Section 1: Cohorts' demographic information

Our discovery and replication cohorts, after quality control (QC), includes 3,134 and 598 adults (age≥19 years old) from Lima, Peru (**Methods**). In both discovery and replication cohorts, height in centimeters was measured by trained healthcare staff upon recruitment of study participants. In addition to height, a number of other varaibles such as sex, age, and socioeconomic factors were also collected (Methods). In this section, we display the height and age distribution for males and females in the discovery and replication cohorts.

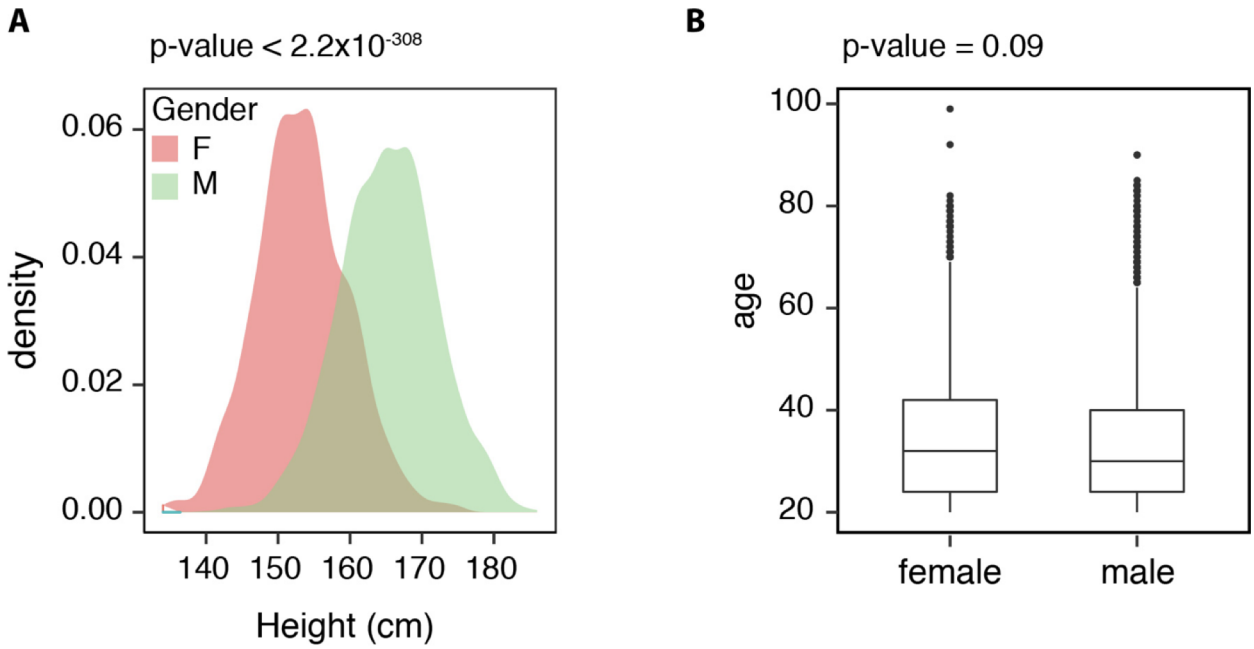

Figure S1.1: **Discovery cohort's demographic information. A)** Density plot of height for all the Peruvian males (N = 1,795 (57%)) and females (N = 1,339 (43%)) included in this study after quality control (e.g. after removing low quality samples, individuals below 19 years old and height outliers (± 3 standard deviations (sd) from the mean). Males were significantly taller than females (Male mean = 165.2 cm (sd = 6.7), Female mean = 153.4 cm (sd = 6.4), t = 50.321, degrees of freedom (df) = 2954.6, unpaired t-test two-sided p-value < $2.2x10^{-203}$). **B**) Age was not significantly different between males and females (t-value = -1.70, df = 2860.2, unpaired t-test two-sided p-value = 0.09). Boxplots show median and interquartile range (IQR).

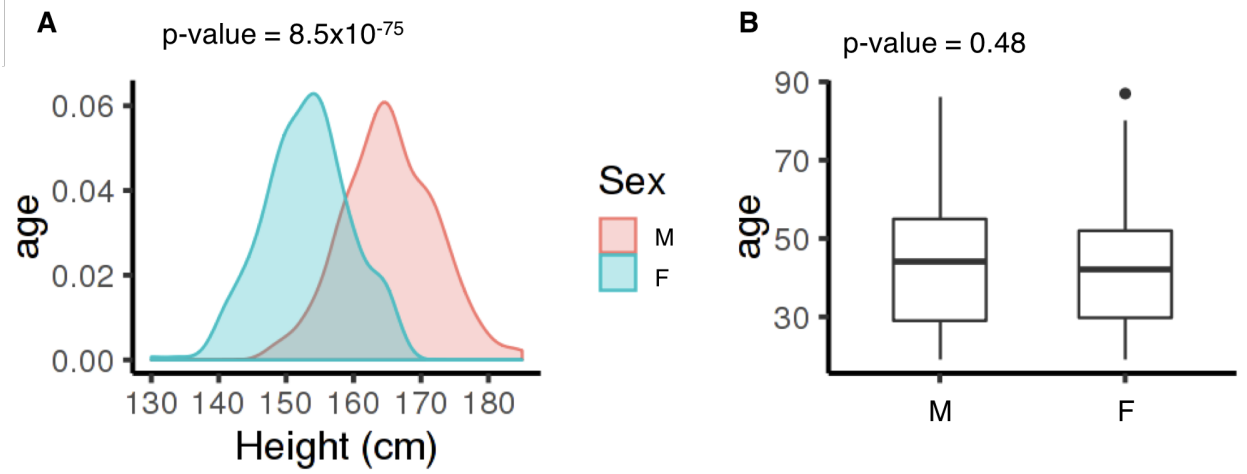

Figure S1.2: **Replication cohort's demographic information. A)** Density plot of height for all the Peruvian males (N = 234 (39%)) and females (N = 364 (61%)) included in the replication cohort after quality control. Males were significantly taller than females (Male mean = 165.4 cm (sd = 6.8), Female mean = 153.1 cm (sd = 6.4), t = 22.063, df = 475.31, unpaired t-test two-sided p-value = 8.5×10^{-75}). **B**) Age was not significantly different between males and females (t = 0.70633, df = 468.43, unpaired t-test two-sided p-value = 0.48). Boxplots show median and interquartile range (IQR).

Section 2: The correlation between height and Native American ancestry proportion

In our cohort, we observed a negative correlation between height and Native American ancestry proportion (Pearson's correlation coefficient (r) = -0.28, p-value = 9.3×10^{-58}, **Figure S2.1**). Native American ancestry remained significantly associated with lower height after including age, sex, African, and Asian ancestry proportions, and a genetic relatedness matrix (GRM) calculated using PC-Relate[1] (**Methods**) to account for relatedness (**Table S2.1**). We repeated this analysis after adding a random effect to account for the individual's household as proxy environmental factors that might not be captured by household-level socioeconomic variables (**Methods**). Native American ancestry remained significantly associated with lower height after including the household random effect (**Table S2.2**). Finally, to ensure adequate control for environmental factors, we randomly assigned height to individuals within each household 10,000 times and recalculated the Native American ancestry effect size using a linear mixed model with age, sex, African, and Asian ancestry proportions, and a GRM calculated using PC-Relate[1] as covariates 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 (**Figure S2.2**).

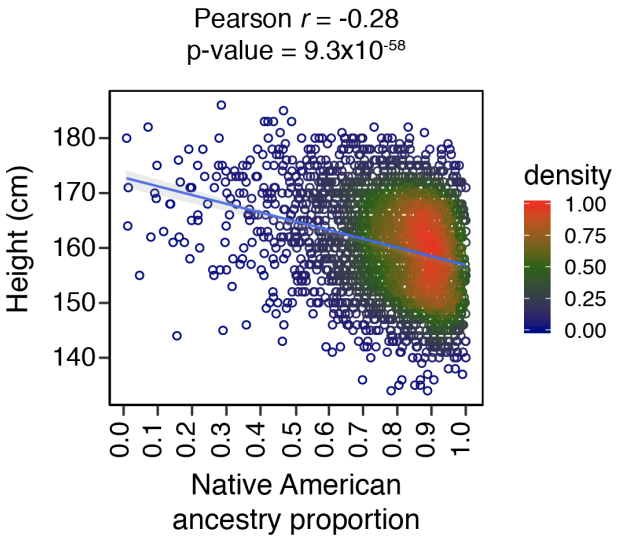

Figure S2.1: Native American ancestry is negatively associated with height. Greater Native American ancestry proportion is associated with lower height (N=3,134 individuals, Pearson's correlation coefficient (r) = -0.28, confidence interval (CI) = -0.31 - -0.25, t-value = -16.36, df = 3132, one-sample t-test two-sided p-value = 9.3×10^{-58}). The x-axis represents Native American ancestry proportion from ADMIXTURE analysis at K = 4 clusters. The y-axis represents the height (cm).

Table S2.1: Native American ancestry is significantly associated with a lower height. The base model is a linear mixed model accounting for age, sex, African and Asian ancestry proportions, and a genetic relatedness matrix (GRM) to account for population structure and genetic relatedness (N = 3,134 individuals). The effect sizes for African, Asian, and Native American ancestry are given relative to European ancestry. For example, the effect size 14.55 cm for the Native American ancestry should be interpreted as being 100 Native American compared to being 100 European decreases height by -14.55 cm. ASI: Asian, AFR: African, EUR: European, NAT: Native American. CI: confidence interval, df: degrees of freedom. For each covariate, we used the χ^2 difference test to compare nested models, p-values are two-sided p-values χ^2 derived from the corresponding χ^2 statistics. Numbers are rounded to two decimal places.

covariate	effect size (cm)	2.5% CI	97.5% CI	χ^2 statistic	p-value χ^2 (df=1)
age	-0.10	-0.12	-0.09	-12.96	1.5×10^{-37}
Gender (male)	11.33	10.90	11.77	50.57	$< 10^{-203}$
AFR proportion	-3.25	-7.46	0.97	-1.51	0.13
ASI proportion	-10.73	-17.34	-4.11	-3.18	0.001
NAT proportion	-14.55	-16.59	-12.52	-14.00	2.4×10^{-43}

Table S2.2: **Native American ancestry remains significantly associated with lower height after the inclusion of a random household effect.** Native American ancestry remained significantly associated with lower height after we included a random household effect as a proxy for socioeconomic and environmental factors in addition to age, sex, African and Asian ancestry proportions, and a GRM (N = 3,134 individuals). The effect sizes for African, Asian, and Native American ancestry are given relative to European ancestry. For example, the effect size -14.75 cm for the Native American ancestry should be interpreted as being 100 Native American compared to being 100 European decreases height by 14.75 cm. ASI: Asian, AFR: African, EUR: European, NAT: Native American. CI: confidence interval, df: degrees of freedom. For each covariate, we used the χ^2 difference test to compare nested models, p-values are two-sided p-values derived from the corresponding χ^2 statistics. Numbers are rounded to two decimal places.

covariate	effect size (cm)	2.5% CI	97.5% CI	χ^2 statistic	χ^2 p-value (df=1)
age	-0.10	-0.12	-0.09	-12.43	1.1×10^{-34}
Gender male	11.47	11.03	11.91	51.20	$< 10^{-203}$
AFR proportion	-3.57	-7.77	0.64	-1.66	0.10
ASI proportion	-11.62	-18.28	-4.95	-3.42	0.001
NAT proportion	-14.75	-16.83	-12.68	-13.94	7.20×10^{-43}
Household*	2.08	1.54	2.53	NA	7.40×10^{-7}

*Household effect size is calculated as the standard deviation (sd) in the model's intercept.

