nature genetics

Many sequence variants affecting diversity of adult human height

Daniel F Gudbjartsson[1], G Bragi Walters[1], Gudmar Thorleifsson[1], Hreinn Stefansson[1], Bjarni V Halldorsson[1,2], Pasha Zusmanovich[1], Patrick Sulem[1], Steinunn Thorlacius[1], Arnaldur Gylfason[1], Stacy Steinberg[1], Anna Helgadottir[1], Andres Ingason[1], Valgerdur Steinthorsdottir[1], Elinborg J Olafsdottir[3], Gudridur H Olafsdottir[3], Thorvaldur Jonsson[4], Knut Borch-Johnsen[5,6], Torben Hansen[5], Gitte Andersen[5], Torben Jorgensen[7], Oluf Pedersen[5,6], Katja K Aben[8], J Alfred Witjes[9], Dorine W Swinkels[10], Martin den Heijer[11], Barbara Franke[12], Andre L M Verbeek[13], Diane M Becker[14], Lisa R Yanek[14], Lewis C Becker[14], Laufey Tryggvadottir[3], Thorunn Rafnar[1], Jeffrey Gulcher[1], Lambertus A Kiemeney[8,9,13], Augustine Kong[1], Unnur Thorsteinsdottir[1] & Kari Stefansson[1]

Adult human height is one of the classical complex human traits[1]. We searched for sequence variants that affect height by scanning the genomes of 25,174 Icelanders, 2,876 Dutch, 1,770 European Americans and 1,148 African Americans. We then combined these results with previously published results from the Diabetes Genetics Initiative on 3,024 Scandinavians[2] and tested a selected subset of SNPs in 5,517 Danes. We identified 27 regions of the genome with one or more sequence variants showing significant association with height. The estimated effects per allele of these variants ranged between 0.3 and 0.6 cm and, taken together, they explain around 3.7% of the population variation in height. The genes neighboring the identified loci cluster in biological processes related to skeletal development and mitosis. Association to three previously reported loci are replicated in our analyses[3–5], and the strongest association was with SNPs in the *ZBTB38* gene.

Adult human height is an easily observable and highly heritable complex continuous trait[6,7]. Because of this, it is a model trait for studying genetic influence on quantitative traits in humans. Most of the genetic variance in height seems to be additive. Environment, and in particular nutrition, has a substantial effect on height, as is evident from the increase in average height observed in most industrialized societies over the last century[8]. There is a large difference between the adult height of males and females; for example, the average heights of US males and females are 176 cm and 163 cm, respectively (s.d. = 6–7 cm)[8]. Longitudinal growth occurs by endochondral ossification, a process that occurs at the growth plate of long bones, where cartilaginous scaffold is replaced by bone[9]. Longitudinal growth rates are high during the first year, followed by modest rates until the onset of puberty, when rates become high again. At the end of puberty, growth velocity rapidly decreases as a result of growth plate maturation and fusion to the shaft, resulting in termination of longitudinal growth[9]. The main factors contributing to the growth rate and final height are genetic, hormonal, environmental and nutritional.

Coding sequence mutations in several genes cause rare syndromes with extreme stature, but these explain only a very small proportion of the normal population variation in height[10]. Numerous genome-wide linkage scans have implicated several regions of the genome[11], but recent genome-wide association scans were the first to identify sequence variants showing strong association with height in multiple studies[4,5].

To search for more genes influencing adult height, we analyzed genome-wide SNP data from 34,450 Icelanders, 25,174 of whom had height measurements available (**Supplementary Table 1** online). The analysis was augmented with data from 49,334 Icelanders with height measurements whose genetic information could be partially inferred from the genotyped individuals. We combined the Icelandic data with data on 2,876 Dutch individuals from a bladder cancer case-control study, 1,770 European Americans and 1,148 African Americans recruited from families of probands with premature coronary disease.

[1]deCODE Genetics, 101 Reykjavik, Iceland. [2]Reykjavik University, 103 Reykjavik, Iceland. [3]Icelandic Cancer Registry, 105 Reykjavik, Iceland. [4]Department of Surgery, University Hospital, 101 Reykjavik, Iceland. [5]Steno Diabetes Center, DK-2820 Copenhagen, Denmark. [6]Faculty of Health Science, University of Aarhus, DK-8000 Aarhus, Denmark. [7]Research Centre for Prevention and Health, Glostrup University Hospital, DK-2600 Glostrup, Denmark. [8]Comprehensive Cancer Centre East, 6500 HB Nijmegen, The Netherlands. [9]Radboud University Nijmegen Department of Urology, 6500 HB Nijmegen, The Netherlands. [10]Radboud University Nijmegen Department of Clinical Chemistry, 6500 HB Nijmegen, The Netherlands. [11]Radboud University Nijmegen Department of Endocrinology, 6500 HB Nijmegen, The Netherlands. [12]Radboud University Nijmegen Department of Genetics, 6500 HB Nijmegen, The Netherlands. [13]Radboud University Nijmegen Medical Center Department of Epidemiology & Biostatistics, 6500 HB Nijmegen, The Netherlands. [14]The Johns Hopkins Sibling and Family Heart Study, The Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA. Correspondence should be addressed to D.F.G. (daniel.gudbjartsson@decode.is) or K.S. (kstefans@decode.is).

