Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

RESULTS

The overall analysis strategy is illustrated in Supplementary Figure 1. We first performed a GWAS meta-analysis of adult height using the summary statistics from 79 studies consisting of 253,288 individuals of European ancestry (Online Methods). We identified 697 SNPs that reached genome-wide significance ($P < 5 \times 10^{-8}$) in this larger study (697 SNPs) clustered in 423 loci, with a number (thousands) of causal variants.

A full list of authors and affiliations appears at the end of the paper.
of the test statistics. We addressed these possibilities by comparing our results with those obtained using more stringent corrections for stratification (linear mixed models; LMMs) and with results obtained in subsets of studies in which a purely family-based analysis was feasible, and by performing a within-family prediction analysis that partitioned the variance in the genetic predictor into the contributions of true associations and population stratification.

Our LMM analyses, performed in a subset of 15 individual studies comprising 59,380 individuals, provided strong evidence that the inflated statistics were driven predominantly by the highly polygenic nature of the trait. This approach uses a genomic relationship matrix (GRM) calculated from genome-wide SNP data to correct for distant relatedness between all pairs of individuals in a study. We obtained a single $\lambda_{GC}$ of 1.20. This value is entirely consistent with the single $\lambda_{GC}$ of 1.94 obtained from the standard GWAS analysis of the same individuals and a single $\lambda_{GC}$ of 1.94 obtained from the analysis of all 253,288 individuals (Supplementary Table 4). Because this approach might be overly conservative for a strongly genetic and highly polygenic trait, each study additionally repeated the analyses for each chromosome using a GRM generated from the remaining 21 chromosomes or, in the case of the largest study (WGHS), repeated the analyses for all odd-numbered chromosomes using a GRM generated from the even-numbered chromosomes and vice versa. The single-$\lambda_{GC}$ inflation factor for this analysis, 1.23, is also entirely consistent with the standard GWAS results (Online Methods, Supplementary Table 4 and Supplementary Note).

Our family-based analyses also provided strong evidence that the inflated statistics were driven predominantly by the highly polygenic nature of height. We assessed whether variants that reached genome-wide significance after single-$\lambda_{GC}$ correction replicated in family-based analyses of up to 25,849 samples (effective sample size of 14,963; using methods that are immune to stratification (Online Methods, Supplementary Tables 5 and 6, and Supplementary Note)). We identified genome-wide significant associations from a meta-analysis that excluded the family-based samples and tested these associations for replication in the family-based samples; a lower rate of replication than expected could be due to inflation of effect sizes in the discovery sample from ‘winner’s curse’ and/or stratification. Of the 416 genome-wide significant SNPs representing multiple signals selected after the exclusion of the family-based studies, 371 SNPs had a consistent direction of effect (in comparison with 208 expected by chance and 400 expected in the absence of any inflation of estimated

Figure 1 Regional association plots for loci with multiple association signals. (a–d) Examples of multiple signals after approximate conditional joint multiple-SNP (GCTA-COJO) analysis. SNPs are represented by colored symbols according to the index SNP with which they are in strongest LD ($r^2 > 0.4$). In some loci, the majority of signals cluster in and around a single gene, for example, at ACAN (a), ADAMTS17 (b) and PTCH1 (c), whereas at some loci multiple signals cluster through proximity (d).
effect sizes) and 142 SNPs replicated with \( P < 0.05 \) (in comparison with 21 expected by chance and 210 expected in the absence of effect size inflation; Supplementary Table 5). These analyses (particularly the directional consistency) show that most of the loci represent true associations but also show that there is a modest inflation in the effect size estimates, due to stratification and/or winner’s curse. To distinguish between these possibilities, we repeated this analysis, substituting for the family-based samples a random set of studies with similar total effective sample sizes. The number of replicating loci was only slightly lower in the family-based cohorts than in the random samples (12–17 fewer replications attributable to stratification at different \( P \)-value thresholds; Supplementary Table 5). This finding indicates that most of the modest inflation in effect estimates is due to winner’s curse, that a small amount of inflation is due to residual stratification and that few (upper limit of ~15–25; Supplementary Table 5 and Supplementary Note) if any of the loci that reach genome-wide significance after single-\( \lambda_{GC} \) correction are likely to be complete false positives due to stratification (that is, with no real association whatsoever with height).

Variance explained by SNPs at different significance levels

Having established that single-\( \lambda_{GC} \) correction is sufficient to identify SNPs that are likely to be truly associated with height, we next performed a series of analyses using the GWAS data from five independent validation studies to quantify the fraction of phenotypic variance explained by SNPs selected from the GCTA-COJO analyses of the meta-analysis data, which excluded data from the validation studies, at a range of statistical thresholds and to quantify the accuracy of predicting height using these selected SNPs (Online Methods). We first developed a new method that uses within-family prediction to partition the variance of the SNP-based predictor into components due to real SNP effects, errors in estimating SNP effects and population stratification (Online Methods), and we applied the method to data on full-sibling pairs from three of the five validation studies (Online Methods). Consistently across the three studies, all the partitioned variance components increased in size as a less stringent significance level was used for SNP selection in the discovery sample, and the error variance increased more dramatically than the genetic variance when more SNPs selected at a less significant level were included...
in the predictor (Fig. 2a–c). We demonstrated the partitioning of variance due to population stratification by the within-family prediction analyses with and without adjusting for principal components (Supplementary Fig. 5). The results again confirmed that the impact of population stratification on the top associated SNPs was minor and demonstrated that the variation in the predictor due to true SNP effect, estimation error and population stratification was quantifiable. We next inferred, using these partitioned variance components from the within-family prediction analysis, how well different selected sets of SNPs would predict height in independent samples. We showed that the observed prediction accuracy (squared correlation between phenotype and predictor, or $R^2$) in five different population-based cohorts was highly consistent with the values inferred from the within-family analyses, with prediction accuracy peaking at ~17% using the within-family prediction analysis. As shown in equation (19), prediction $R^2$ is not equal to the variance explained but is a function of the variance of true SNP effects and the error variance in estimating SNP effects, in the absence of population structure. At thresholds below genome-wide significance, the variance explained is higher than the variance due to population stratification by the within-family prediction analysis. cSNPs from the HapMap 3 project11.

