Discoveries and Lessons from a Genome-Wide Association Study of Human Height in >5 Million Individuals

L. Yengo[1*] on behalf of GIANT Consortium (the height working group)

[1]Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, Brisbane, QLD 4072; Australia; [*]l.yengo@imb.uq.edu.au

Abstract

In this lecture, I will present findings from the largest, to date, genome-wide association study of human height conducted in >5 million individuals of diverse ancestries. This gargantuan effort has led to identify >12,000 independent associations with common SNPs with a frequency >1%. These height-associated variants are clustered within ~7000 genomic loci enriched for genes which mutations cause abnormal skeletal growth. Across five ancestry and ethnic groups, we found that >90% of the common variants SNP-based heritability is explained by variants within these 7000 loci, which cumulatively cover ~21% of the human genome. A genetic predictor combining these 12,000 height-associated variants yields a prediction accuracy of R^2~40%, while combining that predictor with the average height of parents reaches an unprecedented accuracy of R^2~54%. Overall, our study proposes a saturated map of genomic loci likely containing most of the genetic variation responsible for the missing heritability of human height.

Introduction

Over the past 15 years, genome-wide association studies (GWAS) have revolutionised our understanding of the genetic architecture of complex traits by allowing the discovery of tens of thousands of genetic variants associated with thousands of traits. GWAS of human height (a classic polygenic trait) had hitherto identified over 3000 independent SNPs associated with inter-individual differences,[1] explaining about half of the SNP-based heritability.[2] Sample size is the main factor driving GWAS discoveries. However, how large experimental samples need to be before we can account for the totality of the SNP-based heritability has long remained unknown. Here, we perform a GWAS of human height in ~5.4 million individuals with diverse ancestries (Yengo 2022[3]) and show empirically that our sample size is sufficient to map autosomal genomic loci, where genetic variation accounts for 90% to 100% of the SNP-based heritability.

Materials & Methods

We analysed data from 5,380,080 individuals whose ancestries were predominantly European (EUR; 76% of total sample), East-Asian (EAS; 8.8%), Admixed Hispanic ethnicity (HIS; 8.5%), African (AFR; 5.5%) and South-Asian (SAS; 1.4%). We tested association at 1.3 million SNPs catalogued in the HapMap 3 project, which includes over 1 million with a minor allele frequency (MAF) larger than 1% in each of the five ancestral groups. We first meta-analysed cohorts within each ancestral group using a fixed effect meta-analysis then strategy, then meta-analysed summary statistics from all 5 groups of cohorts using fixed effect approach. In total we performed 6 meta-analyses: 5 within-ancestral groups and 1 across-ancestral groups. For each meta-analysis, we identified sets of genome-wide significant (GWS) conditionally independent associations using the approximate conditional and joint multiple-SNP analyses,[4] as implemented in GCTA[5].

We defined GWS loci as genomic regions centred around each GWS SNP and including all SNPs within 35 kb on each side of the lead GWS SNP. Overlapping GWS loci were merged so that the number and cumulative length of GWS loci are calculated on non-overlapping GWS loci. To estimate the total variance explained by SNPs within GWS loci, we (i) partitioned all SNPs from the HapMap 3 panels into two groups (SNPs within those 7209 loci versus SNPs outside those 7209 loci), (ii) calculated genetic relationship matrices based on these two sets of SNPs; and (iii) used these matrices in a GREML analysis to estimate heritability. The cumulative length of non-overlapping GWS loci is ~647 Mb, i.e. ~21% of the genome assuming a genome length of ~3000 Mb.
Finally, we used joint SNP effects from our GWAS meta-analysis (re-estimated using GCTA) to calculate a genetic predictor of height from GWS SNPs only.