Received 13 December 2007; accepted 15 February 2008; published online 6 April 2008; doi:10.1038/ng.122

Table 1 SNPs showing strongest correlation with height in meta-analysis

			Effect						Effect		Effect		
SNP / Allele	Freq.	Region	ICE	DUT	USe	DGI	USa	Scan *P*	DAN	DAN *P*	Comb.	Comb. *P*	Neighboring genes
rs11205277 G	0.44	1q12	4.8	5.5	10.6	–	2.2	7.3×10^{-9}	5.4	0.0052	5.1 (0.8)	1.4×10^{-10}	Histone class 2A, *MTMR11, SV2A, SF3B4*
rs678962 G	0.22	1q24	4.4	11.3	8.0	6.5	0.8	3.2×10^{-8}	–	–	5.4 (1.0)	3.2×10^{-8}	*DNM3*
rs2274432 T	0.37	1q25	5.3	1.2	9.8	–	11.4	9.7×10^{-8}	4.5	0.027	5.3 (0.9)	7.8×10^{-9}	*C1orf19,GLT25D2*
rs3791679 T	0.81	2p16	5.9	4.5	6.6	2.9	18.8	1.1×10^{-7}	9.1	3.0×10^{-5}	5.8 (0.9)	5.9×10^{-11}	*EFEMP1, PNPT1*
rs6763931 A	0.45	3q23	7.4	9.2	6.5	6.4	3.7	2.6×10^{-18}	13.5	9.9×10^{-13}	7.4 (0.7)	1.4×10^{-27}	*ZBTB38*
rs6842303 T	0.26	4p15	3.8	5.0	15.7	–	14.8	5.7×10^{-6}	9.2	2.0×10^{-5}	4.8 (0.8)	4.2×10^{-9}	*LCORL, NCAPG*
rs6830062 T	0.89	4p15	4.8	11.9	21.6	3.6	8.8	7.6×10^{-8}	10.5	0.00018	6.3 (1.0)	1.3×10^{-10}	*LCORL, NCAPG*
rs1812175 C	0.86	4q31	7.9	8.3	7.8	11.0	10.4	2.8×10^{-10}	6.9	0.010	8.3 (1.2)	9.7×10^{-12}	*HHIP*
rs12198986 A	0.50	6p24	7.9	2.7	2.2	–	1.0	6.4×10^{-11}	3.3	0.080	6.8 (1.0)	2.4×10^{-11}	*BMP6*
rs10946808 A	0.70	6p22 B	5.6	6.1	7.7	–	–0.2	1.1×10^{-7}	7.4	0.00085	5.6 (0.9)	5.8×10^{-10}	Histone class 1, Butyrophilin genes
rs2844479 T	0.67	6p21 A	7.0	6.7	4.0	0.7	8.8	8.4×10^{-9}	–	–	6.3 (1.1)	8.4×10^{-9}	HLA class III
rs3130050 G	0.19	6p21 A	8.5	0.3	0.0	–	3.6	5.1×10^{-7}	6.3	0.021	7.4 (1.3)	3.2×10^{-8}	HLA class III
rs185819 T	0.52	6p21 B	5.2	5.2	5.9	–	3.6	1.9×10^{-6}	6.4	0.0048	5.2 (0.9)	3.2×10^{-8}	HLA class III
rs1776897 C	0.07	6p21 D	9.8	2.3	7.3	–	8.4	1.1×10^{-7}	6.9	0.043	8.8 (1.6)	1.4×10^{-8}	*HMGA1, LBH*
rs4713858 G	0.86	6p21 E	6.4	11.2	1.1	8.6	2.5	4.9×10^{-8}	3.9	0.22	6.8 (1.2)	3.5×10^{-8}	*ANKS1, TCP11, ZNF76, DEF6, SCUBE3*
rs3748069 A	0.74	6q24	5.7	9.7	11.4	8.0	3.2	4.9×10^{-11}	8.3	0.00015	6.5 (0.9)	4.5×10^{-14}	*GPR126*
rs798544 G	0.72	7p22	5.0	9.8	13.1	4.8	10.5	4.5×10^{-12}	7.7	0.00029	5.9 (0.8)	6.5×10^{-15}	*GNA12*
rs2282978 C	0.29	7q21	5.5	10.4	–0.5	7.5	1.4	9.8×10^{-9}	–	–	5.8 (1.0)	9.8×10^{-9}	*CDK6, PEX1, GATAD1, ERVWE1*
rs11765954 C	0.24	7q21	6.0	8.3	–0.2	8.9	–0.1	1.2×10^{-7}	2.9	0.17	6.1 (1.1)	6.9×10^{-8}	*CDK6, PEX1, GATAD1, ERVWE1*
rs10958476 C	0.23	8q12	5.5	3.0	9.9	–	4.6	1.4×10^{-6}	5.7	0.015	5.4 (1.0)	6.6×10^{-8}	*PLAG1, MOS, CHCHD7, RDHE2, RPS20, LYN, TGS1, PENK*
rs7846385 C	0.27	8q21	5.1	9.9	2.6	2.7	–4.5	5.3×10^{-7}	4.6	0.030	5.0 (0.9)	4.7×10^{-8}	*PXMP3, ZFHX4*
rs4743034 A	0.23	9q31	4.9	5.2	10.9	7.9	0.0	5.2×10^{-7}	5.8	0.012	5.3 (0.9)	2.1×10^{-8}	*ZNF462*
rs8756 C	0.52	12q14	6.0	5.0	8.8	10.7	9.4	2.2×10^{-13}	7.2	0.00016	6.6 (0.8)	1.8×10^{-16}	*HMGA2*
rs7153027 A	0.52	14q32	4.6	10.1	4.3	6.9	18.1	1.1×10^{-10}	–	–	5.7 (0.9)	1.1×10^{-10}	*TRIP11, FBLN5, ATXN3, CPSF2*
rs4533267 A	0.28	15q26	6.2	7.1	–0.4	2.6	1.5	3.8×10^{-7}	4.8	0.030	5.6 (1.0)	3.3×10^{-8}	*ADAMTS17*
rs3760318 C	0.63	17q11	5.9	6.8	4.1	8.8	–0.8	1.8×10^{-9}	–	–	6.0 (1.0)	1.8×10^{-9}	*CRLF3, ATAD5, CENTA2, RNF135*
rs4794665 A	0.48	17q23 A	3.4	4.3	3.8	3.3	7.8	9.2×10^{-6}	6.2	0.0016	3.6 (0.7)	9.9×10^{-8}	*NOG, DGKE, TRIM25, COIL, RISK*
rs757608 T	0.35	17q23 B	4.7	3.4	10.8	3.4	–7.4	3.0×10^{-7}	3.7	0.073	4.4 (0.8)	6.3×10^{-8}	*BCAS3, NACA2, TBX2, TBX4*
rs4800148 A	0.79	18q11	5.2	10.5	16.3	–	8.7	9.4×10^{-8}	5.9	0.012	6.4 (1.1)	3.7×10^{-9}	*CABLES1, RBBP8,* C18orf45
rs967417 C	0.53	20p12	4.3	4.3	5.5	4.6	1.3	4.6×10^{-6}	7.1	0.00025	4.3 (0.8)	1.5×10^{-8}	*BMP2*

