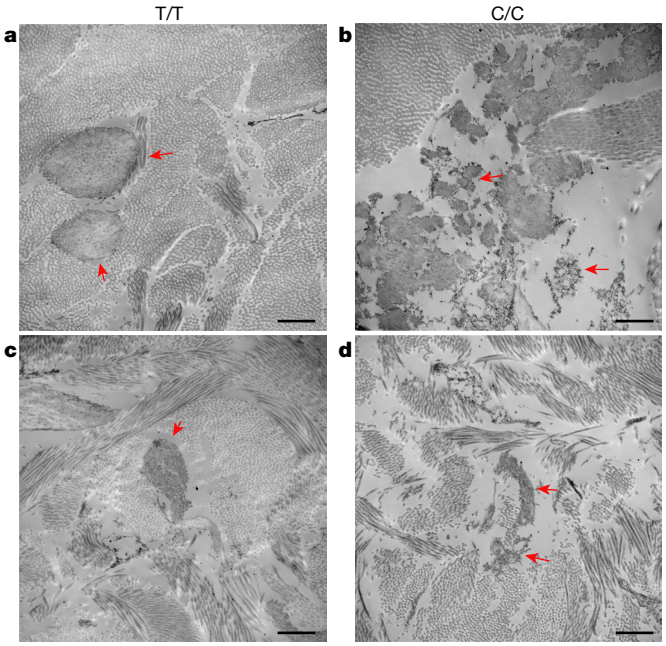


Fig. 3 | Electron microscopy of fibrillin 1 in the skin. a, **c**, Electron microscopy images of the dermal–epidermal junction of samples obtained from two individuals with the rs200342067 T/T genotype (**a**, 60-year-old woman; **c**, 30-year-old woman). **b**, **d**, Electron microscopy images of the dermal–epidermal junction of samples obtained from two individuals with the rs200342067 C/C genotype who were matched for age, sex and ancestry proportions with individuals in **a** and **c**, respectively (**b**, 64-year-old woman; **d**, 35-year-old woman). Individuals with the C/C genotype have short, fragmented and less densely packed microfibrils with irregular edges and their microfibrils are embedded in less dense collagen bundles compared with the individuals with the T/T genotype. Red arrowheads show edges of microfibril bundles. Magnification, 11,000×. Scale bars, 1 µm.

The FBN1(E1297G) mutation is located in the cbEGF domain 17 of fibrillin 1, between a conserved cysteine (FBN1(C1296)) that is involved in forming a disulfide bond with FBN1(C1284) and a conserved asparagine (FBN1(N1298)) that is involved in calcium binding[36] (Extended Data Fig. 7 and Supplementary Information section 8) and may have a role in calcium binding[37]. Calcium binding at the cbEGF domains of fibrillin 1 stabilizes the protein by making the microfibrils more rigid and protecting them from degradation by proteases[38]. The short, fragmented and less packed phenotype seen in the skin of individuals with the rs200342067 C/C genotype compared with individuals with the T/T genotype (Fig. 3 and Extended Data Figs. 8, 9) might reflect the higher susceptibility of mutated fibrillin 1 to proteolysis compared with the wild-type protein. The few previous studies that have reported amino acid changes at positions similar to FBN1(E1297G) in other fibrillin 1 cbEGF-like domains have associated this change with Marfan syndrome[34], highlighting the importance of domain context for studying the molecular effect of *FBN1* mutations[39,40]. Understanding the cellular mechanisms that connect FBN1(E1297G) to microfibril structures and height requires further functional studies (Supplementary Information section 8).

Common variants with large effect sizes on height might increase in frequency in a population as a result of positive selection. A study of height in Sardinian islanders found an intronic variant in *KCNQ1*, which encodes a voltage-gated potassium channel, that reduces height by an average of 1.8 cm (rs150199504, MAF = 7.7%, MAF in central European population = 0.67%); the authors of the study suggested that this variant is positively selected in Sardinians as a result of adaptation to the

island environment[41]. A study of signatures of genetic adaptation in Greenland Inuits found an intronic variant in *FADS3*, which encodes a protein involved in fatty acid metabolism, that reduces height by 1.9 cm, possibly due to the influence of fatty acid composition on the regulation of growth hormones (rs7115739, DAF = 62.7–81.9%, DAF in the central European population = 2.9–3.6%); the authors of the study suggested that this variant is positively selected in Greenland Inuits as a result of adaptation to as fat-rich diet[42]. Similarly, it is plausible that short stature in Peruvian individuals is the result of adaptation to the factors associated with the coastal environment. It is also possible that other *FBN1*-related traits such as changes in the performance of the cardiovascular system have offered an evolutionary advantage in this population. Understanding the exact adaptive processes that could have caused the selection of rs200342067 in the Peruvian population is a challenging task and requires further investigation.

In addition to its implications in medical and population genetics, this study highlights the importance of large-scale genetic studies in underrepresented and founder populations. Our findings show that genetic studies in different populations can uncover previously undescribed trait-associated variants with large effects in functionally relevant genes. Similar studies in diverse populations are required to capture the extent of human genetic diversity and to expand the benefits of genetic research to all human populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2302-0.

1. NCD Risk Factor Collaboration (NCD-RisC). A century of trends in adult human height. *eLife* **5**, e13410 (2016).
2. Homburger, J. R. et al. Genomic insights into the ancestry and demographic history of South America. *PLoS Genet*. **11**, e1005602 (2015).
3. Harris, D. N. et al. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl Acad. Sci. USA* **115**, E6526–E6535 (2018).
4. Ruiz-Linares, A. et al. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet*. **10**, e1004572 (2014).
5. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
6. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
7. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
8. Yengo, L. et al. 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).
9. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet*. **46**, 1173–1186 (2014).
10. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet*. **97**, 576–592 (2015).
11. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet*. **100**, 635–649 (2017).
12. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet*. **51**, 584–591 (2019).
13. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
14. Domingue, B. W. et al. The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health. *Proc. Natl Acad. Sci. USA* **115**, 702–707 (2018).
15. Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Gene–environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS Genet*. **13**, e1006977 (2017).
16. Pelova, N. Considerations on the so-called myelolipoma of the adrenals. *Nauchni Tr. Vissh. Med. Inst. Sofiia* **48**, 31–35 (1969).
17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
18. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol*. **4**, e72 (2006).
19. Johnson, K. E. & Voight, B. F. Patterns of shared signatures of recent positive selection across human populations. *Nat. Ecol. Evol*. **2**, 713–720 (2018).
20. Akbari, A. et al. Identifying the favored mutation in a positive selective sweep. *Nat. Methods* **15**, 279–282 (2018).
21. Sabeti, P. C. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
22. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273 (1979).
23. Arbiza, L., Zhong, E. & Keinan, A. NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics* **13**, 301 (2012).
24. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
25. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol*. **18**, e3000586 (2020).
26. Lamason, R. L. et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782–1786 (2005).
27. Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: a review of recent human adaptation. *Science* **354**, 54–59 (2016).
28. Adhikari, K. et al. A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nat. Commun*. **10**, 358 (2019).
29. Sturm, R. A. & Duffy, D. L. Human pigmentation genes under environmental selection. *Genome Biol*. **13**, 248 (2012).
30. Günther, T. & Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics* **195**, 205–220 (2013).
31. Lasker, G. W. Differences in anthropometric measurements within and between three communities in Peru. *Hum. Biol*. **34**, 63–70 (1962).
32. Sengle, G. & Sakai, L. Y. The fibrillin microfibril scaffold: a niche for growth factors and mechanosensation? *Matrix Biol*. **47**, 3–12 (2015).
33. Schrenk, S., Cenzi, C., Bertalot, T., Conconi, M. T. & Di Liddo, R. Structural and functional failure of fibrillin-1 in human diseases (review). *Int. J. Mol. Med*. **41**, 1213–1223 (2018).
34. Collod-Béroud, G. et al. Update of the UMD-FBN1 mutation database and creation of an FBN1 polymorphism database. *Hum. Mutat*. **22**, 199–208 (2003).
35. Tiecke, F. et al. Classic, atypically severe and neonatal Marfan syndrome: twelve mutations and genotype-phenotype correlations in FBN1 exons 24–40. *Eur. J. Hum. Genet*. **9**, 13–21 (2001).
36. Smallridge, R. S. et al. Solution structure and dynamics of a calcium binding epidermal growth factor-like domain pair from the neonatal region of human fibrillin-1. *J. Biol. Chem*. **278**, 12199–12206 (2003).
37. Booms, P., Tiecke, F., Rosenberg, T., Hagemeier, C. & Robinson, P. N. Differential effect of *FBN1* mutations on in vitro proteolysis of recombinant fibrillin-1 fragments. *Hum. Genet*. **107**, 216–224 (2000).
38. Jensen, S. A., Robertson, I. B. & Handford, P. A. Dissecting the fibrillin microfibril: structural insights into organization and function. *Structure* **20**, 215–225 (2012).
39. Jensen, S. A., Corbett, A. R., Knott, V., Redfield, C. & Handford, P. A. Ca^{2+}-dependent interface formation in fibrillin-1. *J. Biol. Chem*. **280**, 14076–14084 (2005).
40. McGettrick, A. J., Knott, V., Willis, A. & Handford, P. A. Molecular effects of calcium binding mutations in Marfan syndrome depend on domain context. *Hum. Mol. Genet*. **9**, 1987–1994 (2000).
41. Zoledziewska, M. et al. Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet*. **47**, 1352–1356 (2015).
42. Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment unless otherwise noted.

Study participants

Discovery cohort. The individuals in the discovery cohort (the LIMAA cohort) are a subset of 4,002 individuals who were recruited in Lima, Peru, to study the genetics of tuberculosis in the Peruvian population[43]. The catchment area included 12 of the 43 districts of metropolitan Lima, Peru and 3.3 million inhabitants. This catchment area reflects a mix of urban and peri-urban areas and informal settlements[44]. Participants were recruited in any of the 106 public health centres in the catchment area. Informed written consent was obtained from all participants. The study protocol was approved by the Harvard School of Public Health's Institutional Review Board (IRB) and the Research Ethics Committee of the National Institute of Health of Peru.

Replication cohort. We recruited 889 individuals from the same catchment area as the discovery cohort. Similar to the discovery cohort, we followed the guidelines of the Harvard School of Public Health's IRB and the Research Ethics Committee of the National Institute of Health of Peru guidelines and obtained informed consent from all participants.

Phenotype

In both the discovery and replication cohorts, height in centimetres was measured by trained healthcare staff upon recruitment of study participants. We excluded individuals who were under 19 years of age, individuals without height measurements and individuals with a measured height that was ±3.5 s.d. away from the population average from the cohort. In addition to height, the sex, age and tuberculosis status of the individuals were also collected. We also collected household-level socioeconomic factors on housing quality, water supply and sanitation[45] and summarized these factors using principal component analysis (PCA)[45] to calculated household-level composite socioeconomic scores. The continuous socioeconomic scores were then categorized into tertiles corresponding to low, middle and upper socioeconomic groups[45].

