Sizing up human height variation

Peter M Visscher

Genome-wide association studies have identified many variants affecting susceptibility to disease. Now, three studies use this approach to study adult height variation in a combined sample size of ~63,000 individuals and report a total of 54 validated variants influencing this trait.

Human height is a classical quantitative trait. Genetic studies, pioneered by Francis Galton and Ronald Fisher more than 90 years ago[1,2], show a clear pattern of family resemblance, consistent with a polygenic additive model of inheritance. But how many is 'poly', and what are the effect sizes of individual variants? In this issue, three consortia of research groups report a total of 54 loci affecting height variation in the population, identified using genome-wide association studies of hundreds of thousands of genetic markers genotyped on a total of ~63,000 people measured for height[3–5].

Quantitative genetics of height

Height is a highly heritable trait that is easy to measure. It is a 'model' trait for quantitative genetics and developmental biology, it is associated with disease, including cancer[6], and it is a predictor of social outcomes in life. In populations of European descent, the average height is ~178 cm for males and ~165 cm for females, with a s.d. of ~7 cm. Its heritability is ~0.8, which means that within a population, about 80% of the variation in height among individuals is due to genetic factors[7]. If we knew all the variants that are responsible for genetic variation and summed their effects within each person, then the difference between the top 5% and bottom 5% of the population would be about 26 cm, roughly the size of an adult head. Mutations in certain genes can cause extreme short or tall stature, but such mutations are rare and do not contribute to 'normal' variation in the population.

Peter M. Visscher is at the Queensland Institute of Medical Research, Brisbane, Queensland 4029, Australia.
e-mail: peter.visscher@qimr.edu.au

Genetic linkage studies look for a correlation between phenotypic and genotypic similarity within families. Many such studies have been reported for height, implicating many (some would say all) genomic regions, but the resolution of these studies is low, and they have not resulted in the identification of genetic variants that explain the linkage signals. Past studies of candidate genes have also not succeeded in explaining familial resemblance for height. Indeed, similar observations can be made for the studies of many other complex traits, including those of common diseases.

Genome-wide association era

Enter the era of genome-wide association studies (GWAS). These studies take a systematic 'unbiased' approach by interrogating the entire genome for associations between common gene variants (single nucleotide polymorphisms or SNPs) and a phenotype. GWAS have been facilitated by the HapMap project[8], which quantified the total number and genome locations of SNP markers that need to be genotyped in order to detect an association between common genetic variants and a trait in a hypothesis-free genome scan. Further, development of commercial 'SNP chips' that allow rapid genotyping of hundreds of thousands of common SNPs have made this approach possible. The basic design of a GWAS is to associate the trait of interest—for example, disease susceptibility or a quantitative trait—with SNPs on the chip, usually by carrying out a statistical test for each SNP in turn and then following up on the best 'hits' by genotyping them in one or more independent samples for statistical validation. GWAS have been successful in finding previously unknown loci associated with a wide range of diseases[9] and other phenotypes. Two previous GWAS, from two of the three groups that report findings in the current issue, each reported a single newly identified validated variant affecting height[10,11], with each locus explaining a very small proportion of the phenotypic variance (~0.3% to ~0.5%).

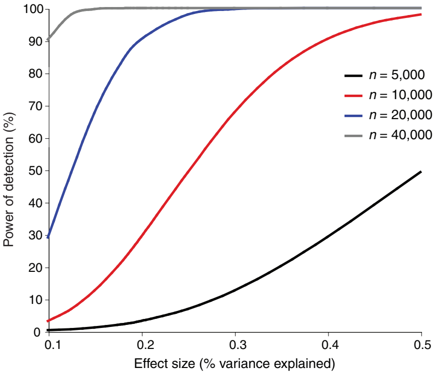

Figure 1 Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5×10^{-7} (calculated using the Genetic Power Calculator[15]). Shown is the power to detect a variant with a given effect size, assuming this type I error rate, which is typical for a GWAS with a sample size of n = 5,000–40,000.

The new studies in this issue[3–5] followed a multistage design in which the first stage was used to select the most promising SNPs and the later stages were used for validation. The validation stage is important because, when over 500,000 variants are tested, as was done in the current studies, many will be statistically 'significant' by chance (about 25,000 if a standard type I error rate of 0.05 is used). The studies employed very large samples sizes, from ~14,000–34,000 in the test stage to ~6,000–20,000 in the validation stages. The total number of SNP chips used across these studies was over 63,000, representing a total investment of roughly $30 million. However, much, if not most, of the genotyping was done by other studies investigating disease, and the researchers cleverly piggy-backed on these studies, taking advantage of the fact that many such studies have included measures on height.

Across the three studies, 95 SNPs were taken forward for validation and 54 survived stringent significance testing. These SNPs were robustly associated with height variation in the general population. Reassuringly, SNPs in the previously associated genes *HMGA2* and *GDF5-UQCC* were again identified. SNPs in three genes (*ZBTB38*, *HHIP*, *HMGA2*) were found to be associated with height in all three studies, and a total of seven genes were implicated in two of the three studies. The three studies found an impressive total of 40 previously unknown variants. The average effect size per 'increasing' allele was ~0.4 cm, or ~0.8 cm between the two homozygous classes. The reason why the overlap in loci is modest is presumably because of the stringent significance testing used by the researchers and because the effect sizes are small. The probability (power) of detecting a variant of small effect when using a genome-wide significance threshold is far from perfect, even when using fairly large sample sizes (**Fig. 1**). Lowering the significance threshold would increase the power but would also yield more false positives. It is a challenge of GWAS designs to get this balance right.