SNPs showing strongest correlation with height in meta-analysis of the Icelandic (ICE, $n = 25{,}174$), Dutch (DUT, $n = 2{,}876$), European Americans (USe, $n = 1{,}770$), DGI ($n = 3{,}024$), African Americans (USa, $n = 1{,}148$) and Danish (DAN, $n = 5{,}517$) samples. Scan *P* refers to the *P* value obtained from combining all the genome-wide association samples (ICE, DUT, USe, DGI and USa). Comb. refers to the results obtained by combining the scan results with the Danish results. Effects are in percentage of standard deviation; s.e.m. is given in brackets for the combined effect estimate. *P* values are corrected using genomic control.

All these samples were genotyped with SNP chips containing a superset of the HapMap panel on the 317K Illumina chip. After controlling for quality, we had 304,226 SNPs available for analysis. All height measurements were adjusted for age and sex. In addition, we combined our results with a previously published analysis of 3,025 Swedish and Finnish individuals from the Diabetes Genetics Initiative (DGI)[2], made up of 1,496 individuals with type 2 diabetes and 1,529 nondiabetic controls typed with the Affymetrix GeneChip Human Mapping 500K platform. On the basis of the HapMap data, we found that 149,004 of the SNPs available for analysis on the Illumina chip had a surrogate SNP ($r^2 > 0.8$) in the DGI dataset[12].

SNPs in 27 regions of the genome satisfied our criteria for genome-wide significance ($P < 1.6 \times 10^{-7} \sim 0.05/304{,}226$; **Table 1** and **Supplementary Fig. 1** online). In addition, 27 secondary signals ($10^{-5} > P > 1.6 \times 10^{-7}$), in regions other than those described in **Table 1**, are summarized in **Table 2** (**Supplementary Fig. 1**; **Supplementary Table 2** online contains all SNPs with $P < 10^{-4}$). If no SNPs were truly correlated with adult height, then only 3.1 SNPs would be expected to have a *P* value $<10^{-5}$. Given this excess of suggestive associations, it is likely that a substantial fraction of the signals in **Table 2** are true associations.

We typed 40 of the SNPs with $P < 10^{-5}$, based on the Illumina and DGI data, on a sample of 5,517 Danes using individual SNP assays. Of the 27 loci that showed significant association with height in genome-wide association analysis, 22 showed one-sided genome-wide significant association at the nominal 0.05 level in the Danish sample. For three of the remaining five loci, no data was available for the Danish sample. Note that all but one of the 40 SNPs tested in the Danish sample showed an effect in the same direction as was observed in the genome-wide association scan (**Tables 1** and **2** and **Supplementary Table 3** online).

Of the 27 significant regions described in **Table 1**, one encompasses and replicates the previously described[4] association to *HMGA2* (effect = 6.6 ± 0.8 percentage of s.d., $P = 1.0 \times 10^{-11}$ for rs8756, the best available surrogate ($r^2 = 0.87$) for the previously associated rs1042725). A SNP in the previously reported region on 20q11 also

Table 2 SNPs with *P* between 10^{-5} and 1.6×10^{-7} in the meta-analysis of association with height