Genotyping and quality control

Discovery cohort. We collected genotyping data for 4,002 individuals from 2,272 households in Lima, Peru, using a customized Affymetrix Axiom array. The array details, as well as the genotyping quality control, phasing and imputation have been described in detail in a previous publication[43]; in brief, we designed an approximately 720,000 marker array based on exome-sequencing data from 116 Peruvian individuals to optimize for population-specific rare and coding variants. Out of 4,002 recruited individuals, 22 individuals were excluded during quality control because there was more than 5% of the genotype data missing, an excess of heterozygous genotypes (±3.5 s.d.), a duplication with identity-by-state of >0.9 or individuals with tuberculosis had an age at diagnosis of over 40 years of age[43]. We further excluded 846 individuals from the analysis: individuals younger than 19 years of age, individuals without height measurements and individuals with a measured height that was ±3.5 s.d. away from the population average. The final cohort for the current study included 3,134 individuals from 1,947 households. We used GRCh37 as the reference genome for all our genetic analyses.

Replication cohort. We collected genotyping data for 889 individuals from 273 households in the same population and catchment area as our discovery cohort. We collected blood using the Whatman protein saver cards (Dried Blood Spot cards) (Sigma-Aldrich, WHA10534320). We extracted genomic DNA from the collected blood and genotyped all samples using the Illumina Multi-Ethnic Genotyping Array (MEGA). rs200342067 is included on the MEGA array and was directly genotyped in all individuals. We used PLINK (version 1.90b3w) to estimate the level of genotyping missingness and the proportion of heterozygous variants per individual. Height data were not available for 127 individuals. Moreover, 164 individuals were excluded as they were under 19 years of age. The final cohort included 598 individuals from 242 households.

Genetic relatedness matrix and kinship estimation

To avoid spurious association results, it is important to account for both recent genetic relatedness, such as family structure (kinship), and more distant genetic relatedness, such as population structure. To this end, we used GEMMA[46] (version 0.96), with default options, to generate a genetic relatedness matrix (GRM) after removing rare variants (MAF ≤ 1%), regions with known long-range linkage disequilibrium[47] and variants in high linkage disequilibrium ($r^2 > 0.2$ in a window of 50 kb and a sliding window of 5 kb). We used PLINK (version 1.90b3w) for pruning the genotypes. We generated a separate GRM following the same steps for the Peruvian individuals that were included in the replication cohort.

Many kinship estimation methods perform under the assumption of sampling from a single population with no underlying ancestral diversity. Kinship estimates are inflated when this assumption is violated[48]. In the presence of population structure and admixture, methods that replace population allele frequencies with ancestry-specific allele frequencies are preferred[48]. We used PC-Relate[49] implemented in the GENESIS R package (version 2.6.1) to estimate the kinship coefficients between individuals. This method uses ancestry representative principal components to correct for population structure before calculating the kinship coefficients. For this analysis, we removed rare variants (MAF < 1%), regions with known long-range linkage disequilibrium[47] and variants in high linkage disequilibrium ($r^2 > 0.2$ in a window of 50 kb and a sliding window of 5 kb). Individuals were considered unrelated if their estimated kinship coefficients were ≤0.0625, corresponding to second-degree genetic relatedness or further. In total, 476 individuals had kinship coefficients of >0.0625.

Next, we inferred pairwise identity-by-descent (IBD) segments between the individuals in our cohort using Refined IBD[5]. Refined IBD uses a haplotype dictionary to find exact short matches between phased haplotypes from different individuals and then expand these matches to identify long, nearly identical IBD segments between these individuals[50]. Refined IBD then evaluates candidate IBD segments with a probabilistic approach to assess the strength of evidence for IBD and reports the segment above a threshold as IBD segments. To calculate IBD segments using Refined IBD, we first used PLINK (version 1.90b3w)[51] on quality-controlled genotypes (n = 677,675 markers) to generate one VCF file per chromosome. We then used the Refined IBD function[5] implemented in Beagle (version 4.1) to phase these genotypes and to calculate IBD segments in our cohort (n = 3,134). We used Refined IBD with nthreads = 8, overlap = 3000, default options for other parameters and genetic maps from HapMap II (build GRCh37/hg19) (provided on the Beagle website: https://faculty.washington.edu/browning/beagle/b4_1.html). Finally, we calculate the proportion of IBD by dividing the length of IBD segments by the length of diploid GRCh37 autosomal chromosomes excluding GRCh37 gap regions such as the centromere (also called the accessible genome, 5.7×10^9 bp). We used the Pearson correlation coefficient in R (version 3.4) to compare the GENESIS and Refined IBD results.

The PGP

In some analyses, we used whole-genome sequencing data from the PGP[3] cohort. This previously described cohort[3] includes 150 Peruvian individuals who were recruited separately from our cohort from three different geographical regions in Peru: coast (n = 46), Amazon

($n = 28$) and Andes ($n = 76$). Individuals were assigned to different Native American groups from the three geographical regions, as described previously[3], as follows: "Native American population cohort participants were recruited from the Matzes, Uros, Afroperuvians, Chopccas, Moches, Q'eros, Nahuas, and Matsiguenka populations. We applied three criteria to optimize individuals to best represent the Native American populations: (i) the place of birth of the participant and that of his or her parents and grandparents, (ii) their last names (only those corresponding to the region), and (iii) age (eldest to mitigate effects of the last generation). Participants of the mestizo population cohorts were recruited from the cities Iquitos, Puno, Cusco, Trujillo, and Lima and were randomly selected. The Afroperuvians were sampled as a Native American population; however, for all analyses, we treated them as a mestizo group due to their expected admixture between multiple ancestries."

Difference in allele frequency of rs200342067 between the coastal and non-coastal regions

We compared the extent of the difference in allele frequency between individuals from the coastal regions in Peru ($n = 46$) and individuals that were not from the coastal regions in Peru ($n = 104$) using a two-sided Fisher's exact test ($n = 9{,}381{,}550$ variants). Next, to ensure we adequately controlled for population structure, we used Bayenv2 (version 2.0)[30] to calculate a covariance matrix between the coastal and non-coastal populations using 63,758 SNPs with MAF > 10% in the PGP cohort[3] using the default options. We then used Bayenv2 (version 2.0)[30] to calculate standardized allele frequencies and XTX statistics, a population differentiation statistic that is designed to detect variant level deviations from the neutral patterns of population structure while correcting for population structure[30], for rs200342067 as well as the 2,062 randomly selected SNPs that were matched in MAF and local recombination rate to rs200342067 (described in detail in 'Selecting variants in the same DAF and recombination bin as rs200342067') and using the default options in Bayenv2.

PCA

To obtain principal components within the LIMAA cohort, we merged our genotype data with data from the continental populations of phase 3 of the 1000 Genomes Project[17] and genotype data from Siberian and Native American populations from a previously published study[52] by matching chromosomes, positions and reference and alternate alleles. After merging the datasets, variants with an overall MAF < 1% were excluded. We used GCTA[53] (version 1.26.0) to perform the PCA. We used PLINK (version 1.90b3w)[51] for linkage disequilibrium pruning, merging and quality control. The merged dataset included 34,936 variants.

To ensure we adequately controlled for population structure and differences in ancestry that might exist within the different geographical regions of Peru, we also calculated the principal components of the LIMAA cohort ($n = 3{,}134$) projected into the principal component space of the PGP cohort[3] ($n = 150$). To do this, we selected 247,940 common (MAF ≥ 5%) variants that were shared between the PGP and LIMAA cohorts. We then calculated the principal components in the PGP cohort using the EIGENSOFT (version 6.1.4)[54] smartpca function. Finally, we used the SNP loadings from the smartpca analysis to project the individuals from the LIMAA cohort to the principal component space of the PGP cohort using the SNPWEIGHTS (version 2.1) software[55]. We used ANOVA (R version 3.4) to test the association of the first ten principal components of the PGP cohort with coast–non-coast status. We tested the association between the principal components of the LIMAA cohort with the Native American ancestry proportion, height, or rs200342067 minor allele count using the linear mixed model implemented in lme4qtl[56] (R version 3.4), with age and sex as fixed effects and a genetic relatedness matrix to account for genetic relatedness (calculated using GEMMA[46] version 0.96) as random effect.

Global ancestry inference

We used ADMIXTURE[57] (version 1.3) at $K = 4$ clusters for global ancestry inference. The choice of four ancestral populations for ADMIXTURE analysis was based on the demographic history of Peru and previous studies of Peruvian population structure[2–4]. We used the merged dataset described above as input for the ADMIXTURE analysis. We used PLINK (version 1.90b3w)[51] to exclude variants with a genotyping missingness rate of >5% and to perform linkage disequilibrium pruning by removing the markers with $r^2 > 0.1$ with any other marker within a sliding window of 50 markers per window and an offset of 10 markers.

Correlation between global ancestry proportions and height

We used the R package lme4qtl[56], a linear mixed model framework, to measure the correlation between global ancestry proportions and height. We included the following covariates in the base model: age, sex, African and Asian ancestry proportions, as well as a GRM to account for population structure and genetic relatedness generated using PC-Relate[49], which is implemented in the GENESIS R package (version 2.6.1). We repeated this analysis after adding a random effect to account for the individual's household as a proxy for unmeasured environmental factors. Finally, to ensure we adequately controlled for environmental factors, we randomly assigned height to individuals within each household 10,000 times and recalculated the Native American ancestry effect size using the base model to generate an empirical null distribution. We compared the null distribution with the observed Native American ancestry effect size from the original data to generate an empirical permutation *P* value.

Common variant association analysis

In the discovery cohort, imputed SNPs with an imputation quality score $r^2 < 0.4$, Hardy–Weinberg equilibrium $P < 10^{-5}$ or a missing rate per SNP > 5% were excluded. After filtering, 7,756,401 SNPs were left for further association analyses. We used GEMMA[46] (version 0.96) to perform the single variant GWAS, with age, sex and a GRM generated using GEMMA[46] (version 0.96) as covariates. We repeated the association for chromosome 15 by adding one or more of the following covariates: 10 principal components, 20 principal components, socioeconomic status, African global ancestry proportion, Asian global ancestry proportion and European global ancestry proportion. To ensure we adequately controlled for population structure, we also repeated the association test between height (cm) and rs200342067 with age, sex, 10 principal components derived from projecting the LIMAA cohort into the principal component space of the PGP cohort[3] (see 'PCA' for details), and a GRM generated using GEMMA[46] (version 0.96). To ensure that our choice of GRM did not affect the association between rs200342067 and height, we repeated the association analysis using two new GRMs. First, a GRM calculated using PC-Relate[49], with age, sex and 10 principal components as fixed covariates. Second, a GRM calculated using Refined IBD[5] with age, sex and 10 principal components as fixed covariates. All association *P* values are reported as two-sided Wald test *P* values.

To ensure that local (per chromosome) relatedness patterns such as autozygosity segments did not bias the relatedness, we generated 23 GRMs, leaving one chromosome out in each GRM using PC-Relate[49] and repeated the association for all post-quality-control imputed variants on each chromosome using a GRM generated without that chromosome. Age, sex and 10 principal components were included as additional covariates in this analysis.

For the replication analysis in the Peruvian population, we used the minor allele count information of rs200342067, directly genotyped on the Illumina MEGA array, from 598 Peruvian individuals (see 'Replication cohort' for details). Similar to the discovery cohort, we tested the association of rs200342067 with height (cm) using a linear mixed model framework implemented in GEMMA[46] (version 0.96) with age, sex and a GRM (calculated using GEMMA) as covariates.