Results
We identified 9863, 1888, 918, 493 and 69 independent genome-wide significant (GWS; $P<5\times10^{-8}$) associations in the EUR, HIS, EAS, AFR and SAS groups, respectively (Table 1). We found a strong genomic co-localisation of associations identified in each population and a correlation of estimated SNP effects >0.64 across all pairs of ancestral groups. Next, we detected 12,111 conditionally independent associations in a multi-ancestry meta-analysis including all study participants. We identified various genomic regions with elevated density of associations. For example, we found on chromosome 15 near the *ACAN* gene a 200 kb-long genomic segments containing 25 independent associations. More generally, high density of associations within 100 kb-long genomic window is associated with the presence of an autosomal gene curated from the Online Mendelian Inheritance in Man (OMIM) database[6], which pathogenic mutations are known to cause syndromes of abnormal skeletal growth (>2.5x enrichment; $P<0.001$).
The 12,111 GWS associations were clustered within 7209 non-overlapping GWS loci) covering ~21% of the human genome. Using a partitioned GREML analysis we found that >90% of the common variants SNP-based heritability in 5 ancestry groups can be explained by SNPs located within GWS loci. For European ancestries, which constitutes the majority of our discovery sample, that proportion reaches 100% of the SNP-based heritability. Importantly, these observations suggest that although the majority of HapMap3 SNPs within GWS loci are not GWS themselves (only 12,111 out of ~300,000 SNPs; Figure 1), yet-to-be-discovered height-associated variants are most likely located within these GWS loci.
Finally, we used estimated SNP effects from our GWAS meta-analysis at these 12,111 height-associated SNPs to calculate a polygenic score (PGS) of height (Methods). This PGS explains 40% of height variance in European ancestries individuals (76% of our discovery sample; Methods) but shows a reduced accuracy in individuals whose ancestries are more genetically distant to EUR. In particular, the prediction accuracy of the height PGS is only ~10% in individuals with predominantly African ancestries. Note that a 4-fold reduction in prediction accuracy is consistent with previous studies[7,8] and largely explained by linkage disequilibrium and allele frequency differences between European and African ancestries.[8]

Table 1. [Extracted and simplified from Yengo (2022)] Summary of results from within-ancestry and multi-ancestry GWAS meta-analyses. GWS: Genome-Wide Significant ($P<5\times10^{-8}$). COJO SNPs: near independent GWS SNPs identified using an approximate conditional and Joint analysis implemented in the GCTA software.

Cohort Main Ancestry or Ethnicity	Number of studies	Sample Size	Number of GWS COJO SNPs	Number of GWS loci
European	173	4,080,687	9,863	6,386
East-Asian	56	472,730	918	821
Hispanic	11	455,180	1,888	1,599
African	29	293,593	493	436
South Asian	12	77,890	69	66
Multi-ancestry meta-analysis	281	5,314,291*	12,111	7,209

