University of Pittsburgh | School of Public Health

A Saturated map of common genetic variants associated with human height -Yengo et al. (2022)

Published in Nature

Presentation by
Dylan Maher
HUGEN 2029
27Sep2023

Article

A saturated map of common genetic variants associated with human height

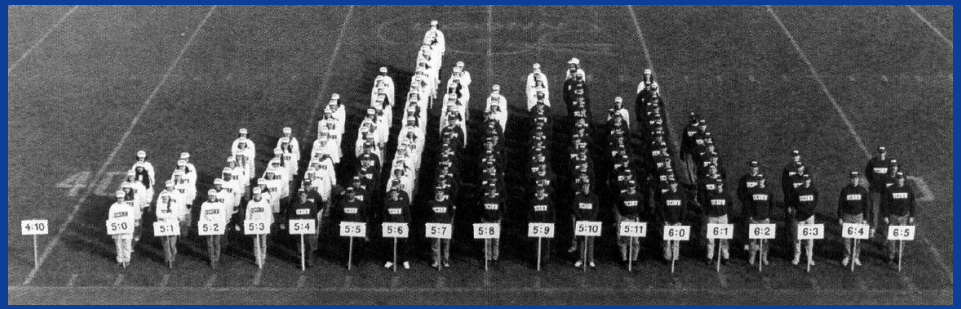
Author contributions Steering committee: G.R.A., T.L.A., S.I.B., M.B., D.I.C., Y.S.C., T.E., T.M.F., I.M.H., J.N.H., G.L., C.M.L., A.E.L., R.J.F.L., M.I.M., K.L.M., M.C.Y.N., K.E.N., C.J.O., Y.O., F. Rivadeneira, Y.V.S., E.S.T., C.J.W., U.T., P. M. Visscher and R.G.W. Conveners of GIANT working groups: S.I.B., P. Deloukas, J.N.H., A.E.J., G.L., C.M.L., R.J.F.L., E.M., K.L.M., K.E.N., Y.O., C.N.S., R.G.W., C.J.W., A. R. Wood and L. Yengo. Writing group (drafted, edited and commented on manuscript): E. Bartell, J.N.H., G.L., E.M., Y.O., S. Raghavan, S. Sakaue., S. Vedantam, P. M. Visscher, A. R. Wood and L. Yengo. Coordinated or supervised data collection or analysis specific to manuscript: A. Auton, P. Deloukas, T.E., T.M.F., S.E.G., J.N.H., A.E.J., G.L., A.E.L., P.-R.L., Y.O., K.S., U.T., P. M. Visscher, R.G.W., A. R. Wood, Jian Yang and L. Yengo. Data preparation group (checked and prepared data from contributing cohorts for meta-analyses): J. D. Arias, S.I.B., S.-H.C., T.F., S.E.G., M. Graff, H.M.H., Y. Ji, A.E.J., T. Karaderi, A.E.L., K. Lüll, D.E.M., E.M., C.M.-G., M.Mo., A. Moore, S. Rüeger, X.S., C.N.S., S. Vedantam, S. Vrieze, T.W.W., X.Y. and K.L.Y. Meta-analysis working group: J.N.H., E.M., S. Vedantam and L. Yengo. Primary height analysis working group (post meta-analysis): E. Bartell, A.D.B., M. Graff, Y. Jiang, M. Kanai, K. Lin, J. Miao, E.M., R. E. Mukamel, S. Raghavan, S. Sakaue, J. Sidorenko, S. Vedantam, A. R. Wood and L. Yengo. All other authors were involved in the design, management, coordination or analysis of contributing studies.

University of Pittsburgh | School of Public Health

Motivating Questions

Height is “ideal” trait

- Easily measured (negligible measurement error)
- Commonly gathered on questionnaires
- Highly polygenic, highly heritable (h^2~0.8) [as estimated from family-based studies]
- Approximately normally distributed



To what extent can (molecular) genetic variation account for phenotypic variation?

How much additional information can be expected as GWAS sample sizes scale up?

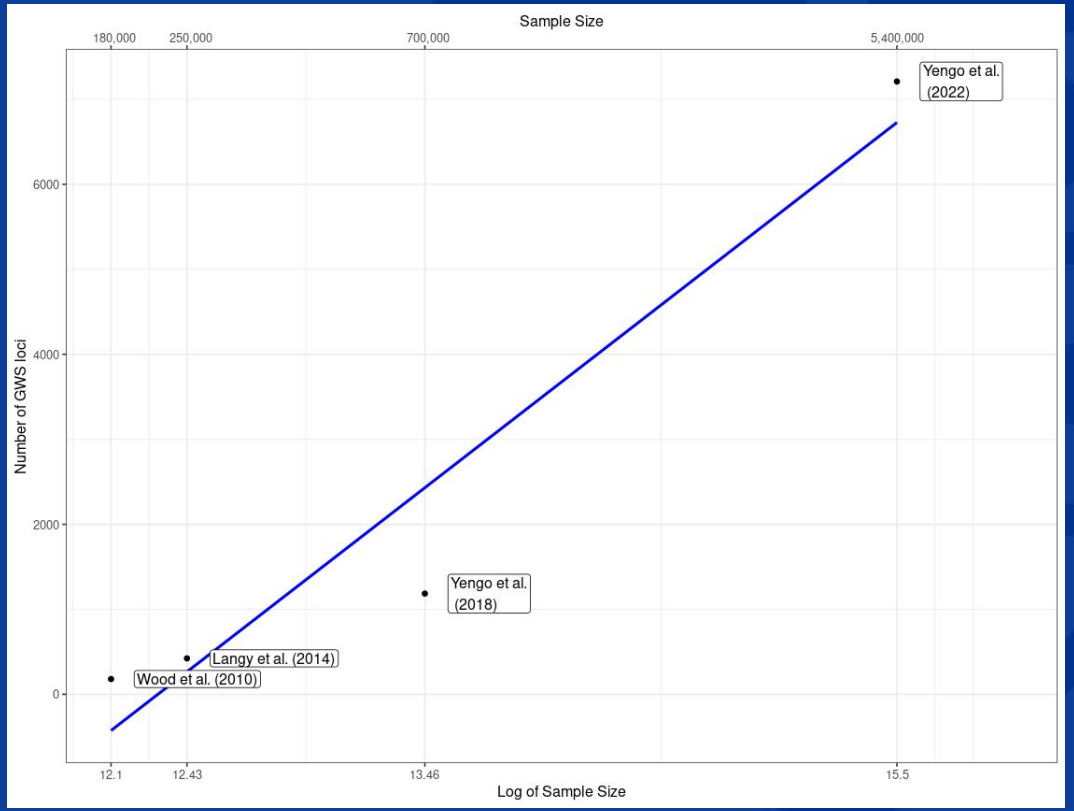
How accurate is cross-ancestry prediction (and why)?

University of Pittsburgh | School of Public Health

Existing Literature

Year	Study	Sample Size	Loci	SNPs
2010	Lango et al.	180,000	180	“hundreds”
2014	Wood et al.	250,000	423	697
2018	Yengo et al.	700,000	1185	3290
2022	Yengo et al.	5,400,000	7209	12,111

“...nearly linear relationship between the number of GWS loci and the logarithm of the sample size…”
-Yang et al. (2017)



University of Pittsburgh | School of Public Health

Background

(Imprecise) definition of heritability:

$$\text{heritability} \approx \frac{\text{genotypic variation}}{\text{phenotypic variation}}$$

In 2010, Yang et. al showed that ~45% of variation in height can be explained by common-SNP variation.

This was done by fitting ~295K SNPs simultaneously in a linear model framework.

However, the identity of the SNPs was still unknown.

In 2022, Yengo et al. identified the SNPs and showed 45% of variation in height was attributable to 12,111 SNPs.

ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

Yang et al. (2010)

University of Pittsburgh | School of Public Health

Study Outline

Meta-analysis identifies 12,111 height-associated SNPs

↓

Genomic distribution of height-associated SNPs

↓

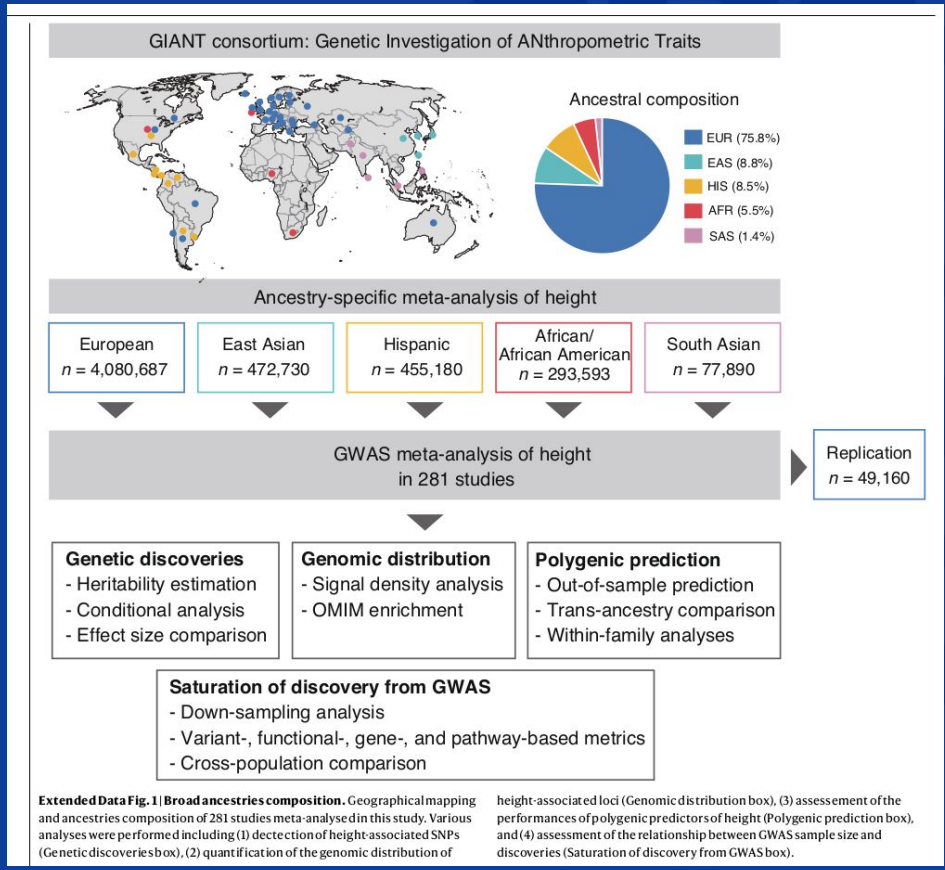
Variance explained by SNPs within identified loci

↓

Out-of-sample prediction accuracy

↓

GWAS discoveries, sample size and ancestry diversity



Extended Data Fig. 1 | Broad ancestries composition. Geographical mapping and ancestries composition of 281 studies meta-analysed in this study. Various analyses were performed including (1) dectection of height-associated SNPs (Genetic discoveries box), (2) quantification of the genomic distribution of height-associated loci (Genomic distribution box), (3) assessement of the performances of polygenic predictors of height (Polygenic prediction box), and (4) assessment of the relationship between GWAS sample size and discoveries (Saturation of discovery from GWAS box).

University of Pittsburgh | School of Public Health

QC/Data Prep

Quality Control checks performed in EasyQC (adapted from RVTESTS)

Checks performed for

- Allele frequency differences with ancestry-specific reference panels
- Total number of markers not present in reference panels
- Imputation quality
- Genomic inflation factor

Thresholds used (each study)

- r^2_{INFO}>0.3
- p_{HW}>1e-08
- MAC > 5

Two studies excluded

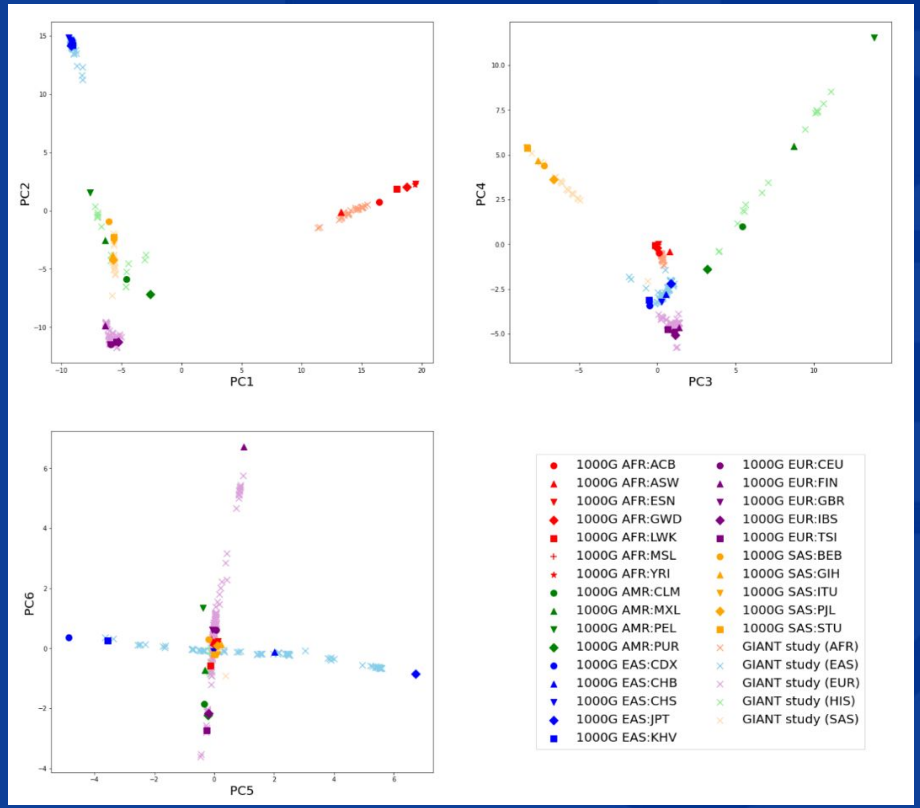
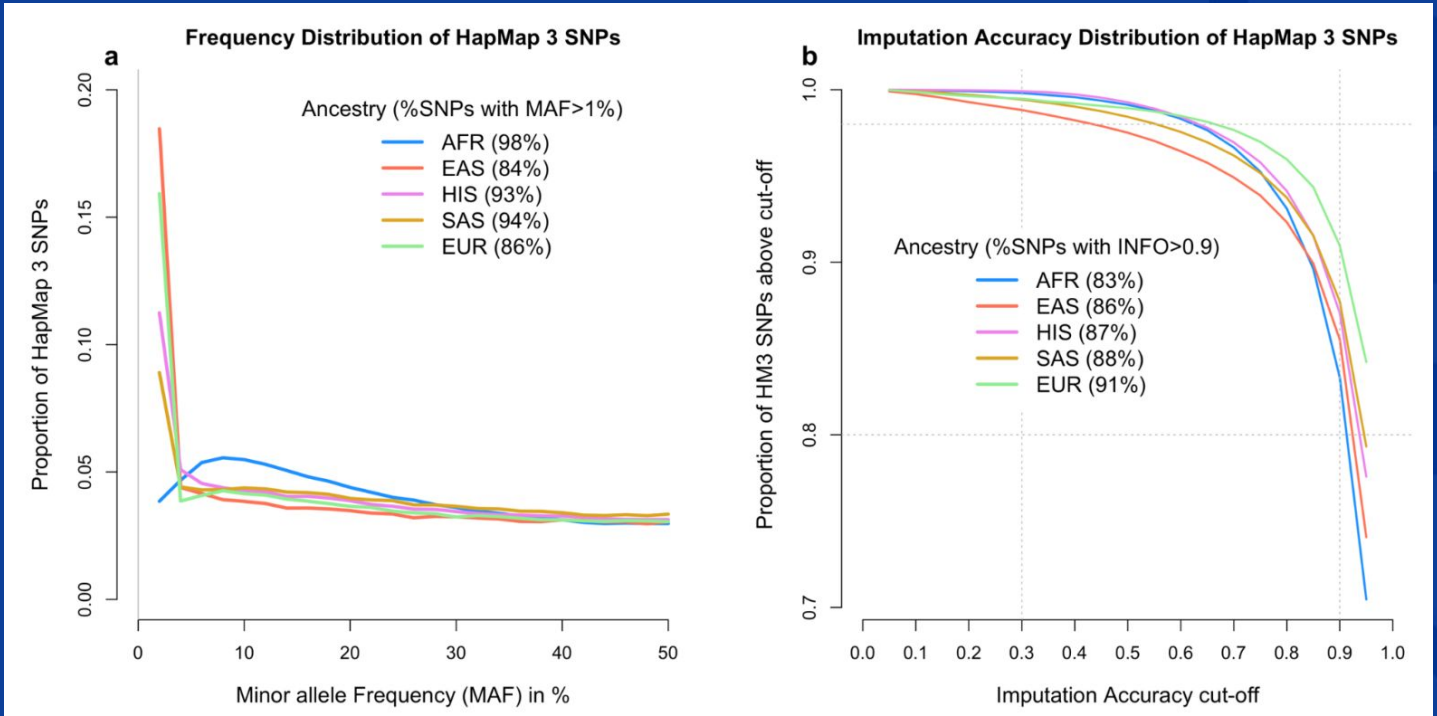