Table 1

| Threshold | Number of SNPs | $n_g^2$ | SE   | $n_g^2$ | SE   | $n_g^2$ | SE   | $n_g^2$ | SE   |
|----------|---------------|--------|------|--------|------|--------|------|--------|------|
| 5 x 10^{-8} | 675           | 0.164  | 0.016 | 656.0  | 0.040 | 670.0  | 0.013 | 679.0  | 0.025 |
| 5 x 10^{-7} | 887           | 0.187  | 0.017 | 862.0  | 0.045 | 866.0  | 0.017 | 890.0  | 0.028 |
| 5 x 10^{-6} | 1,245         | 0.196  | 0.018 | 1,202  | 0.050 | 1,186  | 0.014 | 1,256  | 0.030 |
| 5 x 10^{-5} | 1,950         | 0.212  | 0.020 | 1,891  | 0.060 | 1,918  | 0.017 | 1,985  | 0.037 |
| 5 x 10^{-4} | 3,754         | 0.248  | 0.024 | 3,671.0| 0.080 | 3,689  | 0.017 | 3,771  | 0.047 |
| 5 x 10^{-3} | 9,693         | 0.297  | 0.035 | 9,403  | 0.126 | 9,548  | 0.025 | 9,677  | 0.070 |
| HM3c      | 1.08           | 0.473  | 0.086 | 1.06   | 0.313 | 1.22   | 0.097 | 1.09   | 0.126 |

SNPs were selected by an approximate conditional and joint multiple-SNP analysis (GCTA-COJO) of the summary statistics from the meta-analysis. The target cohort for variance estimation was excluded from the meta-analysis. $R^2$, heritability, SE, standard error. Values in bold highlight the averages of five studies.

The larger GWAS identifies new biologically relevant genes and pathways

Having shown that ~1% of variants can account for the majority of heritability attributable to common variation, we next considered whether the expanded set of height-associated variants could be used to identify the genomic features and biological pathways of most relevance to normal variation in adult height. To test whether our GWAS could implicate new biology, we used established and novel approaches to test whether the height-associated loci were enriched for functionally relevant variants, genes, pathways and tissues.

Similar to the 180 variants identified in our previous analysis, the 697 variants were non-randomly distributed with respect to functional and putatively functional regions of the genome (Online Methods). We observed that the height-associated variants were enriched for nonsynonymous SNPs (empirical enrichment of 1.2-fold; $P = 0.02$), cis-regulatory effects in blood (empirical enrichment of 1.5-fold; $P = 0.03$), overlap with a curated list of genes that underlie monogenic syndromes of abnormal skeletal growth12 (empirical enrichment of 1.4-fold; $P = 0.013$), associations with apparently unrelated complex traits in the National Human Genome Research Institute (NHGRI) GWAS catalog (empirical enrichment of 2.6-fold; $P < 1 \times 10^{-4}$) and functional chromatin annotations in multiple tissues and cell types (empirical enrichment of 1.8-fold; $P < 1 \times 10^{-3}$) (Supplementary Tables 7–11 and Supplementary Note).

The greater resolution for the height-associated variants provided by increased sample size, in combination with improved gene prioritization and gene set enrichment approaches, resulted in the identification of multiple new tissues, gene sets and specific genes that were highly likely to be involved in the biology of skeletal growth. Specifically, using a variety of established and novel pathway methods, we identified ~3 times as many enriched pathways and prioritized ~5 times as many genes (including genes newly prioritized in previously identified loci) in comparison to the results derived from identical pathway methods applied to the previous GWAS of 133,000 individuals (Table 2).

We first focused on existing pathway and gene prioritization methods: (i) MAGENTA13, a method designed to identify gene sets enriched in GWAS data, and (ii) GRAIL14, which uses the published literature to highlight connections between likely relevant genes within GWAS-identified loci. As expected, the GRAIL and MAGENTA analyses confirmed several previously identified gene sets and pathways clearly relevant to skeletal growth, but in the larger sample they also provided evidence for additional known and new
We found that genes within loci associated with height were enriched and new loci provided the additional power needed to identify new growth-associated gene sets, and the combined analysis of both old and new loci provided the additional power needed to identify new gene sets (Table 3 and Supplementary Table 14).

The DEPICT analysis also prioritized tissues and individual genes. Associated genes, gene sets and protein complexes not identified in our previous smaller study (for example, fibroblast growth factor (FGF) signaling, WNT signaling, osteoglycin and other genes related to bone metabolism) (Supplementary Table 14). DEPICT strongly prioritized genes that did not have published annotations related to growth pathways but were predicted to be in gene sets that were both enriched in the associated loci and clearly connected to growth. These included genes newly predicted to be in pathways related to cartilage or bone development (Supplementary Table 16). We also showed that a subset of the 697 height-associated SNPs that represented lead cis expression quantitative trait loci (eQTLs) in blood defined 75 genes that were collectively enriched for expression in cartilage (P = 0.008) (Supplementary Table 8 and Supplementary Note).

We used DEPICT to prioritize 649 genes (at FDR < 0.05) within height-associated loci (Table 3 and Supplementary Table 16). Of these 649 genes, 202 genes (31%) were significant in the GRAIL analysis (Supplementary Tables 13 and 16) and/or overlapped with a list of genes involved in abnormal skeletal growth syndromes that we assembled from the Online Mendelian Inheritance in Man (OMIM) database (n = 40; Supplementary Tables 9 and 16). Many other newly prioritized genes had additional supporting evidence (Supplementary Table 16), including specific expression in the growth plate and/or connections to relevant pathways (for example: GLI2 and LAMA5 (Hedgehog signaling); FRS2 (FGF signaling); AXIN2, NATC1, C11N81, FBXW11, WNT4, WNT5A and VANGL2 (WNT/β-catenin signaling); SMAD3 and MTOR (transforming growth factor (TGF)-β and/or mTOR signaling); WWP2-MIR140, IBSP, SHOX2 and SP3 (the corresponding genes are required in mice for proper bone and cartilage formation); CHSY1, DSE and PCOLCE2 (glycosaminoglycan/collagen metabolism); and SCARA3, COP2Z, TBX18, CRISPLD1 and SLIT3 (differential expression in growth plate and predicted to be in highly relevant pathways)).