			Effect						Effect		Effect		
SNP / Allele	Freq.	Region	ICE	DUT	USe	DGI	USa	Scan *P*	DAN	DAN *P*	Comb.	Comb. *P*	Neighboring genes
rs6733301 G	0.87	2p23	6.4	12.7	17.1	–	2.7	8.2×10^{-7}	–	–	7.5 (1.5)	8.2×10^{-7}	*ADCY3, RBJ, POMC, DNMT3A, DTNB*
rs1052483 C	0.91	2q35	6.5	2.4	8.2	13.1	8.6	1.2×10^{-6}	–	–	6.9 (1.4)	1.2×10^{-6}	*IHH, CRYBA2, FEV, SLC23A3, TUBA1*
rs749052 A	0.94	2q37	8.5	12.4	9.0	8.0	3.6	1.4×10^{-6}	–	–	8.7 (1.8)	1.4×10^{-6}	*NPPC, DIS3L2, COPS7B, PDE6D, PTMA*
rs4345115 T	0.63	3q26	4.0	7.4	–2.6	9.2	0.0	6.7×10^{-6}	–	–	4.4 (1.0)	6.7×10^{-6}	*GOLIM4, SERPINI1*
rs710841 A	0.27	4q21	5.3	3.8	5.3	3.8	4.7	1.9×10^{-6}	–	–	5.0 (1.0)	1.9×10^{-6}	*BMP3, PRKG2, RASGEF1B*
rs31198 T	0.75	5q31	4.5	9.5	5.0	5.1	–4.6	8.2×10^{-6}	–	–	4.8 (1.1)	8.2×10^{-6}	*PITX1, PCBD2, CATSPER3 TXNDC15, DDX46, CAMLG*
rs12199222 T	0.33	6p22 A	4.6	0.3	7.9	5.8	6.2	5.0×10^{-6}	4.1	0.048	4.4 (0.9)	6.5×10^{-7}	*NUP153, CAP2, KIF13A*
rs9395066 C	0.48	6p21 F	3.2	4.7	4.6	4.9	0.2	7.5×10^{-6}	–	–	3.5 (0.8)	7.5×10^{-6}	*SUPT3H, RUNX2*
rs314268 C	0.34	6q16	4.3	6.9	7.4	3.0	4.2	7.7×10^{-7}	–	–	4.6 (0.9)	7.7×10^{-7}	*LIN28B, HACE1, BVES, POPDC3*
rs9487094 G	0.69	6q21	4.4	8.8	1.2	5.3	1.8	4.1×10^{-6}	–	–	4.7 (1.0)	4.1×10^{-6}	*PPIL6, CD164, SMPD2,MNICAL1, ZBTB24*
rs1490388 T	0.42	6q22	4.7	6.2	2.4	7.0	–2.8	7.1×10^{-7}	2.3	0.25	4.8 (1.0)	5.6×10^{-7}	*C6orf173*
rs6899976 G	0.28	6q23	3.1	5.1	4.6	8.2	4.5	1.1×10^{-6}	–0.4	0.85	3.8 (0.8)	5.6×10^{-6}	*L3MBTL3, SAMD3*
rs2814828 T	0.25	9q22	5.4	2.8	13.5	6.5	–1.0	3.7×10^{-7}	1.4	0.53	5.4 (1.1)	9.3×10^{-7}	*SPIN1, CCRK*
rs946053 T	0.52	9q32	4.2	5.1	2.3	–	9.5	1.2×10^{-6}	3.8	0.052	4.4 (0.8)	1.7×10^{-7}	*COL27A1*
rs2187642 A	0.39	12p13	4.4	5.1	13.1	0.1	8.9	3.5×10^{-6}	2.8	0.16	4.6 (1.0)	1.5×10^{-6}	*ETV6*
rs11611208 A	0.06	12p12	11.4	4.9	10.6	–	36.9	2.2×10^{-6}	–	–	11.4 (2.4)	2.2×10^{-6}	*PDE3A, SLCO1C1, SLCO1B3*
rs11177669 A	0.31	12q15	3.7	8.2	9.3	–	3.9	3.2×10^{-6}	–	–	4.5 (1.0)	3.2×10^{-6}	*LYZ, YEATS4, FRS2, CPSF6, CCT2, LRRC10*
rs3825199 C	0.24	12q22	6.5	3.7	4.6	7.9	2.9	4.6×10^{-7}	3.9	0.16	6.2 (1.2)	1.8×10^{-7}	*SOCS2, MRPL42, CRADD, UBE2N*
rs1239947 G	0.35	13q14	3.6	5.1	0.2	–	11.6	7.6×10^{-6}	–	–	3.8 (0.8)	7.6×10^{-6}	*DLEU7*
rs10132817 G	0.57	14q13	5.6	2.0	0.5	0.5	6.6	9.9×10^{-6}	–	–	4.5 (1.0)	9.9×10^{-6}	*NKX2-1, MBIP, NKX2-8, PAX9*
rs2554380 T	0.78	15q25	4.6	8.0	8.0	8.3	–27.3	8.6×10^{-7}	–	–	4.5 (0.9)	8.6×10^{-7}	*ADAMTSL3, SH3GL3*
rs2326458 C	0.26	16q24	5.4	6.2	3.8	–	–3.4	7.7×10^{-7}	–	–	5.1 (1.0)	7.7×10^{-7}	*ZDHHC7, CRISPLD2, USP10*
rs7209435 C	0.27	17q23 C	5.0	7.1	3.4	1.6	2.9	7.1×10^{-7}	–	–	4.8 (1.0)	7.1×10^{-7}	*MAP3K3, WDR68, LYK5, MT1F*
rs7249094 G	0.59	19p13	4.1	2.3	11.5	–	4.0	1.3×10^{-6}	–	–	4.3 (0.9)	1.3×10^{-6}	*ADAMTS10, MYO1F, PRAM1, OR2Z1*
rs6088792 T	0.26	20q11	5.0	5.5	5.3	4.2	–4.5	3.9×10^{-6}	3.7	0.078	4.7 (1.0)	8.0×10^{-7}	*UQCC, GDF5, CEP250, EIF6, MMP24*
rs5751614 A	0.49	22q11	4.2	5.8	3.9	–	3.2	6.4×10^{-6}	–	–	4.3 (1.0)	6.4×10^{-6}	*BCR, GNAZ, RTDR1, IGLL1*
rs1474563 T	0.58	Xq21	4.9	0.4	–2.0	1.1	–6.4	3.1×10^{-6}	–	–	3.5 (0.8)	3.1×10^{-6}	*ITM2A*

SNPs with *P* between 10^{-5} and 1.6×10^{-7} in the meta-analysis of association with height in the Icelandic (ICE, $n = 25{,}174$), Dutch (DUT, $n = 2{,}876$), Europeans Americans (USe, $n = 1{,}770$), DGI ($n = 3{,}024$), African Americans (USa, $n = 1{,}148$) and Danish (DAN, $n = 5{,}517$) samples. Scan *P* refers to the *P* value obtained from combining all the genome-wide association samples (ICE, DUT, USe, DGI and USa). Comb. refers to the results obtained by combining the scan results with the Danish results. Effects are in percentage of standard deviation; s.e.m. is given in brackets for the combined effect estimate. *P* values are corrected using genomic control.

replicates the previously reported[5] association with height (effect = 4.7 ± 1.0, $P = 8.0 \times 10^{-7}$).

The most significant associations we observed were with variants in the *ZBTB38* (zinc finger and BTB domain containing 38) gene on 3q23, which encodes a protein that has been shown to repress transcription by specifically binding to methylated DNA[13]. The rat homolog of ZBTB38, Zenon, regulates transcription (independent of methylation) of the tyrosine hydroxylase gene, the rate-limiting enzyme of catecholamine biosynthesis[14]. The strongest association is with the A allele of rs6763931 (effect = 7.4 ± 0.7, $P = 1.4 \times 10^{-27}$), which is also the SNP in the region with the most significant correlation with the expression of *ZBTB38* in blood and in adipose tissue (see Methods), the same allele being positively correlated to height and expression ($P = 5.2 \times 10^{-13}$ in blood and 5.6×10^{-6} in adipose tissue). A SNP in *ZBTB38* showed suggestive association in a previous genome-wide association scan for height[5], but none of the other regions reported in that study, apart from the region on 20q11 mentioned above, showed suggestive association in our analysis.