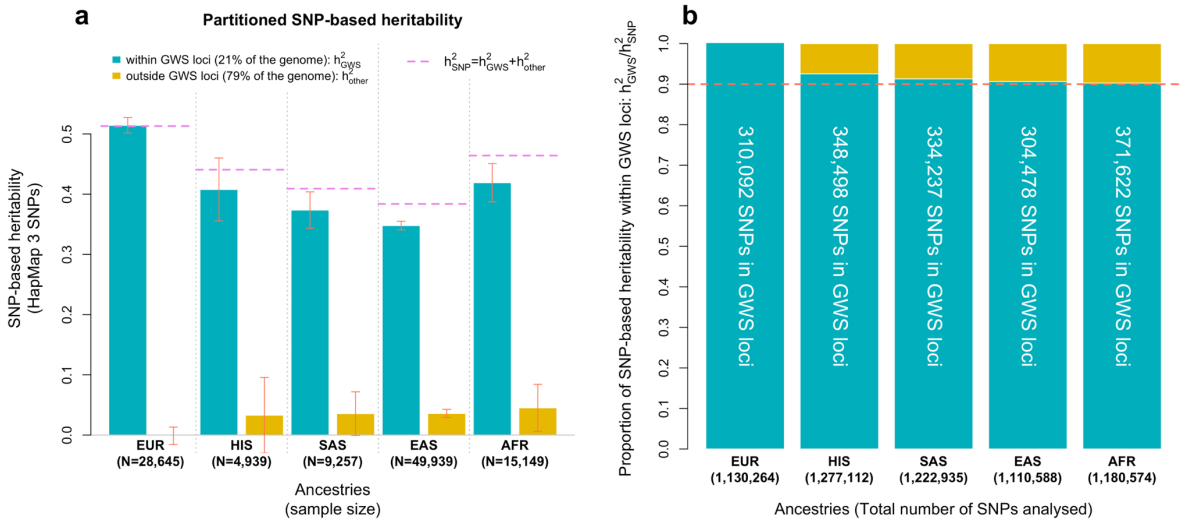

Figure 1. Variance of height explained by HapMap 3 SNP within genome-wide significant (GWS) loci. [Extracted and simplified from Ref.[3]]. **Panel a** shows stratified SNP-based heritability (h^2_{SNP}) estimates obtained after partitioning the genome into SNPs within 35 kb of a GWS SNP ("GWS loci" label) vs. SNPs >35 kb away from any GWS SNP. Analyses were performed in samples of five different ancestry/ethnic groups: European (EUR: meta-analysis of UK Biobank (UKB) + Lifelines study), African (AFR: meta-analysis of UKB + PAGE study), East-Asian (EAS: meta-analysis of UKB + China Kadoorie Biobank), South-Asian (SAS: UKB) and Hispanic group (HIS: PAGE). **Panel b** shows that >90% of ☐ in all ancestries is explained by SNPs within GWS loci identified in this study.

Discussion

In this study, we performed the largest, to date, GWAS of any human trait and thereby identified >12,000 associations explaining >90% of the common variants SNP-based heritability of height. A genetic predictor of height derived from these findings reaches an accuracy of 40% in European ancestry individuals, which rivals that of the average height of parents (i.e. 43%). Interestingly, we showed that these two predictors can be linearly combined to reach an accuracy of 54% and that the optimal weights only depend on the accuracy of the SNP-based predictor, the narrow sense heritability and the spousal correlation for traits, like height, subjected to assortative mating in the population.

By quantifying the relationship between sample sizes and GWAS discoveries, we found that prioritisation of the relevant biological pathways and gene sets is possible at relatively small sample sizes (N<250,000), while identification of variants almost linearly increases with sample size, thereby contributing to increase the density of associations within previously identified loci and reflecting the large allelic heterogeneity for height-associated haplotypes.

Our study has a few limitations, many of which are being currently addressed. The first one is the over-representation of European ancestries in our discovery samples, which limits the transferability of some of our findings to other human populations. Another limitation is that our study focused on common genetic variation (>1%), while our large sample size could, in principle, yield enough power to detect associations involving lower frequency SNPs. However, our large sample is in fact an aggregation of many small studies, which creates a challenge to distinguish true signals from study-specific artefacts.

Altogether, our study reveals that extremely large sample sizes can resolve the missing heritability problem by mapping specific genomic regions likely containing most of the causative genetic variation for complex traits like human height.

References

1. Yengo, L. *et al*. (2018) *Hum. Mol. Genet.* 27(20):3641–3649.
2. Yang, J. *et al.* (2010) *Nat Genet* 42(7):565–569.
3. Yengo, L. *et al.* (2022) doi:10.1101/2022.01.07.475305.
4. Yang, J. *et al.* (2012) *Nat Genet* 44 (4):369–375.
5. Yang, J. *et al*. (2011) *AJHG* 88(1):76–82.
6. Lui, J. C. *et al.* (2012) *Hum Mol Genet* 21(23):5193–5201.
7. Martin, A. R. *et al.* (2019) *Nat Genet* 51(4):584–591.
8. Wang, Y. *et al.* (2020) *Nat Comm* 11(1):3865.