DEPICT also prioritized the genes that were new candidates for having a role in skeletal growth. The genes newly and strongly implicated in this study included not only genes with obvious relationships to skeletal biology, such as SOX5 and collagen genes, but also genes that have no clear published connection to skeletal growth and likely represent as-yet-unknown biology (Supplementary Table 16).
Table 3 Significantly prioritized new human growth-associated genes

| Locus (height SNP) | Gene | New locus | Prioritization P value | Lines of supporting evidence | Top-ranking reconstituted gene sets |
|-------------------|------|-----------|------------------------|----------------------------|----------------------------------|
| rs10748128        | FRS2 | No        | 1.0 × 10^{-16}         | 7                           | PI3K cascade (REACTOME, P = 6.2 × 10^{-13})  
|                   |      |           |                        |                             | Chronic myeloid leukemia (KEGG, P = 1.6 × 10^{-12})  
|                   |      |           |                        |                             | Response to fibroblast growth factor stimulus (GO, P = 5.4 × 10^{-11})  
|                   |      |           |                        |                             | Growth factor binding (GO, P = 2.6 × 10^{-14})  
|                   |      |           |                        |                             | Regulation of osteoblast differentiation (GO, P = 2.3 × 10^{-11})  
|                   |      |           |                        |                             | WNT protein binding (GO, P = 1.9 × 10^{-12})  
| rs2166898         | GLI2 | Yes       | 4.4 × 10^{-10}         | 7                           | Short mandible (MP, P = 3.3 × 10^{-19})  
|                   |      |           |                        |                             | Respiratory system development (GO, P = 3.1 × 10^{-17})  
|                   |      |           |                        |                             | Abnormal ulna morphology (MP, P = 1.9 × 10^{-15})  
| rs526896-rs9327705| TBX4 | No        | 9.9 × 10^{-9}          | 7                           | Small thoracic cage (MP, P = 6.9 × 10^{-14})  
|                   |      |           |                        |                             | Short ribs (MP, P = 2.7 × 10^{-8})  
|                   |      |           |                        |                             | Short sternum (MP, P = 6.5 × 10^{-7})  
|                   |      |           |                        |                             | Partial lethality throughout fetal growth and development (MP, P = 1.2 × 10^{-18})  
| rs16860216        | SOX8 | No        | 0.016                  | 7                           | Growth factor binding (GO, P = 2.6 × 10^{-14})  
|                   |      |           |                        |                             | TGFB1 protein complex (InWeb, P = 6.3 × 10^{-12})  
|                   |      |           |                        |                             | Chromatin binding (GO, P = 6.4 × 10^{-17})  
|                   |      |           |                        |                             | Nuclear hormone receptor binding (GO, P = 2.4 × 10^{-12})  
|                   |      |           |                        |                             | RBBP4 protein complex (InWeb, P = 1.3 × 10^{-11})  
|                   |      |           |                        |                             | WNT16 protein complex (InWeb, P = 1.9 × 10^{-8})  
| rs6746356         | SP3  | Yes       | 1.0 × 10^{-16}         | 6                           | Partial lethality throughout fetal growth and development (MP, P = 1.2 × 10^{-18})  
| rs3923086         | AXIN2 | Yes       | 2.2 × 10^{-10}         | 6                           | Tissue morphogenesis (GO, P = 4.1 × 10^{-20})  
| rs3790086         | LTB1 P | No       | 1.3 × 10^{-13}         | 6                           | Abnormal skeleton morphology (MP, P = 1.1 × 10^{-25})  
|                   |      |           |                        |                             | TGF β signaling pathway (KEGG, P = 3.8 × 10^{-13})  
| rs2034172         | WNT5A | Yes       | 4.3 × 10^{-13}         | 6                           | Complete embryonic lethality during organogenesis (MP, P = 4.9 × 10^{-21})  
| rs3915129         | CTNNB1 | Yes      | 3.5 × 10^{-12}         | 6                           | Short mandible (MP, P = 3.3 × 10^{-19})  
| rs12330322        | BMP2 | No        | 5.6 × 10^{-10}         | 6                           | Small basiphemid bone (MP, P = 8.9 × 10^{-17})  
| rs10958476-rs6999671| BMP6 | No        | 2.9 × 10^{-8}          | 6                           | TGFB β signaling pathway (KEGG, P = 3.8 × 10^{-15})  
| rs564914          | SOX5 | Yes       | 4.6 × 10^{-7}          | 6                           | Growth factor binding (GO, P = 2.6 × 10^{-14})  
| rs17807185        | WNT4 | Yes       | 4.6 × 10^{-7}          | 6                           | Abnormal cartilage morphology (MP, P = 1.9 × 10^{-13})  
|                   |      |           |                        |                             | Short limbs (MP, P = 2.8 × 10^{-13})  
|                   |      |           |                        |                             | Morphogenesis of an epithelium (GO, P = 2.3 × 10^{-17})  
|                   |      |           |                        |                             | Gland development (GO, P = 5.4 × 10^{-16})  
|                   |      |           |                        |                             | Basal cell carcinoma (KEGG, P = 1.5 × 10^{-12})  
| rs8042424         | CHSY1 | No        | 1.0 × 10^{-16}         | 7                           | Abnormal cartilage morphology (MP, P = 1.9 × 10^{-13})  
|                   |      |           |                        |                             | Abnormal bone ossification (MP, P = 2.1 × 10^{-12})  
|                   |      |           |                        |                             | Signaling by transforming growth factor β (KEGG, P = 9.9 × 10^{-10})  
|                   |      |           |                        |                             | Abnormal spongiospinothoblast layer morphology (MP, P = 3.2 × 10^{-16})  
|                   |      |           |                        |                             | Decreased length of long bones (MP, P = 2.7 × 10^{-12})  
| rs7652177         | FND3C8 | No        | 1.0 × 10^{-16}         | 5                           | TGFB1 protein complex (InWeb, P = 5.2 × 10^{-8})  
| rs7284476         | TRIOB P | Yes      | 1.0 × 10^{-16}         | 5                           | Negative regulation of cell proliferation (GO, P = 4.3 × 10^{-17})  
| rs2149163-rs3927536| BNC2 | No        | 1.0 × 10^{-16}         | 5                           | Abnormal vitelline vasculature morphology (MP, P = 1.7 × 10^{-15})  
| rs3790086         | WWP2 | Yes       | 1.0 × 10^{-16}         | 5                           | β-catenin binding (GO, P = 3.0 × 10^{-5})  