Figure: Pairs of first 6 PCs

University of Pittsburgh | School of Public Health

QC/Data Prep



Left: MAF distribution for each ancestry

Right: >98% HM3 SNPs (INFO>0.3) >80% (INFO>0.9)

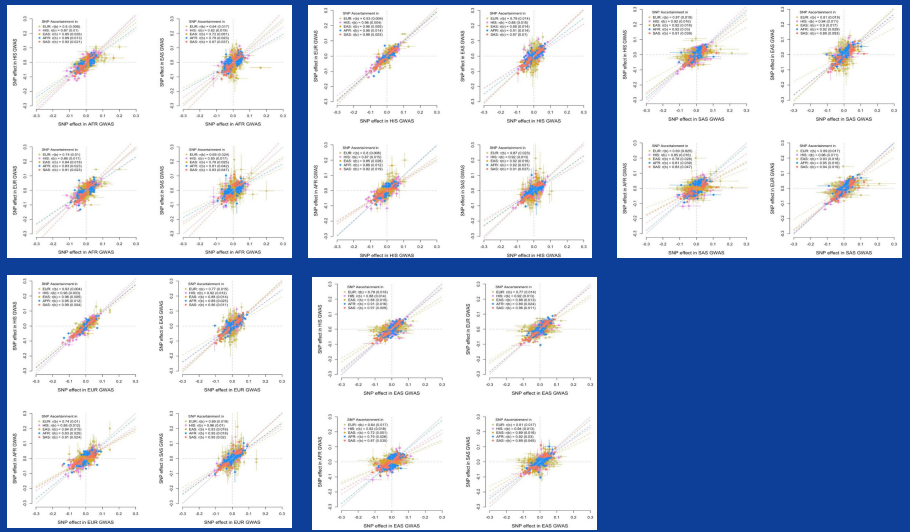
University of Pittsburgh® School of Public Health

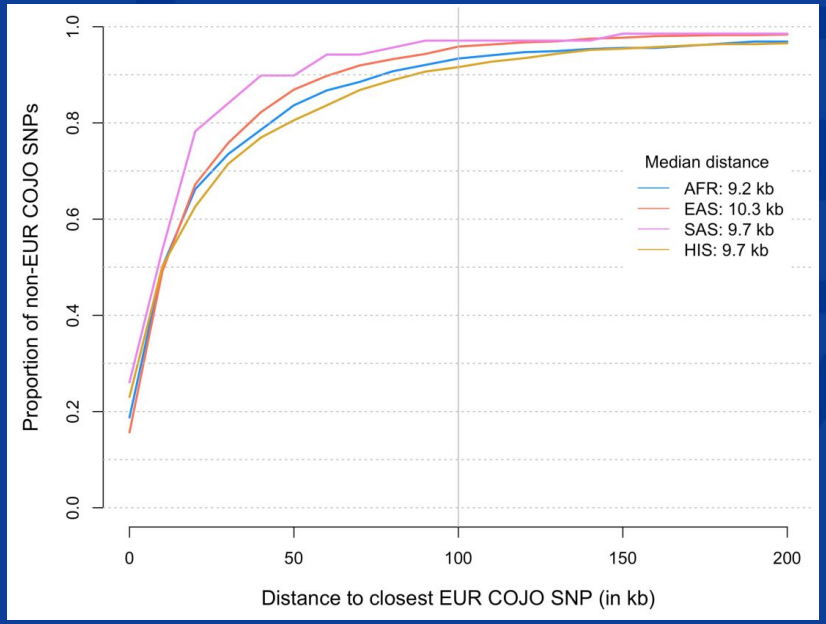
Within-Ancestry Meta-Analysis

- Investigated colocalization & correlation of effects
- Significant overlap–justification for fixed-effect meta-analysis

Table 1 Summary of results from within-ancestry and trans-ancestry GWAS meta-analyses

Cohort ancestry or ethnic group	Number of studies	Max *n* (mean *n*)	Number of GWS COJO SNPs ($P_{GWAS}<5\times10^{-8}$)	Number of GWS loci (35 kb)	Cumulative length of non-overlapping GWS loci in Mb (% of genome)
European (EUR)	173	4,080,687 (3,612,229)	9,863 (8,382)	6,386	552.5 (18.4%)
East Asian (EAS)	56	472,730 (320,570)	918 (807)	821	60.5 (2.0%)
Hispanic (HIS)	11	455,180 (431,645)	1,511 (1,195)	1,373	101.0 (3.3%)
African (AFR)	29	293,593 (222,981)	453 (404)	412	30.4 (1.0%)
South Asian (SAS)	12	77,890 (59,420)	69 (65)	66	4.7 (0.2%)
Trans-ancestry meta-analysis ($META_{FE}$)	281	5,314,291* (4,611,160)	12,111 (9,920)	7,209	647.5 (21.6%)



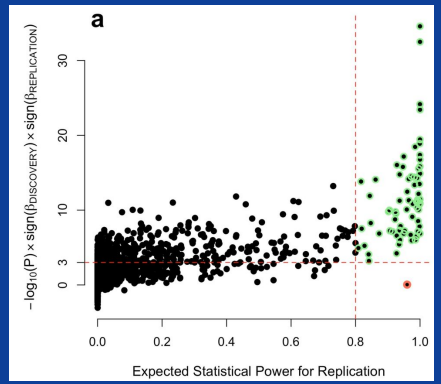


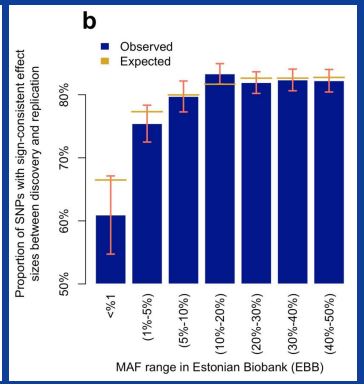
University of Pittsburgh | School of Public Health

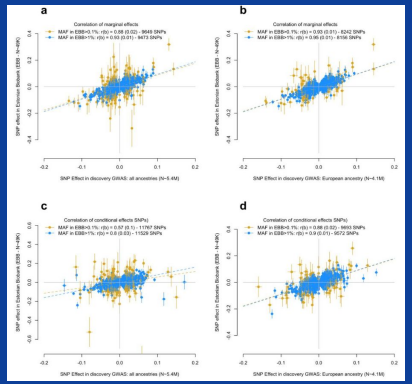
Between-Ancestry Meta-Analysis

Table 1 Summary of results from within-ancestry and trans-ancestry GWAS meta-analyses

Cohort ancestry or ethnic group	Number of studies	Max *n* (mean *n*)	Number of GWS COJO SNPs ($P_{GWAS}<5\times10^{-8}$)	Number of GWS loci (35 kb)	Cumulative length of non-overlapping GWS loci in Mb (% of genome)
European (EUR)	173	4,080,687 (3,612,229)	9,863 (8,382)	6,386	552.5 (18.4%)
East Asian (EAS)	56	472,730 (320,570)	918 (807)	821	60.5 (2.0%)
Hispanic (HIS)	11	455,180 (431,645)	1,511 (1,195)	1,373	101.0 (3.3%)
African (AFR)	29	293,593 (222,981)	453 (404)	412	30.4 (1.0%)
South Asian (SAS)	12	77,890 (59,420)	69 (65)	66	4.7 (0.2%)
Trans-ancestry meta-analysis ($META_{FE}$)	281	5,314,291* (4,611,160)	12,111 (9,920)	7,209	647.5 (21.6%)