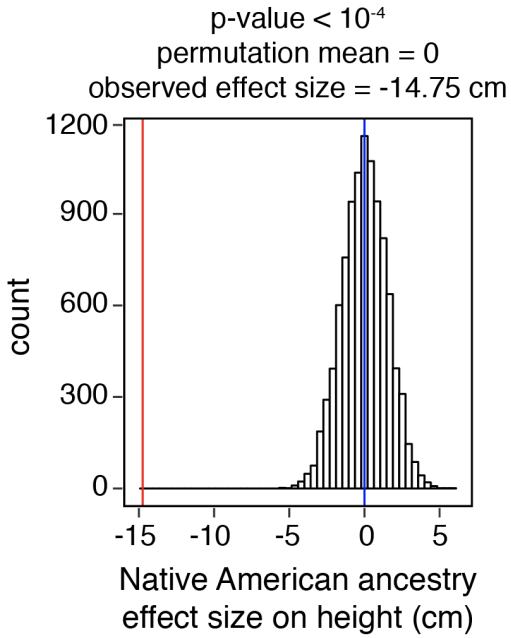

Figure S2.2: Permuting height within households. To ensure adequate control for environmental factors that might affect the correlation between height and Native American ancestry, we randomly reassigned individuals' height values within each household while keeping all the other covariates untouched in our cohort (N = 3,134). We then tested the association of Native American ancestry with height using a linear mixed model with age, sex, African, and Asian ancestry proportions, and a GRM calculated using PC-Relate[1] as covariates and calculated the effect size for Native American ancestry in this model. We repeated the permutation and association testing 10,000 times to derive an empirical null distribution of effect sizes. None of the permutations resulted in a greater effect size than that of the original data (permutation effect size ranging from -5.62 cm to 5.85 cm, permutation mean effect size = 0 cm, observed effect size = -14.75 cm, one-sided permutation test p-value < 10^{-4}). The Native American ancestry effect sizes are given relative to European ancestry. For example, the effect size 14.75 cm for the Native American ancestry should be interpreted as being 100 Native American compared to being 100 European decreases height by 14.75 cm.

Section 3: Accounting for population structure and identity-by-descent in association analysis

In GWAS presence of both recent genetic relatedness such as family structure or more distant genetic relatedness such as population structure can lead to biased estimation of allele frequencies and spurious association results[2]. In our original single variant association analysis, we included age and sex as fixed covariates and included a GRM generated using GEMMA[3] to correct for both relatedness and population structure (**Methods**). To ensure adequate control for population structure we included additional covariates such as population PCs, ancestry proportions, or household level socioeconomic scores (to capture confounding socioeconomic factors that might be related to indigeneity) in our model and repeated the associated analysis. Inclusion of these additional covariates did not affect the effect size or the strength of the association between rs200342067 and height (**Table S3.1**). Suggesting that our association results are not affected by population structure.

Genetic relatedness due to structure, such as recent admixture, can manifest itself as increased allele sharing between individuals. As a result, using relatedness estimation methods developed for non-admixed populations can lead to inflated estimation of genetic relatedness in admixed populations. To ensure that our choice of GRM has not biased our association results, we repeated our height using PC-Relate[1] GRM. PC-Relate[1] accounts for population structure in calculating relatedness between admixed individuals and correct for this structure using PCs derived from unrelated individuals [1,4]. Moreover, to ensure local (chromosome level) allele sharing between individuals does not bias the relatedness estimation, we generated 23 GRMs using PC-Relate[1] leaving one chromosome out each time[5] and tested the association of variants on each chromosome with height using the GRM that did not include that chromosome. We observed similar results to our original GWAS using the PC-Relate GRMs confirming that our choice of GRM or biased estimation of relatedness estimations does not derive our findings (**Figure S3.1**).

Another potential source of bias in recently admixed populations is the possible non-random mating in recent ancestral generations which can lead to the presence of autozygosity segments in the genome of individuals that are not related by pedigree[2]. The presence of these autozygosity segments can lead to biased estimates of allele frequencies and makes individuals within the same subpopulation appear more related than they truly are[2]. To ensure that the presence of autozygosity segments does not bias our relatedness estimates, we first used Refined IBD [6] to estimate the level of autozygosity in our cohort. We found autozygosity segments in 1,927 of the 3,134 individuals (61%), with lengths ranging from 0.2 Mb to 575.3 Mb (interquartile range (IQR) = 1.7 - 12.9 Mb). Autozygosity segments accounted for less than 1% and 5% of the accessible genome ($5.7x10^9$ bp) in 97% and 99.7% of the individuals respectively (**Methods, Figure S3.2**). Next, we inferred the amount of pairwise IBD sharing between the individuals in our cohort (N = 3,134 individuals) using Refined IBD[6] and calculated the proportion of pairwise IBD by dividing the total length of IBD segments by the length of the accessible genome (**Methods**). Finally, we compared the pairwise IBD sharing proportions calculated in this way with pairwise kinship coefficient estimations using PC-Relate[1] on all chromosomes (**Methods**). Overall, the two methods produced relatedness estimates that were highly concordant (Pearson's $r = 0.55$, p-value < $2.2x10^{-203}$, **Figure S3.3**). This result is in line with previous studies comparing the performance of variant-based and haplotype-based relatedness inference methods[4]. We also repeated the single variant association analysis at rs200342067 locus using a genetic relatedness matrix generated using either PC-Relate[1] or Refined IBD[6] as the random effect with age, sex, and 10 principal components as fixed effects and observed similar association results at this locus regardless of the choice of GRM (**Table S3.1**). Collectively these results indicate that the association between rs200342067 and height is not the result of the presence of autozygosity in our cohort.

Table S3.1: **Additional correction for population structure or the choice of GRM does not affect the association between rs200342067 and height**. Inclusion of additional covariates such as population PCs or ancestry proportions or repeating the association analysis using different GRMs did not change the effect size or strength of the association between rs200342067 and height (N = 3,134 individuals): Numbers are rounded to two decimal places. Association p-values are two-sided Wald test p-values. se: standard error, SES: socioeconomic status.

Covariates	effect size (cm)	se	z-score	Wald p-value
Age, gender, GEMMA GRM	-2.22	0.36	-6.17	6.8×10^{-10}
Age, gender, 10 PCs, GEMMA GRM	-2.22	0.36	-6.17	6.8×10^{-10}
Age, gender, 10 PCs, SES, GEMMA GRM	-2.22	0.36	-6.17	6.8×10^{-10}
Age, gender, 20 PCs, GEMMA GRM	-2.22	0.36	-6.16	7.3×10^{-10}
Age, gender, ASI, AFR, EUR, GEMMA GRM	-2.22	0.36	-6.17	6.8×10^{-10}
Age, gender, 10 PCs, PC-Relate GRM	-2.22	0.36	-6.19	6.0×10^{-10}
Age, gender, 10 PCs, Refined IBD GRM	-2.10	0.36	-5.79	7.0×10^{-9}

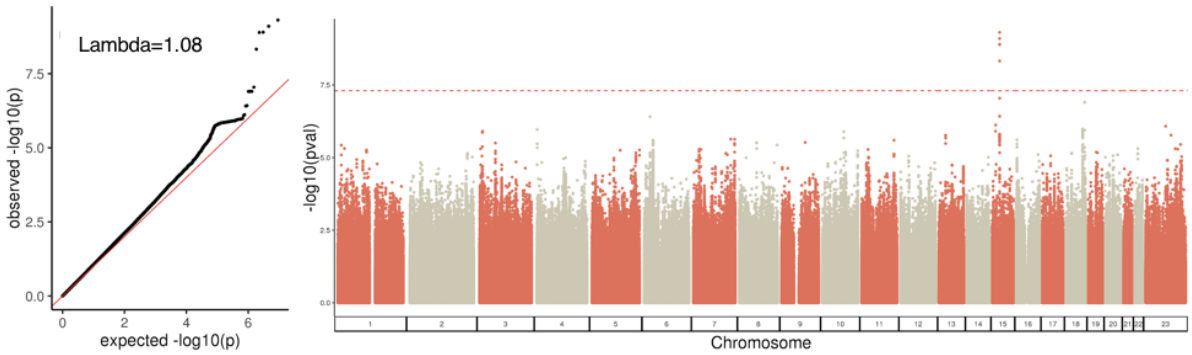

Figure S3.1: Single variant association analysis using PC-Relate[1] GRM. We generated 23 GRMs, leaving one chromosome out each time and tested the association for variants on each chromosome using the PC-Relate GRM generated without that chromosome (N = 3,134 individuals and 7,756,401 variants). Five SNPs overlapping the coding sequence of *FBN1* passed the genome-wide significance threshold (two-sided p-value < 5×10^{-8}, dotted red line). Of these one variant, rs200342067, is a missense variant in *FBN1* and the other four are intronic variants. We did not observe any inflation in test statistics ($\lambda = 1.08$). Association p-values are two-sided Wald test p-values.