Additional replication cohorts

PAGE. The PAGE study is a meta-analysis of multiple existing major population-based cohorts[7]. The cohorts included in PAGE height study include the BioMe biobank (BioMe), the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), The Multiethnic Cohort (MEC) and the Women's Health Initiative (WHI)[7]. Height in centimetres was measured by trained clinic staff in the SOL and WHI studies at the time of enrolment. In the MEC and BioMe studies, height was self-reported by questionnaire. Individuals with height measurements that were ±6 s.d. from the mean (based on sex and race), individuals who were younger than 18 years of age and women who were pregnant were excluded from the height GWAS analysis in PAGE. To get comparable phenotypes between different cohorts, PAGE uses inverse normal transformation of sex-specific height residuals adjusted for age as the dependent variable in a linear mixed model that includes self-identified ancestry, study, study centre and 10 principal components (measured from unrelated individuals) as fixed effects and a genetic relatedness matrix (using GENESIS[58]) as a random effect[7]. Detailed information about the phenotype and statistical analysis of the PAGE cohort has been published previously[7]. We used the summary statistics from 31,214 Hispanic and Latino individuals from the PAGE study in our replication analysis.

GIANT. The exome array study of the GIANT consortium is a meta-analysis of 147 studies comprising 458,927 adult individuals[6]. Height in centimetres was corrected for age and the genomic principal components, as well as any additional study-specific covariates (for example, recruiting centre) in a linear regression separately by sex. For family-based studies, sex was included as a covariate in the model. In addition, residuals for case–control studies were calculated separately. Similar to the PAGE cohort, GIANT uses the inverse normal transformation of calculated residuals as the dependent variable in an ancestry-specific linear mixed model that corrects for cryptic relatedness using a kinship matrix in each cohort separately followed by a meta-analysis of the results. Detailed information about the phenotype and statistical analysis of the GIANT cohort has been published previously[6]. We used the summary statistics from 10,776 Hispanic individuals from the GIANT study in our replication analysis.

Meta-analysis

We used the meta R package[59] (version 4.9-3) to perform an inverse variance-based meta-analysis using summary statistics from the height GWAS in the LIMAA discovery cohort and the Peruvian replication cohort in which the measured phenotype was the height (cm). To perform the meta-analysis using the GIANT[6] and PAGE[7] cohorts, we repeated our association analysis in both the discovery and the replication cohorts as described above with age, sex and a GRM generated using GEMMA[46] (version 0.96) as covariates and inverse normally transformed height as the dependent variable. We used summary statistics from these analyses as well as summary statistics form the GIANT[6] and PAGE[7] cohorts to perform an inverse variance-based meta-analysis using the meta R package[59].

Heritability analysis

We used GREML analysis in GCTA[60] (version 1.26.0) to calculate the amount of variance in height explained by all common variants (MAF > 1%). We included 423,108 variants from 2,667 unrelated individuals in this analysis with age, sex and the first 10 principal components as covariates in the analysis. To calculate height heritability for each sex, we repeated the heritability analysis in men and women separately.

PRS analysis

Out of 3,290 previously reported independent genome-wide significant variants[8], 2,993 were present in our cohort. We constructed PRSs for each individual using height-increasing effect sizes of these 2,993 previously published height-associated variants[8] as follows:

$$\mathrm{PRS}_j = \sum_{i-1}^{m} n_{ij}\beta_i,$$

in which β_i is the reported conditional effect size for variant i in the European population, n_{ij} is the allele count of variant i in individual j in our Peruvian cohort, and m is the total number of variants used in the construction of the PRS. We calculated the amount of variance explained (r^2) using lm function in R (version 3.4) with height residuals adjusted for age, sex and a GRM generated using GEMMA[46] (version 0.96) to account for relatedness and cryptic population structure as the dependent variable and PRS as the explanatory variable. Out of 3,290 previously reported independent genome-wide significant variants[8], 2,388 reached genome-wide significance in an unconditional analysis. We repeated the PRS calculation using the unconditional effect sizes of 2,195 of these variants that were also present in our cohort. We used the cocor package in R (version 1.1-3) to test the significance of the difference between the amount of variance explained using different PRS.

For the sex-specific analysis, we first calculated height residuals in each sex separately after adjustment for age and a GRM generated using GEMMA[46]. We then calculated the r^2 using lm(), with height residuals as the dependent variable and PRS as the explanatory variable for each sex separately. For the analysis of individuals with high or low European ancestry proportions, we separated the cohort into individuals with high European ancestry proportions (top quartile) and low European ancestry proportions (first, second and third quartiles) and calculated height residuals after adjustment for age, sex and a GRM generated using PC-Relate[49] to account for relatedness but not population structure. We then calculated the r^2 using the lm function, with height residuals as the dependent variable and PRS as the explanatory variable in each group separately.

Gene-based association analysis

We used SKAT[61] (version 1.3.2.1) for gene-based association testing of rare (MAF < 1%) variants. Null distributions were generated using SKAT_NULL_emmaX, which incorporates kinship structure in the calculation of SKAT parameters and residuals. Age and sex were included as covariates. The statistical significance threshold was set at $P < 2.5 \times 10^{-6}$, which is the Bonferroni correction threshold for 20,000 protein-coding genes. For common variants (MAF ≥ 1%), we used fastBAT analysis in GCTA[62] to perform gene-based association testing using GWAS summary statistics.

Positive selection analyses

iSAFE analyses. We used SHAPEIT2 (version v2.r837) to phase the imputed genotypes for chromosome 15 for all the individuals in our cohort (n = 3,134). We then used iSAFE[20] (version v1.0.4) software, available at https://github.com/alek0991/iSAFE with the following options: MaxRank = 300, MaxFreq = 0.85, and enabling the IgnoreGaps flag to detect variants under positive selection in a 1.2-Mb window around rs200342067.

EHH analyses. We used selscan[63] (version 1.2.0a) to calculate EHH[21] in our cohort (n = 3,134) or in the simulated data. The analysis was restricted to variants with MAF > 1%. For all variants, including rs200342067, we calculated EHH in a 2-Mb window around the variant. For EHH, we interpolated the genetic position based on the average recombination rate of the chromosome to get a comparable measure of haplotype length between positively selected regions, regions under neutral selection and simulated data. To ensure that the EHH calculation at rs200342067*C is not biased due to selection at the nearby selected locus rs12441775*G, we repeated the EHH calculation for at rs200342067*C after removing the nine haplotypes that had both rs200342067*C and rs12441775*G

(updated MAF for rs200342067*C = 4.6%). For integrated EHH values, we calculated the area under the EHH curve. The global map of rs12441775*G was generated using the Geography of Genetic Variants (GGV) browser[64] (http://www.popgen.uchicago.edu/ggv).

Comparing EHH of rs200342067 with similar variants under neutral selection. We selected 2,380 variants that overlapped the previously published putative neutral regions of the genome[23] and had a similar DAF to the rs2003420678*C allele in our cohort (4.7 ± 1%). We calculated EHH for these variants using selscan[63] (version 1.2.0a) and compared the EHH decay plots as well as the integrated EHH values for rs2003420678*C and these variants. In a second step, we removed the nine haplotypes that carried rs12441775*G from our cohort and repeated the EHH analysis using 2,309 variants that overlapped the previously published putative neutral regions of the genome[23] and had a similar DAF to the updated frequency of the rs2003420678*C allele (AF = 4.6 ± 1%).

Selecting variants in the same DAF and recombination bin as rs200342067. We restricted the analysis to biallelic variants, the ancestral allele was assigned using the 'ancestral allele' information provided in the 1000 Genomes Project[17]. We calculated the derived allele frequency of each common variant (MAF > 1%) in the Peruvian individuals from the 1000 Genomes Project[17] (n = 85). We also interpolated the genetic position of each common variant (MAF > 1%) using the 1000 Genomes Project[17] phase 3 genetic maps. The recombination rate was calculated as follows: genetic position (cM)/physical position (Mb). Variants on each chromosome were divided into 100 DAF bins and 20 recombination bins. The DAF for rs200342067 in the Peruvian individuals from the 1000 Genomes Project[17] is 4.1% (DAF bin 4) and its recombination rate is 1.4 (recombination bin 5). For comparison with rs200342067, we selected 2,062 variants that were in the same DAF and recombination bin as rs200342067 and that were at least 1 Mb apart (that is, independent).

iHS analyses. iHS[18] values for the Peruvian individuals and other populations from the 1000 Genomes Project[17] were obtained from a previously published study[19] (http://coruscant.itmat.upenn.edu/data/JohnsonEA_iHSscores.tar.gz).

Testing the extent of rs200342067 MAF difference between the coastal and non-coastal regions

Fisher's exact test. We used minor allele counts for rs200342067 as well as 2,062 independent variants matched in DAF and local recombination rates to rs200342067 (see 'Selecting variants in the same DAF and recombination bin as rs200342067') in populations from the coastal regions (n = 46) or non-coastal regions (that is, the Andes and Amazon, n = 104) of the Peruvian Genome Project cohort[3], to perform Fisher's exact tests in R (version 3.4).

XTX analysis. We used Bayenv 2.0 (version 2.0)[30] to calculate a covariance matrix between the coastal and non-coastal populations using 63,758 SNPs with MAF > 10% in the PGP cohort[3]. We then used Bayenv 2.0 (version 2.0)[30] to calculate standardized allele frequencies and XTX statistics for rs200342067 as well as all the 2,062 SNPs randomly selected SNPs that were matched in DAF and local recombination rate to rs200342067.

Simulation of haplotypes under a neutral demographic model

We used msprime (version 0.7.3)[65], a coalescent model with recombination, to simulate 2,000 Peruvian individuals with the recombination map from HapMap Project[66] 1,000 times. We adapted and constructed the population structures from the previously proposed out-of-Africa model[67] with parameters previously inferred from the 1000 Genomes Project[17]. To mirror the Peruvian migration history, we created a three-way admixture event around 500 years (25 generations) ago. We used the 1000 Genomes Project[17] phase 3 genetic maps for chromosome 15 to interpolate chromosomal recombination rate in our simulation. We set the admixture to have 80% Native American, 16% European and 4% African ancestry, inferred from the LIMMA cohort. We compared the integrated EHH values for 1,000 simulated variants that had similar DAF to rs200342067 (DAF = 4.7 ± 1%) and overlapped the same region on the simulated chromosome (physical position 48,773,926 ± 20 kb) with the integrated EHH value of rs200342067 in our cohort (n = 6,628 haplotypes). We repeated the analysis for two putative neutral regions on chromosome 15 in the simulated data (n = 2,000 haplotypes) and compared the integrated EHH values with the integrated EHH values for two variants, rs17580697 (DAF = 4.6%) and rs305008 (DAF = 4.6%), which overlapped these neutral regions of chromosome 15[23] in our cohort (n = 6,628 haplotypes).

Mutation age

We used the pre-calculated mutation age estimates based on the 1000 Genomes Project populations[17] from the human genome dating server (https://human.genome.dating/)[25].

Three-dimensional structure of the *FBN1* cbEGF domain 17

The three-dimensional structure was obtained based on homology with fibrillin 1 cbEGF domains 12 and 13 (Protein Data Bank (PDB) 1LMJ)[36].