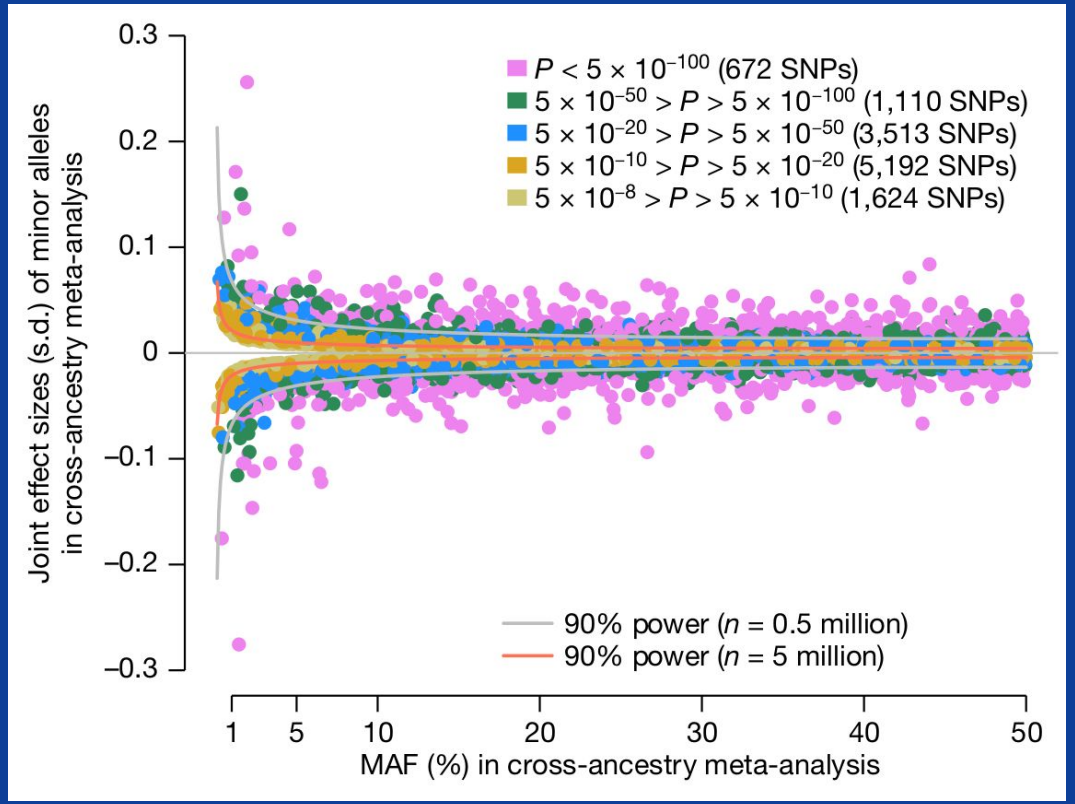


- 12,111 associations obtained (COJO)
- Replicated* with Estonian Biobank (EBB)

*Due to sample size discrepancy, direct replication not possible. Instead, replicability assessed by correlation of SNP effects

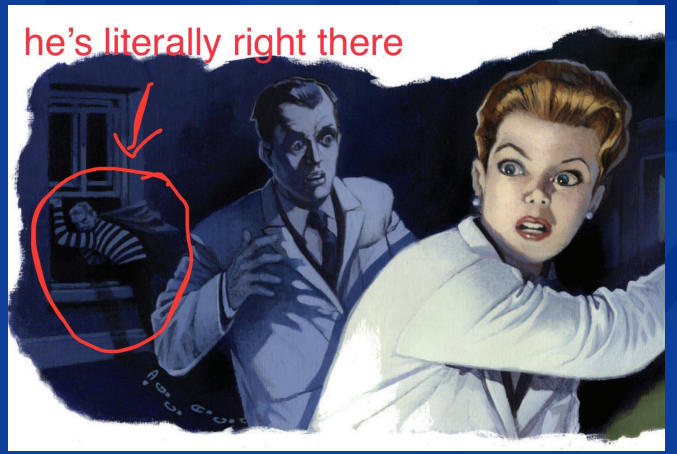
University of Pittsburgh | School of Public Health

Meta-Analysis



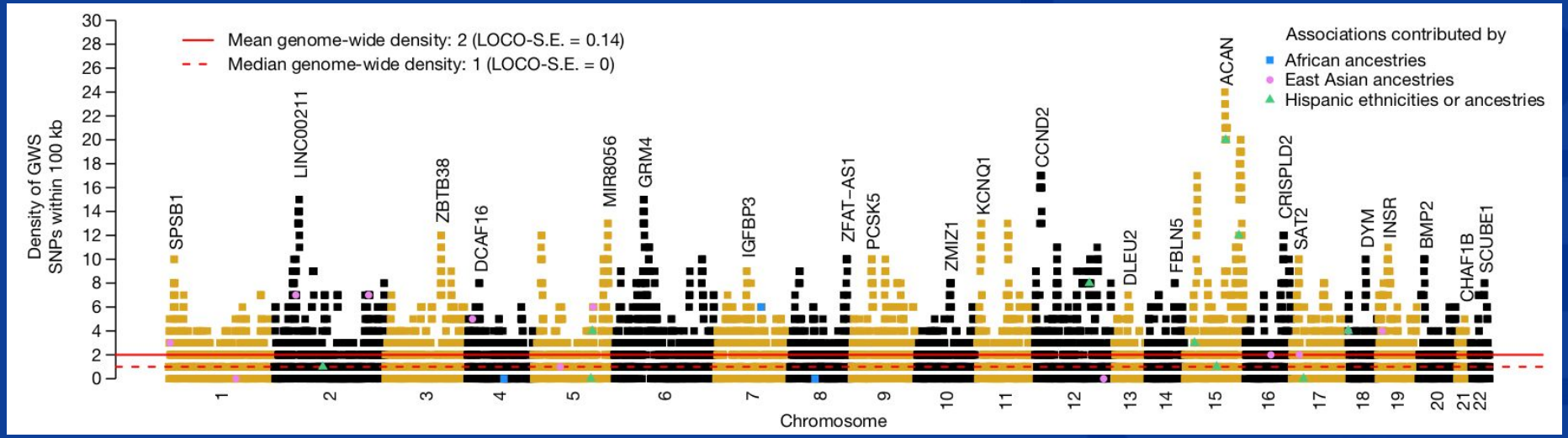
Genetic architecture of height (n=5.4M)

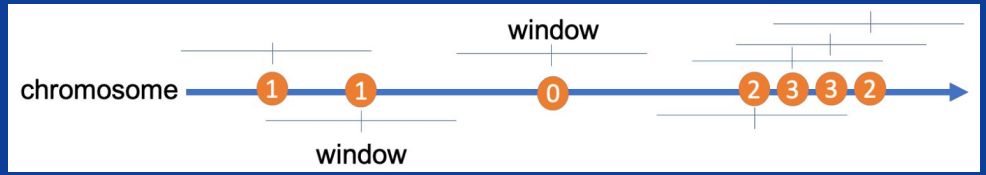
…or the “““missing heritability problem””””, solved



University of Pittsburgh | School of Public Health

Genomic Distribution

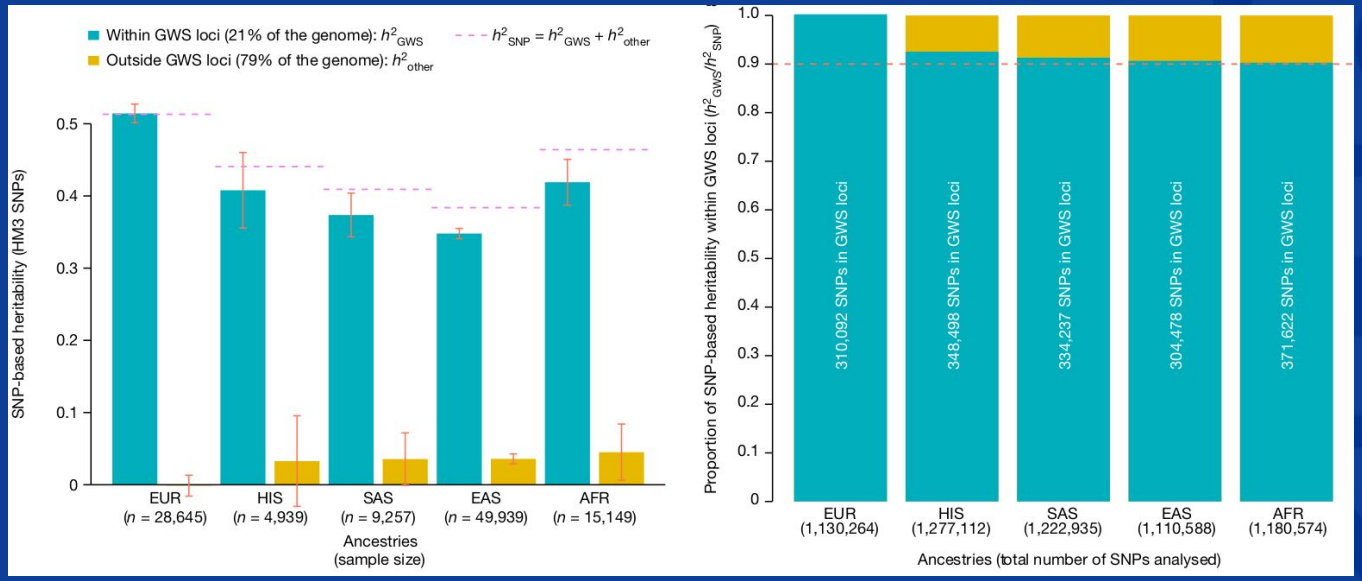




- 69% of GWS SNPs co-localized with other (conditionally indep.) associated SNPs
- 12,111 SNPs cluster non-randomly across genome