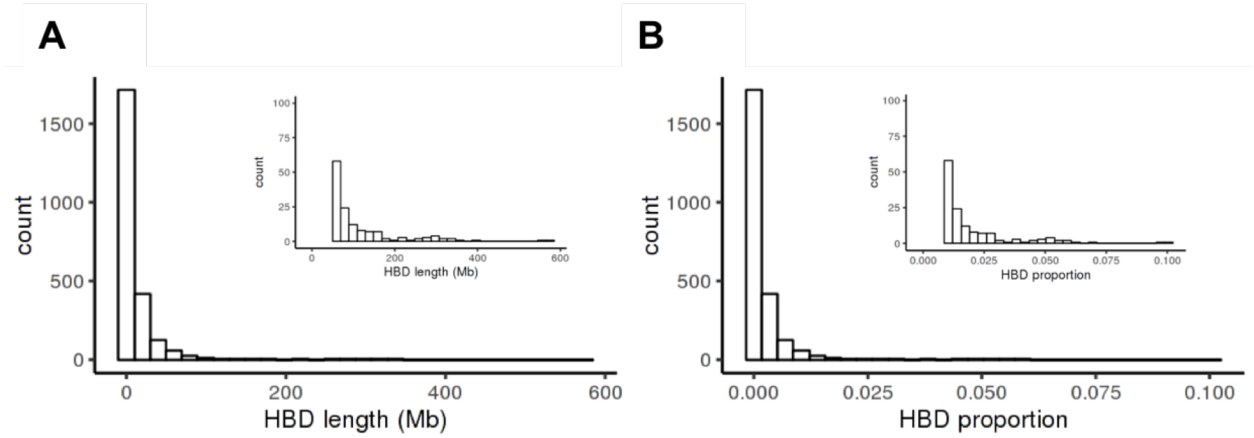

Figure S3.2: Distribution and the genomic proportion of autozygosity (homozygousity-by-decent, HBD) segments in our cohort (N = 3,134 individuals). A) Segment lengths range from 0.2 Mb to 575.3 Mb (IQR = 1.7 - 12.9 Mb). **B)** Autozygosity segments accounted for less than 1% and 5% of the accessible genome in 97% and 99.7% of the individuals respectively. Zoomed in facets are shown within each plot.

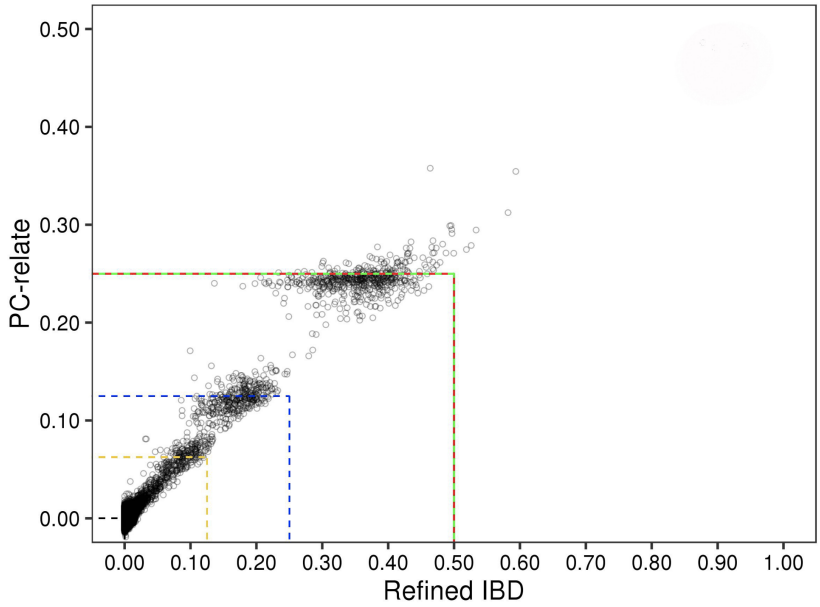

Figure S3.3: Comparison between PC-Relate kinship coefficient estimates and IBD proportion calculated using Refined IBD[6], IBD segments. The relatedness estimates were highly concordant between the two methods (Pearson's r = 0.56 (CI = 0.55 - 0.56), t-value = 1478.70, df = 4,909,400, one-sample t-test two-sided p-value < 2.2×10^{-203}). Each dot represents the pairwise kinship estimates between two individuals (N = 3,134 individuals and 4,909,411 pairwise combinations). X-axis: Refind IBD, IBD proportions. Y-axis: PC-Relate kinship coefficients. Red line: parent-offspring, green line: full siblings, blue line: second-degree relative, yellow line: third-degree relative, black: unrelated.

Section 4: Gene-based association analysis

In this section, we display the gene-based association analysis (N = 3,134 individuals and 25,000 genes) results for both rare (MAF < 1%) and common (MAF ≥ 1%) variants.

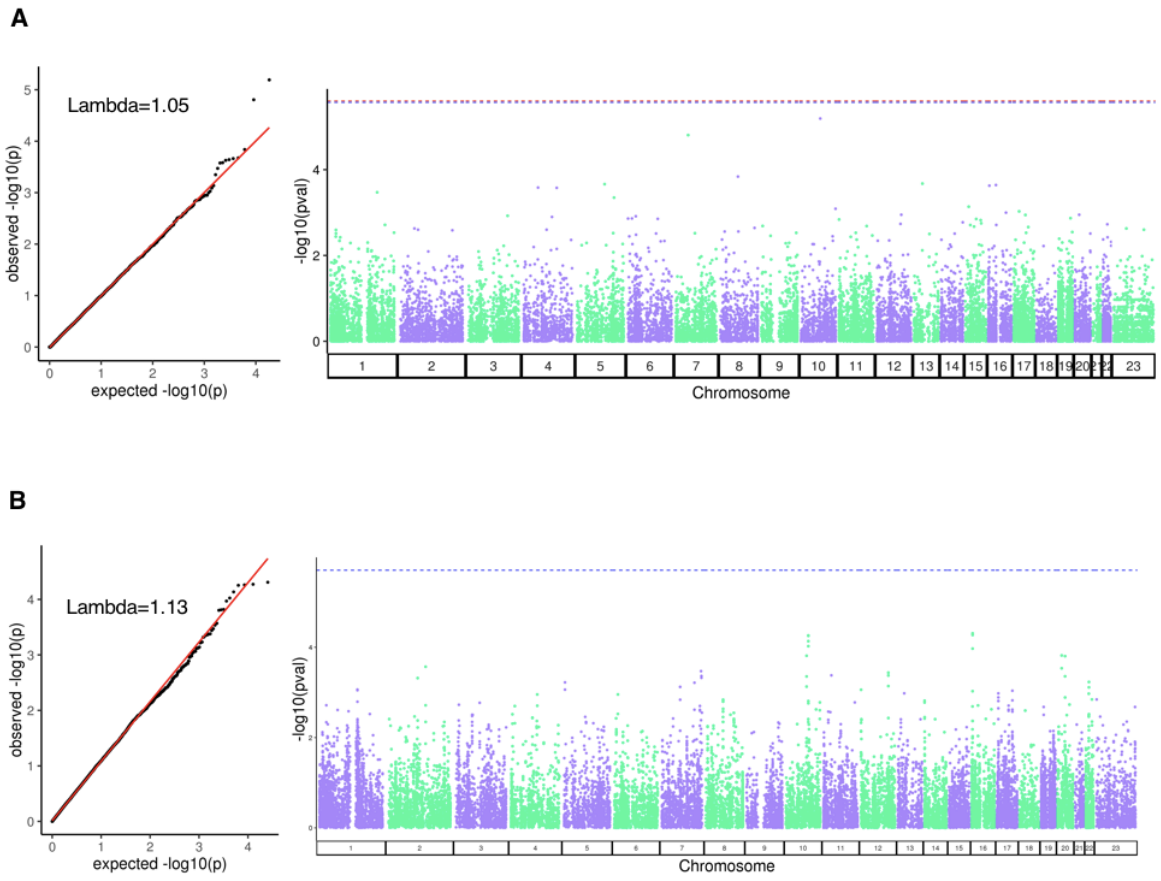

Figure S4.1: Gene-based association analysis. A) Rare (MAF < 1%) variants gene-based analysis using SKAT[7] (N = 3,134 individuals). The dotted red line corresponds to the genome-wide significance threshold of $2x10^{-6}$ for 25,000 tested genes. No genes reached the genome-wide significance threshold. Association p-values are two-sided Wald test p-values. **B)** Gene-based meta-analysis of common (MAF ≥ 1%) variants using GCTA fastBAT[8] (N = 3,134 individuals). The dotted red line corresponds to the genome-wide significance threshold of $2x10^{-6}$ for 25,000 tested genes. No genes reached the genome-wide significance threshold. Association p-values are two-sided Wald test p-values.

Section 5: Polygenic Risk Score analysis

In our initial PRS analysis, we calculated PRS using the reported conditional effect sizes of 2,993 common height-associated variants found in a European height meta-analysis (N ~ 700,000 individuals) and were present in our Peruvian cohort (N = 3,134 individuals, **Figure S5.1**, **Methods**). After considering the possibility that the choice to use conditional effect sizes may have mitigated heritability explained, we repeated the PRS analysis using the unconditional effect sizes of 2,195 height-associated variants that reached genome-wide significance before conditional analysis in the original study and were also present in our cohort (**Methods**). The PRS calculated in this way explained 7.2% (CI = 5.6 – 9.1, p-value = 4.0×10^{-53}**, Figure S5.2**) of height phenotypic variance in our cohort. The PRS calculated using unconditional and conditional effect sizes are highly correlated (Pearson's r = 0.77, and the amount of height variance explained by them is not statistically different (z-score = -1.85, two-sided p-value = 0.06, **Figure S5.1E**).

Previous studies [9–11] suggested that the lower predictive power of European-biased PRSs in non-European populations might reflect that different variants are responsible for the height variance in non-European populations or that the lead European variants do not tag the same causal variants in non-European populations. To test whether the PRS calculated using European effect size explain a higher proportion of height variance in individuals with higher European ancestry proportion, we divided our cohort into two groups with high (≥ 0.22, top quartile, N = 784 individuals) and low (< 0.22, N = 2,350 individuals) proportions of European ancestry. PRS explained a significantly higher proportion of height phenotypic variance (z-score = 2.27, p-value = 0.02) in individuals with high European ancestry proportion (Pearson's r^2 = 9.8% (CI = 6.2 - 14.1), p-value = 2.2×10^{-19}) compared to the individuals with low European ancestry proportion (r^2 = 5.1% (CI = 3.5 - 7.0), p-value = 9.0×10^{-29}, **Figure S5.3**). Altogether these results suggest that the reduced effect of PRS in Peruvians may be at least in part related to genetic differences.

Differences in the sex composition of different cohorts might also affect the predictive power of PRS [12]. In our cohort, PRS explained similar (z-score = -0.72, p-value = 0.47) proportions of height phenotypic variance in men (N = 1,795, Pearson's r^2 = 7.2% (CI = 5.1 – 9.7), p-value = 5.6×10^{-31}) and women (N = 1,339, Pearson's r^2 = 6.0% (CI = 3.7 – 8.6), p-value = 1.4×10^{-19}, **Figure S5.4**).