To investigate how the associated loci might affect height, we looked for clustering of the genes neighboring (within 200 kb) the loci with association with height using the biological process classification of the Panther database[15] (**Table 3** and **Supplementary Tables 4** and **5** online). This database currently consists of 242 overlapping classes of biological processes and classifications of 14,177 genes. The association signals were divided into three classes for this analysis: those of genome-wide significance $P < 1.6 \times 10^{-7}$, those with $P < 10^{-5}$ and those with $P < 10^{-4}$.

A significant excess of markers was close to genes having the 'skeletal development' classification, a subset of the 'mesoderm development' class, which also had a significant excess of loci. The genes in the skeletal development class (*BMP2, BMP6, EFEMP1, FBLN5* and *NOG* were all genome-wide significant) encode proteins known to be involved in the local regulation of the growth plate[9]. Another class of genes important in skeletal development are those encoding enzymes from the ADAMTS (a disintegrin-like and metalloproteinase with thrombospondin) family, which mediate cartilage aggrecan loss[16]. We observed a significant association between SNPs in *ADAMTS17* and height, and SNPs in *ADAMTS10* showed suggestive association with height. In addition, we observed a significant association of SNPs in *ADAMTS3* with height when the African American samples were excluded (**Supplementary Table 2**). The FIBULIN and FIBRILLIN glycoproteins encoded by *FBLN5* and *EFEMP1* affect skeletal development through elastic fiber formation. FIBRILLIN 1 (encoded by *FBN1*) has been linked to Marfan syndrome[17], which is

Table 3 Number of genes from different biological process categories within 200 kb of SNPs with $P < 1.6 \times 10^{-7}$, $P < 10^{-5}$ or $P < 10^{-4}$

	$P < 1.6 \times 10^{-7}$ ($n = 26$)			$P < 10^{-5}$ ($n = 54$)			$P < 10^{-4}$ ($n = 120$)		
Biological process	Obs.	Exp.	*P*	Obs.	Exp.	*P*	Obs.	Exp.	*P*
Mesoderm development	10	2.1	1.9×10^{-5}	14	4.4	7.5×10^{-5}	20	9.7	0.0020
Skeletal development	6	0.5	9.6×10^{-6}	8	1.1	1.1×10^{-5}	9	2.4	0.00066
Mitosis	5	1.1	0.0042	10	2.3	7.8×10^{-5}	14	5.1	0.00057
Nucleoside, nucleotide and nucleic acid metabolism	14	7.6	0.0088	26	15.8	0.0039	54	35.1	0.00027
Intracellular signaling cascade	5	3.1	0.23	14	6.5	0.0049	28	14.4	0.00059

For all loci with at least one SNP with $P < 1.6 \times 10^{-7}$, $P < 10^{-5}$ or $P < 10^{-4}$, the observed (Obs.) and expected (Exp.) number of genes from each biological process (BP) that are within 200 kb from these SNPs are given, as well as a *P* value for comparing the two counts. Indented BPs are children (subsets) of the unindented BPs above them. We tested 242 BPs, and those with $P < 0.001$ are shown. The genes observed to overlap with the observed association loci are listed in **Supplementary Table 5**.

characterized by very tall, slender build and loose joints. Recently, a SNP (rs8033037) within the *FBN1* gene was found to have a nominally significant association with height in the Japanese population[3]. We examined a correlated SNP, rs8043050 (the two SNPs are equivalent in the nine genotyped European individuals available in dbSNP[18]), and observed a nominally significant association in the same direction as previously reported (effect $= 6.3 \pm 2.7$, $P = 0.019$). Both SNPs are rare in Europeans (minor allele frequency for rs8043050 is 5.8% in the Icelandic samples).

We also observed a significant excess of loci neighboring genes with the 'mitosis' classification. Most of this excess was driven by genes having the 'chromosome segregation' classification (*HMGA1*, *HMGA2*, *NCAPG*, *CDK6* and *COIL* were all genome-wide significant). *HMGA1* has a biological function similar to *HMGA2* and has been linked to adipose mass variation in pigs[19]. Although *HMGA1* is located near the HLA region, the association is tagged by a SNP that is not in strong LD with the SNPs in the HLA region that also associated to height ($r^2 < 0.003$ between the *HMGA1* SNP and all SNPs in the HLA region that associate with height with $P < 10^{-5}$).

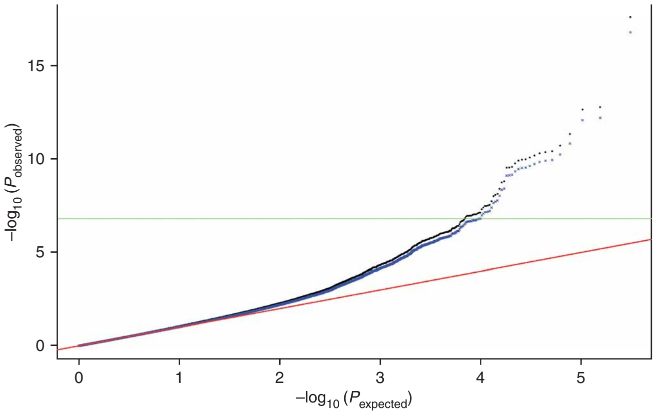

Figure 1 Quantile-quantile plot of 304,226 SNPs in the genome-wide association scan for height. The black dots represent observations, and the blue 'x's represent the same data scaled down by an inflation factor of 1.051. The diagonal red line represents where the dots are expected to fall under the null hypothesis of no association. The horizontal green line represents the threshold for genome-wide significance.