University of Pittsburgh | School of Public Health

Variance Explained



Figures: h^2_{SNP} (left), and proportion of h^2_{SNP} (right)

Takeaway: SNPs grouped in loci account for **100% (!!!)** of the expected h^2_{SNP}

12,111 SNPs group into 7209 non-overlapping loci, lengths 70kb-711kb (mean = 90kb, total length 647kb, 21% of genome)

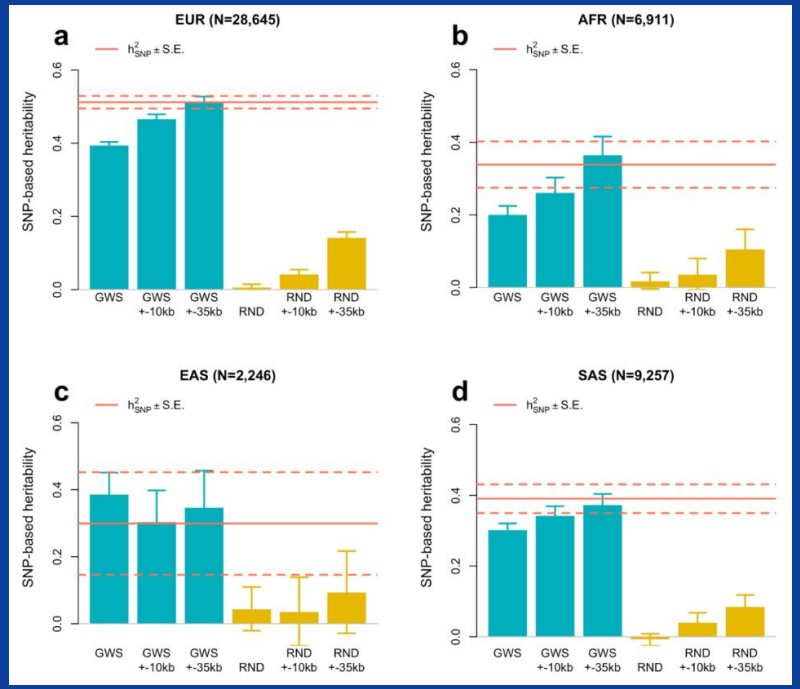
Two GRMs constructed

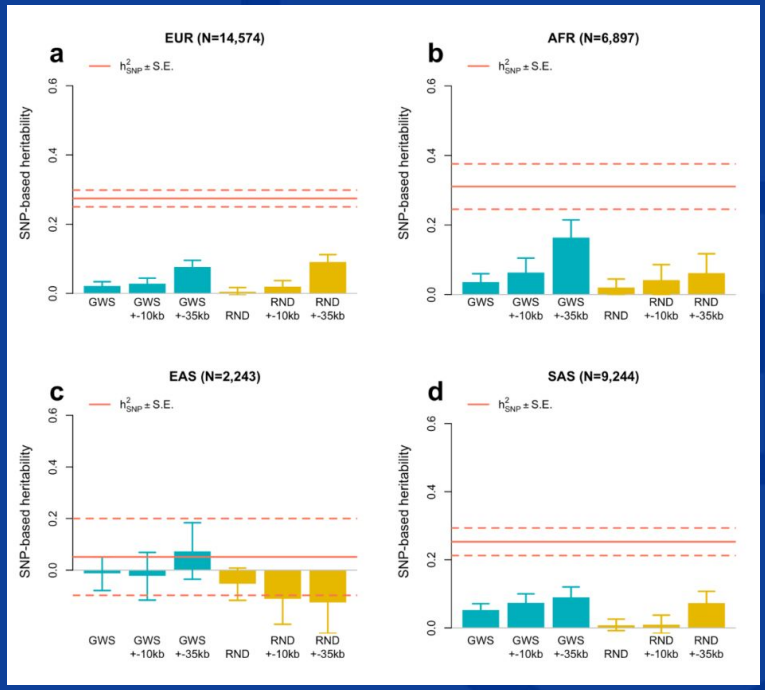
1. SNPs inside loci
2. SNPs outside loci

Estimated stratified h^2_{SNP} in of all five population groups

University of Pittsburgh | School of Public Health

Variance Explained (Robustness Check)

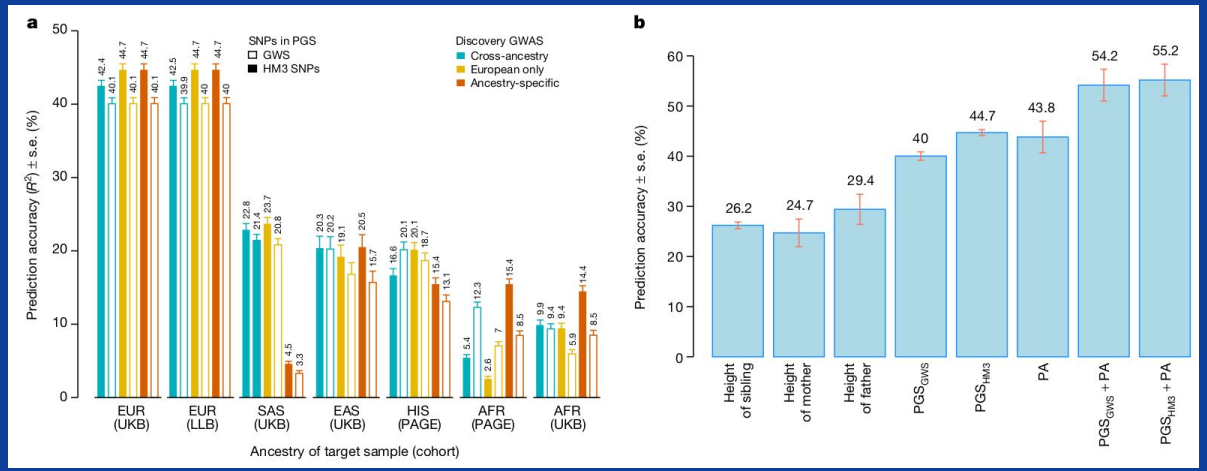




Turquoise: GWS SNPs, Gold: Randomly drawn SNPs. Left: Height, Right: BMI

University of Pittsburgh | School of Public Health

Prediction Accuracy



$$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{\pi m}\right), & \text{with probability of } \pi \\ 0, & \text{with probability of } 1-\pi \end{cases}$$

Model outline of SBayesC (m=# of variants)

Two sets of PRS constructed for each ancestry group:

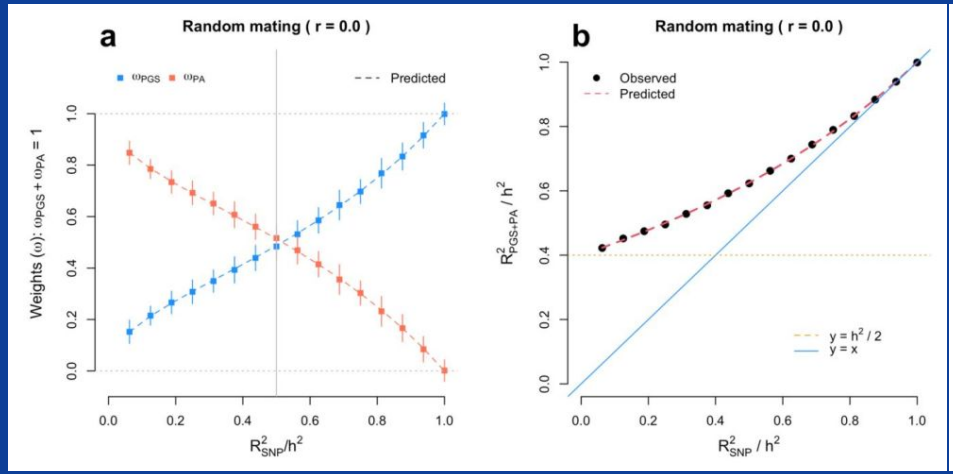
1. Meta-analysis SNPs (COJO)
2. HapMap3 SNPs (SBayesC)

