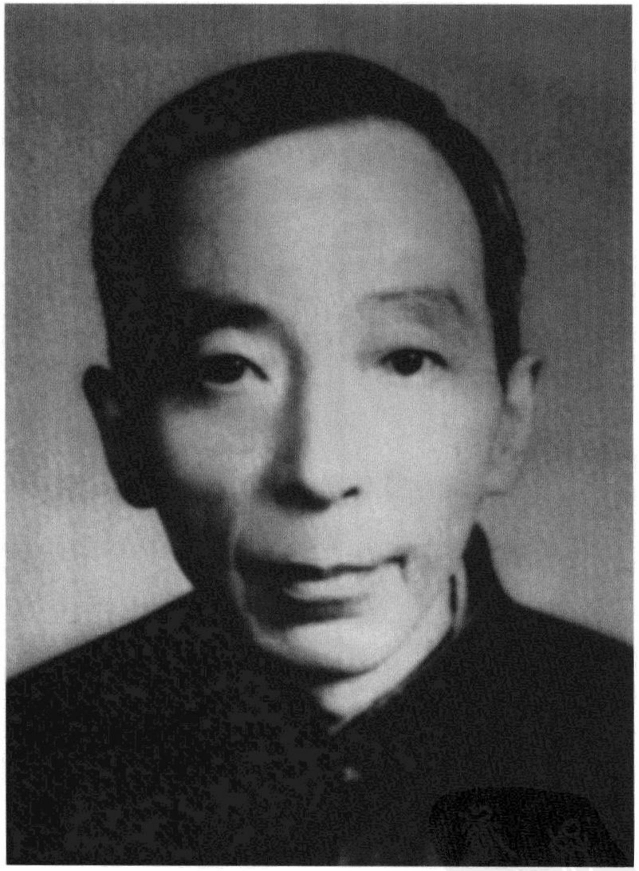
Fisher’s 1918 Quantitative Genetics Model In the Genomics Era

Hongyu Zhao
Department of Biostatistics
Yale School of Public Health
12/20/2019

Pao-Lu Hsu Award Lecture
The 11th ICSA International Conference

Yale SCHOOL OF PUBLIC HEALTH

MDCCCXLVII



Pao-Lu Hsu was born in Beijing on September 1, 1910, with his ancestral home in **Hangzhou**, Zhejiang Province.

1933: Graduated from Tsinghua University

1933-1936: Taught at Peking University (PKU)

1936-1940: University College London

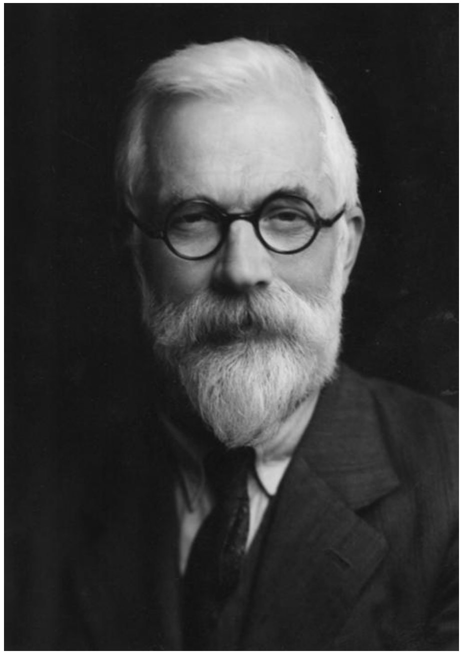
1938: Ph.D.