Clinical examination

Clinical examination was approved by the local IRB. Individuals with the T/T genotype (controls) were matched with cases (individuals with the C/C and C/T genotypes) for sex, age ±5 years, Native American ancestry proportion ±0.05 and European ancestry proportion ±0.05. A board-certified rheumatologist performed a musculoskeletal exam and history, including a detailed musculoskeletal history with review of systems, past medical history, medication history, social history and family history; vital signs; range of motion on knees, wrists, elbows, index fingers, middle fingers and hips; joint exam of hands for bony changes, synovitis or other abnormalities; joint exam of knees, feet and spine for instability, bony changes, inflammation or other abnormalities. A board-certified dermatologist performed a standardized total body skin exam. This includes an examination of the skin of the face, eyelids, ears, scalp, neck, chest, axillae, abdomen, back, buttocks, genitalia, upper extremities, lower extremities, hands, feet, digits, nails, lips, mouth, mucosae and lymph nodes. We also obtained skin biopsies for two individuals with the C/C genotype and two age-, sex- and ancestry-matched individuals with the T/T genotype. Biopsies were obtained 5 cm lateral to the umbilicus (in clinically normal skin) to assess histological differences associated with genotype.

Histology

Following Harvard Medical School IRB approval, samples were obtained by skin punch biopsy as routinely performed by a Massachusetts General Hospital (MGH) dermatologist and placed inside specimen jars containing 10% neutral buffered formalin. The specimens were shipped by courier to the Massachusetts General Hospital at ambient temperature, placed into tissue cassettes, processed routinely and paraffin-embedded tissue blocks were prepared at the Histopathology Research Core of the Massachusetts General Hospital. One glass slide stained with haematoxylin and eosin was prepared for each block and additional unstained 5-µm thick sections cut from the tissue blocks were placed onto Fisher superfrost slides (protein-coated).

Immunohistochemical analysis

Anti-fibrillin-1 antibody staining was performed using the citrate buffer antigen retrieval technique. Appropriate negative-control sections (primary antibody omitted to monitor for background staining) and positive-control sections (human placental tissue known to express the

antigen, as recommended by the manufacturer) were evaluated. Tissue sections were manually stained with rabbit polyclonal anti-fibrillin-1 antibody (FBN1, dilution 1:250, HPA021057, MilliporeSigma) and counterstained with haematoxylin following deparaffinization of 5-µm cut sections. Antigen expression in dermal fibroblasts was assessed by a board-certified pathologist for each specimen in a blinded fashion.

Electron microscopy on formalin-fixed paraffin-embedded tissues

Areas of interest were identified on slides stained with haematoxylin and eosin and matched to the corresponding paraffin blocks. Under a dissecting microscope, these areas were cut out using a sharp razor blade and placed into glass vials containing 100% xylene. The vials were left overnight at room temperature and the xylene was changed the following morning. The vials were then left gently rotating for an additional 3 h before rehydrating for 1 h each in a series of ethanol (100%, 95%, 70%, 50% and 25%) solutions. Tissues were then rinsed in sodium cacodylate buffer and fixed for 1.5 h with our routine glutaraldehyde fixative (2.5% GTA, 2.0% PFA, 0.025% calcium chloride, in a 0.1M sodium cacodylate buffer pH 7.4). Tissues were further processed in a Leica Lynx automatic tissue processor. In brief, tissues were post-fixed with osmium tetroxide, dehydrated in a series of ethanol solutions, en bloc stained during the 70% ethanol dehydration step for 1 h, infiltrated with propylene oxide epoxy mixtures, embedded in pure epoxy and polymerized overnight at 60 °C. Thick sections were cut and stained with toluidine blue and examined with a light microscope. Thin sections were cut from representative areas, stained with lead citrate and examined with an FEI Morgagni transmission electron microscope. Images were captured with an AMT (Advanced Microscopy Techniques) 2K digital CCD camera.

FBN1 and *SLC24A5* Hi-C data

To investigate whether rs200342067 or the other four variants that were linked to rs200342067 in our cohort can act as an enhancer for *SLC24A5*, we investigated the H3K27ac marks from the ENCODE dataset to search for active enhancer that overlap these variants (data were obtained from the ENCODE portal) as well as Hi-C data in published cell types[68–70] for evidence of a physical interaction between these variants and *SLC24A5* (3D Genome Browser, http://promoter.bx.psu.edu/hi-c/view.php).

Reporting summary

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

Data availability

Genotyping data are available through dbGAP, under accession number phs002025.v1.p1.

Code availability

No custom code was used to draw the central conclusions of this work. All the software and packages used in this work are included and referenced in the manuscript.

43. Luo, Y. et al. Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. *Nat. Commun.* **10**, 3765 (2019).
44. Zelner, J. L. et al. Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. *J. Infect. Dis.* **213**, 287–294 (2016).
45. Odone, A. et al. Acquired and transmitted multidrug resistant tuberculosis: the role of social determinants. *PLoS ONE* **11**, e0146642 (2016).
46. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
47. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. Am. *J. Hum. Genet.* **83**, 132–135 (2008).
48. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
49. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
50. Gusev, A. et al. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
51. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
52. Reich, D. et al. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
53. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
54. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
55. Chen, C.-Y. et al. Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).
56. Ziyatdinov, A. et al. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* **19**, 68 (2018).
57. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
58. Schick, U. M. et al. Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. *Am. J. Hum. Genet.* **98**, 229–242 (2016).
59. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. *Evid. Based Ment. Health* **22**, 153–160 (2019).
60. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
61. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
62. Bakshi, A. et al. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci. Rep.* **6**, 32894 (2016).
63. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).
64. Marcus, J. H. & Novembre, J. Visualizing the geography of genetic variants. *Bioinformatics* **33**, 594–595 (2017).
65. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent simulation and genealogical analysis for large sample sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016).
66. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
67. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
68. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
69. Lin, D. et al. Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat. Genet.* **50**, 754–763 (2018).
70. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).

Acknowledgements We thank D. B. Moody for discussions, T. Horn for his feedback on optimizing skin immunohistochemistry and J. N. Katz for advising us on a structured clinical assessment of the musculoskeletal system. The study was supported by the National Institutes of Health (NIH) TB Research Unit Network, grants U19-AI111224-01 and U01-HG009088. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. S.A. was supported by the Swiss National Science Foundation (SNSF) postdoctoral mobility fellowships P2ELP3_172101 and P400PB_183823.

Author contributions S.R. and M.B.M. designed the study. S.A. analysed and interpreted the data. S.A. and S.R. drafted the manuscript. Y.L., G.M.B., E.E.K., J.N.H., E.B., K.S., H.G., T.D.O., A.A., D.N.H. and X.L. performed statistical analysis. M.B.M., L.L., R.C., J.M.C., C.C., R.Y., J.T.G., J.J., J.M.C. and C.F. recruited patients and obtained samples for this study. S.R., E.E.F., H.C.D., R.M.N. and M.S. conducted clinical assessment. All authors discussed the results and commented on the manuscript.

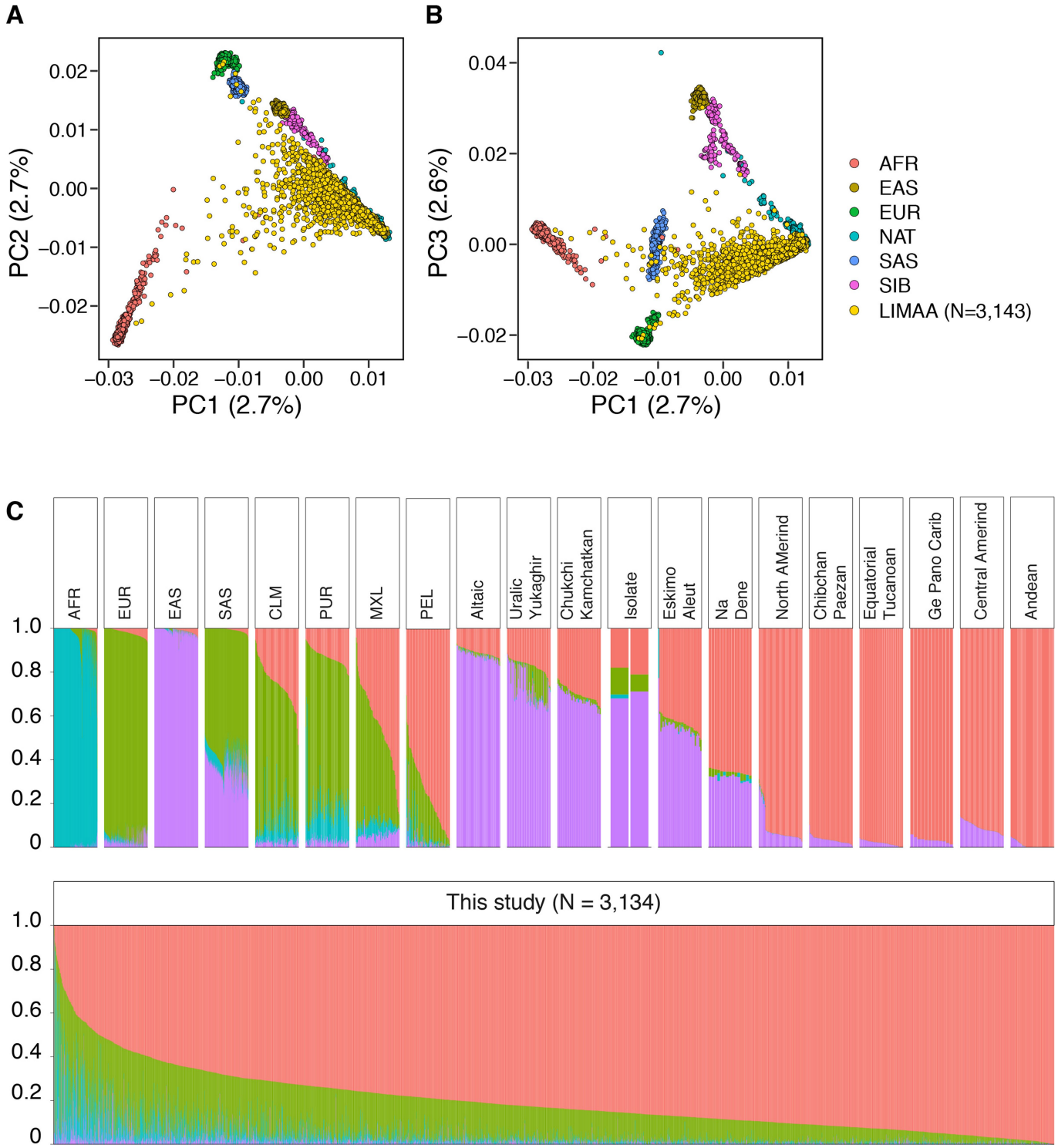
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2302-0.