The table lists 20 genes prioritized by DEPICT. Genes are ranked by the number of lines of supporting evidence and the DEPICT P value (Supplementary Table 16). Lines of supporting evidence for a gene included (1) annotation by GRAIL, (2) differential expression within the growth plate, (3) specific expression within the growth plate, (4) a mouse skeletal growth phenotype, (5) DEPICT FDR < 0.05, (6) membership in a DEPICT-prioritized gene set, (7) being the nearest gene to the lead SNP, (8) the lead SNP being an eQTL for the gene and (9) the lead SNP being a missense SNP for the gene. Because 20 of the 30 top-ranked genes were in a curated list of genes known to cause syndromes of abnormal skeletal growth, these ‘OMIM genes’ are not shown here. The top 15 genes with previous literature support (based on GRAIL) are shown, followed by the top 5 new genes. Each gene is accompanied by the significantly enriched reconstituted gene sets in which it appears (DEPICT gene set enrichment analysis). GO, Gene Ontology; MP, Mice Phenotypes from Mouse Genome Informatics database; InWeb, protein-protein interaction complexes; KEGG and REACTOME databases.
DISCUSSION

By performing a large GWAS on adult height, a highly heritable polygenic trait, we have provided answers to several current questions of relevance to the genetic study of polygenic diseases and traits. First, we showed that, by conducting larger GWAS, we can identify SNPs that explain a substantial proportion of the heritability attributable...
to common variants. As hypothesized by Yang et al.\textsuperscript{5}, the heritability directly accounted for by variants identified by GWAS and inferred by whole-genome estimation approaches are converging with increasing sample size. The variance explained by genome-wide significant SNPs has increased from 3–5\% with discovery samples of ~25,000 (ref. 19) to 10\% with a discovery sample size of ~130,000 (ref. 6) and 16\% with a discovery sample size of 250,000 (this study), and the variance explained from all captured common SNPs is ~50\% (refs. 4, 5). The variance explained by genome-wide significant SNPs on a chromosome is also proportional to chromosome length, consistent with the conclusion made by Yang et al.\textsuperscript{5} using all SNPs (Supplementary Fig. 9).

Our new results show that ~21\%, ~24\% and ~29\% of phenotypic variance in independent validation samples is captured by the best ~2,000, ~3,700 and ~9,500 SNPs, respectively, selected in the discovery samples (Table 1), and that the correlation between actual and predicted height in independent samples from the same population has increased to 0.41 (maximum prediction $R^2 = 0.41^2 = 0.17$; Fig. 2d).

The results are consistent with a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants, located throughout the genome but clustered in both a biological and genomic manner. Such a genetic architecture may be described as pseudo-infinite and may characterize many other polygenic traits and diseases. There is also strong evidence of multiple alleles at the same locus segregating in the population and for associated loci overlapping with mendelian forms, suggesting a large but finite genomic mutational target for height and effect sizes ranging from minute (<1 mm; ~0.01 s.d.) to gigantic (>300 mm; >3 s.d.) in the case of monogenic mutations.

It has been argued that the biological information emerging from GWAS will become less relevant as sample sizes increase because, as thousands of associated variants are discovered, the range of implicated genes and pathways will lose specificity and cover essentially the entire genome\textsuperscript{20}. If this were the case, then increasing sample sizes would not help to prioritize follow-up studies aimed at identifying and understanding novel biology and the associated loci would blanket the entire genome. Our study provides strong evidence to the contrary: the identification of many hundred and even thousand associated variants can continue to provide biologically relevant information. In other words, the variants identified in larger sample sizes both display a stronger enrichment for pathways clearly relevant to skeletal growth and prioritize many additional new and relevant genes. Furthermore, the associated variants are often non-randomly and tightly clustered (typically separated by <250 kb), resulting in the frequent presence of multiple associated variants in a locus. The observations that genes and especially pathways are now beginning to be implicated by multiple studies suggests that the larger set of results retain biological specificity but that, at some point, a new set of associated variants will emerge as relevant to skeletal growth and prioritize many additional new and relevant genes. Further study of the associated variants is likely to reveal new biological hypotheses and motivate future research into the basis of human biology and disease.

URLs. Genetic Investigation of Anthropometric Traits (GIANT) Consortium, http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium; Mouse Genome Informatics, http://www.informatics.jax.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

A full list of acknowledgments appears in the Supplementary Note.

AUTHOR CONTRIBUTIONS

Steering committee overseeing the consortium: G.R.A., T.L.A., I.B., S.I.B., M. Boehnke, I.B.D., P.D., C.S.F., T.M.F., L.C.G., I.M.H., J.N.H., D.J.H., E.I., R.C.K., R.J.F.L., M.I.M., K.L. Mohlke, K.E.N., J.R.O., D. Schlessinger, D.P.S., U.T., and C.M.v.d. Writing group (wrote, edited and commented on the manuscript): S.I.B., D.J.C., A.Y.C., T.E., T.M.F., J.N.H., E.I., T.H.P., S.V., P.M.V., M.N.W., A.R.W. and J.Y. Data preparation group (checked and prepared data from contributing cohorts for meta-analyses): D.C.C.–C., F.R.D., T.E., F.T., F. Ferreira, S.G., I.M.H., Z.K., C.M.L., A.E.L., R.J.F.L., J. Luan, R.M., I.C.R., A. Scherag, E.K.S., S.V., T.W.R., A.R.W. and T. Workalemahu. Height meta-analyses group (GWAS and Metabochip) (analyses specific to the manuscript): T.E., T.M.F. (chair), S.V., P.M.V., A.R.W (lead meta-analyses) and J.Y. (lead joint-effects and approximate conditional analyses). Mixed linear model analyses: J.S.B., M. Boehnke, D.I.C., A.Y.C., K.E., T.M.F. (chair), S.G., J.N.H., I.H.Z., E.I., A.U.J., Z.K., R.J.F.L., J. Luan, A. Metspalu, E.M., I.R.O., A.L.P., A.G.U., S.V., P.M.V., M.N.W. (A.R.W) (lead) and J.Y. Large $\lambda$ group: T.M.F., J.N.H., P.M.V., M.E. Goddard, A.L.P., M.N.W., J.Y. and G.R.A. Family transmission analyses: G.R.A., N.A., I.B.D., Y.D., C.M.V.D., J.N.H. (chair), E.I., J.R.O., E.P., S.V. (lead). P.M.V. and J.Y. Variance, heritability and parameter estimation analyses: K.E., M.E. Goddard, M.I.M., A.A.E.V., P.M.V. (chair), M.N.W., A.R.W and J.Y. (lead). Biological enrichment and pathway analyses: T.E. (lead biological enrichment analyses), J.N.H. (chair) and T.H.P. (lead pathway analyses), ENCODE working group: M.I.B., G.L. (chair) and K.S.L. Gene expression (eQTL) working group: J. Baron, T.E. (chair), I.L. Franke, J. Karjalainen, J.C.L., A. Metspalu, E.R., J.E.P. and H.-J.W. (lead). Other contributions: (DEPICT) R.F., I.L. Franke, J. Karjalainen and T.H.P.