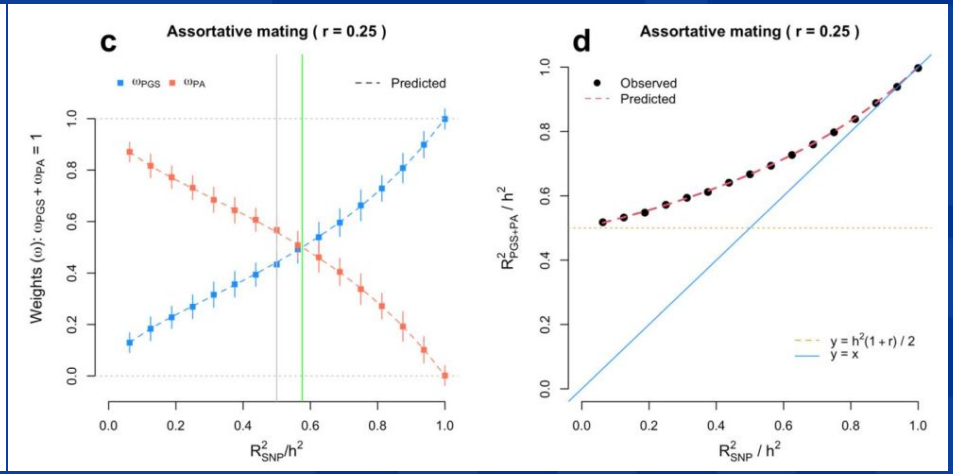
PRS most accurate for ancestry trained on (consistent with previous empirical work as well as theory)

Additional analysis for prediction when compared with phenotypic measurements–PRS outperform family info

University of Pittsburgh | School of Public Health

Prediction Accuracy (interesting theoretical note)





we show that the optimal weighting between parental average and PGS can be predicted theoretically as a function of the prediction accuracy of the PGS, the full narrow sense heritability and the phenotypic correlation between spouses (Supplementary Note 4 and Supplementary Fig. 20).

Overview of theory and theoretical results proved in Supplementary Note 4.

University of Pittsburgh | School of Public Health

Further Analysis

Table 2 Overview of five European-ancestry GWASs re-analysed in our study to quantify the relationship between sample size and discovery

Down-sampled GWAS	Max *n* (mean *n*)	Number of GWS COJO SNPs	Percentage of the genome covered by GWS loci (35 kb) (%)
Lango Allen et al. (2010)[19a]	130,010 (128,942)	240	0.5
Wood et al. (2014)[20]	241,724 (239,227)	633	1.4
Yengo et al. (2018)[3]	695,648 (688,927)	2,794	5.8
GIANT-EUR (no 23andMe)	1,632,839 (1,502,499)	4,867	9.7
23andMe-EUR	2,502,262 (2,498,336)	7,020	13.6

Summary statistics from the three published GWASs were imputed using the ImpG-Summary software to maximize the coverage of HM3 SNPs (Methods). GWS loci are defined as in the legend of Table 1.

[a]Summary statistics from the Lango Allen et al. study[19], initially over-corrected for population stratification using a double genomic control correction, were re-inflated such that the LD score regression intercept estimated from re-inflated test statistics equals 1.

Goal: Asses how much increasing sample size affects discovery

Re-analyzed three previous GWAS and down-sampled meta analysis into 4 subsets

Quantified with 8 metrics

Variant/locus based:

1. GWS SNPs
2. GWS loci
3. R^2_{GWS}
4. Proportion of genome covered by GWS loci

Functional/annotation based:

5. Enrichment statistics (stratified LDSC)

Gene based:

6. Genes priorities by SMR
7. Proximity of variants with OMIM genes

Gene-set based:

8. Enrichment within gene set/pathway clusters

University of Pittsburgh | School of Public Health

LDSC, briefly

At SNP level, GWAS is essentially *t*-test (difference between null=0 and estimated effect)

n = number of SNPs (large), so *t* distribution can be approximated by a Z (standard normal)

$$\hat{\beta} \sim t(n-2) \xrightarrow{d} N(0,1)$$

Square a Z distribution and you get a chi-squared distribution (1 degree of freedom)

$$X \sim N(0,1) \implies X^2 \sim \chi^2(1)$$

Expectation of a chi-squared = its degree(s) of freedom, this gives Y|x=0 (intercept value). Asses stratification.

$$E\left[\chi^2(1)\right] = 1$$

Regress chi-squared stats onto LD scores (slope estimates heritability)

$$y \sim ax + b + \epsilon$$

$$y \sim \text{intercept} + b + \epsilon$$

$$\chi^2 \sim \text{intercept} + \text{LD Scores} + \epsilon$$

Annotate SNPs and compare to look for enrichment

=

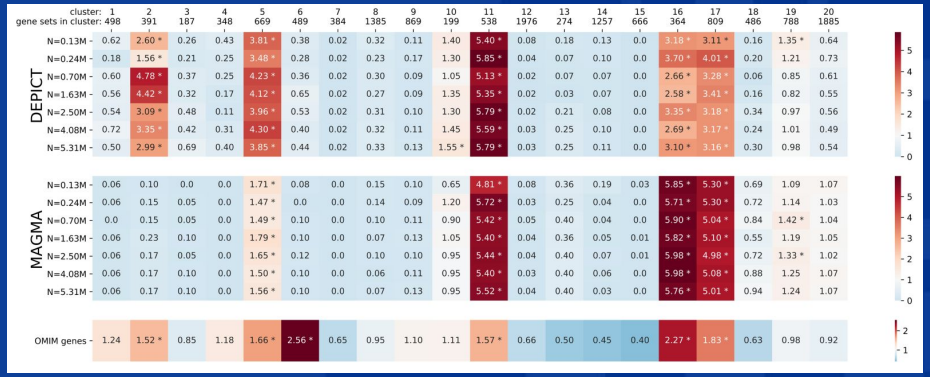
Stratified LDSC

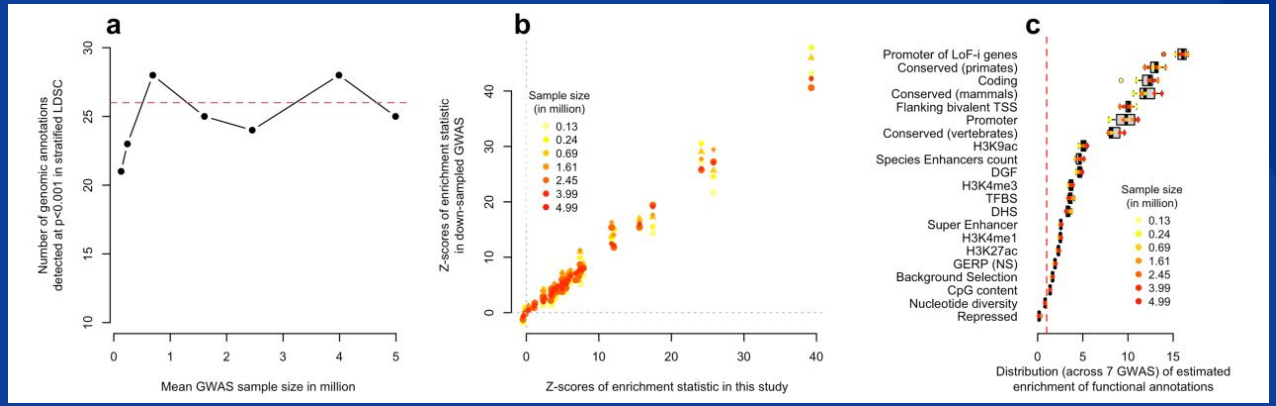
University of Pittsburgh | School of Public Health

Further Analysis

Strongest saturation observed for

- **Gene-set results**
  - DEPICT
  - MAGMA
- **Functional annotation results**
  - Stratified LDSC (heritability enrichment)





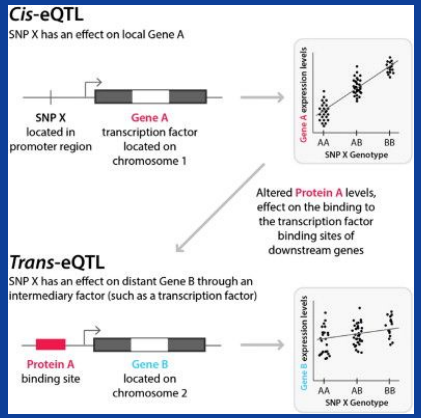
Additionally, no significant improvement in metrics from adding non-EUR samples, implying similar biological mechanism(s).