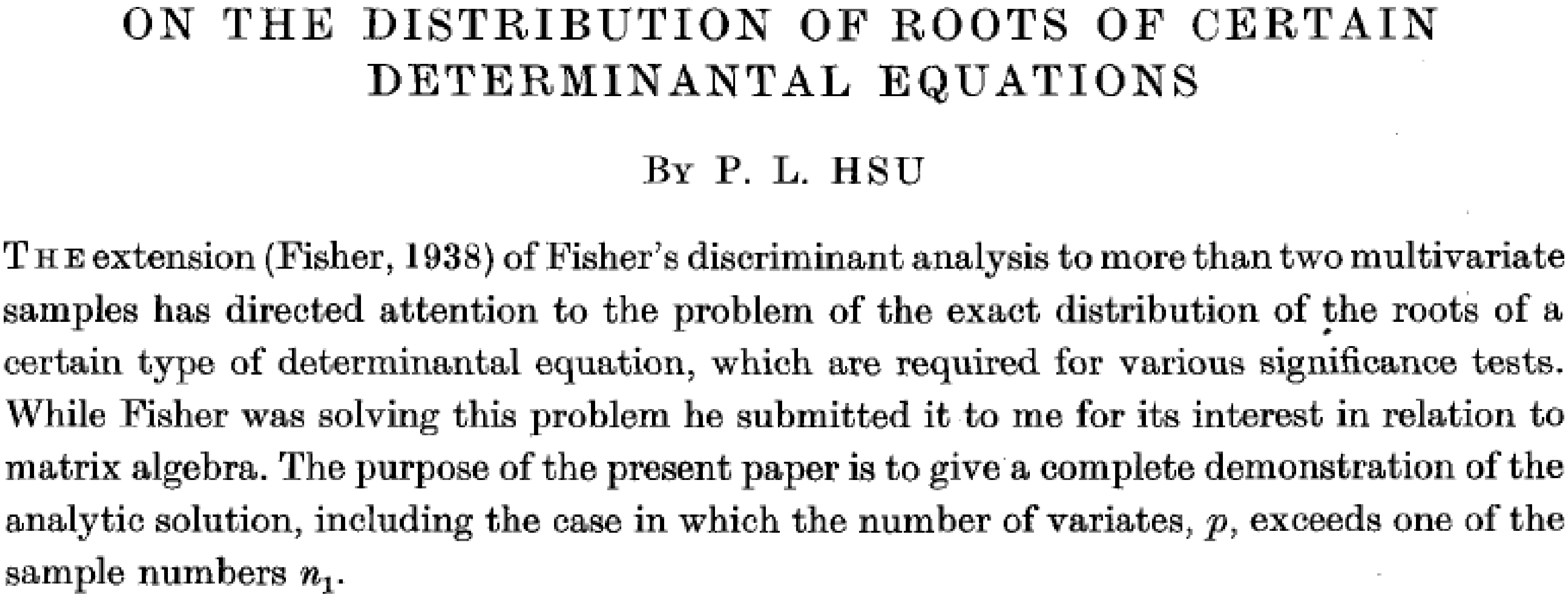
1940: Sc.D.

R. A. Fisher

University College London, 1933–39

The Design of Experiments
Null hypothesis
Fiducial inference
Discriminant analysis
Fisher–Kolmogorov equation