Project design, management and coordination of supporting studies

Previous GWAS: (AGES) V. Gudnason, T.R.H.; (AMISH) A.R.S.; (ARIC) K.E.N.; (B358 T1D CONTROLS) P.D.S.; (B358 WTCCC) D.P.S.; (BRIGHT) M.J.B., N.J.S.; (CAPS) E.I.; (CHS) J.I.R.; (COLAUS) J.S.B., S. Bergmann; (CROATIA-Vis HR; (deCODE) K. Stefansson, U.T.; (DGI) L.C.G.; (EGCUT) A. Metspalu; (EPIC-Norfolk) N.I.W.; (FENLAND) N.I.W.; (Finnish Twin Cohort) J. Kaprio, K. Silventoinen; (FRAM) L.A.C.; (FUSION) R.N.B., M. Boehnke; (GerMIFS II) J.E., C. Hengstenberg; GerMIFS II) H. Schunkert; (H2000) S. Koskenniemi; (HPFS) D.J.H.; (KORA S4) C.G., A.P.; (MICROS) A.A.H., P.P.P.; (NFBfC6) M.-R.I., S. Sebert; (NHS) D.J.H.; (NSPHS) U.G.; (NTRNEda) D.J.; (ORCADES) H.C.; (PLCO) S.I.B., S.J.C.; (RS I) C.M.v.D., A. Hofman, M. Kayser, F. Rivadeneira, A.G.U.; (RUNMC) L.A.K.; (SardiNIA) G.R.A.; (SASABAC) E.I.; (SHIP) R.B., H.V.; (WGHS) P.M.R.; (WTCCC-2CAD) A.S.H., N.S.; (WTCCC-T2D) C.M.L., M.I.M.; (Young Finns Study (YFS)) T.L.; O.T.R.

New GWAS: (ASCOT) M.J.C., P.S.; (ATGC) P.I.W.d.B., D.W.H.; (Athero-Express Biobank Studies) F.W.A., H.M.D.R., F.L.M., G.P.; (B-PROOF) R.D.-R., L.C.P.M.d.G., N.M.V.s., N.v.d.V.; (BLSA) L. Ferrucci; (CHLNS) K.L. Mohlke; (CHOROGENE) M.P., J. Sinisalo; (DGEB) S.C., P.F.; (DNBS) M. Melbye, J.C.M.; (EGCUT) A. Metspalu; (eMERGE) M.G.H.; (ERF) B.A.O., C.M.v.D.; (FamHS) I.B.R.; (FINGESTURE) J.C.T.; (GOOD) C.O.; (HBCS) J.G.E.;
Phenotype coordination of contributing studies

Previous GWAS: (AGES) A.V. Smith; (ASCOT) A.L. Kondo, K.E.N.; (ATCG) D.W. H.; (Athero-Express Biobank Studies) T. Ferreira, A. Mahajan, R.M.; (B58C) A.V. Smith; (B58C T1D CONTROLS) D.P.S.; (B58C WTCDD) D.P.S.; (B58C WTCCC) D.P.S.; (B58C WTCCC-T2D) A.B., A.T.H.

New GWAS: (ATCG) P.I.W.d.B., P.J.M., S.R.; (Athero-Express Biobank Studies) T. Ferreira; (ARIC) K.L. Monda, K.E.N.; (B1958C) T. K. Silventoinen, K.

1. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. 52, 399–433 (1918).
2. Silventoinen, K. et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 6, 399–408 (2003).
3. Visscher, P.M. et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2, e41 (2006).

4. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).

5. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43, 519–521 (2011).

6. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832–838 (2010).

7. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. 44, 369–375 (2012).

8. 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).

9. Voight, B.F. et al. The Metaiod, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 8, e1002793 (2012).

10. Yang, J. et al. Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. 19, 807–812 (2011).

11. Altshuler, D.M. et al. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58 (2010).

12. Lui, J.C. et al. Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. Hum. Mol. Genet. 21, 5193–5201 (2012).

13. Segel, A.V., Groop, L., Mootha, V.K., Daly, M.J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6, e1001058 (2010).

14. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 5, e1000534 (2009).

15. de Pontual, L. et al. Germ-line deletion of the miR-17–92 cluster causes skeletal and growth defects in humans. Nat. Genet. 43, 1026–1030 (2011).

16. Morales, J. et al. Homozygous mutations in ADAMTS10 and ADAMTS17 cause lenticular myopia, ectopia lentis, glaucoma, spherophakia, and short stature. Am. J. Hum. Genet. 85, 558–568 (2009).

17. Li, Y. et al. Tentamy premial brachydactyly syndrome is caused by loss-of-function mutations in chorionb1 synthe 1, a potential target of BMP signaling. Am. J. Hum. Genet. 87, 757–767 (2010).

18. Tan, J. et al. Loss of CHSY1, a secreted FRINGE enzyme, causes syndromic brachydactyly in humans via increased NOTCH signaling. Am. J. Hum. Genet. 87, 768–778 (2010).

19. Visscher, P.M. Sizing up human height variation. Nat. Genet. 40, 489–490 (2008).

20. Goldstein, D.B. Common genetic variation and human traits. N. Engl. J. Med. 360, 1696–1698 (2009).

21. Nösslein-Volhard, C., Wieschaus, E. & Klüdgen, H. Mutations affecting the pattern of the larval cuticle in Drosphiphila melanogaster: zygotic loci on the second chromosome. Roux's Arch. Dev. Biol. 193, 267–282 (1984).
ONLINE METHODS

A summary of the methods, together with a full description of genome-wide association analyses and follow-up analyses, can be found here and in the Supplementary Note. Written informed consent was obtained from every participant in each study, and the study was approved by relevant ethics committees.