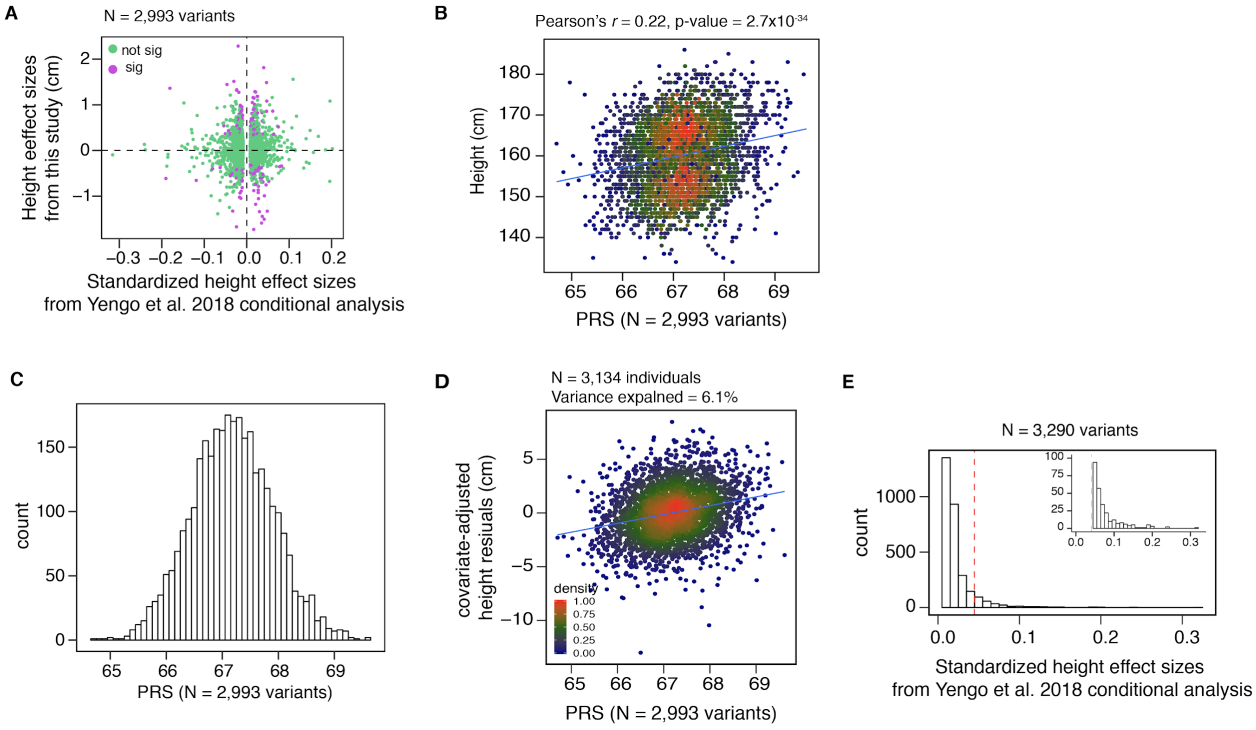

Figure S5.1: Polygenic risk score (PRS) analysis using conditioned effect sizes of variants that reached genome-wide significance before or after conditional analysis. We used conditional effect sizes from 2,993 independent, common height-associated variants that reached genome-wide significance before or after conditional analysis in the Yengo et al 2018 meta-analysis (N ~ 700,000 European individuals)[13] and were present in our cohort (N = 3,134 Peruvian individuals) to derive the PRS. **A)** Out of 2,993 variants, 1,519 (51%) show directionally consistent effects, and 199 (7%) had p-value < 0.05 in our Peruvian GWAS. **B)** Higher PRS values are associated with increased height (Pearson's $r = 0.22$ (CI = 0.18 - 0.25), t-value = 12.36, df = 3132, one-sample t-test two-sided p-value = 2.7×10^{-34}). **C)** Histogram showing the PRS distribution. **D)** Previously identified height-associated variants explained only 6.1% of height phenotypic variance in our cohort (Pearson's $r^2 = 0.061$ (CI = 0.046 - 0.078), t-value = 14.29, df = 3132, one-sample t-test two-sided p-value = 6.8×10^{-45}), x-axis: PRS, y-axis: height residuals after adjustments for age and sex as fixed effects and a GRM as a random effect. **E)** The majority (99%) of previously identified common height-associated variants (N = 3,290) have effects less than 5 mm per allele (dashed red line: cutoff corresponding to 5 mm effect size, the smaller plot shows the zoomed-in tail of the main plot).

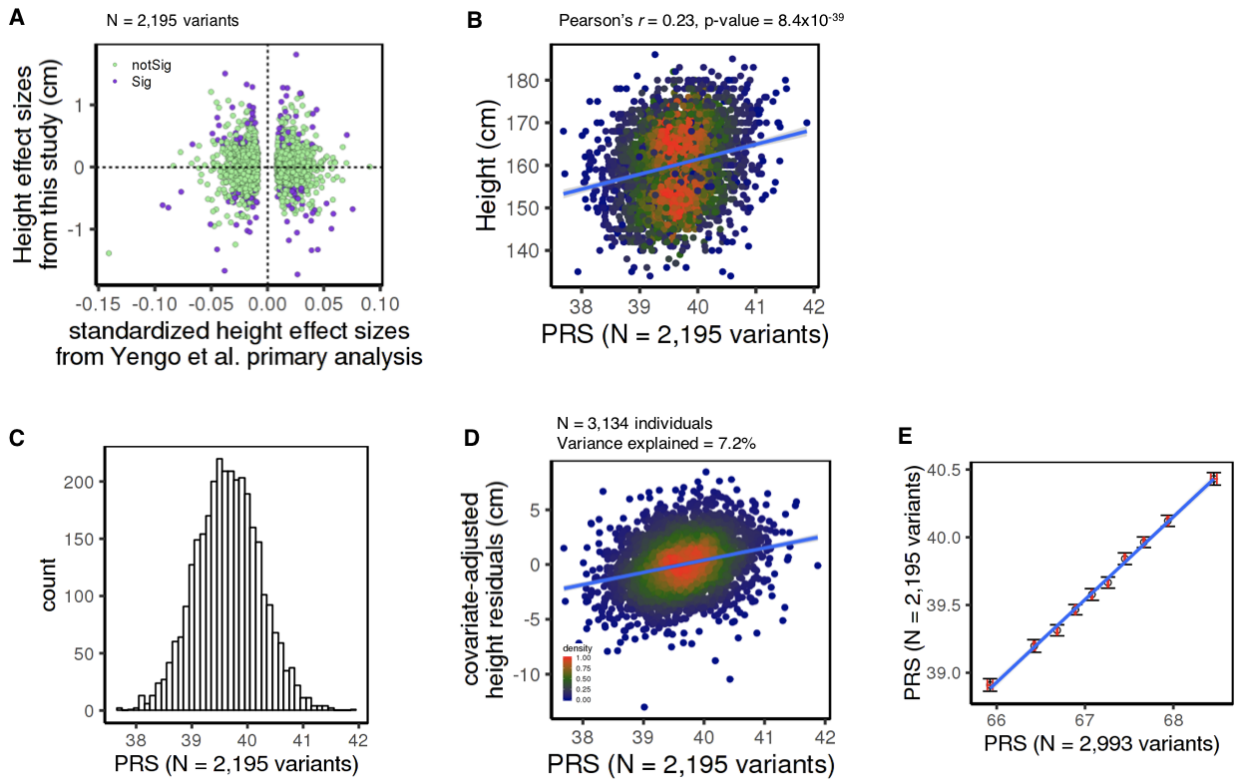

Figure S5.2: Polygenic risk score (PRS) analysis using unconditioned effect sizes of variants that reached genome-wide significance before conditional analysis. We used effect sizes from 2,195 common height-associated variants that reached the genome-wide significance threshold in the Yengo et al 2018 primary analyses (e.g before conditional analysis, N ~ 700,000 European individuals)[13] and were present in our cohort (N = 3,134 Peruvian individuals) to derive the PRS. **A)** Out of 2,195 variants, 1,101 (50%) showed directionally consistent effects, and 155 (7%) had p-value < 0.05 in our Peruvian GWAS. **B)** Higher PRS values are associated with increased height (Pearson's $r = 0.23$, (CI = 0.20 - 0.26), t-value = 13.21, df = 3132, one-sample t-test two-sided p-value = 8.4×10^{-39})**. C)** Histogram showing the distribution of PRS. **D)** The 2,195 variants that were included in the analysis explained 7.2% of height phenotypic variance in our cohort (Pearson's $r^2 = 0.072$ (CI = 0.055 - 0.091), t-value = 15.64, df = 3132, one-sample t-test two-sided p-value = 4.0×10^{-53}), x-axis: PRS, y-axis: height residuals after adjustments for age and sex as fixed effects and a GRM as a random effect. **E)** The PRS values calculated using the primary effect sizes of 2,195 primary height-associate variants (y-axis) and the PRS calculated using the conditional effect sizes of primary and conditional height-associate variants (x-axis) are highly correlated (Pearson's $r = 0.77$ (CI = 0.75 - 0.80), t-value = 66.49, df = 3132, one-sample t-test two-sided p-value < 2.2×10^{-203}). Each point represents the mean for a PRS decile (calculated using conditional effect sizes, x-axis) and the average of PRS generated using primary effect sizes for that decile (y-axis). The red (x-axis) and black (y-axis) error bars are confidence intervals.

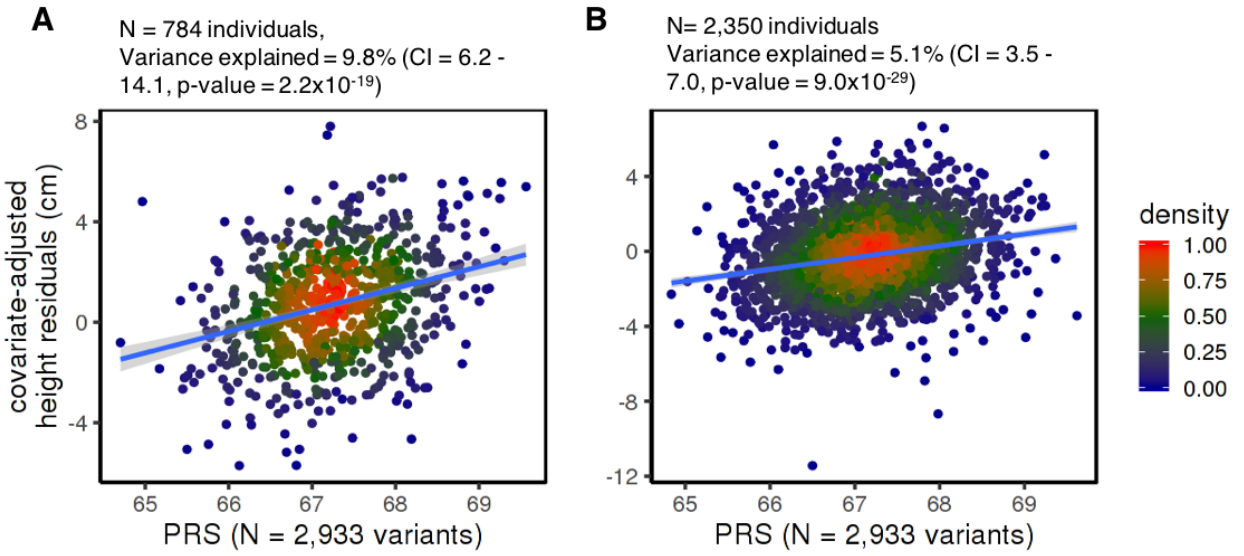

Figure S5.3: PRS analysis in individuals with a high or low proportion of European ancestry. A) PRS based on previously identified height-associated variants in the European population explain a higher proportion of height phenotypic variance in individuals with high European ancestry proportion (≥ 0.22, N = 784, Pearson's r^2 = 9.8% (CI = 6.2 - 14.1), t-value = 9.24, df = 782, one-sample t-test two-sided p-value = 2.2×10^{-19}) compared to **B)** the individuals with low European ancestry proportion (< 0.22, N= 2,350, Pearson's r^2 = 5.1% (CI = 3.5 - 7.0), t-value = 11.30, df = 2,348, one-sample t-test two-sided p-value = 9.0×10^{-29}). In both analyses Height was adjusted for age, sex, genetic relatedness but not population structure (PC-Relate GRM). x-axis: PRS, y-axis: covariate-adjusted height residuals.

