ARTICLE

doi:10.1038/nature21039

Rare and low-frequency coding variants alter human adult height

A full list of authors and affiliations appears in the online version of the paper.

Height is a highly heritable, classic polygenic trait with approximately 700 common associated variants identified through genome-wide association studies so far. Here, we report 83 height-associated coding variants with lower minor-allele frequencies (in the range of 0.1–4.8%) and effects of up to 2 centimetres per allele (such as those in *IHH*, *STC2*, *AR* and *CRISPLD2*), greater than ten times the average effect of common variants. In functional follow-up studies, rare height-increasing alleles of *STC2* (giving an increase of 1–2 centimetres per allele) compromised proteolytic inhibition of PAPP-A and increased cleavage of IGFBP-4 *in vitro*, resulting in higher bioavailability of insulin-like growth factors. These 83 height-associated variants overlap genes that are mutated in monogenic growth disorders and highlight new biological candidates (such as *ADAMTS3*, *IL11RA* and *NOX4*) and pathways (such as proteoglycan and glycosaminoglycan synthesis) involved in growth. Our results demonstrate that sufficiently large sample sizes can uncover rare and low-frequency variants of moderate-to-large effect associated with polygenic human phenotypes, and that these variants implicate relevant genes and pathways.

Human height is a highly heritable, polygenic trait[1,2]. The contribution of common DNA sequence variation to inter-individual differences in adult height has been systematically evaluated through genome-wide association studies (GWAS). This approach has thus far identified 697 independent variants located within 423 loci that together explain around 20% of the heritability of height[3]. As is typical of complex traits and diseases, most of the alleles that affect height that have been discovered so far are common (with a minor allele frequency (MAF) > 5%) and are mainly located outside coding regions, complicating the identification of the relevant genes or functional variants. Identifying coding variants associated with a complex trait in new or known loci has the potential to help pinpoint causal genes. Furthermore, the extent to which rare (MAF < 1%) and low-frequency (1% < MAF ≤ 5%) coding variants also influence complex traits and diseases remains an open question. Many recent DNA sequencing studies have identified only a few of these variants[4–8], but this limited success could be due to their modest sample size[9]. Some studies have suggested that common sequence variants may explain the majority of the heritable variation in adult height[10]. It is therefore timely to assess whether and to what extent rare and low-frequency coding variations contribute to the genetic landscape of this model polygenic trait.

In this study, we used an ExomeChip[11] to test the association between 241,453 variants (of which 83% are coding variants with a MAF ≤ 5%) and adult height variation in 711,428 individuals (discovery and validation sample sizes were 458,927 and 252,501, respectively). The ExomeChip is a genotyping array designed to query in very large sample sizes coding variants identified by whole-exome DNA sequencing of approximately 12,000 participants. The main goals of our project were to determine whether rare and low-frequency coding variants influence the architecture of a model complex human trait (in this case, adult height) and to discover and characterize new genes and biological pathways implicated in human growth.

Coding variants associated with height

We conducted single-variant meta-analyses in a discovery sample of 458,927 individuals, of whom 381,625 were of European ancestry. We validated our association results in an independent set of 252,501 participants. We first performed standard single-variant association analyses (Extended Data Figs 1–3 and Supplementary Tables 1–11; technical details of the discovery and validation steps are presented in the Methods). In total, we found 606 independent ExomeChip variants at array-wide significance ($P < 2 \times 10^{-7}$), including 252 non-synonymous or splice-site variants (Methods and Supplementary Table 11). Focusing on non-synonymous or splice-site variants with a MAF < 5%, our single-variant analyses identified 32 rare and 51 low-frequency height-associated variants (Extended Data Tables 1, 2). To our knowledge, these 83 height variants (MAF range of 0.1–4.8%) represent the largest set of validated rare and low-frequency coding variants associated with any complex human trait or disease to date. Among these 83 variants, there are 81 missense, one nonsense (in *CCND3*), and one essential acceptor splice site (in *ARMC5*) variants.

We observed a strong inverse relationship between MAF and effect size (Fig. 1). Although power limits our capacity to find rare variants with small effects, we know that common variants with effect sizes comparable to the largest seen in our study would have been easily discovered by prior GWAS, but were not detected. Our results agree with a model based on accumulating theoretical and empirical evidence that suggest that variants with strong phenotypic effects are more likely to be deleterious, and therefore rarer[12,13]. The largest effect sizes were observed for four rare missense variants, located in the androgen receptor gene *AR* (NCBI single nucleotide polymorphism (SNP) reference ID: rs137852591; MAF = 0.21%, $P_{combined} = 2.7 \times 10^{-14}$), in *CRISPLD2* (rs148934412; MAF = 0.08%, $P_{combined} = 2.4 \times 10^{-20}$), in *IHH* (rs142036701, MAF = 0.08%, $P_{combined} = 1.9 \times 10^{-23}$), and in *STC2* (rs148833559, MAF = 0.1%, $P_{combined} = 1.2 \times 10^{-30}$). Carriers of the rare *STC2* missense variant are approximately 2.1 cm taller than non-carriers, whereas carriers of the remaining three variants (or hemizygous men that carry a rare X-linked *AR* allele at rs137852591) are approximately 2 cm shorter than non-carriers. By comparison, the mean effect size of common height alleles is ten times smaller in the same dataset. Across all 83 rare and low-frequency non-synonymous variants, the minor alleles were evenly distributed between height-increasing and height-decreasing effects (48% and 52%, respectively) (Fig. 1 and Extended Data Tables 1, 2).

Coding variants in new and known height loci

Many of the height-associated variants discovered in this study are located near common variants previously associated with height.