Genome-wide association study meta-analysis. We combined the height summary association statistics from 79 GWAS in a meta-analysis of 253,288 individuals using the same methods and studies as previously described[6] and additional studies as described in Supplementary Tables 17–19. Meta-analysis was performed on a total of 2,550,858 autosomal single nucleotide polymorphisms (SNPs) using the inverse variance fixed-effects method with METAL[22].

GCTA-COJO: conditional and joint multiple-SNP analysis. We used GCTA-COJO analysis[7,8] to select the top associated SNPs. This method uses the summary statistics from the meta-analysis and LD correlations between SNPs estimated from a reference sample to perform a conditional association analysis[7]. The method starts with an initial model of the SNP that shows the strongest evidence of association across the whole genome. It then implements the association analysis conditioning on the selected SNP(s) to search for the top SNPs iteratively one by one via a stepwise model selection procedure until no SNP has a conditional P-value that passes the significance level. Finally, all the selected SNPs are fitted jointly in the model for effect size estimation. We used 6,654 unrelated individuals from the ARIC cohort as the reference sample for LD estimation. There were ~3.0 million SNPs included in the original meta-analysis. We included in this analysis only the SNPs (~2.48 million) in HapMap 2 for which we had a sample size of >50,000. We used the genome-wide significance level \( P < 5 	imes 10^{-8} \) (as reported in Supplementary Table 1).

Metabochip replication. We combined the height summary association statistics from 37 independent studies genotyped using the Illumina Metabochip array[9] in a meta-analysis of 80,067 individuals of European ancestry (Supplementary Tables 20–22). Each study tested the association for each genotyped SNP using the same quality control procedures, height transformation, adjustment and inheritance model as described for the GWAS analysis. Genomic control correction was applied to the results for each study before meta-analysis, using a set of 4,427 SNPs associated with QT interval to control study-specific inflation factors. We used the inverse variance fixed-effects meta-analysis method.

Validation: linear mixed model–based association analysis. Each of 15 studies (59,380 individuals) used genome-wide SNP information to calculate a GRM for all pairs of individuals and used this matrix to correct association statistics for cryptic relatedness and population stratification. Each study used an LMM for all pairs of individuals and used this matrix to correct association statistics (59,380 individuals) used genome-wide SNP information to calculate a GRM.

Validation: within-family (transmission) association analyses. A pure transmission-based analysis was performed in 7 cohorts for SNPs representing 416 signals of association (Supplementary Note), selected after repeating meta-analysis excluding these studies with single- \( \lambda_{GC} \) correction. The filtering out of SNPs with low imputation quality in the studies was followed by applying the inverse variance method of meta-analysis to the family-based results. Because of the presence of related individuals, family-based studies have lower power at a given sample size. For each study, we calculated the effective sample size (the size of a sample of unrelated individuals that would have equivalent power; see the Supplementary Note and Winkler et al.[14]). Estimation of winner’s curse in our data set was performed by repeating the meta-analysis excluding either the family-based studies or random sets of studies from GIANT matched by effective sample size to the family-based studies. Independent genome-wide significant loci were selected from each meta-analysis. Power for replication in the excluded samples was estimated at different \( P \)-value thresholds, and the deficit in replication (number of replications expected minus the number observed replications) was calculated. The contribution of winner’s curse to the deficit in replication was estimated as the average deficit across the three sets of random non-family-based cohorts. By subtracting this deficit from the deficit observed for the family-based cohorts, we estimated the lack of replication that could be attributed to stratification (either inflation of effect size for true associations or false positive associations).

Variance and heritability explained. We used GCTA-COJO analysis to select the top associated SNPs at a range of stringent significance levels (\( 5 	imes 10^{-3}, 5 	imes 10^{-4}, 5 	imes 10^{-5}, \ldots, 5 	imes 10^{-8} \)) for estimation and prediction analyses. We then quantified the variance explained by the selected SNPs using a three-stage analysis—within-family prediction, GCTA-GREML analysis and population-based prediction—in five validation studies (B-PROOF, FRAM, QIMR, TwinGene and WTCCC-T2D). To avoid sample overlap, we repeated the main GWAS meta-analysis and the multiple-SNP analysis five times, each time excluding one of the five validation studies. This approach ensured complete independence between the data used to discover SNPs and the data used to estimate how much variance in height these SNPs explained and how well they predicted height. For the within-family prediction analyses, we selected 1,622, 2,758 and 1,597 pairs of full siblings from the QIMR, TwinGene and FRAM cohorts, respectively, with 1 sibling pair per family. For the whole-genome estimation and prediction analyses, we used GCTA-GREML[8] to estimate the genetic relatedness between individuals and selected unrelated individuals with pairwise genetic relatedness of <0.025 in each of the 5 studies: B-PROOF \((n = 2,555)\), FRAM \((n = 1,145)\), QIMR \((n = 3,627)\), TwinGene \((n = 5,668)\) and WTCCC-T2D \((n = 1,914)\).

Within-family prediction analysis. We used the SNPs selected from GCTA-COJO analysis to create a genetic predictor (also called a ‘genetic profile score’) for each of the full-sibling pairs using PLINK[25]. We then adjusted the genetic predictor by the first 20 principal components generated from principal-component analysis (PCA)[26]. By comparing the predictors within and between families, we partitioned the variance in the predictor analysis into components due to real SNP effects \( (V_p) \), errors in estimating SNP effect \( (V_e) \) and population structure \( (C_p + C_e) \).

We calculated the weighted average of each of the four \( (v) \) variance components over the three cohorts by their sample size:

\[
\sum_i V_g(i) n_i / \sum_i (n_i)
\]

with the subscript \( i \) indicating the cohort and \( n_i \) indicating the sample size. From the results of these partitioning analyses within families, we can infer what the prediction \( R^2 \) value (equation (19)) and what the proportion of the variance explained by SNPs \( (V_p/V_g) \), with \( V_p \) being the phenotypic variance) would be in a sample of unrelated individuals when using the same set of SNPs. We then tested these inferred values in unrelated samples.