Sample size (range: 130K-5.3M) made no difference

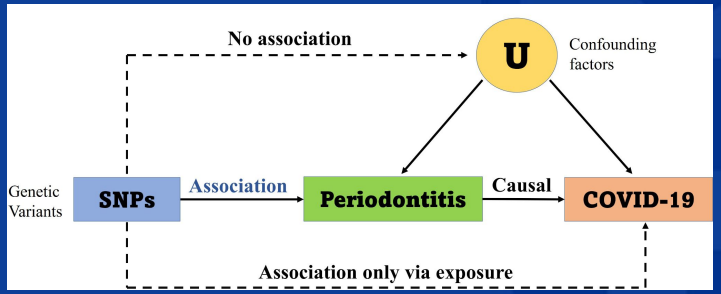
University of Pittsburgh | School of Public Health

SMR, briefly

Find SNPs associated with gnxp (eQTL)



MR to see if SNP also associated with trait



All can be performed with summary statistics

Posit: SNP → GNXP → Trait

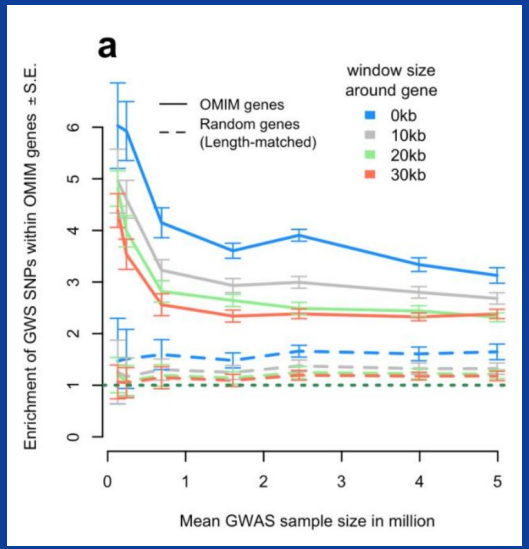
University of Pittsburgh | School of Public Health

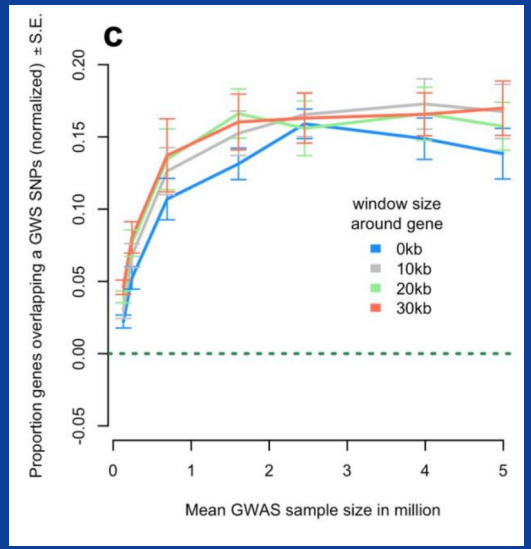
Further Analysis

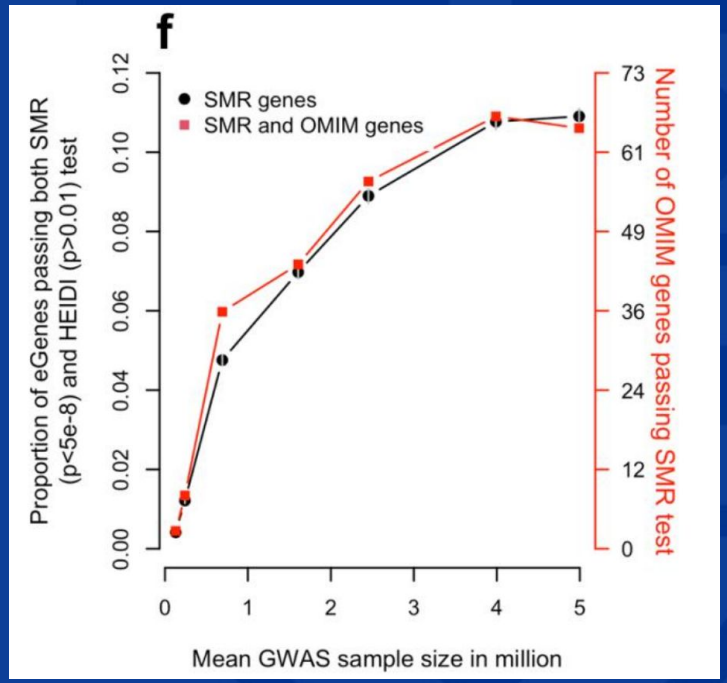
Gene-level metrics (compared with random SNPs)

- Enrichment of GWS SNPs in OMIM genes
- Proportion of genes overlapping GWS SNPs

Both plateau ~1.5M samples



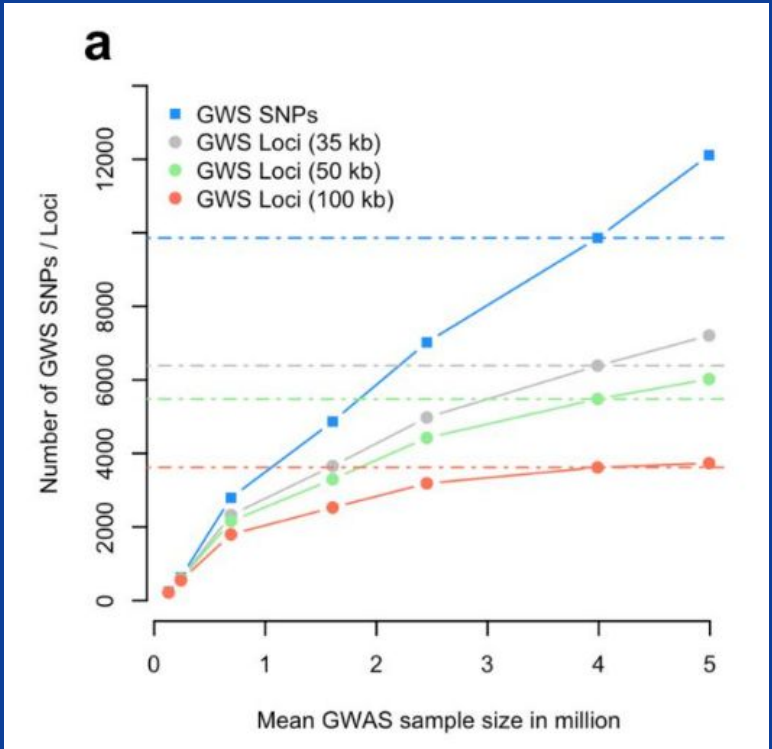




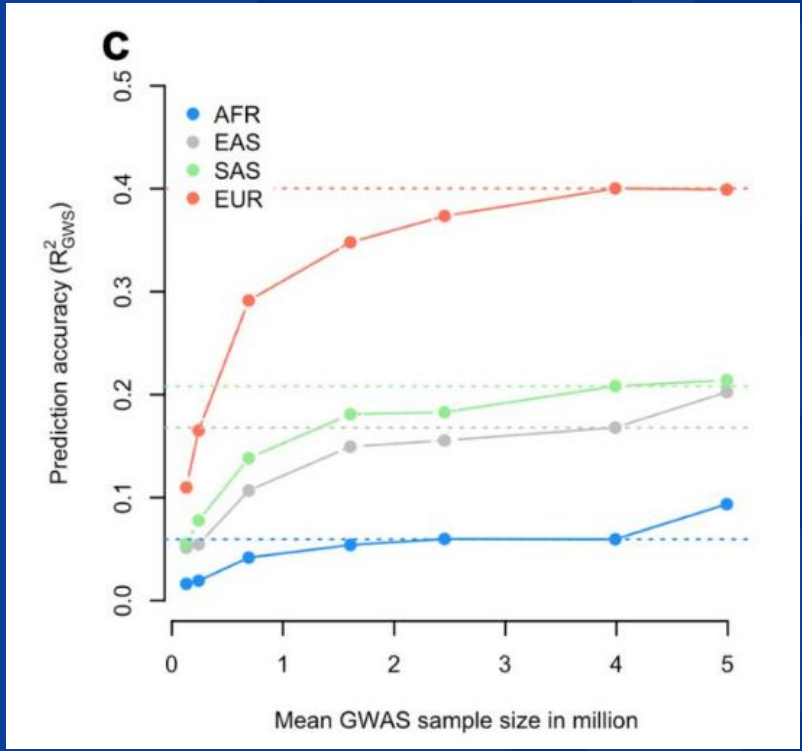
SMR genes showed increase
(may be artifact of power)

University of Pittsburgh | School of Public Health

Further Analysis



Gene/variant level analysis saw greatest discovery advantage from increasing sample size (SNPs roughly linear)



Marked saturation for E(R^2)