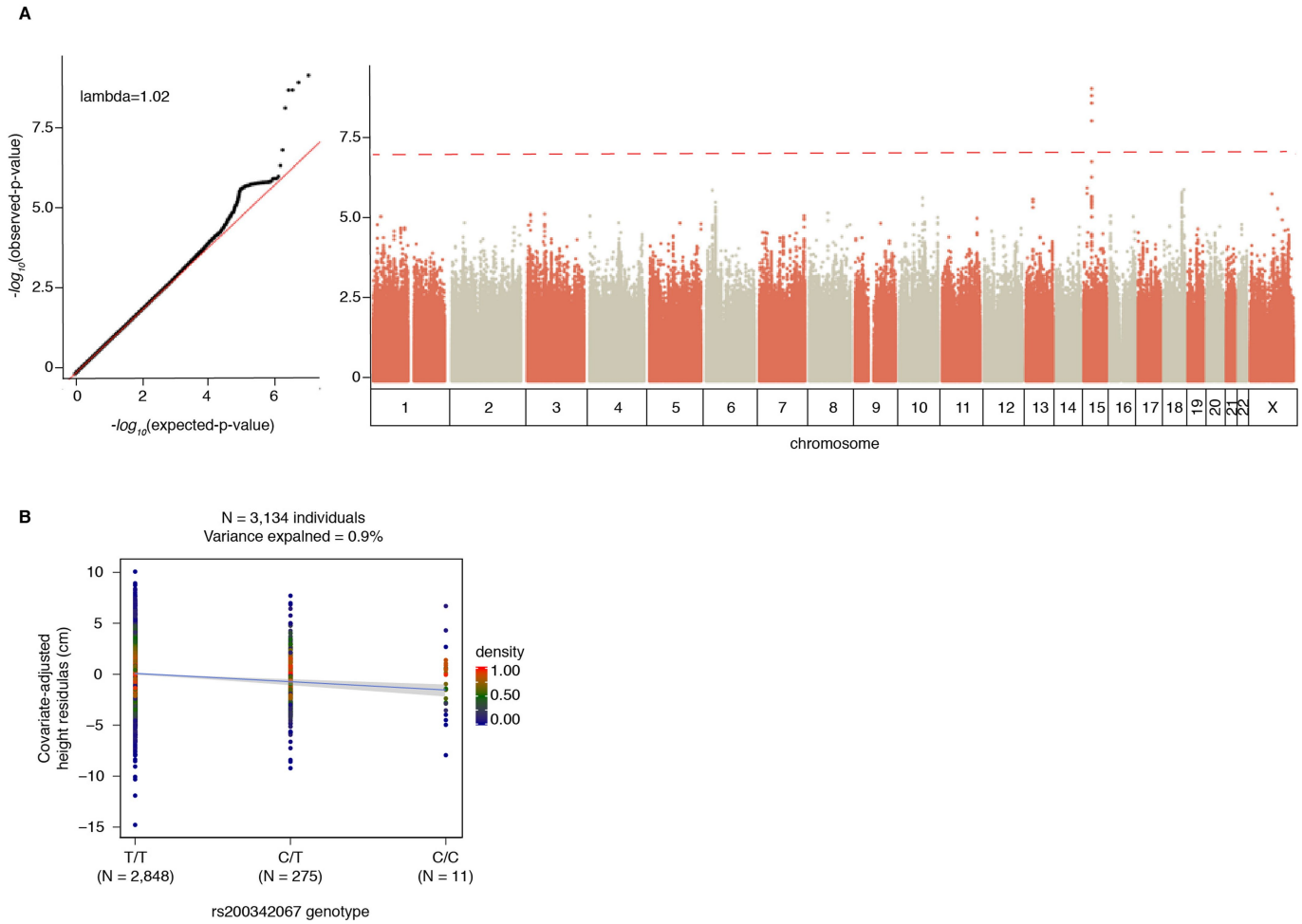
Correspondence and requests for materials should be addressed to S.R.

Peer review information *Nature* thanks Guillaume Lettre, Ben Voight and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

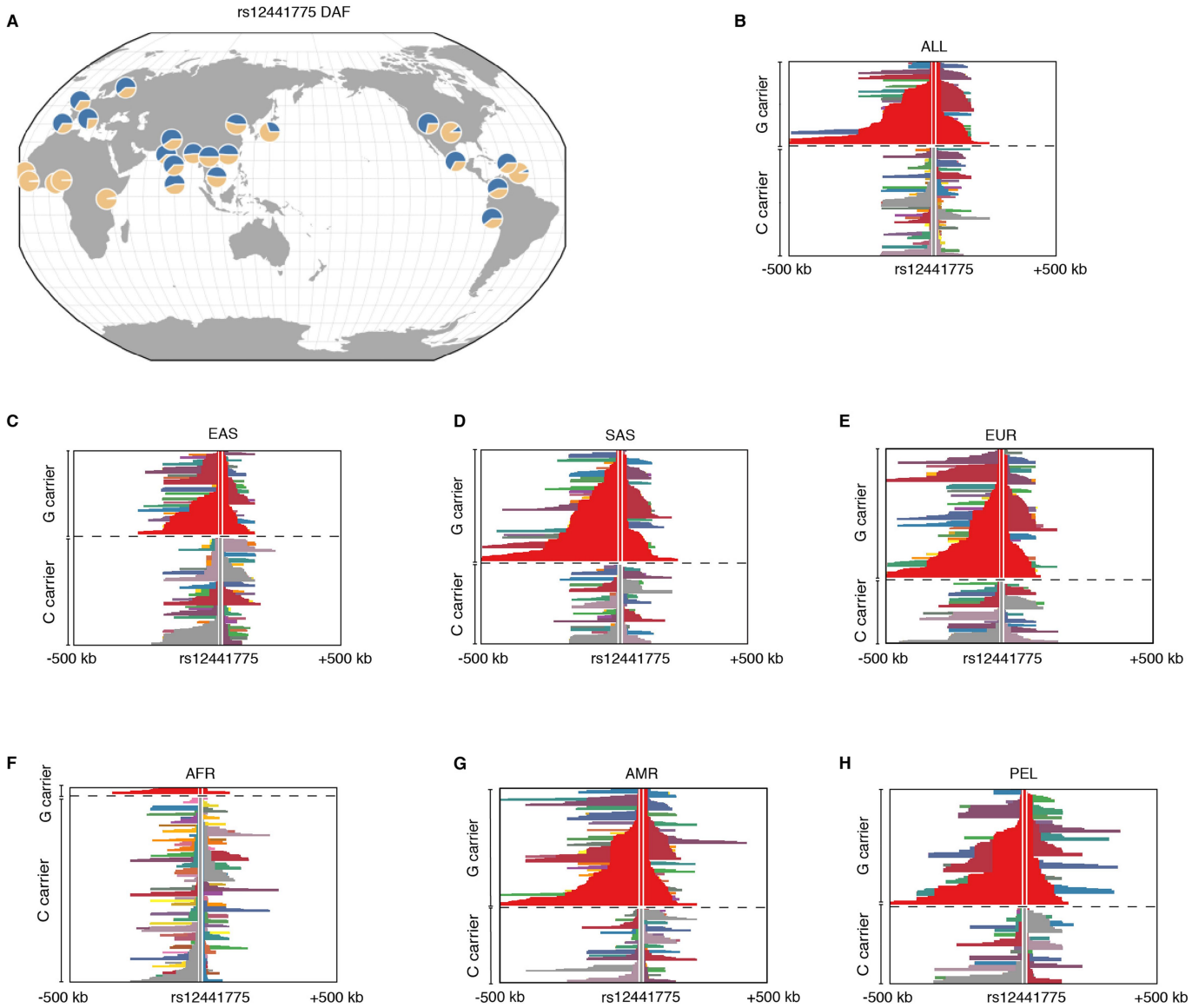


Extended Data Fig. 1 | See next page for caption.

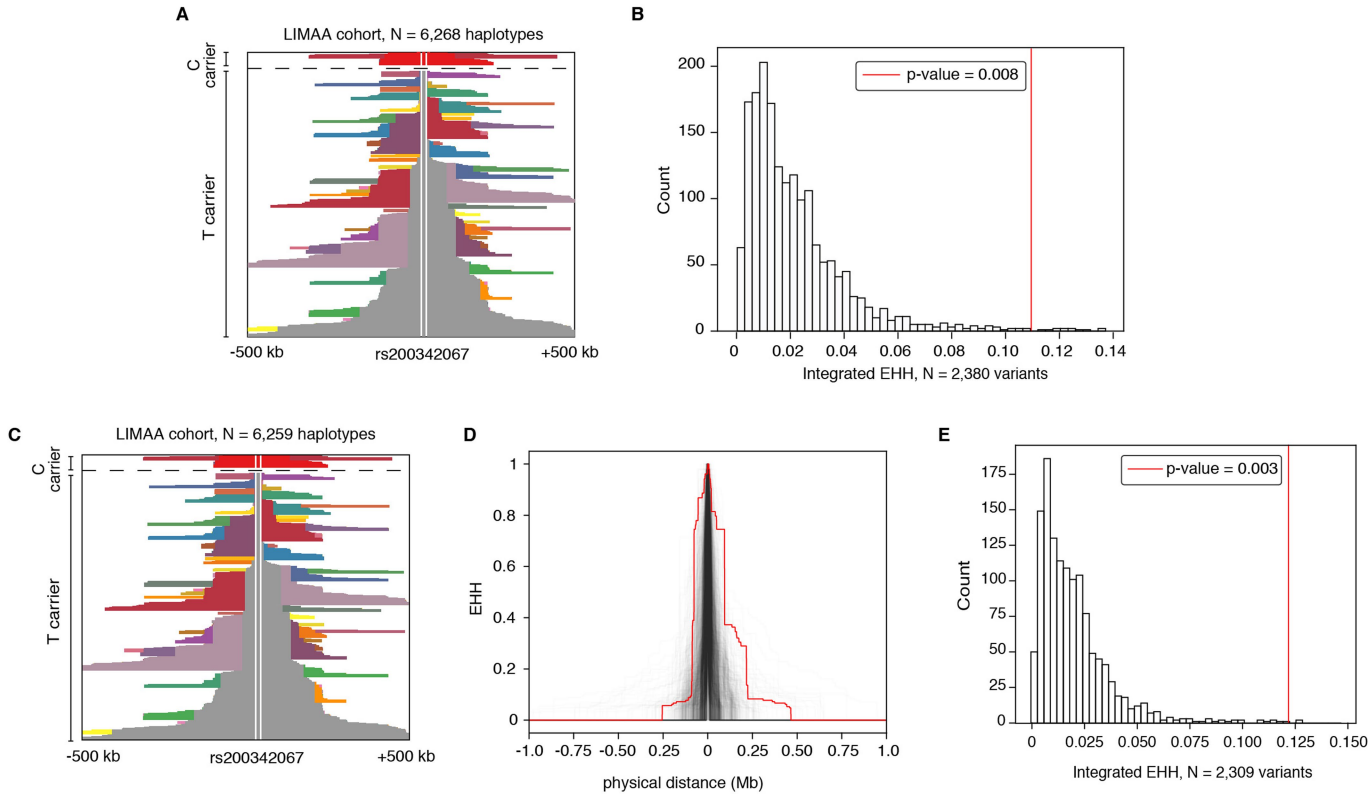
Extended Data Fig. 1 | Peruvian population structure. a, **b**, PCA of genotyping data from Peruvian individuals included in this study (n = 3,134 individuals) merged with the data from continental populations from phase 3 of the 1000 Genomes Project (n = 3,469 individuals) as well as the data from Siberian and Native American populations from the previously published study[52] (n = 738 individuals), which were used as a reference panel (number of variants, 34,936). Dots, individuals; colour, populations (AFR, African; AMR, South American; EAS, east Asian; SAS, south Asian; EUR, European; SIB, Siberian; NAT, Native American). **c**, Global ancestry analysis using ADMIXTURE (K = 4). We observed varying levels of European, African and Asian admixture in our cohort (n = 3,134 individuals) with a median proportion of Native American, European, African and Asian ancestry per individual of 0.83 (IQR = 0.72–0.91), 0.14 (0.08–0.21), 0.01 (0.003–0.03) and 0.003 (10^{-5}–0.01), respectively. Vertical lines, individuals; colours, genomic proportion of a given ancestry in the genome of each individual. ADMIXTURE analysis (K = 4) is done using all populations in phase 3 of the 1000 Genomes Project as well as the Siberian and Native American populations from the previously published study[52], which were used as a reference. African (AFR) ancestry includes Yoruba in Ibadan, Nigeria, Luhya in Webuye, Kenya, Gambian in Western Divisions in the Gambia, Mende in Sierra Leone, Esan in Nigeria, Americans of African Ancestry in southwest United States. European (EUR) ancestry includes central European, Utah residents (CEPH) with northern and western European ancestry (USA), Toscani in Italy, Finnish in Finland, British in England and Scotland, Iberian population in Spain. East Asian (EAS) ancestry includes Han Chinese in Beijing, China, Japanese in Tokyo, Japan, Southern Han Chinese, Chinese Dai in Xishuangbanna, China, Kinh in Ho Chi Minh City, Vietnam. South Asian (SAS) ancestry includes Gujarati Indian from Houston, Texas (USA), Punjabi from Lahore, Pakistan, Bengali from Bangladesh, Sri Lankan Tamil from the United Kingdom, Indian Telugu from the United Kingdom. Puerto Ricans (PUR) from Puerto Rico. Colombians (CLM) from Medellin, Colombia. Mexicans (MXL) from Los Angeles, California (USA). Peruvian individuals (PEL) from Lima, Peru. Altic, Altaic language family, which includes Yakut, Buryat, Evenki, Tuvinians, Altaian, Mongolian, Dolgan. North Amerind, northern Amerindian language family, which includes Maya, Mixe, Kaqchikel, Algonquin, Ojibwa and Cree. Central Amerind, central Amerindian language family, which includes Pima, Chorotega, Tepehuano, Zapotec, Mixtec and Yaqui. Andean, Andean language family, which includes Quechua, Aymara, Inga, Chilote, Diaguita, Chono, Hulliche and Yaghan. A full list of all populations in all language groups has been published previously[52].