GCTA-GREML analysis. We performed GREML analysis[4] in GCTA[8] to estimate the variance explained by the selected SNPs \( (h^2_g) \) in each of the five validation studies. This method fits the effects of a set of SNPs simultaneously in a model as random effects and estimates the genetic variance captured by all the fitted SNPs without testing the significance of the association of any single SNPs. We combined the estimates of \( h^2_g \) for each of the five studies by the inverse variance approach:

\[
\sum_i (h^2_g(i)/\text{SE}(i))^2 / \sum_i (1/\text{SE}(i)^2)
\]

Population-based prediction analysis. We created a genetic predictor using the selected SNPs for the unrelated individuals in each of the five validation studies. We then calculated the squared correlation \( (R^2) \) between the phenotype and predictor in each validation study and calculated the average prediction \( R^2 \) value weighted by the sample size across the five studies:

\[
\sum_i (R^2(i)n_i) / \sum_i (n_i)
\]
Theory and method to partition the variance in a genetic predictor. Under the assumption of an additive genetic model, the phenotype of a quantitative trait can be written as:

\[ y = g + \epsilon = \sum_i x_i h_i + \epsilon \]  

where \( y \) is the trait phenotype, \( g \) is the total genetic effect of all SNPs, \( x \) is an indicator variable for SNP genotypes, \( h \) is the SNP effect and \( \epsilon \) is the residual.

From this model, the additive genetic variance is:

\[ \text{var}(g) = \sum_i \text{var}(x_i) h_i^2 + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) h_i h_j \]  

with the first component being the expected value of additive genetic variance under linkage equilibrium (LE) and the second component being deviation from the expected value that could be caused by LD, population structure or selection\(^{27}\).

Considering a pair of full siblings in a family, the additive genetic covariance between the siblings is:

\[ \text{cov}(g_1, g_2) = \text{cov}\left(\sum_i x_i h_i, \sum_j x_j h_j\right) = \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) h_i h_j \]  

For full siblings:

\[ \text{cov}(x_i, x_j) = 1/2 \text{var}(x_i) \]

for SNPs that are in LD and

\[ \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) h_i h_j = \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) h_i h_j \]

for SNPs that are not in LD (as shown by both empirical and simulation results).

Let:

\[ V_g = \sum_i \text{var}(x_i) h_i^2 + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) h_i h_j \]  

\[ C_g = \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) h_i h_j \]

where \( g \) is the total genetic effect of all SNPs, \( h \) is the SNP effect and \( \epsilon \) is the residual in estimating \( g \).

Then, the genetic variance is:

\[ \text{var}(g) = V_g + C_g \]  

The genetic covariance between a pair of full siblings is:

\[ \text{cov}(g_1, g_2) = 1/2 V_g + C_g \]  

If we take a set of SNPs with their effects estimated from GCTA-COJO analysis and create a predictor using these SNPs in an independent validation sample, we can write the predictor as:

\[ \hat{g} = \sum_i x_i \hat{h}_i \]

where \( \hat{h} \) is the estimate of \( h \), with \( \hat{h} = b + e \) where \( e \) is the error in estimating \( b \).

If we assume \( b \) and \( e \) to be independent and denote

\[ V_e = \sum_i \text{var}(x_i) e_i^2 \]

and

\[ C_e = \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) e_i e_j \]

the variance of the predictor is:

\[ \text{var}(\hat{g}) = \text{var}\left(\sum_i x_i \hat{h}_i\right) = \sum_i \text{var}(x_i) \hat{h}_i^2 + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) \hat{h}_i \hat{h}_j \]

\[ = \sum_i \text{var}(x_i) \hat{h}_i^2 + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) \hat{h}_i \hat{h}_j + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) e_i e_j \]

\[ = V_g + V_e + C_g + C_e \]  

The covariance between the predictors of a pair of full siblings is:

\[ \text{cov}(\hat{g}_1, \hat{g}_2) = \text{cov}\left(\sum_i x_i \hat{h}_i, \sum_j x_j \hat{h}_j\right) = 1/2 \sum_i \text{var}(x_i) \hat{h}_i^2 + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) \hat{h}_i \hat{h}_j \]

\[ = 1/2 V_g + 1/2 V_e + C_g + C_e \]  

The covariance between the true phenotype and the predictor for the same individual is:

\[ \text{cov}(y, \hat{g}) = \text{cov}(g + e, b + e) = \text{var}(g) = V_g + C_g \]

The covariance between the true phenotype of one sibling and the predictor of the other sibling is:

\[ \text{cov}(y_1, \hat{g}_2) = \text{cov}(g_1 + e_1, \hat{g}_2) = \sum_i \text{var}(x_i) \hat{h}_i^2 + \sum_i \sum_{j(i \neq j)} \text{cov}(x_i, x_j) \hat{h}_i \hat{h}_j \]

\[ = 1/2 V_g + 1/2 V_e + C_g + C_e \]

If we define \( \Delta g = \hat{g}_1 - \hat{g}_2 \) and \( \Delta y = y_1 - y_2 \),

\[ \text{var}(\Delta g) = \text{var}(\hat{g}_1) + \text{var}(\hat{g}_2) - 2\text{cov}(\hat{g}_1, \hat{g}_2) = V_g + V_e \]

\[ \text{cov}(\Delta y, \Delta g) = \text{cov}(y_1, \hat{g}_1) + \text{cov}(y_2, \hat{g}_2) - \text{cov}(y_1, \hat{g}_2) - \text{cov}(y_2, \hat{g}_1) = V_g \]

We then can calculate the four parameters as:

\[ V_g = \text{cov}(\Delta y, \Delta g) \]

\[ V_e = \text{var}(\Delta g) - V_g \]

\[ C_g = \text{cov}(y, \hat{g}) - V_g \]

\[ C_e = 2\text{cov}(\hat{g}_1, \hat{g}_2) - \text{var}(\hat{g}) - C_g \]

where \( V_g \) can be interpreted as the variance explained by real SNP effects, \( C_g \) is the covariance between predictors attributed to the real effects of SNPs that are not in LD but are correlated owing to population stratification. \( V_e \) is the accumulated variance due to errors in estimating SNP effects and \( C_e \) is the covariance between predictors attributed to errors in estimating the effects of SNPs that are correlated owing to population stratification.