Figure 1 shows a quantile-quantile plot for our genome-wide association scan based on the combination of the results from all the samples described above, after correction of the results from each sample using genomic controls[20]. Even after these individual corrections, the combined results are still estimated to be inflated by a factor of 1.051. Around the threshold for genome-wide significance, this inflation factor corresponds roughly to multiplying *P* values by a factor of 2, which would move one locus (17q23 A) above our threshold for genome-wide significance. The remaining inflation factor must be due to positive correlation between association tests in different samples, probably caused either by an abundance of real signals or by stratification. Given the precautions we have taken to reduce the effect of stratification, the fact that stratification is unlikely to be a large issue in our largest sample (Iceland), and the strength and consistency of the Danish replication results, we feel that a plethora of real signals of modest effect is the most likely cause of the inflation, and that further correcting by 1.0508 would be an over-correction. The quantile-quantile plot indicates that there is a substantial excess of signals, even at the 0.01 level of significance.

The loci listed in **Table 1** explain around 3.7% of the population diversity of height. The largest fraction is explained by the *ZBTB38* locus (0.27%), and the smallest fraction is explained by the 17q23 A locus (0.06%). Our estimate of the fraction of variance explained by the *HMGA2* variant is 0.22%, slightly less than the previously reported estimate of around 0.3%. Our emphasis here is on discovery rather than estimation, and it should be noted that estimates of the effect sizes of variants close to the threshold for discovery are biased upwards.

We were able to compare the effect on height between the sexes in all the samples, except for DGI, and we did not find any locus to have a greater effect on the height of one sex over the other (all $P > 0.002$; see **Supplementary Table 6** online).

In summary, we have described sequence variants at 27 loci, 26 of which are previously undescribed, that associate significantly with adult human height. A large proportion of the loci cluster around genes involved in skeletal development and mitosis. This is in harmony with previous studies and suggests that genetic differences in adult human height stem from variations in local regulation of the growth plate, rather than systemic regulation. The Danish replication of the results from the scan, the shape of the quantile-quantile plot and the residual inflation factor that remains after all samples have been corrected individually all suggest that many more variants associating with height remain to be found.

METHODS

Icelandic sample. Height measurements from 75,121 Icelanders were collected in our studies of obesity and cancer (**Supplementary Table 1**). Of these individuals, 25,174 have been genotyped on an Illumina 317K SNP chip in one of several genome-wide association studies. These studies were all approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. Written informed consent was obtained from all participants. Personal identifiers associated with phenotypic information and blood samples were encrypted using a third-party encryption system as previously described[21]. We included only individuals with a genotype yield over 96% in the study. Height was measured using a stadiometer with the subjects wearing no shoes for 57,404 individuals and was self-reported on questionnaires by 17,717 individuals.

Dutch sample. Lifestyle information was collected from Dutch individuals with bladder cancer ($n = 1{,}164$) who were recruited as a part of the Polygene study from the population-based cancer registry held by the Comprehensive Cancer Centre East in Nijmegen. All individuals were invited to participate in the study by their treating physicians. Informed consent was obtained for the collection of questionnaire data on lifestyle, medical history, and family history, the collection of two 10 ml blood samples, linkage to population and disease registries (cancer registry, mortality registry, hospital information systems and the Dutch demographic register), collection of additional clinical data from their medical records and the keeping of identifying information for a duration of 25 years.

The Nijmegen Biomedical Study was a survey of the general population done in 2002–2003 by the Radboud University Nijmegen Medical Centre. This survey was based on an age-stratified random sample of the population of Nijmegen. Lifestyle information, family history of cancer, reproductive and medical history as well as blood samples were available from a group of 6,700 individuals. From this group, 1,942 individuals were randomly selected as a control group for the Polygene study. Similar informed consent as described above was obtained from these individuals. Self-reported height was used for the Dutch sample.

All individuals were fully informed about the goals and the procedures of the studies. All study protocols were approved by the Institutional Review Board of the Radboud University Nijmegen Medical Centre.

US sample. Subjects consisted of families identified from 1983–2006 from probands with a premature coronary disease event before 60 years of age. This ongoing prospective family study was designed to determine the environmental and genetic causes of premature chronic and cardiovascular diseases. Probands with documented coronary artery disease (CAD) before the age of 60 were identified at the time of hospitalization in any of ten Baltimore area hospitals. Their apparently healthy 30- to 59-year-old siblings without known CAD were recruited. In 2002, adult offspring over 21 years of age of all participating siblings and probands were recruited and underwent risk factor measurement and phenotypic characterization. In addition, the spouse was recruited for all participants for which at least one offspring was recruited. Height was measured using a stadiometer with the subjects wearing no shoes.

Danish sample. The Danish Inter99 cohort is a population-based sample of 30- to 60-year-old individuals living in the greater Copenhagen area, sampled at the Research Centre for Prevention and Health in the years 1999–2000. All individuals were drawn from the Civil Registration System. All study participants had undergone a thorough phenotype characterization based on a standardized questionnaire and interview as well as physical examination including standardized measurements of height. Informed written consent was obtained from all subjects before participation. The study was approved by the Ethical Committee of Copenhagen County and was in accordance with the principles of the Helsinki Declaration.

Genotyping. The Icelandic study participants were genotyped using genotyping systems and specialized software from Illumina (Human Hap300 and Human Hap300-duo+ Bead Arrays)[22]. In total, 311,388 SNP markers, distributed across the human genome, were common to both platforms. For the association analysis, we used 304,226 SNP markers, as 7,162 were deemed unusable because of low yield (<95% in at least one sample), deviations from Hardy-Weinberg expectations ($P < 10^{-5}$ in at least one sample) or discrepancies in genotype frequencies between the arrays. The Dutch study participants were genotyped with the Illumina Human Hap300-duo+ Bead Arrays. The American study participants were genotyped with the Human Hap1000 Bead Arrays.

Single SNP genotyping was carried out on the Centaurus (Nanogen) platform[23]. We evaluated the quality of each Centaurus SNP assay by genotyping each assay in the CEU HapMap samples and comparing the results with the HapMap data. Assays with >1.5% mismatch rate were not used.