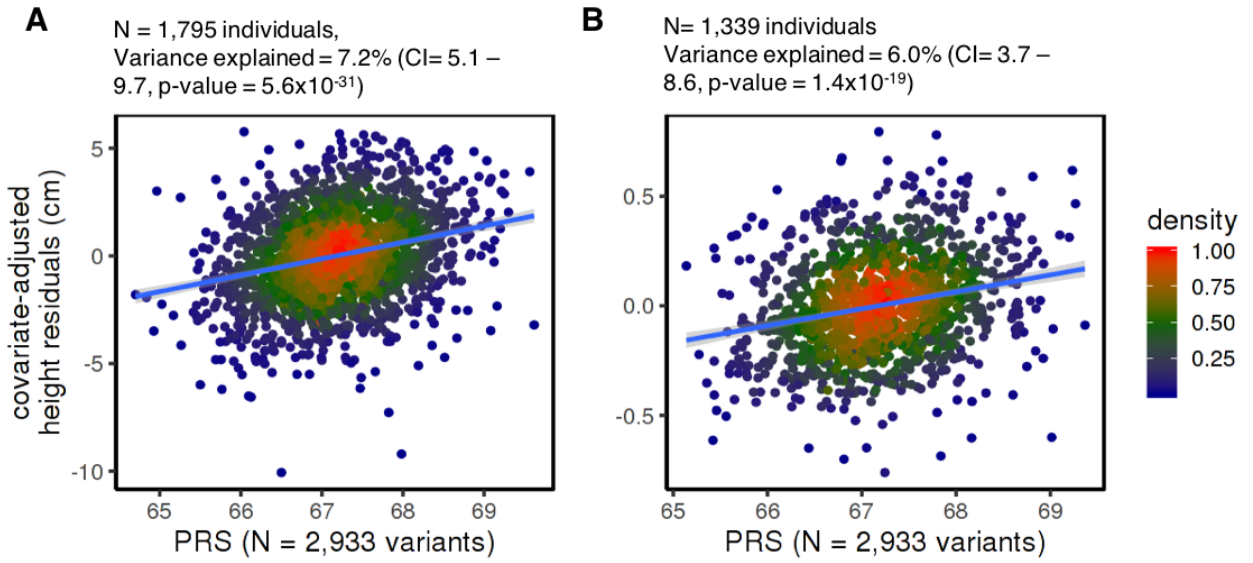

Figure S5.4: PRS analysis in men and women. A) PRS based on previously identified height-associated variants in the European population explain similar proportions of height phenotypic variance in men (N = 1,795, Pearson's r^2 = 7.2% (CI= 5.1 – 9.7), t-value = 11.80, df = 1,793, one-sample t-test two-sided p-value = 5.6×10^{-31}) and **B)** women (N= 1,339, Pearson's r^2 = 6.0% (CI= 3.7 – 8.6), t-value = 9.20, df = 1,337, one-sample t-test two-sided p-value = 1.4×10^{-19}). Sex-specific height values are adjusted for age, genetic relatedness, and cryptic population structure (GEMMA GRM). x-axis: PRS, y-axis: covariate-adjusted height residuals.

Section 6: Positive selection at rs200342067 locus in Peruvians form the 1000 Genomes Project

In the 1000 Genomes Project[14] data, the genomic region overlapping rs200342067 shows absolute integrated Haplotype Score (|iHS|) values > 2 [15,16] in certain European, South Asian, East Asian, and South American populations suggesting that this region is is under positive selection in these populations (**Table S6.1**). To test whether variants overlapping this region are also under positive selection in the Peruvian population, we used the integrated Selection of Allele Favored by Evolution (iSAFE) [17] to search for variants under positive selection in a 1.2Mb region around rs200342067. Using iSAFE, the top positive selection signal in this locus (15q21.1) comes from rs12441775 an intronic variant overlapping *FBN1*. This allele shows evidence of positive selection (iHS < -2) in certain European, South Asian, and South American populations including the Peruvian population [15,16] (**Table S6.2**). Since rs12441775 is located 77kb upstream of rs200342067, we considered the possibility that positive selection at rs12441775 led to an increased frequency of rs200342067 in the Peruvian population. However, rs12441775*G (derived/major) and rs200342067*C (derived/minor) alleles are out of phase with each other, for example in the Peruvian individuals from the 1000 Genomes Project[14] rs12441775*G and rs200342067*C do not co-occur on the same extended haplotypes (**Figure S6.1**) suggesting that positive selection at rs12441775*G does not derive the increased allele frequency of rs200342067*C in the Peruvian population. We also checked the haplotype structure of rs200342067*C and rs1426654*A allele in *SLC24A5* (located 266kb upstream of *FBN1*), an allele that is known to be under strong positive selection [18]. We observed that rs200342067*C and rs1426654*A alleles are out of phase with each other for example in the Peruvian individuals from the 1000 Genomes Project[14] these two alleles do not co-occur on the same extended haplotypes (**Figure S6.1**). Moreover, *FBN1* and *SLC24A5* are in different topologically associating domains (TADs, **Figure S6.2**) suggesting that rs200342067 or other variants in *FBN1* are unlikely to have been selected due to their regulatory effect on *SLC24A5* suggesting that positive selection at rs1426654*A or other *SLC24A5* variants is unlikely to derive the increased allele frequency of rs200342067*C in the Peruvian population.

Table S6.1: **Positive selection at *FBN1* locus.** *FBN1* overlaps genomic regions that are under putative hard selective sweeps (|iHS| > 2) in East Asian (EAS), European (EUR), American (AMR), and South Asian populations (SAS) from the 1000 Genomes Project[14]. A subset of these intervals (highlighted in green) overlap rs200342067 specifically. This table is a subset of supplementary table 3 from Johnson and Voight, 2018, Nature Evolution and Ecology, study[15,16]. pop: the 1000 Genomes project population and (super population), start pos: start position of the interval on chromosome 15 (GRCh37), stop pos: stop position of the interval on chromosome 15 (GRCh37), start rs: id of the first variant in the interval, stop rs: id of the last variant in the interval, tag rs: id of the variant with the lowes |iHS| in the interval, tag pos: position of the variant with the lowes |iHS| on chromosome 15 (GRCh37), tag stdiHS: standardized iHS value of the tag variant. bp: base pair.

pop	start rs	start pos (bp)	stop rs	stop pos (bp)	tag rs	tag pos (bp)	tag stdiHS
PUR (AMR)	rs184055877	48255052	rs4775762	48744005	rs10152385	48544388	4.803
PJL (SAS)	rs75043581	48281768	rs56286136	48755814	rs8030205	48540936	4.956
GBR (EUR)	rs76770579	48285747	rs35716640	48800163	rs191970530	48540022	6.622
CLM (AMR)	rs1869456	48287801	rs79973522	48696347	rs77929857	48659007	5.286
IBS (EUR)	rs139688845	48314143	rs1820488	48713996	rs8025278	48595192	4.35
ITU (SAS)	rs1377686	48321081	rs363836	48722884	rs8030205	48540936	4.517
TSI (EUR)	rs12907018	48519932	rs116868609	48816785	rs1872303	48658712	4.484
CEU (EUR)	rs17350938	48589981	rs57829342	48808830	rs74961364	48730562	3.569
JPT (EAS)	rs4775750	48652764	rs11854943	48861287	rs143594551	48818908	-3.473
PJL (SAS)	rs75227249	48763008	rs16961323	48959082	rs16961125	48841044	4.821
ITU (SAS)	rs17363371	48847615	rs10851470	48970280	rs12101348	48941369	3.227

Table S6.2: **iHS results for rs12441775 in the 1000 Genomes Project[14] data**. This table is a subset of the standardized iHS values for all 26 populations in the 1000 Genomes Project[14] provided by Johnson and Voight, 2018 35 study[15,16]. African populations are highlighted in purple. pop: the 1000 Genomes project population and (super population), DAF: derived allele frequency (rs12441775*G is the derived allele), stdiHS: standardized iHS value.

pop	DAF (%)	stdiH
ACB (AFR)	6.77	-1.556
ASW (AFR)	10.66	-1.851
ESN (AFR)	0.51	NA
GWD (AFR)	3.10	-0.693
LWK (AFR)	3.03	-1.091
MSL (AFR)	1.76	NA
YRI (AFR)	0.46	NA
CLM (AMR)	57.98	-2.119
MXL (AMR)	67.97	-2.09
PEL (AMR)	60.59	-2.16
PUR (AMR)	61.54	-2.264
CDX (EAS)	50	-0.983
CHB (EAS)	45.63	-0.837
CHS (EAS)	47.62	NA
JPT (EAS)	30.29	-1.07
KHV (EAS)	46.46	-1.777
CEU (EUR)	71.72	-2.077
FIN (EUR)	60.61	-1.329
GBR (EUR)	65.93	-1.273
IBS (EUR)	66.36	-1.308
TSI (EUR)	73.36	-2.122
BEB (SAS)	51.74	-1.405
GIH (SAS)	61.65	-1.44
ITU (SAS)	62.75	-3.017
PJL (SAS)	59.9	-3.021

STU (SAS)	55.39	-2.032

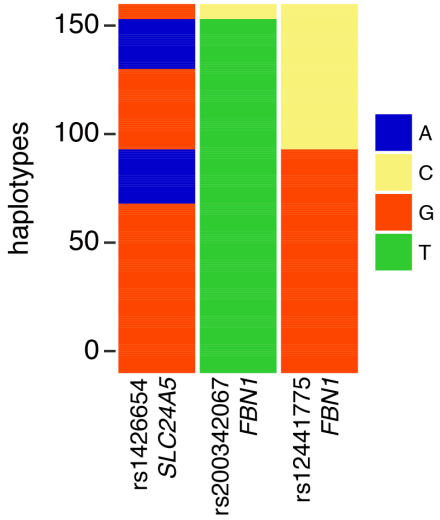

Figure S6.1: Extended haplotypes around rs200342067 in the Peruvians from the 1000 genomes project[14]**.** Side-by-side stacked barplot of haplotypes carrying rs1426654*(A/G), rs200342067*(C/T), and rs12441775*(C/G) (N = 85). None of the haplotypes carrying rs200342067*C allele (AF = 4.1%) carries rs1426654*A allele (AF in PEL = 28%) or rs12441775*G allele (AF in PEL = 61%)**.** X-axis: count of derived and alternate alleles for rs200342067 and rs1426654. Y-axis: individual haplotypes.