In the three height studies[3–5], the sample sizes of the test stages were approximately 14,000, 16,000 and 34,000. As shown in **Figure 1**, a GWAS with a sample size of 10,000, with an assumed type I error rate of 5×10^{-7}, has 29% power to detect a variant that explains 0.2% of the variance. Therefore, the chance that two independent studies of this size both detect this variant is only $0.29^2 = 0.08$. For sample sizes of 20,000, the probability of detecting the variant in two studies is 0.81. Therefore, one explanation for why there was only modest overlap in the detected loci between the studies is statistical power at the chosen stringent type I error rates.

What have we learned about the nature of quantitative trait variation for height from these studies? At a first glance it looks quite simple: variation is explained by many variants of small effects, with no evidence for interactions between alleles, either within loci (dominance) or between loci (epistasis), and there are no strong differences in effects between males and females. These observations are consistent with patterns of familial resemblance for height. However, given the design and analysis used, there was little statistical power to find evidence for departures from this simple model. Not surprisingly, given the small effect sizes found, there was no significant overlap between the location of the associated variants and previously reported loci from linkage studies. It remains a challenge to reconcile the findings of GWAS and linkage studies, because the former suggest individual variants with small effects, whereas the latter suggest genomic regions with large effects within pedigrees.

A return to candidate genes?

The GWAS approach, in general, is set to identify associations with individual variants that are unlikely to be causal. Nevertheless, by looking at the genes in the vicinity of the associated SNPs, one can look for biological pathways that are overrepresented. Taking this approach, the height studies report that their validated SNPs are in or near genes involved in pathways related to mesoderm development, skeletal development, mitosis, cancer, Hedgehog signaling and chromatin remodeling. Variants in one of the genes (*ZBTB38*) that was detected in all three studies were shown to be associated with gene expression of that same gene in blood and adipose tissue[5], suggesting a possible biological mechanism.

The authors also found a significant number of the validated SNPs in genes that could be considered 'candidate genes' on the basis of known mutant phenotypes in humans or mice. But wasn't one of the rationales of the GWAS approach that previous candidate gene studies hadn't worked? One explanation for this (perceived) failure of earlier candidate gene studies is that they were severely underpowered, both in terms of experimental sample size and SNP coverage. It is only by performing large, unbiased GWAS that the relative importance of candidate genes can be quantified.

Looking to the future

The main conclusion emerging from the current studies is that GWAS are able to robustly identify common variants that are associated with height but that the effect sizes of individual variants are small, so that very large sample sizes are needed to detect associations reliably. Single laboratories are unlikely to have sufficient sample sizes to do powerful studies on their own, and the trend in human complex trait mapping has been to create consortia of research groups and even consortia of consortia.

It remains unclear at this stage how much genetic variation can be explained through the GWAS approach. However, if the samples in these three studies were combined together with other datasets that have been collected on height and genome-wide SNP data, then this question could be answered empirically. Genome-wide studies on, say, 100,000 individuals, unthinkable only a few years ago, will be soon be a reality. From the limited amount of overlap in the genes reported across the three studies, and the fact that all three studies were very stringent in their hypothesis testing, it seems that much more variation is likely to be explained by common variants of small effect.

There is still a long way to go from associated SNPs to causal variants. Resequencing will uncover all variants in a given region, not just the common ones, but it will unearth so many that picking out the responsible variant(s) will be a real challenge. Statistical association alone will not provide the answer, and other sources of information, such as predicted effects of variants on proteins and evolutionary conservation[12,13], may help.

Looking further to the future, if most genetic variation for height can be explained by association, even without knowing causal variants, how can this knowledge be used? Apart from gaining insights into the genetic architecture of continuous traits and biological pathways involved in human height, associations between variants and height can in principle be used to predict height from genetic data alone[14], that is, without having observed the phenotype. Such predictions are useful in artificial selection programs in agriculture and may also prove useful in forensics and human medicine.

1. Fisher, R.A. *Trans. Roy. Soc. Edin.* **52**, 399–433 (1918).
2. Galton, F. *Nature* **33**, 295–298 (1886).
3. Weedon, M.N. *et al. Nat. Genet.* **40**, 575–583 (2008).
4. Lettre, G. *et al. Nat. Genet.* **40**, 584–591 (2008).
5. Gudbjartsson, D.F. *et al. Nat. Genet.* **40**, 609–615 (2008).
6. Gunnell, D. *et al. Epidemiol. Rev.* **23**, 313–342 (2001).
7. Visscher, P.M., Hill, W.G. & Wray, N.R. *Nat. Rev. Genet.* **9**, 255–266 (2008).
8. The International HapMap Consortium. *Nature* **437**, 1299–1320 (2005).
9. The Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
10. Sanna, S. *et al. Nat. Genet.* **40**, 198–203 (2008).
11. Weedon, M.N. *et al. Nat. Genet.* **39**, 1245–1250 (2007).
12. Easton, D.F. *et al. Am. J. Hum. Genet.* **81**, 873–883 (2007).
13. Rebbeck, T.R., Spitz, M. & Wu, X. *Nat. Rev. Genet.* **5**, 589–597 (2004).
14. Wray, N.R., Goddard, M.E. & Visscher, P.M. *Genome Res.* **17**, 1520–1528 (2007).
15. Purcell, S., Cherny, S.S. & Sham, P.C. *Bioinformatics* **19**, 149–150 (2003).