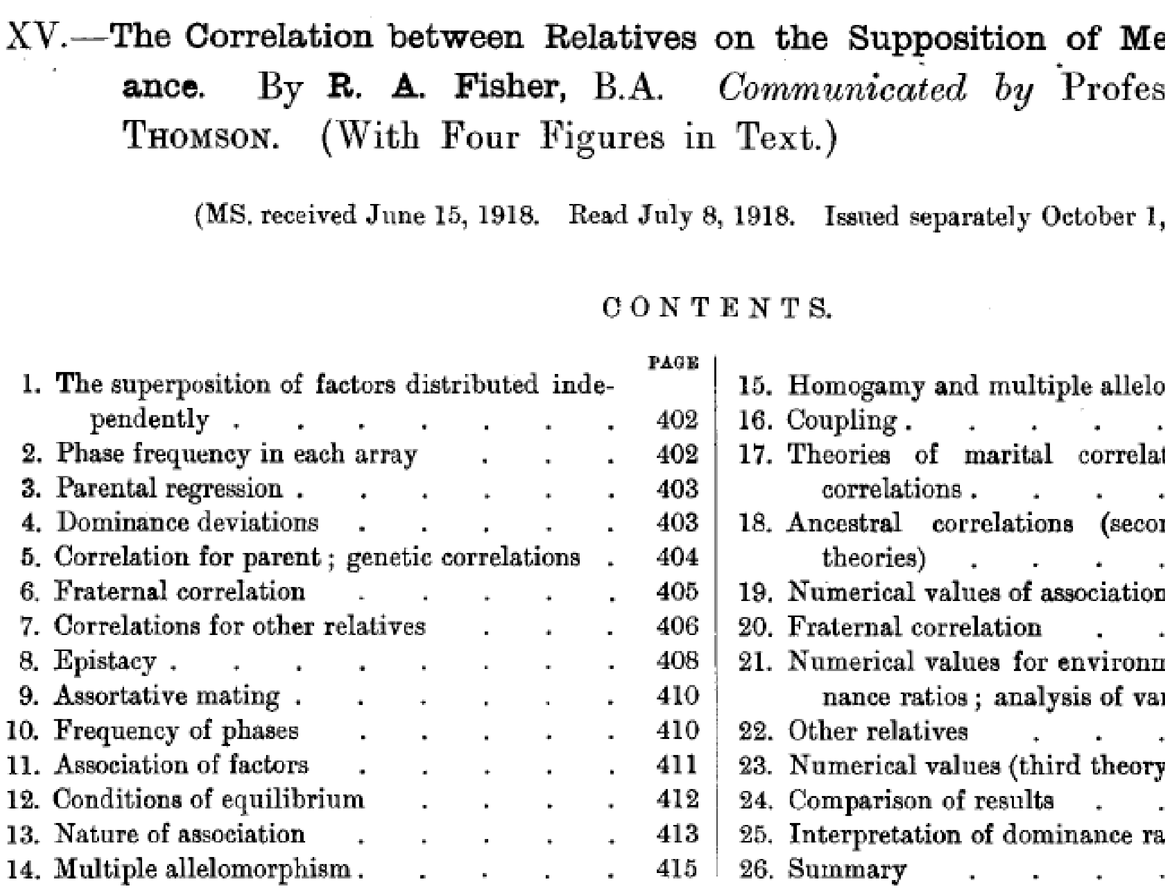


ON THE DISTRIBUTION OF ROOTS OF CERTAIN DETERMINANTAL EQUATIONS

By P. L. HSU

THE extension (Fisher, 1938) of Fisher's discriminant analysis to more than two multivariate samples has directed attention to the problem of the exact distribution of the roots of a certain type of determinantal equation, which are required for various significance tests. While Fisher was solving this problem he submitted it to me for its interest in relation to matrix algebra. The purpose of the present paper is to give a complete demonstration of the analytic solution, including the case in which the number of variates, p, exceeds one of the sample numbers n_1.

Annals of Eugenics, 1939, 9: 250-258.