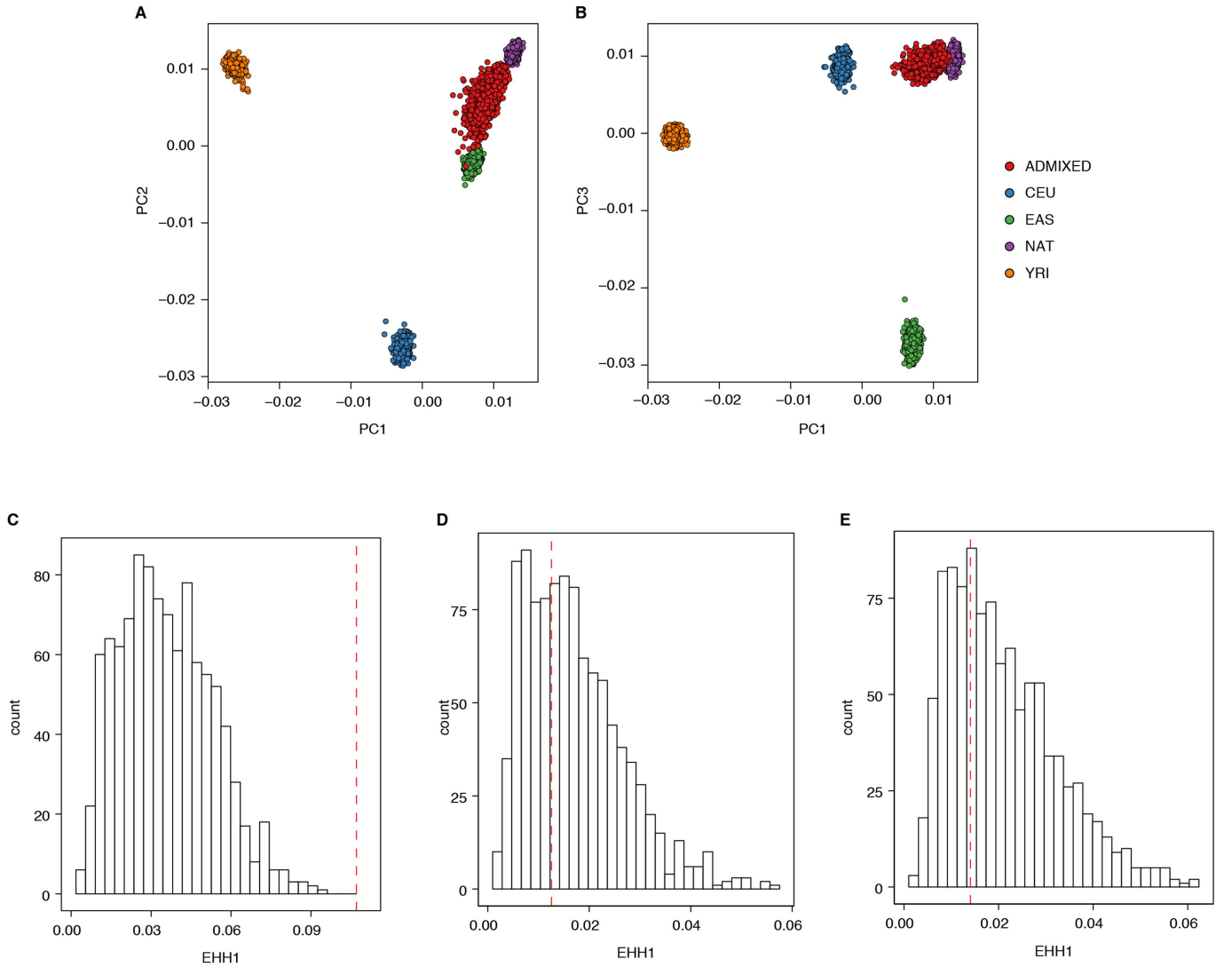
Extended Data Fig. 2 | Association of rs200342067 and height. **a**, Single-variant association analysis ($n = 3{,}134$ individuals and 7,756,401 variants). Dotted red line, genome-wide significance threshold of 5×10^{-8}. Five SNPs that overlap the coding sequence of *FBN1* passed the genome-wide significance threshold. We did not observe any inflation in test statistics ($\lambda = 1.02$). Association *P* values are from two-sided Wald tests. **b**, rs200342067 in heterozygous individuals reduces height by 2.2 cm (4.4 cm in homozygous individuals, including 11 individuals with the C/C genotype, 275 the C/T genotype and 2,848 the T/T genotype) and could explain 0.9% of the phenotypic variance in height in our cohort ($n = 3{,}143$ individuals). The *x* axis shows the rs200342067 genotype; the *y* axis shows the height residuals after adjustments for age, sex and a GRM as random effect.



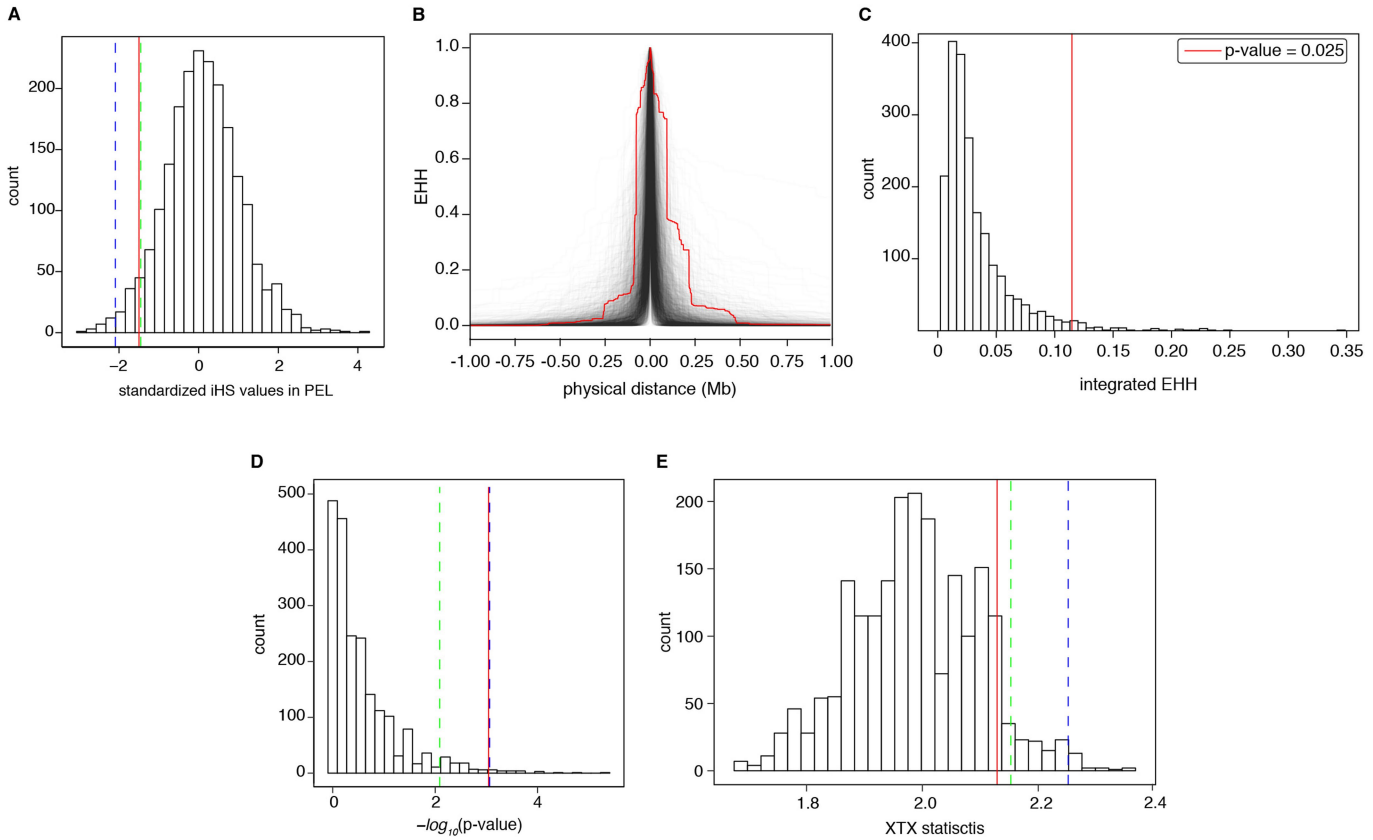
Extended Data Fig. 3 | rs12441775 DAF (rs12441775*G) and extended haplotype structure in the 1000 Genomes Project. a, The derived allele, rs12441775*G, has a high frequency in all non-African populations in the 1000 Genomes Project (average DAF in non-Africans = 58% (IQR = 51–64) and in Africans = 4% (IQR = 1–5)). The map is generated using the GGV browser[64] (http://www.popgen.uchicago.edu/ggv). **b**–**h**, Haplotypes that carry the rs12441775*G (major/derived) allele are longer than haplotypes that carry the rs12441775*C (minor/ancestral) allele in non-African populations. Horizontal lines, haplotypes; the position of rs12441775 is marked below the haplotype. At any given position, adjacent haplotypes with the same colour carry identical genotypes between the core SNP (rs12441775) and that site, dashed line separates the haplotypes that carry the derived (above the line) and ancestral (below the line) alleles.