Quality control of samples. Genotypes of markers on the X chromosome were compared to reported sex, and individuals with inconsistent sex were removed from the study. We removed 70 individuals from the Icelandic sample, 10 from the Dutch sample and 20 from the US sample. Only height measurements between 130 cm and 250 cm were used.

Genealogical information in Iceland is very nearly complete; thus, individuals with missing mothers are likely to be foreigners (missing paternity is likely due to other factors). Hence, we removed 208 individuals with missing mothers.

We applied the STRUCTURE software package[24] to all our genome-wide samples individually and used the HapMap data to train the classification of individuals into those of European, Asian and African ancestry. Individuals with less than 95% European ancestry were removed from the Icelandic ($n = 37$), Dutch ($n = 67$) and European American ($n = 107$) samples, and individuals with less than 30% African ancestry were removed from the African American sample ($n = 6$).

Association testing. All adult height measurements were corrected for year of birth and standardized to have a standard normal distribution, within both sexes, for each study population separately. For the Iceland population, we rounded year of birth to five years and used it as a factor variable in the correction, but for the other populations, we used a linear term in year of birth to correct the measurements. Measured and self-reported heights were corrected separately in the Iceland population. For each SNP, a classical linear regression, using the genotype as an additive covariate and height as a response, was fit to test for association.

We employed the EIGENSTRAT[25] method to detect and correct for possible population stratification in the Dutch (corrected for the 10 greatest eigenvalues) and the two US samples (corrected for the 30 greatest eigenvalues). Previous efforts have suggested that the EIGENSTRAT method does not detect meaningful stratification in the Icelandic population[26].

We scaled all test statistics by the method of genomic control[20] using an estimate of the inflation factor obtained by comparing the observed median of all χ^2-test statistic to the value predicted by theory (0.675^2). The estimated inflation factors were 0.959 in the Icelanders, 1.046 in the Dutch, 1.498 in the European Americans, 1.548 in the African Americans and 1.077 in the DGI data. The negative inflation in the Iceland population is due to the negative correlation caused by the family correction described below, and the high inflation factors in the US samples are due to several close relatives included in these samples.

We combined data from all the samples by using an estimate of the effective sample size of each population. For the non-Icelandic samples, the actual sample size divided by the inflation factor estimated using the method of genomic control[20] was used as an estimate of the effective sample size: 2,750.6 for the Dutch, 1,181.6 for the European Americans, 741.7 for the African Americans and 2,550.1 for the DGI sample. For the Icelandic sample, a rough estimate of 20,000 was used as an estimate of the effective sample size because of the complexities introduced by the family-based association described below. Inappropriate effective sample size estimates will reduce power, but will not affect the validity of the analysis. An overall z score was calculated by summing the z scores weighted by the square root of the effective sample size over all populations, and dividing by the square root of the sum of the sample sizes. An overall estimate of the effect per allele was calculated by weighting together the effects in each population by the population's effective sample size.

We calculated the fraction of variance explained using the formula $2f(1 - f)a^2$, where f is the frequency of the variant and a is its additive effect.

For the calculation of the fraction of variance explained by the SNPs in **Table 1**, we chose only one SNP per locus.

Use of genealogical information to test for association in Iceland. The Icelandic genealogy database currently contains 739,291 Icelanders and is quite complete, with 650,692 of the Icelanders in the database having a father in the database and 651,821 having a mother in the database. It is recorded that 363,960 of the individuals were born in the twentieth century, and of these, 94.7% have both parents in the database, 0.2% have a father but no mother in the database, 1.8% have a mother but no father in the database and 3.2% have no parent in the database. We verified the relatedness of closely related genotyped individuals in the Icelandic genealogy, and we removed links between individuals or between individuals and genotypes as appropriate.

Height measurements exist for many more Icelandic individuals than those who had been genotyped. We used the available genotype information on the relatives of individuals who had not been genotyped in order to extract more information on association from our data. A distribution over the possible genotype configurations of each individual was obtained through peeling[27]. For example, if the genotypes of an individual's parent are available, we can either infer one of his alleles with full certainty (if the parent is homozygous) or know that the individual has one of two alleles with equal probabilities (if the parent is heterozygous). Given the probability distribution of possible genotypes, the standard linear regression model can be implemented by calculating a weighted sum over the possible genotypes. We used only relatives within two meiotic events of the individual, which includes parents, offspring, siblings (full and half), grandparents and grandchildren, as well as spouses of the probands and the probands' children. Height measurements from a total of 49,334 ungenotyped individuals with at least one close genotyped relative were used. Estimates of effects in the Icelandic sample were based only on unadjusted height values for the individuals with both genotypes and height measurements, that is, ignoring the imputed data and the familial adjustment.

It has long been speculated that the high heritability of height is due to multiple small genetic components[1]. Hence, we carried out a simple correction procedure in which a generalized least-squares method was used to predict the height of each individual from the height of his relatives, and we subtracted this predicted value from the individual's height (after correcting for sex and year of birth). The rationale is that the loss of power, a result of partially correcting away the effect of the locus being investigated, is small compared to what is gained by reducing variance by correcting for the multitude of other genetic factors. In the generalized least-squares model, a constant variance is assumed, and the covariance between individuals is assumed to be directly proportional to the kinship coefficient[28] between the individuals. The correlation between individuals was estimated to be 1.604 times the kinship coefficients between individuals, based on relatives up to four meioses apart. Then, assuming the height measurements have a multivariate normal distribution and given the height measurements of an individual's relatives, the height of the individual will have a normal distribution with an easily obtainable mean. We adjusted the height measurements by subtracting 0.8 times this mean value from the observed value. This reduces the correlation between related individuals, and in fact makes the adjusted traits of related individuals slightly negatively correlated. This procedure has similar motivation as that previously described for the GRAMMAR procedure[29], and our implementation of the scheme was similar to that described previously; the most substantial difference between the two methods is that we did not use the trait value of the individual to adjust itself.