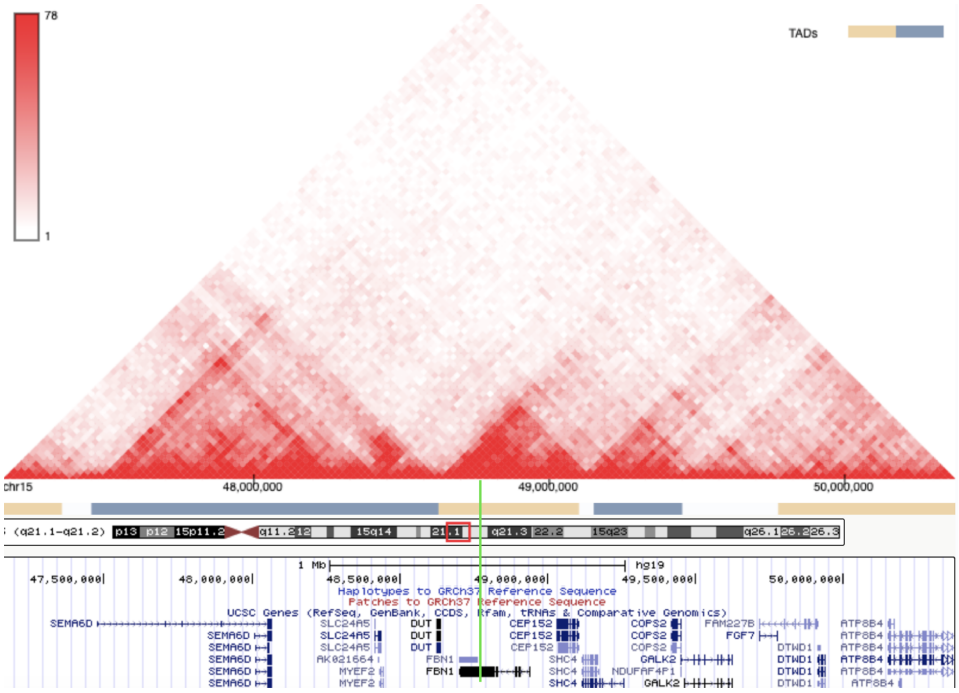

Figure S6.2: Hi-C results in HUVE cell line for *FBN1* locus. *FBN1* and *SLC24A5* are in different topologically associating domains (TADs) and there is no evidence of physical interaction between the two genes. rs200342067's position is shown with a vertical green line.

Section 7: The coast-non-coast axis within the Peruvian population

In the Peruvian Genome Project[19] (PGP) cohort (N = 150), the rs200342067 variant is more frequent in the individuals from coastal regions compared to the individuals from the Andes or the Amazon (MAF = 9.7%, 1.7%, and 0% for Coast, Andes and Amazon respectively, coast vs. non-coast two-sided Fisher's exact test p-value = 0.0005, **Table S7.1**). To alleviate any concerns regarding the possible confounding effect of population structure on the observed association between rs200342067 and height in the Peruvian population, we first performed PCA analysis in the PGP cohort using 247,940 common (MAF $\geq$ 5%) variants that were shared between the PGP and LIMAA (N = 3,134) cohorts (**Figure S7.1A-B**). We then tested the association of the first ten PCs with coast-non-coast status in the PGP cohort. We observed that the first three PCs were significantly associated with coast-non-coast status in the PGP cohort (p-value < 0.005, Bonferroni correction for ten tests, **Table S7.2**) showing that a coast-non-coast axis, captured by the first three PCs, is present in the Peruvian population

We then used the SNP loadings from the PGP PCA analysis described above to infer population PCs in the LIMAA cohort (**Figure S7.1C-D**). Next, we tested the association between the first 3 PCs of LIMAA cohort (calculated using the SNP loadings in PGP, N = 150) and Native American ancestry proportion, height, or rs200342067 minor allele count using a linear mixed model with age, and sex, as fixed effects and a genetic relatedness matrix to account for genetic relatedness as random effect (Methods, significance threshold < 0.016, Bonferroni correction for three tests). The first three PCs were significantly associated with Native American ancestry (p-value < 5.7×10^{-7}, **Table S7.3)**. PC1 and PC2 were significantly associated with rs200342067 (p-value < 8.6×10^{-4},effect size = 0.02 (se = 0.006), p-value = 0.009 for PC1 and effect size = 0.02 (se = 0.006), p-value = 0.001 for PC2, **Table S7.3)** This result is in line with the observed higher frequency of rs200342067 in populations from the coastal regions in Peru (**Figure S7.1, Table S7.1**).
PC1 and PC3 were significantly associated with height (p-value < 8.1×10^{-6}, **Table S7.3)**. To ensure that the observed association between rs200342067 and height is independent of the population structure within

Peru, we repeated this association using a linear mixed model with following covariates: age, sex, the first 10 PCs (calculated as described above using the variant weights in PGP, N = 150), and a genetic relatedness matrix to account for genetic relatedness and population structure (calculated using GEMMA[3]). Inclusion of these PCs did not affect the effect size or the strength of the observed association between rs200342067 and height (N= 3,134, MAF = 4.7%, effect size = -2.3 cm (se = 0.36), p-value = 3.0×10^{-10}, **Table S7.4**).

Collectively these results suggest that, as described before[19], (1) a cost-non-coast population substructure exists within the Peruvian population, (2) this structure is consistent with the observed higher frequency of rs200342067 in populations from the coastal regions in Peru, and (3) this structure does not explain the association between rs200342067 and height.

Table S7.1: **Comparison of rs200342067 minor/major allele count between populations from different geographical regions in Peru.** In the PGP cohort[19] (N = 150), rs200342067 is significantly more frequent in Coastal populations than in populations from the Andes and the Amazon.

region	Gender	rs200342067 C/C	rs200342067 C/T	rs200342067 T/T	total
Amazon-PGP	F	0	0	15	15
	M	0	0	13	13
Andes-PGP	F	0	1	45	46
	M	0	1	29	30
Coast-PGP	F	0	7	27	34
	M	0	2	10	12
total		0	11	130	150

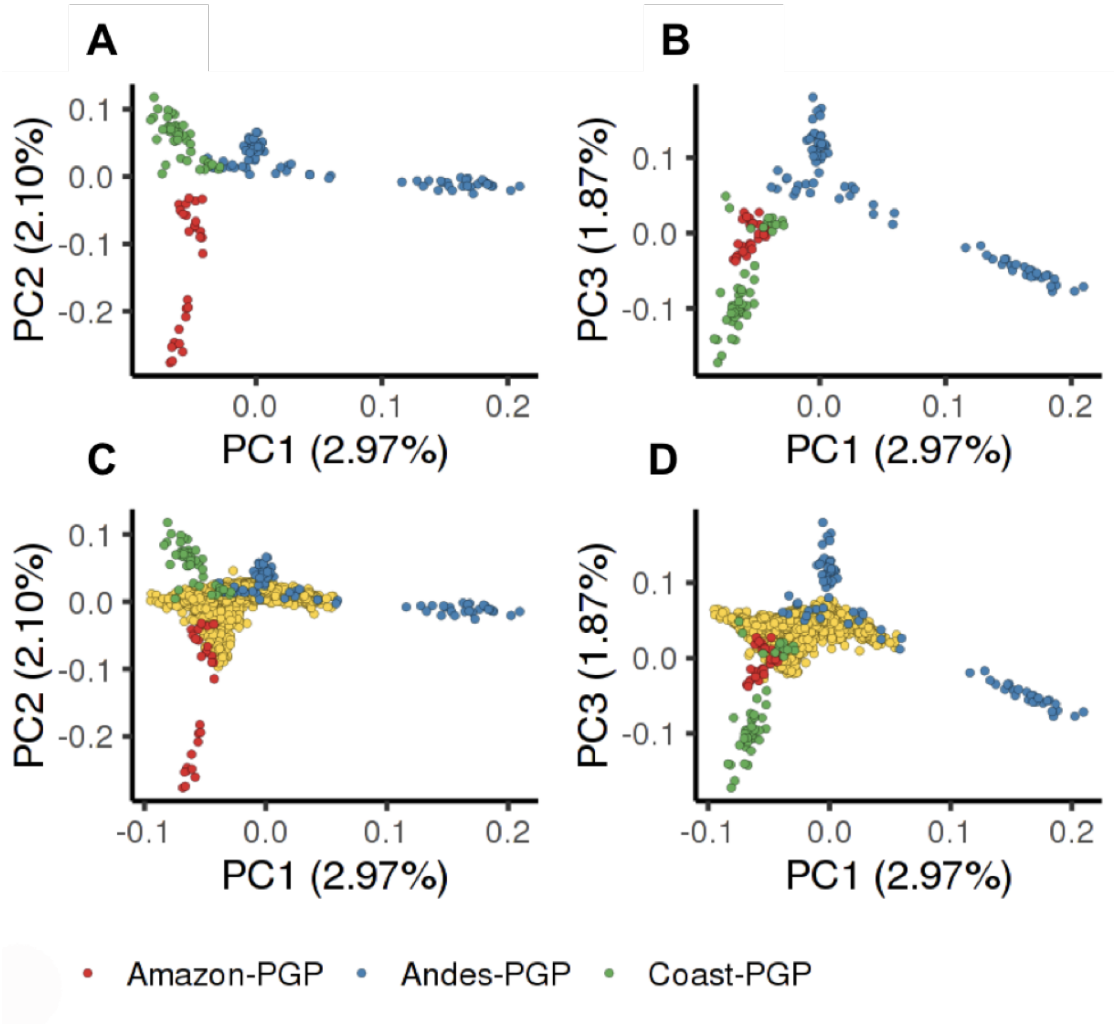

Figure S7.1: Association of population PCs with coast-non-coast status. PCA analysis of genotyping data from 150 Peruvians from three geographical regions in Peru (PGP cohort, N= 150 individuals)[19] only common (MAF 5%) variants that were also present in our cohort (LIMAA cohort, N= 3,134 individuals) were included in the analysis (N = 247,940 variants). Principal components are inferred using SNP weights calculated in the PGP cohort. The amount of variance explained by each PC is shown in the parentheses. Each dor represents one individual. **A-B)** PC1, PC2, and PC3, for better visualization only individuals from the PGP cohort are shown **C-D)** PC1, PC2, and PC3 individuals from the LIMAA cohort (yellow dots) as well as the individuals from the PGP cohort are shown together.