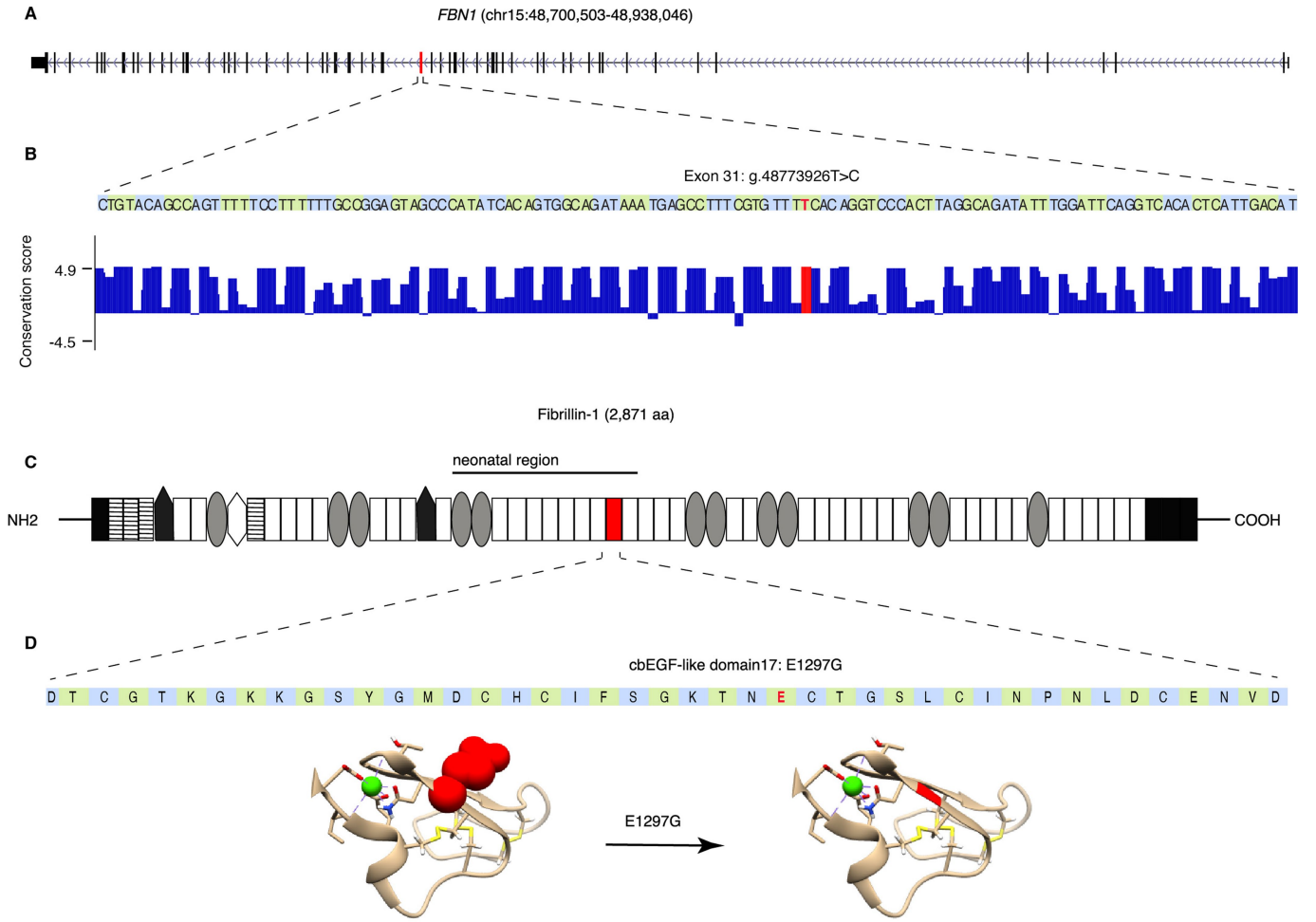
Extended Data Fig. 4 | Haplotypes that carry the rs200342067 allele are longer than what is expected under neutral selection. a, Haplotype decay around rs200342067 in our cohort (n = 3,134 individuals and 6,268 haplotypes). The position of rs200342067 is marked below the haplotypes. Haplotypes above the dashed line carry rs200342067*C allele (derived/minor, n = 297 haplotypes) and haplotypes below the dashed line carry the rs200342067*T allele (ancestral/major, n = 5,971 haplotypes). **b**, Integrated EHH of haplotypes carrying the rs200342067*C allele (n = 297 haplotypes) compared with the integrated EHH of haplotypes carrying 2,380 variants with similar DAF (4.7 ± 1%) that overlap the neutral regions of the genome in our cohort (n = 3,134 individuals). Haplotypes that carry the rs200342067*C allele are taller than 99.2% of the haplotypes carrying similar variants in neutral regions of the genome. Vertical red line, integrated EHH of haplotypes carrying the rs200342067*C allele (integrated EHH = 0.115). **c**, The same as **a**, but excluding the nine haplotypes that carry both rs200342067*C and rs12441775*G alleles. **d**, EHH decay curves for haplotypes carrying the rs200342067*C allele excluding the nine haplotypes that carry both rs200342067*C and rs12441775*G alleles (n = 288 haplotypes) compared with haplotypes carrying 2,309 variants that have a similar DAF to the updated frequency of rs200342067*C (4.6 ± 1%) and that overlap the neutral regions of the genome in our cohort (n = 3,134 individuals). Haplotypes with the rs200342067*C allele are longer than 99.7% of the haplotypes carrying similar variants in the neutral genomic regions. **e**, Integrated EHH for data shown in **d**. Vertical red line, integrated EHH for haplotypes carrying the rs200342067*C but not the rs12441775*G allele (integrated EHH = 0.124).



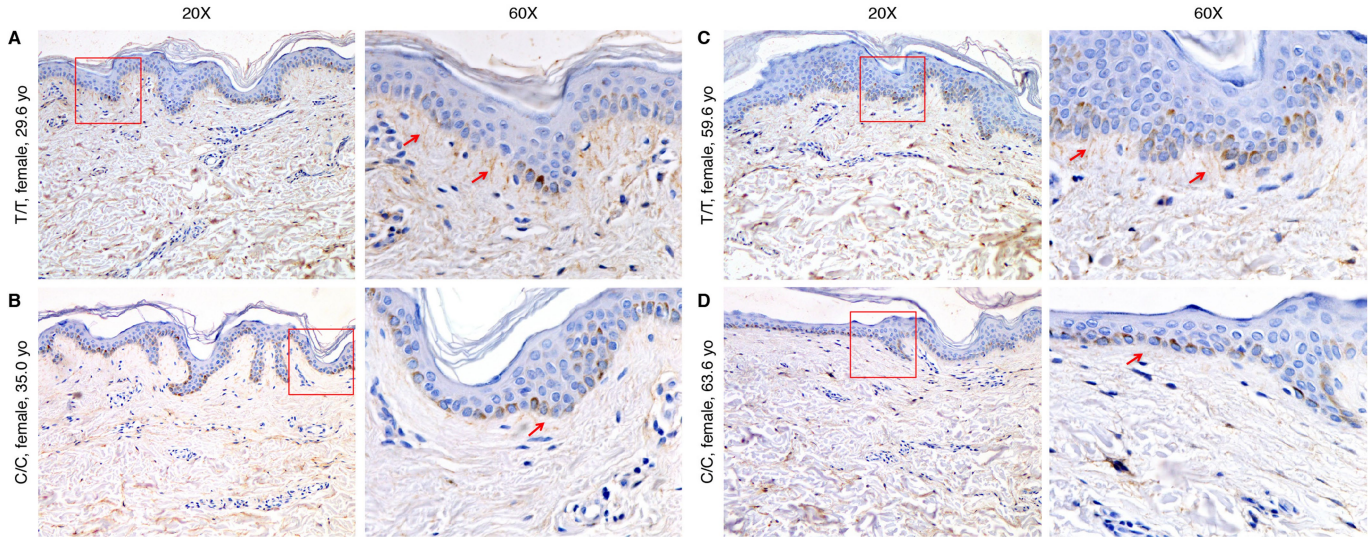
Extended Data Fig. 5 | Simulation of haplotypes under the neutral demographic model. a, PCA plot of principal component (PC)2 versus PC1 for simulated individuals (n = 1,000 simulated individuals and 2,000 simulated haplotypes). Individuals were simulated using a demographic model matching the population history of Peru and under neutral selection. Red dots, simulated individuals; other dots, reference populations from the 1000 Genomes Project. **b**, PCA plot of PC3 versus PC1 as described for **a**. **c**, We compared the integrated EHH of rs200342067*C with the integrated EHH of 1,000 variants that had a similar DAF to rs200342067 (DAF = 4.7 ± 1%) and that overlapped the same genomic region as rs200342067 on a simulated chromosome 15 (physical position, 48,773,926 ± 20 kb). The integrated EHH of rs200342067 is more extreme than the integrated EHH observed for any of the variants in the simulated data. The x axis shows the integrated EHH; the distribution is the integrated EHH of variants in simulated haplotypes (n = 2,000 haplotypes); the vertical red line shows the integrated EHH value of rs200342067 in our cohort (n = 6,628 haplotypes, integrated EHH = 0.115). **d**, **e**, Similar to **c** for two different neutral regions on chromosome 15. Vertical red lines, integrated EHH of rs17580697 (**d**, integrated EHH = 0.012, 76th percentile) and rs305008 (**e**; integrated EHH = 0.010, 74th percentile) in our cohort (n = 6,628 haplotypes).