Testing biological processes. The Panther database[15] defines sets of genes participating in 242 biological processes. The classification of genes is largely based on the Gene Ontology database[30]. To look for patterns in the results of the genome-wide association scan, we compared the observed number of loci close to genes in each class to what would be expected by chance as described below.

Given a threshold on P values, p, and a window radius, w, we constructed a set, L, of loci by going through the set of SNPs, ordered in ascending order according to the P value attached to the SNP. If the P value of the SNP was less than p and the closest SNP in L was further away than w, then the SNP was added to L. This procedure guarantees that the minimum physical distance between SNPs in L is w. For a given set of genes, G, we defined the observed number of SNPs in L close to genes in G as the number of SNPs that have at least one gene in G within a distance of $w/2$. In our case, we fixed w at 400 kb and p at 1.6×10^{-7}, 10^{-5} or 10^{-4}. The only instance of SNPs in L being close was in the HLA region, and no gene was close to two SNPs in L. To test whether the observed number of SNPs in L close to genes in G was consistent with that expected by chance, we compared it to the expected number of SNPs close to genes in G from the genome-wide set of SNPs, using Fisher's exact test. The expected number of SNPs close to genes in G is the number of SNPs within a distance of $w/2$ of at least one gene in G divided by the total number of SNPs.

Correlation between genotype and expression. We collected blood in the morning, between 8 and 10 a.m., after subjects had fasted overnight (from 9 p.m.), and we extracted RNA within 2 h from phlebotomy from 1,002 individuals. Subcutaneous fat samples (5–10 cm^3) were removed through a 3 cm incision at the bikini line (always from the same site to avoid site-specific variation) after local anaesthesia using 10 ml of lidocaine-adrenalin (1%) from 673 individuals. Each labeled RNA sample including reference pools—1,765 samples in total—was hybridized to a Human 25K array (Agilent). The hybridizations went through a standard quality control process; we checked signal to noise ratio, reproducibility and accuracy at spike-in compounds by comparing Cy3 to Cy5 intensities. Individuals for whom this expression data existed were genotyped with the Illumina 317K chip, and expression was correlated with SNP data.

Note: Supplementary information is available on the Nature Genetics website.

AUTHOR CONTRIBUTIONS

D.F.G., G.B.W., G.T., H.S., P.S. and K.S. wrote the first draft of the paper. G.B.W., S.T., E.J.O., G.H.O., T. Jonsson, L.T. and T.R. participated in the collection of Icelandic data. V.S., K.B.-J., T.H., G.A., T. Jorgensen and O.P. collected the Danish data. K.K.A., J.A.W., D.W.S., M.H., B.F., A.L.M.V. and L.A.K. collected the Dutch data. D.M.B., L.R.Y. and L.C.B. collected the US data. D.F.G., G.T., H.S., B.V.H., P.Z., P.S., A.G., S.S. and A.I. analyzed the data. G.B.W., A.H. and U.T. carried out the genotyping. D.F.G., J.G., A.K., U.T. and K.S. planned and supervised the work. All authors contributed to the final version of the paper.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Fisher, R.A. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinburgh* **52**, 399–433 (1918).
2. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
3. Mamada, M. *et al.* Fibrillin I gene polymorphism is associated with tall stature of normal individuals. *Hum. Genet.* **120**, 733–735 (2007).
4. Weedon, M.N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nat. Genet.* **39**, 1245–1250 (2007).
5. Sanna, S. *et al.* Common variants in the *GDF5-UQCC* region are associated with variation in human height. *Nat. Genet.* **40**, 198–203 (2008).
6. Carmichael, C.M. & McGue, M. A cross-sectional examination of height, weight, and body mass index in adult twins. *J. Gerontol. A Biol. Sci. Med. Sci.* **50**, B237–B244 (1995).
7. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.* **6**, 399–408 (2003).
8. Ogden, C.L., Fryar, C.D., Carroll, M.D. & Flegal, K.M. Mean body weight, height, and body mass index, United States 1960–2002. *Advance data from vital and health statistics* **347** (2004).
9. van der Eerden, B.C., Karperien, M. & Wit, J.M. Systemic and local regulation of the growth plate. *Endocr. Rev.* **24**, 782–801 (2003).
10. Palmert, M.R. & Hirschhorn, J.N. Genetic approaches to stature, pubertal timing, and other complex traits. *Mol. Genet. Metab.* **80**, 1–10 (2003).
11. Perola, M. *et al.* Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet.* **3**, e97 (2007).
12. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
13. Filion, G.J. *et al.* A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol. Cell. Biol.* **26**, 169–181 (2006).

14. Kiefer, H. *et al.* ZENON, a novel POZ Kruppel-like DNA binding protein associated with differentiation and/or survival of late postmitotic neurons. *Mol. Cell. Biol.* **25**, 1713–1729 (2005).
15. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284–D288 (2005).
16. Kuno, K. *et al.* ADAMTS-1 cleaves a cartilage proteoglycan, aggrecan. *FEBS Lett.* **478**, 241–245 (2000).
17. Milewicz, D.M. *et al.* A mutation in *FBN1* disrupts profibrillin processing and results in isolated skeletal features of the Marfan syndrome. *J. Clin. Invest.* **95**, 2373–2378 (1995).
18. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
19. Kim, K.S. *et al.* Investigation of obesity candidate genes on porcine fat deposition quantitative trait loci regions. *Obes. Res.* **12**, 1981–1994 (2004).
20. Devlin, B., Bacanu, S.A. & Roeder, K. Genomic Control to the extreme. *Nat. Genet.* **36**, 1129–1130 (2004).
21. Gulcher, J.R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
22. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659–662 (2006).
23. Kutyavin, I.V. *et al.* A novel endonuclease IV post-PCR genotyping system. *Nucleic Acids Res.* **34**, e128 (2006).
24. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
25. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
26. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39**, 1443–1452 (2007).
27. Elston, R.C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).
28. Malecot, G. *Les Mathématiques de l'Héredité* (Masson, Paris, 1948).
29. Aulchenko, Y.S., de Koning, D.J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577–585 (2007).
30. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).