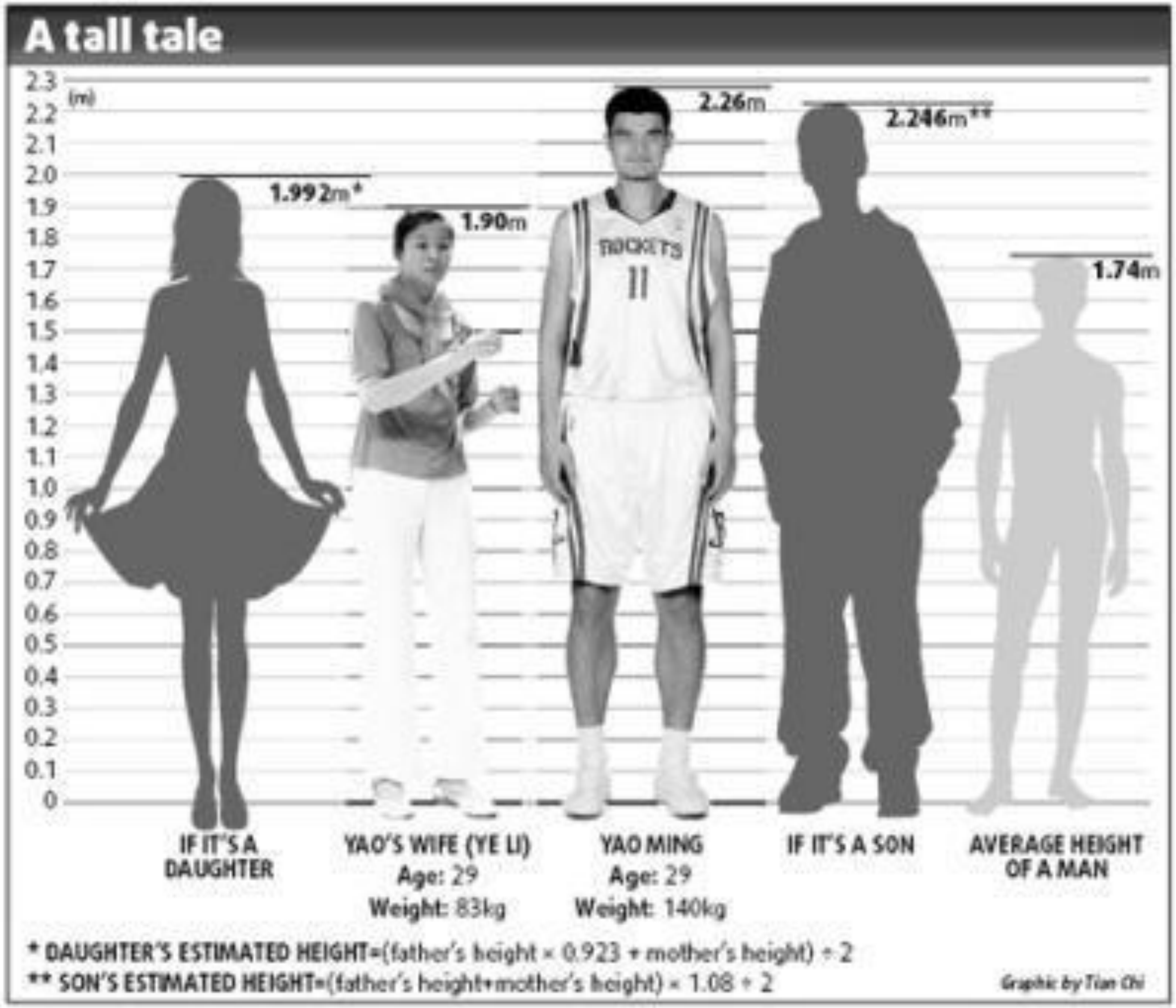


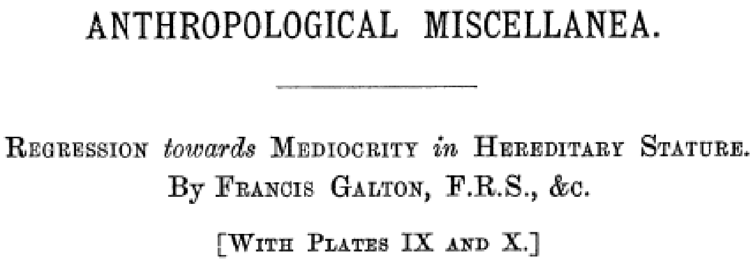
XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. By R. A. Fisher, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