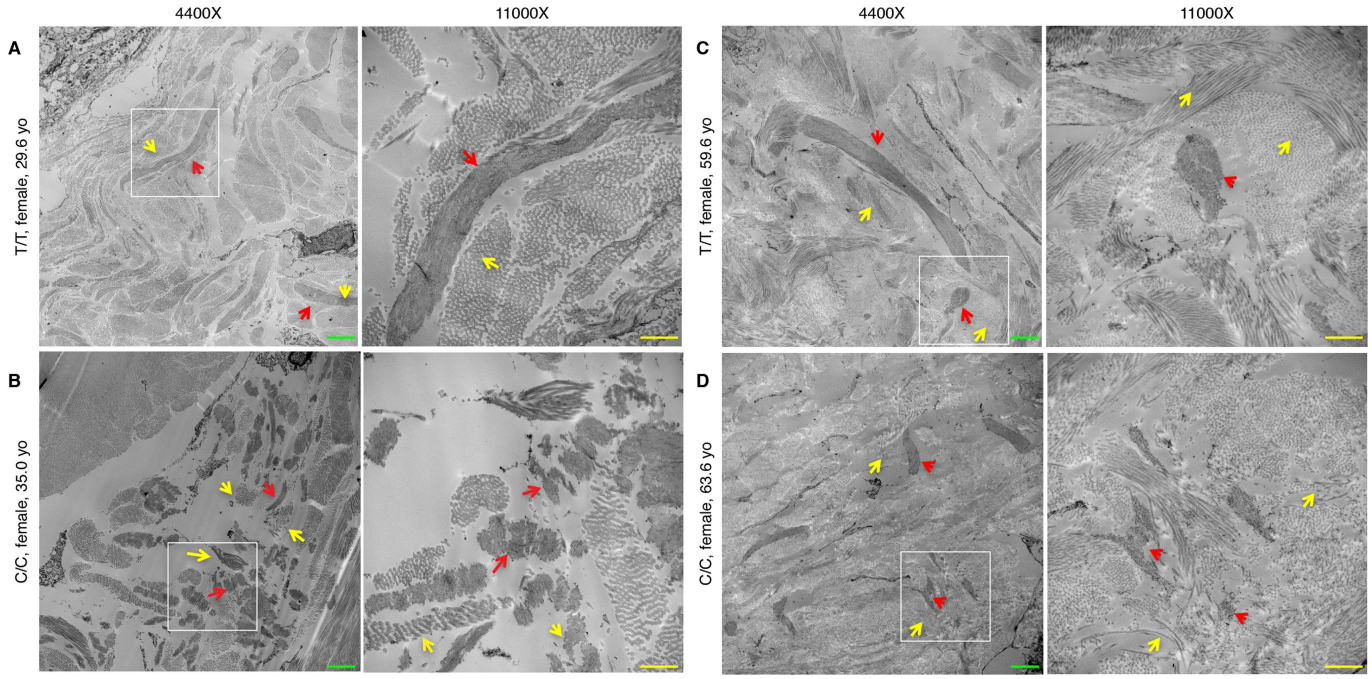
Extended Data Fig. 6 | Comparison of different selection statistics for rs200342067 and other variants with a similar DAF and recombination rate. a, Distribution of iHS for 2,062 independent variants (that are at least 1 Mb apart) matched in DAF and local recombination rate to rs200342067. iHS values are calculated for Peruvian individuals in the 1000 Genomes Project (n = 85 individuals) and were obtained from a previously published study[19]. Red line, iHS of rs200342067 (iHS = −1.5; 4.7th percentile); green and blue lines, fifth and first percentile of the iHS distribution. **b**, EHH decay curves for rs200342067 (red line) as well as haplotypes that carry 2,062 independent variants (at least 1 Mb apart) matched in DAF and local recombination rate to rs200342067 in our cohort (n = 6,268 haplotypes (grey lines)). **c**, Distribution of integrated EHH for haplotypes shown in **b**, haplotypes carrying the rs200342067*C allele are longer than 97.5% of haplotypes that carry similar variants. The x axis shows the integrated EHH; the red line indicates the integrated EHH of the rs200342067*C allele (integrated EHH = 0.115). **d**, Histogram of Fisher's exact test results comparing the extent of allele frequency differences between coastal (n = 46 individuals) and non-coastal (n = 104 individuals) regions in Peru for 2,062 independent variants that were matched in DAF and local recombination rate to rs200342067. the x axis shows the $-\log_{10}$-transformed P values from the two-sided Fisher's exact test; the dashed blue and green vertical lines show the 99th and 95th percentiles, respectively; the solid red line indicates the $-\log_{10}$-transformed P value of the two-sided Fisher's exact test (P = 0.0005) for rs200342067 (1.1% percentile). **e**, Bayenv2 XTX statistics, a measure of deviation from neutral patterns of population structure, for 2,062 independent variants that were matched in DAF and local recombination rate to rs200342067. The x axis shows the XTX statistics; the red line indicates the XTX value for rs200342067 (XTX = 2.13; 8.3th percentile); the green and blue lines show the fifth and first percentile of the XTX distribution, respectively.



Extended Data Fig. 7 | Genomic context of rs200342067 FBN1(E1297G). **a**, Schematic of *FBN1*, exons are shown as black bars. Exon 31 (ENSE00001753582) is shown in red. **b**, The *FBN1* exon 31 sequence and PhyloP per-nucleotide conservation score based on multiple sequence alignment of 100 vertebrate species (obtained using the GRCh37 assembly conservation track of the UCSC genome browser). The T>C change due to rs200342067 occurs in a conserved nucleotide. **c**, Schematic of fibrillin 1 (ENST00000316623.5). Fibrillin 1 consists of the following domains: N- and C-terminal domains (black rectangles), EGF-like domains (stripped rectangles), hybrid domains (black pentagons), TGFβ-binding domains (grey ovals), a proline-rich domain (white hexagon) and 43 calcium-binding cbEGF-like domains (white rectangles). cbEGF domain 17, which is affected by rs200342067 FBN1(E1297G), is shown in red; E1297G is located between a conserved cysteine FBN1(C1296) involved in forming a disulfide bond with FBN1(C1284) and a conserved asparagine FBN1(N1298) involved in calcium binding. **d**, The sequence of FBN1(cbEGF) domain 17 of fibrillin 1 and the three-dimensional structures of cbEGF domains 17 and 18 (the three-dimensional structure was obtained based on homology with the previously published[36] cbEGF domains 12 and 13 of fibrillin 1 (PDB 1LMJ). rs200342067 changes the glutamic acid, a large amino acid with a negatively charged side chain, to glycine, the smallest amino acid with no side chain (shown in red). The side chains are shown for rs200342067 (red spheres), as well as the calcium-interacting residues (beige sticks) and the cysteine residues involved in disulfide bonds (yellow sticks). A calcium ion is shown in green.



Extended Data Fig. 8 | Immunohistochemical staining of fibrillin 1. **a**, **b**, Fibrillin 1 staining of skin biopsies from two individuals with the rs200342067 C/C genotype. **c**, **d**, Fibrillin 1 staining of skin biopsies from two individuals with the T/T genotype matched for age, sex and ancestry proportions. Individuals with the C/C genotype have less fibrillin 1 deposition in the dermal extracellular matrix and shorter microfibrillar projections from the dermal–epidermal junction into the superficial (papillary) dermis (red arrows, 20×) as well as less fibrillin 1 deposition in the deeper dermis. Two magnification are shown, the red rectangles in the first column (20× magnification) are magnified in the second column (60×).



Extended Data Fig. 9 | Electron microscopy of fibrillin 1 in skin. a, **c**, Electron microscopy images of the dermal–epidermal junction in samples from two individuals with the rs200342067 T/T genotype. **b**, **d**, Electron microscopy images of the dermal–epidermal junction in samples from two individuals with the rs200342067 C/C genotype who are matched for age, sex and ancestry proportions. Individuals with the C/C genotype have short, fragmented and less densely packed microfibrils with irregular edges (red arrows) and their microfibrils are embedded in less dense collagen bundles (yellow arrows) compared with individuals with the T/T genotype. Two magnification are shown, the white rectangles in the first column (4,400× magnification; green scale bars, 2 µm) are magnified in the second column (11,000× magnification; yellow scale bars, 1 µm).

Extended Data Table 1 | SNPs that overlap the 15q15–21.1 locus

rs	position	allele1	allele2	MAF (%)	effect size (cm)	se	z-score	Wald p-value
rs193211234	48752674	A	T	4.66	-2.37	0.38	-6.24	$4.4x10^{-10}$
rs200342067	48773926	C	T	4.72	-2.22	0.36	-6.17	$6.8x10^{-10}$
rs544786245	48822780	T	G	4.46	-2.32	0.38	-6.11	$1x10^{-9}$
rs143730951	48858921	T	C	4.72	-2.27	0.37	-6.14	$8.2x10^{-10}$
rs180913076	48928052	C	A	4.56	-2.22	0.38	-5.84	$5.2x10^{-9}$

In our height GWAS (n = 3,134 individuals), one locus reached the genome-wide significance threshold ($P < 5 \times 10^{-8}$). This locus overlaps with *FBN1* on chromosome 15 and includes five tightly linked SNPs. One SNP, rs200342067, is a missense variant and the other four are intronic variants. Association P values are from two-sided Wald tests. Numbers are rounded to two decimal places. se, standard error.

nature research

Corresponding author(s): Soumya Raychaudhuri

Last updated by author(s): Feb 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Genotyping of the discovery cohort samples was performed using our customized Affymetrix LIMAArray. Genotypes were called in a total of 4,002 samples using the apt-genotype-axiom (Luo Y et al, Nat Commun, 2019). Genotyping of the replication cohort samples was performed using the Illumina MEGA array. Genotypes of the replication cohort were called in a total of 789 samples using the GenomeStudio software.
Data analysis	GENESIS R package (version 2.6.1), GEMMA (version 0.96), PLINK (version 1.90b3w), GCTA (version 1.26.0), ADMIXTURE (version 1.3), SHAPEIT2 (version v2.r837), R (version 3.4), R meta, metafor, dyplyr, ggplot2, broom, lm and lme4qtl packages, R cor.test, round functions, SKAT (version 1.3.2.1), selscan (version 1.2.0a), Bayenv2 (version 2.0), Beagle (version 4.1), EIGENSOFT (version 6.1.4), SNPweights (version 2.1), iSAFE (version v1.0.4)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Genotyping data will be made available through dbGAP.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Discovery cohort: We collected genotyping data for 4,002 individuals from 2,272 households. The sample size is provided in the corresponding figure captions in the main manuscript and supplementary information files. Replication cohort: We collected genotyping data for 789 individuals from 273 households. he sample size is provided in the corresponding figure captions in the main manuscript and supplementary information files. Genetic studies of complex traits, including height, have identified novel associations with sample sizes ranges from a few hindered to hundreds of thousands. However, As far as we are aware, this is the largest study of height in the Peruvian population.
Data exclusions	Discovery cohort: Out of 4002 recruited individuals, 22 individuals were excluded during quality control due to missing more than 5% of the genotype data, excess of heterozygous genotypes (±3.5 SD), duplicated with identity-by-state > 0.9, or TB cases with aga-at-diagnosis above 40. We further excluded 846 individuals from the analysis: individuals below 19 years of age, individuals without height measurement, and individuals with a measured height ±3.5 SD away from the population average. The final cohort for the current study included 3,134 from 1,947 households. Replication cohort: Height data were not available for 27 individuals. Moreover, 164 individuals were excluded due to age < 19 years old. The final cohort included 598 individuals from 242 households. The exclusion criteria based on genotyping quality is per-established by previous genetic studies of complex traits, also it is customary in genetic studies of anthropometric traits to include only adults as children and adolescents measurmenst are subject to future change.
Replication	Replication cohort: Height data were not available for 27 individuals. Moreover, 164 individuals were excluded due to age < 19 years old. The final cohort included 598 individuals from 242 households. We also tested the association of rs200342067 in two publicly available datasets of Hispanic/Latino individuals, PAGE and GIANT, to replicate our association signal. Our replication attempts were successful.
Randomization	We used permutation analysis to test the association between the Native American ancestry and height. As individuals within the same household share the same environment, household serves as a proxy for unmeasured environmental factors. to correct for these factors, randomly reassigned heights within each household 10,000 times, and recalculated the effect size for Native American ancestry in each round, to make an empirical null distribution.
Blinding	blinding was not relevant as no subset of individulas received any different treatment

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	rabbit polyclonal anti-fibrillin 1 antibody (FBN1, dilution 1:250, HPA021057, MilliporeSigma, St. Louis, MO)
Validation	antibody was validated by the Human Protein Atlas (HPA) project (www.proteinatlas.org) usig orthogonal RNAseq. Orthogonal validation is an enhanced validation method where the antibody staining is verified by a non-antibody based method. Here, the antibody staining was compared to RNA-Seq data for the same samples. This antibody has also been used successfully in prior publications: https://www.sigmaaldrich.com/catalog/product/sigma/hpa021057?lang=en&region=US

Human research participants

Policy information about studies involving human research participants

Population characteristics	Height in centimeters, gender, age, socioeconomic status were collected. Also as the recruitment was originally for a TB study individuals' TB status were collected. Please see methods for details of collected covariates.
Recruitment	Participants were collected in any of the 106 public health centers. Blood samples were collected from individuals following institutional IRB guidelines and with informed consent from participants. Recruitment site was a large catchment area of Lima, Peru that included 20 urban districts and approximately 3.3 million residents. Clinical examination was approved by the local IRB committee.
Ethics oversight	Harvard Medical School

Note that full information on the approval of the study protocol must also be provided in the manuscript.