University of Pittsburgh | School of Public Health

Prediction Accuracy Sidenote

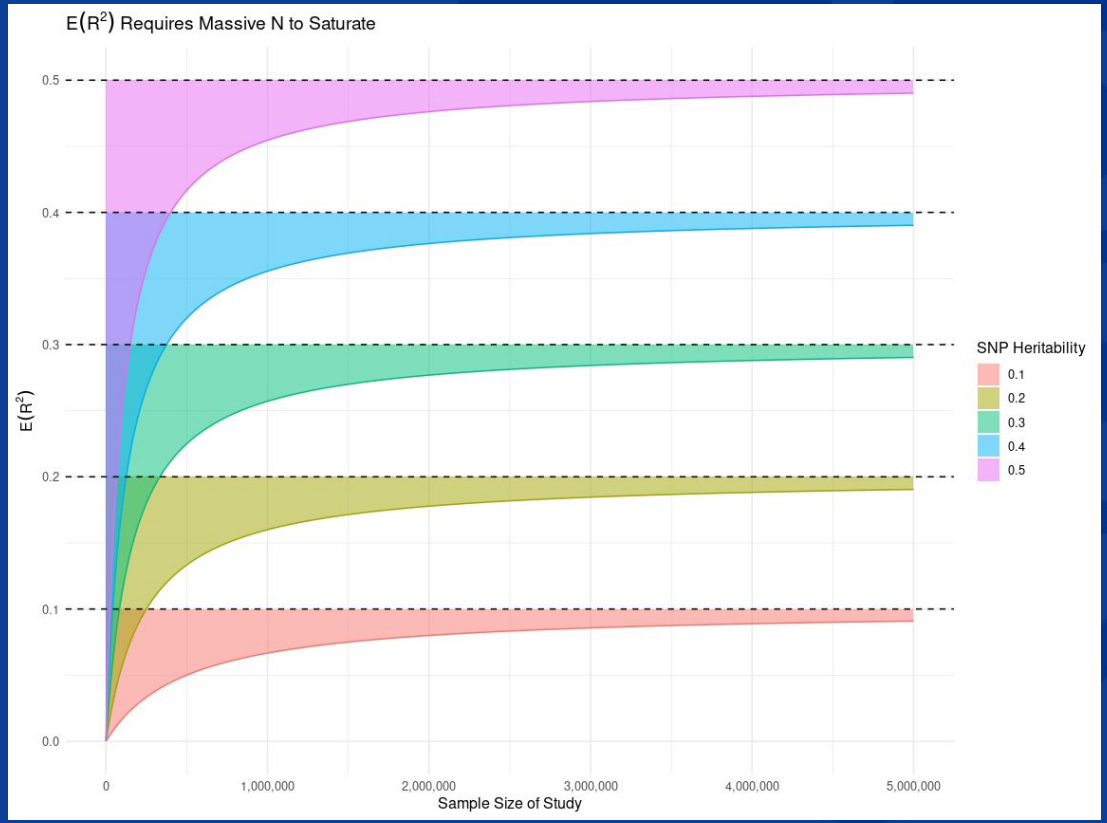
$E(R^2)$ is a function *not just* of the true SNP heritability, but of the effective number SNPs

Effective number of SNPs:
Total number of SNPs divided by mean LD score (for humans ~50,000)

Takeaway: even with high h^2_{SNP}, $E(R^2)$ <<< h^2_{SNP} unless sample sizes are massive.

$$E(R^2) \approx \frac{h_M^2}{1 + M \Big/ \left(N h_M^2\right)}$$

Daetwyler Equation
(m=effective SNPs)



University of Pittsburgh | School of Public Health

Discussion

Some Key Takeaways:

- Signal density not randomly distributed, clustered in “height genes”
- Strong genetic overlap for various ancestries (range of effect size correlations: 0.64-0.99)
- h_{SNP}^2 ~ 100% explained EUR, 90% for non-EUR
- LD & MAF explain ~84% (SE: 1.5%) of prediction accuracy loss
- Large N needed to pinpoint variants, but gene sets feasible for much smaller N

Implications for other traits

EUR individuals. Discrepancies between observed and predicted levels of saturation could be explained by several factors, such as (i) heterogeneity of SNP effects between cohorts and background ancestries, which may have reduced the statistical power of our study as compared to a homogenous sample like UKB; (ii) inconsistent definitions of GWS SNPs (using COJO in this study versus standard clumping in ref. [52]); and, most importantly, (iii) misspecification of the SNP-effects distribution assumed to make these predictions. Nevertheless, if these predictions reflect proportional levels of saturation between traits, then we could expect that two- to tenfold larger samples would be required for GWASs of inflammatory bowel disease (×2, that is, $n = 10$ million), schizophrenia (×7; $n = 35$ million) or BMI (×10; $n = 50$ million) to reach a similar saturation of 80–90% of SNP-based heritability.

University of Pittsburgh | School of Public Health

Limitations

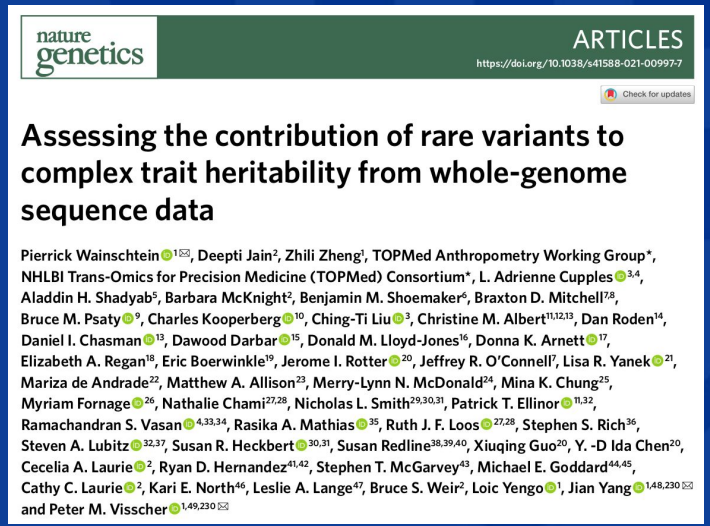
As summarized by authors:

1. Focus on SNPs from HM3 panel
2. COJO used EUR LD
3. Proper replication not done

All essentially a result of unavailability of better options

have identified in the present study. A question for the future is whether rare genetic variants associated with height are also concentrated within the same loci. We provide suggestive evidence supporting this hypothesis from analysing imputed SNPs with 0.1% < MAF < 1% (Supplementary Note 6, Extended Data Fig. 10 and Supplementary Fig. 25). Our results are consistent with findings from a previous study[45], which showed across 492 traits a strong colocalization between common and rare coding variants associated with the same trait. Nevertheless, our conclusions remain limited by the relatively low performances of imputation in this MAF regime[46,47]. Therefore, large samples with whole-genome sequences will be required to robustly address this question. Such datasets are increasingly becoming available[48–50]. Separately,

Conclusions could have detailed contemporaneous research?



nature genetics ARTICLES

https://doi.org/10.1038/s41588-021-00997-7

Check for updates

Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data

Pierrick Wainschtein[1], Deepti Jain[2], Zhili Zheng[1], TOPMed Anthropometry Working Group*, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium*, L. Adrienne Cupples[3,4], Aladdin H. Shadyab[5], Barbara McKnight[2], Benjamin M. Shoemaker[6], Braxton D. Mitchell[7,8], Bruce M. Psaty[9], Charles Kooperberg[10], Ching-Ti Liu[3], Christine M. Albert[11,12,13], Dan Roden[14], Daniel I. Chasman[13], Dawood Darbar[15], Donald M. Lloyd-Jones[16], Donna K. Arnett[17], Elizabeth A. Regan[18], Eric Boerwinkle[19], Jerome I. Rotter[20], Jeffrey R. O'Connell[7], Lisa R. Yanek[21], Mariza de Andrade[22], Matthew A. Allison[23], Merry-Lynn N. McDonald[24], Mina K. Chung[25], Myriam Fornage[26], Nathalie Chami[27,28], Nicholas L. Smith[29,30,31], Patrick T. Ellinor[11,32], Ramachandran S. Vasan[4,33,34], Rasika A. Mathias[35], Ruth J. F. Loos[27,28], Stephen S. Rich[36], Steven A. Lubitz[32,37], Susan R. Heckbert[30,31], Susan Redline[38,39,40], Xiuqing Guo[20], Y. -D Ida Chen[20], Cecelia A. Laurie[2], Ryan D. Hernandez[41,42], Stephen T. McGarvey[43], Michael E. Goddard[44,45], Cathy C. Laurie[2], Kari E. North[46], Leslie A. Lange[47], Bruce S. Weir[2], Loic Yengo[1], Jian Yang[1,48,230] ✉ and Peter M. Visscher[1,49,230] ✉

University of Pittsburgh® | School of Public Health

FIN.