Table S7.2: **Association of the first ten PCs in the PGP cohort (N = 150) with coast-non-coast status.** The first three PCs were significantly associated with coast-non-coast status (ANOVA, Bonferroni correction threshold for ten tests: p-value < 0.005). df: degrees of freedom; PC: principal component, r^2: the proportion of variance explained, P-values are two-sided p-values. Numbers are rounded to two decimal places.

PC	categories	F-value	df	r^2	ANOVA p-value
1	coast vs. non-coast	47.19	1	0.24	1.7×10^{-10}
2	coast vs. non-coast	38.99	1	0.21	4.3×10^{-9}
3	coast vs. non-coast	76.03	1	0.34	5.3×10^{-15}
4	coast vs. non-coast	0.02	1	0.0001	9.0×10^{-1}
5	coast vs. non-coast	0.14	1	0.001	7.1×10^{-1}
6	coast vs. non-coast	0.68	1	0.005	4.1×10^{-1}
7	coast vs. non-coast	0.35	1	0.002	5.6×10^{-1}
8	coast vs. non-coast	0.18	1	0.001	6.7×10^{-1}
9	coast vs. non-coast	0.17	1	0.001	6.8×10^{-1}
10	coast vs. non-coast	0.39	1	0.003	5.4×10^{-1}

Table S7.3: **Association between height (cm), Native American ancestry proportion, or rs200342067 minor allele count and the first three PCs.** We calculated the association between the first three PCs and each dependent variable (e.g height (cm), Native American ancestry proportion, or rs200342067 minor allele count) using a linear mixed model with the first three PCs of LIMAA cohort (N = 3,134) calculated using SNP loading in PGP cohort[19] (N = 150), age, sex, and the PC-Relate[1] GMR to account for relatedness as covariates. For comparison, we repeated the same analysis using the PCs calculated LIMAA cohort directly. P-values are two-sided Wald test p-values. PC: principal component, se: standard error, NAT: Native American ancestry, MAC: minor allele count. Numbers are rounded to two or the closet non-zero decimal places.

dependant varaible	PC	PCs calculated using PGP SNP weights		PCs calculated using LIMAA genotypes directly	
		Effect size (se)	Wald p-value	Effect size (se)	Wald p-value
Height (cm)	PC1	-1.51 (0.12)	2.6×10^{-36}	-1.99 (0.11)	3.8×10^{-73}
	PC2	-0.21 (0.13)	0.11	-0.30 (0.11)	0.006
	PC3	0.58 (0.13)	8.1×10^{-6}	-0.26 (0.11)	0.02
NAT (% increase)	PC1	0.13 (0.002)	$< 2.2\times10^{-203}$	0.15 (0.0004)	$< 2.2\times10^{-203}$
	PC2	0.01 (0.002)	5.7×10^{-7}	0.01 (0.0004)	6.1×10^{-138}
	PC3	-0.04 (0.002)	5.5×10^{-89}	0.006 (0.0004)	7.3×10^{-51}
rs200342067 (MAC)	PC1	0.02 (0.006)	8.6×10^{-4}	0.02 (0.006)	8.6×10^{-4}
	PC2	0.02 (0.006)	8.6×10^{-4}	0.005 (0.005)	0.32
	PC3	-0.01 (0.006)	0.09	-0.002 (0.005)	0.70

Table S7.4: **Correction for population PCs derived from the PGP cohort does not affect the association between rs200342067 and height**. Inclusion of population PCs inferred using SNP weights calculated in the PGP cohort[19] (N = 150), as covariates did not change the effect size or strength of the association between rs200342067 and height in the LIMAA cohort (N = 3,134 individuals): Numbers are rounded to two decimal places. Association p-values are two-sided Wald test p-values. se: standard error.

Covariates	effect size (cm)	se	z-score	Wald p-value
Age, gender, GEMMA GRM	-2.22	0.36	-6.17	6.8×10^{-10}
Age, gender, 10 PCs, GEMMA GRM	-2.22	0.36	-6.17	6.8×10^{-10}
Age, gender, 10 PCs inferred using SNP loadings in the PGP cohort, GEMMA GRM	-2.30	0.36	-6.39	1.7×10^{-10}

Section 8: Clinical and molecular context of rs20034206

The majority of disease-causing mutations previously reported in *FBN1* are mutations that lead to Marfan or Marfan-like syndromes and are associated with a taller stature [20]. Moreover, the majority of these mutations are either loss-of-function mutations or missense mutations that lead to reduced protein function, gain-of-function mutations are not common in this gene [20]. However, *FBN1* mutations that reduce protein function can also lead to clinical phenotypes opposite to what is observed in Marfan Syndrome [20] (**Table S8.1**). For example, loss-of-function *FBN1* mutations that lead to defective interaction between microfibrils and cell surface can lead to acromelic dysplasia syndromes, a group of syndromes characterized by short stature, short hands and feet, stiff joints, and a hypermuscular build [20]. Similarly, deletions in *FBN1* transforming growth factor-β (TGFβ)-binding protein-like domain 5 (TB5) cause dominant forms of Weill–Marchesani syndrome a Mendelian disorder characterized by short stature[21]. To investigate the possible clinical effects of rs200342067 we performed dermatological and rheumatological clinical exams on 11 individuals from our cohort: 2 homozygous (C/C) cases, 2 heterozygous (C/T) cases, and 7 matched controls with reference (T/T) genotype (**Table S8.2, Methods**).

Musculoskeletal examination on these individuals did not reveal any obvious differences in the range of motion of knees, hips, wrists, and proximal interphalangeal and metacarpophalangeal joints of the second and third digits (**Table S8.2**). One C/C genotype individual had notably thicker skin upon a total body skin examination and appeared much older than the stated age. The other C/C genotype individual had no clinically abnormal cutaneous findings and none of the C/T or T/T individuals had an abnormal skin exam (**Table S8.2**).

The rs200342067 variant changes the conserved T (major/ancestral) allele to a C (minor/derived) allele in *FBN1* (g.48773926T>C, **Figure S8.1**). This mutation leads to an amino acid change (E1297G) in fibrillin-1 calcium binding epidermal growth factor domain 17 (cbEGF-domain 17). Previous mechanistic studies of missense mutations in fibrillin-1 cbEGF domains have primarily focused on the six highly conserved

cysteines or residues that are involved in disulfide bond formation or the calcium binding consensus sequence (**Figure S8.2**). rs200342067 (fibrillin-1 E1297G) is located between a conserved cysteine residue and a conserved calcium-binding asparagine residue, both of these residues are identical in all cbEGF domains in fibrillin-1 (**Figure S8.2)**. It is believed that E1297 in cbEGF domain 17 is involved in calcium-binding[22]. Calcium binding at fibrillin-1 cbEGF domains stabilizes the protein by making the microfibrils more rigid and protecting them from degradation by proteases[23]. The short fragmented and less packed phenotype seen in the skin of rs200342067 C/C individuals compared to T/T individuals might reflect the higher susceptibility of mutated fibrillin-1 to proteolysis compared to the wild-type protein.

cbEGF domains are structurally conserved and have high sequence similarity [24] (**Figure S8.2)** as a result it might be expected that missense mutations at structurally similar positions in other fibrillin-1 cbEGF domains should lead to a similar phenotype as E1297G. However, the few previous studies that have reported amino acid changes at similar positions in other fibrillin-1 cbEGF-like domains have associated this change with Marfan syndrome [25]. In line with previous reports [26,27], this observation highlights the importance of domain context for studying the molecular effect of fibrillin-1 mutations [26,27]. For example, mutations that change a calcium-binding asparagine to serine in cbEGF33 (N2183S) lead to increased proteolytic susceptibility whereas the same mutation in the same position in cbEGF32 (N2144S) does not affect proteolytic susceptibility[27]. Moreover, different cbEGF domains in fibrillin-1 are involved in the interaction with different ECM molecules[28], it is thus possible that similar mutations in different fibrillin-1 cbEGF domains affect interaction with different molecules and lead to different molecular phenotypes. The points discussed here emphasize the need for future functional studies to understand the mechanism via which E1297G affect fibrillin-1 deposition in skin and height.

Table S8.1: Disease, phenotypes, and traits caused by mutations in *FBN1*. Mutations in FBN1 can lead to clinical phenotypes that, among other symptomes, show abnormally tall or short staturt.

Disease	OMIM ID	Inheritance	Most common height phenotype (if any)
Acromicric dysplasia	102370	AD	short statuture
Ectopia lentis	129600	AD	
Geleophysic dysplasia	614185	AD	short statuture
Marfan lipodystrophy syndrome	616914	AD	tall statuture
Marfan syndrome	154700	AD	tall statuture
MASS syndrome	604308	AD	tall statuture
Stiff skin syndrome	184900	AD	short statuture
Weill-Marchesani syndrome	608328	AD	short statuture
Shprintzen-Goldberg syndrome	182212	AD	tall statuture

Table S8.2: Demographic information of clinical examination participants. Skin biopsies were obtained from 11 including: 2 with C/C, 2 with C/T, and 7 with T/T genotypes at rs200342067. rheum: rheumatological, derm: dermatological, MAC: minor allele count, EUR: European ancestry proportion, AFR: African ancestry proportion, NAT: Native American ancestry proportion, ASI: Asina ancestry proportion.

	age	**gender**	**height (cm)**	**EUR (%)**	**AFR (%)**	**NAT (%)**	**ASI (%)**	**rs200342067* C (MAC)**	**Skin biopsy**	**rheum exam**	**derm exam**
Individual 1	64	F	146	2.0	0.3	97.5	0.2	2	yes	none	thick skin
Individual 2	35	F	144	21.3	0.0	78.2	0.5	2	yes	none	none
Individual 3	30	F	146	5.2	0.3	93.2	1.3	1	no	none	none
Individual 4	60	M	164	16.6	2.4	81.0	0.1	1	no	none	none
Individual 5	56	M	164	12.5	0.3	86.5	0.8	0	no	none	none
Individual 6	37	F	160	20.1	1.6	78.2	0.1	0	yes	none	none
Individual 7	30	F	167	7.2	0	92.8	0	0	no	none	none
Individual 8	60	F	157	4.7	0	95.3	0	0	yes	none	none
Individual 9	46	F	153	7.1	0	92.9	0	0	no	none	none
Individual 10	44	F	150	3.2	0.2	95.2	1.5	0	no	none	none
Individual 11	36	F	154	8.7	0.3	67.8	23.2	0	no	none	none