To assess prediction accuracy, we usually perform a regression analysis of the real phenotype against the predictor:

\[ y = b \hat{b} + \hat{g} \hat{b}_1 + \epsilon \]

so that the regression slope is actually

\[ \beta = \text{cov}(y, \hat{g})/\text{var}(\hat{g}) = (V_g + C_g)/(V_g + V_e + C_g + C_e) \]

with the regression \( R^2 \) being

\[ R^2 = \frac{(V_g + C_g)^2}{(V_g + V_e + C_g + C_e)} \]
In the absence of population structure:

\[ r^2 = \frac{V_g^2}{(V_g + V_e)} \]  

(20)

Variance explained by SNPs in proximity to the top associated SNPs. We performed analyses to quantify the variance explained by SNPs in close physical proximity to the top associated SNPs in 9,500 unrelated individuals (pairwise genetic relatedness < 0.025) from a combined data set of the QIMR and TwinGene cohorts. As in previous analyses, to avoid sample overlap between the discovery and validation studies, we repeated the discovery meta-analysis excluding the QIMR and TwinGene cohorts and identified 643 genome-wide significant SNPs from the GCTA-COJO analysis of the summary statistics using ARIC data for LD estimation. We used GCTA-GREML analysis to quantify the phenotypic variance explained by all the common SNPs (minor allele frequency (MAF) > 0.01) within 100 kb, 500 kb or 1 Mb of the 643 significant SNPs. There are 104,000, 423,000 and 745,000 SNPs within 100 kb, 500 kb and 1 Mb of the top associated SNPs, respectively, which explain 20.8% (SE = 1.3%), 25.7% (SE = 1.8%) and 29.5% (SE = 2.2%) of phenotypic variance (Supplementary Fig. 6a). We then applied a regression-based approach to adjust for LD between SNPs. The estimates of variance explained after LD adjustment were slightly higher than those without adjustment, and the ratio of the estimates with and without LD adjustment was consistently \( r^2 > 0.1 \), regardless of the window size (Supplementary Fig. 6a). However, this difference is small.

We sought to investigate whether there was an enrichment of associated signals at the top associated loci. We varied the window size, using windows of 20 kb, 50 kb, 100 kb, 150 kb, 200 kb, 300 kb, 400 kb, 500 kb, 750 kb and 1 Mb, and fitted a two-component model in GCTA-GREML analysis, with the first component being the top associated SNPs and the second component being the rest of the SNPs in the window. We found that the per-SNP variance explained excluding the top SNPs (variance explained by the second component divided by the number of SNPs included in this component) decreased with the size of the window (Supplementary Fig. 6b), implying that SNPs in closer physical proximity to the top associated SNPs tend to explain disproportionally more variance.

Enrichment of associated SNPs in ENCODE regions, loci containing OMIM genes, eQTLs and nonsynonymous SNPs. To identify putative causal variants among the height-associated markers, we explored whether the height-associated SNPs were in strong LD \( (r^2 > 0.8) \) with nonsynonymous coding variants in 1000 Genomes Project CEU Phase 1 data (Utah residents of Northern and Western European ancestry), showed an effect on whole-blood gene expression levels, were located within Encyclopedia of DNA Elements ( ENCODE)-annotated regions, were within loci harboring monogenic growth genes or had previously been associated with other complex traits in the NHGRI GWAS catalog \((P < 5 \times 10^{-8})\) (Supplementary Tables 7–11). To estimate the empirical assessment of enrichment for listed features, we used 10,000 permutations of random sets of SNPs matched to the LD-pruned \((r^2 > 0.1)\) 628 height-associated SNPs by the number of nearby genes (within an LD distance of \( r^2 > 0.5 \)), the physical distance to the nearest gene and MAF.

Enrichment of genes in associated loci in known and new pathways. Data-Driven Expression-Prioritized Integration for Complex Traits (DEPICT) analysis. The DEPICT method (T.H.P., J. Karjalainen, Y. Chan, H. Westra and A.R.W. et al., unpublished data; see Geller et al. for an earlier application of DEPICT) relies on precomputed predictions of gene function based on a heterogeneous panel of 77,840 expression arrays (Fehrmann et al., unpublished data; ref. 30), 5,984 molecular pathways (based on 169,810 high-confidence experimentally derived protein-protein interactions), 2,473 phenotypic gene sets (based on 211,882 gene-phenotype pairs from the Mouse Genome Informatics; see URLs), 737 REACTOME pathways, 5,083 Gene Ontology terms and 184 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The method leverages these predictions to extend the functional annotations of genes, including genes that previously had only a few or no functional annotations.

DEPICT facilitates the analysis of GWAS data by (i) assessing whether genes in associated loci are enriched in tissue-specific expression, (ii) identifying reconstituted gene sets that are enriched in genes from associated loci and (iii) systematically identifying the most likely causal gene(s) at a given locus (see the Supplementary Note for a more detailed description of DEPICT). To run DEPICT, we first clumped the summary statistics from the meta-analysis using 500-kb flanking regions with \( r^2 > 0.1 \) and excluded SNPs with \( P \geq 5 \times 10^{-8} \), which yielded 628 SNPs. We then mapped genes to each of the 628 most strongly associated SNPs. For a given SNP, this mapping was accomplished by including all genes that resided within the boundaries of \( r^2 > 0.5 \) of that SNP and always including the nearest gene to its locus gene set. We used a locus definition that was calibrated using the GWAS data for height levels presented in this report and optimized the capture of known monogenic genes for those traits. We merged overlapping loci and excluded loci that mapped near or within the major histocompatibility complex locus (chromosome 6, 20–40 Mb), which resulted in a list of 566 non-overlapping loci that were used as input for DEPICT. HapMap Project Phase 2 CEU genotype data were used for all LD calculations.

GRAIL and MAGENTA analyses. The GRAIL algorithm was run using the LD-pruned \((r^2 > 0.1)\) 628 SNPs, without correcting for gene size and using text-mining data available up to December 2006 (default setting). MAGENTA was run with the adjusted summary statistics with single \( \lambda_{\text{UR}} \) as input using default settings and excluding the human leukocyte antigen (HLA) region.

22. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
23. Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–354 (2010).
24. Winkler, T.W. et al. Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. 9, 1192–1212 (2014).
25. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
26. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909 (2006).
27. Lynch, M. & Walsh, B. Genetics and Analysis of Quantitative Traits (Sinauer Associates, Inc., 1998).
28. Gusev, A. et al. Quantifying missing heritability at known GWAS loci. PLoS Genet. 9, e1003993 (2013).
29. Geller, F. et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. Nat. Genet. 46, 957–963 (2014).
30. Cvejic, A. et al. SMIM21 undulates the Veil blood group and influences red blood cell traits. Nat. Genet. 45, 542–545 (2013).
31. Lage, K. et al. A human pheno-encode interaction network of protein complexes implicated in genetic disorders. Nat. Biotechnol. 25, 309–316 (2007).
32. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–D697 (2011).
33. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–D114 (2012).