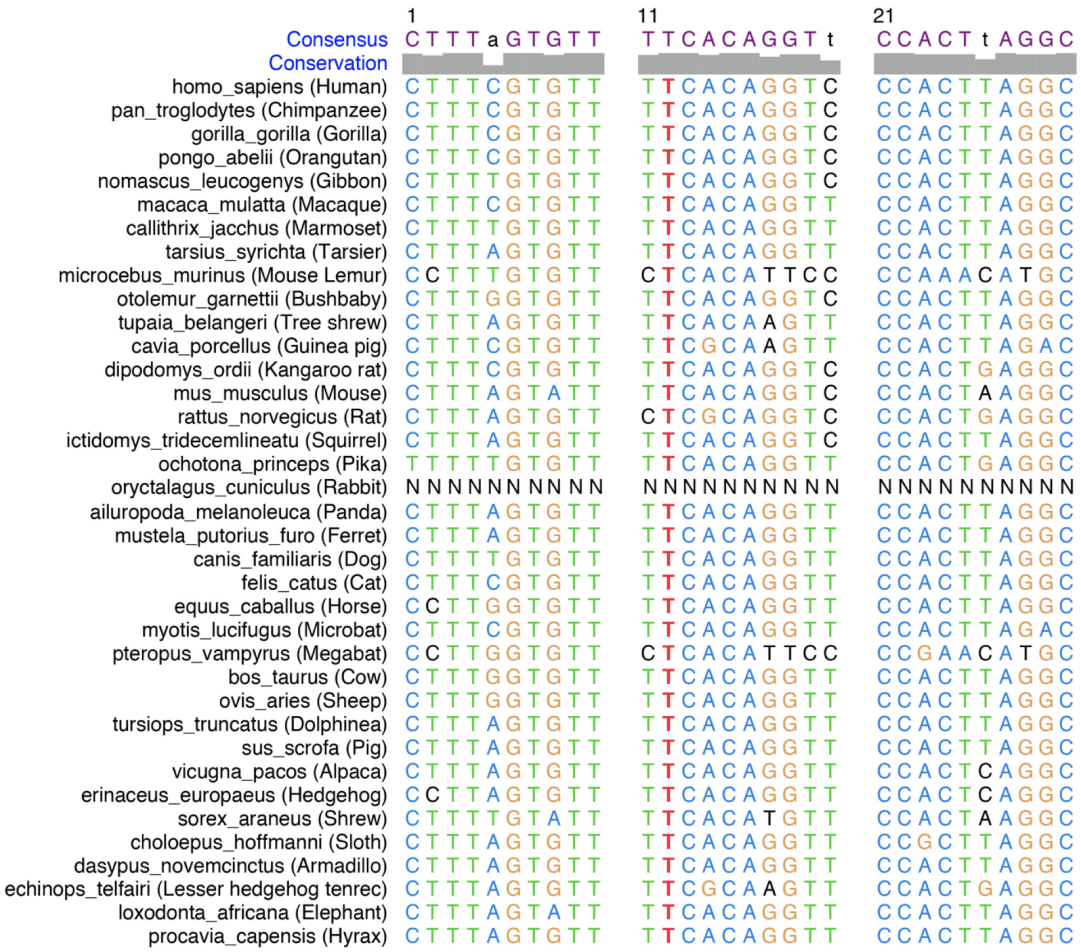

Figure S8.1: Multiple sequence alignment around rs20034206 in 37 eutherian mammals. rs200342067 changes a conserved T allele (ancestral, shown in red) to a C allele (derived). Sequence alignments were obtained from Ensembl GRCh37 release 95.

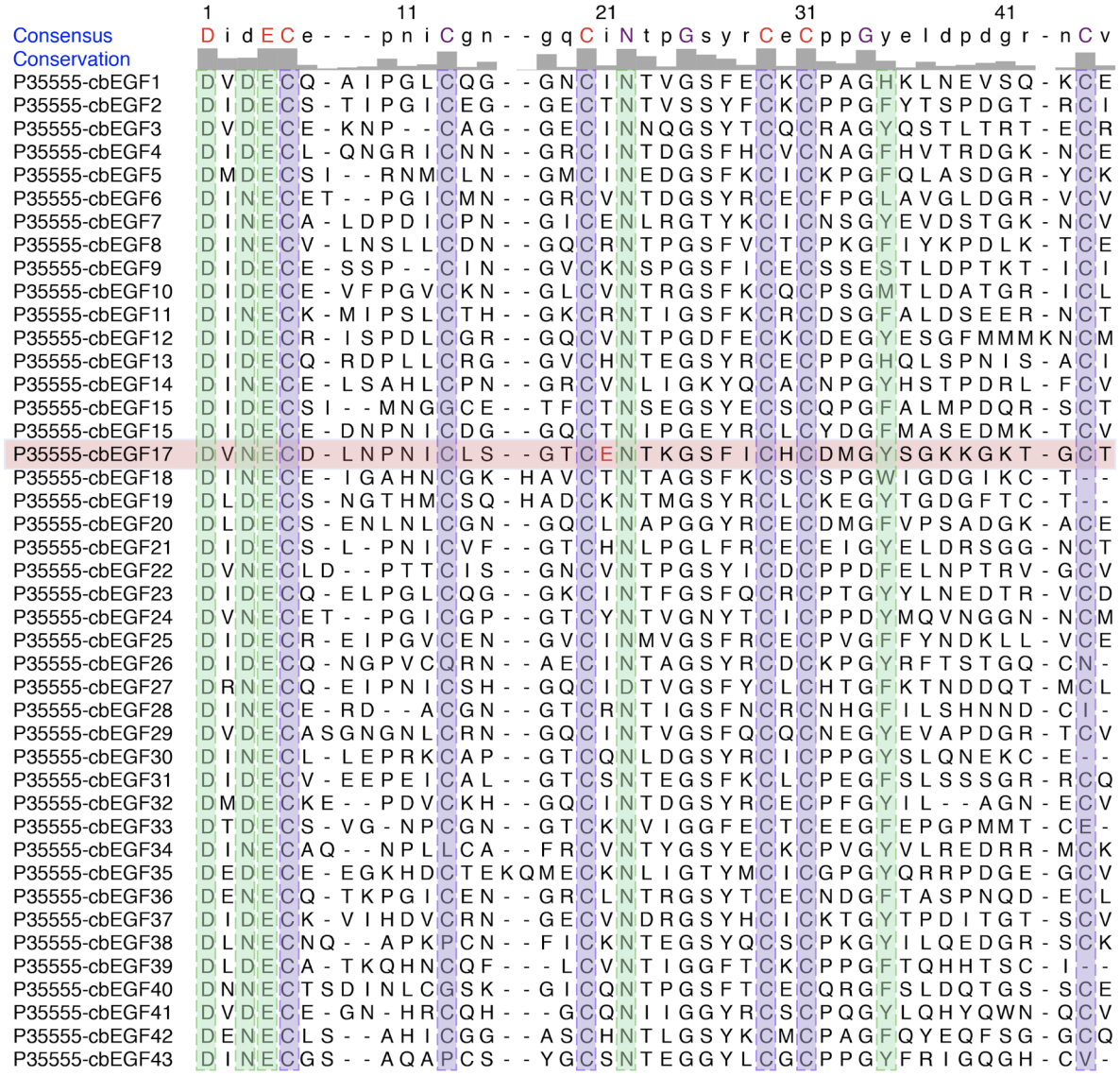

Figure S8.2: Multiple amino acid sequence alignment of fibrillin-1's cbEGF domains. The primary structure of fibrillin-1 cbEGF domains consists of six cysteine residues (highlighted purple) that are involved in forming three disulfide bonds and five conserved residues involved in calcium binding (highlighted green). fibrillin-1 E1297G (shown in red) is located in the cbEGF domain 17 (red rectangle) and is surrounded by a conserved cysteine and a conserved asparagine residue. Protein sequences were obtained from Uniprot (Uniprot ID = P35555).

Supplementary References:

1. 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).
2. Sethuraman, A. Estimating Genetic Relatedness in Admixed Populations. *G3* **8**, 3203–3220 (2018).
3. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
4. Ramstetter, M. D. *et al.* Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics* **207**, 75–82 (2017).
5. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
6. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
7. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. **89**, 82–93 (2011).
8. 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).
9. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
10. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
11. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
12. Mostafavi, H., Harpak, A., Conley, D., Pritchard, J. K. & Przeworski, M. Variable prediction accuracy of polygenic scores within an ancestry group. *bioRxiv* 629949 (2019) doi:10.1101/629949.
13. 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.* (2018) doi:10.1093/hmg/ddy271.

14. Consortium, 1000 *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
15. 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).
16. Johnson, K. E. & Voight, B. F. Patterns of shared signatures of recent positive selection across human populations. *Nat Ecol Evol* **2**, 713–720 (2018).
17. Akbari, A. *et al.* Identifying the favored mutation in a positive selective sweep. (2018) doi:10.1038/nmeth.4606.
18. Sturm, R. A. & Duffy, D. L. Human pigmentation genes under environmental selection. *Genome Biol.* **13**, 248 (2012).
19. Harris, D. N. *et al.* Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6526–E6535 (2018).
20. Sakai, L. Y., Keene, D. R., Renard, M. & De Backer, J. FBN1: The disease-causing gene for Marfan syndrome and other genetic disorders. *Gene* **591**, 279–291 (2016).
21. Faivre, L. *et al.* In frame fibrillin-1 gene deletion in autosomal dominant Weill-Marchesani syndrome. *J. Med. Genet.* **40**, 34–36 (2003).
22. 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).
23. Jensen, S. A., Robertson, I. B. & Handford, P. A. Dissecting the fibrillin microfibril: structural insights into organization and function. *Structure* **20**, 215–225 (2012).
24. 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).
25. 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).

26. Jensen, S. A., Corbett, A. R., Knott, V., Redfield, C. & Handford, P. A. Ca2+-dependent interface formation in fibrillin-1. *J. Biol. Chem.* **280**, 14076–14084 (2005).
27. 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).
28. Jensen, S. A. & Handford, P. A. New insights into the structure, assembly and biological roles of 10–12 nm connective tissue microfibrils from fibrillin-1 studies. *Biochem. J* **473**, 827–838 (2016).