Variance, ANOVA
Assortative Mating

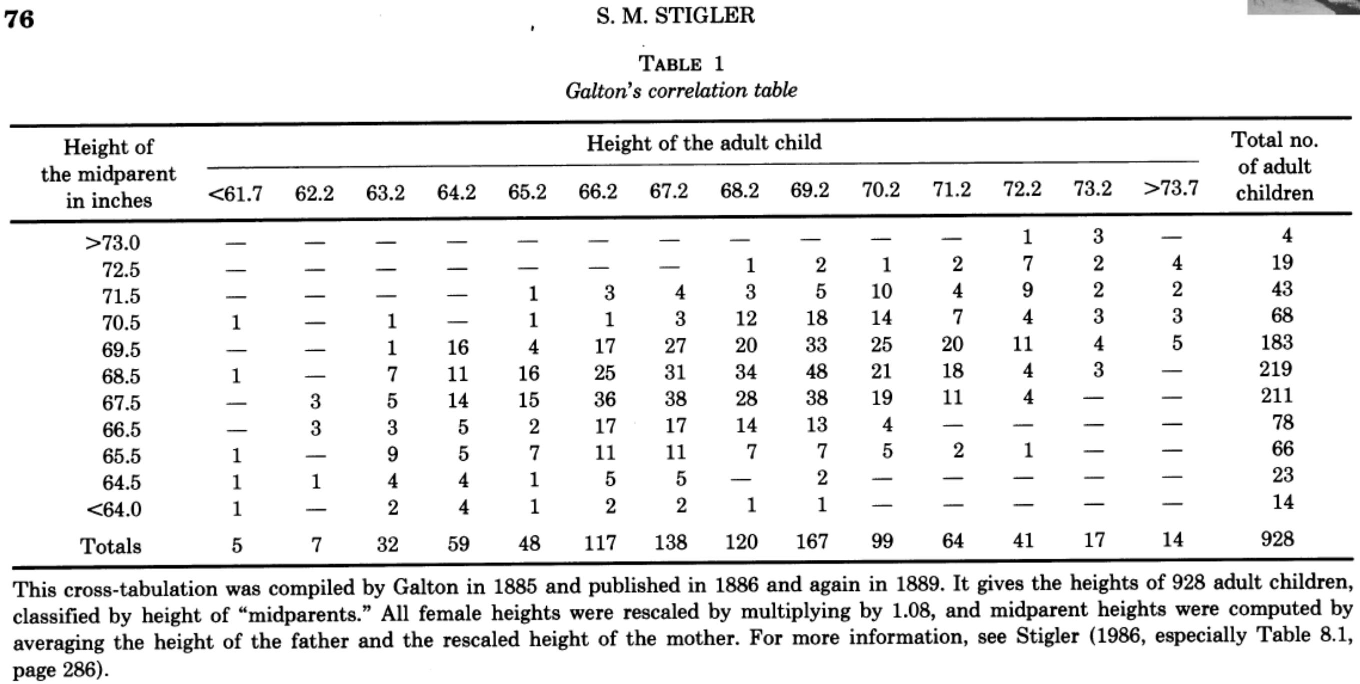

A tall tale
(m)
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.992m*
1.90m
2.26m
2.246m**
1.74m
ROCKETS
11
IF IT'S A DAUGHTER
YAO'S WIFE (YE LI)
Age: 29
Weight: 83kg
YAO MING
Age: 29
Weight: 140kg
IF IT'S A SON
AVERAGE HEIGHT OF A MAN
* DAUGHTER'S ESTIMATED HEIGHT=(father's height × 0.923 + mother's height) ÷ 2
** SON'S ESTIMATED HEIGHT=(father's height+mother's height) × 1.08 ÷ 2
Graphic by Tian Chi



ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.
By Francis Galton, F.R.S., &c.

[With Plates IX and X.]



76 S. M. STIGLER

Table 1
Galton's correlation table

Height of the midparent in inches	Height of the adult child														Total no. of adult children
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7	
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928

This cross-tabulation was compiled by Galton in 1885 and published in 1886 and again in 1889. It gives the heights of 928 adult children, classified by height of "midparents." All female heights were rescaled by multiplying by 1.08, and midparent heights were computed by averaging the height of the father and the rescaled height of the mother. For more information, see Stigler (1986, especially Table 8.1, page 286).

A Simple Model for Quantitative Traits

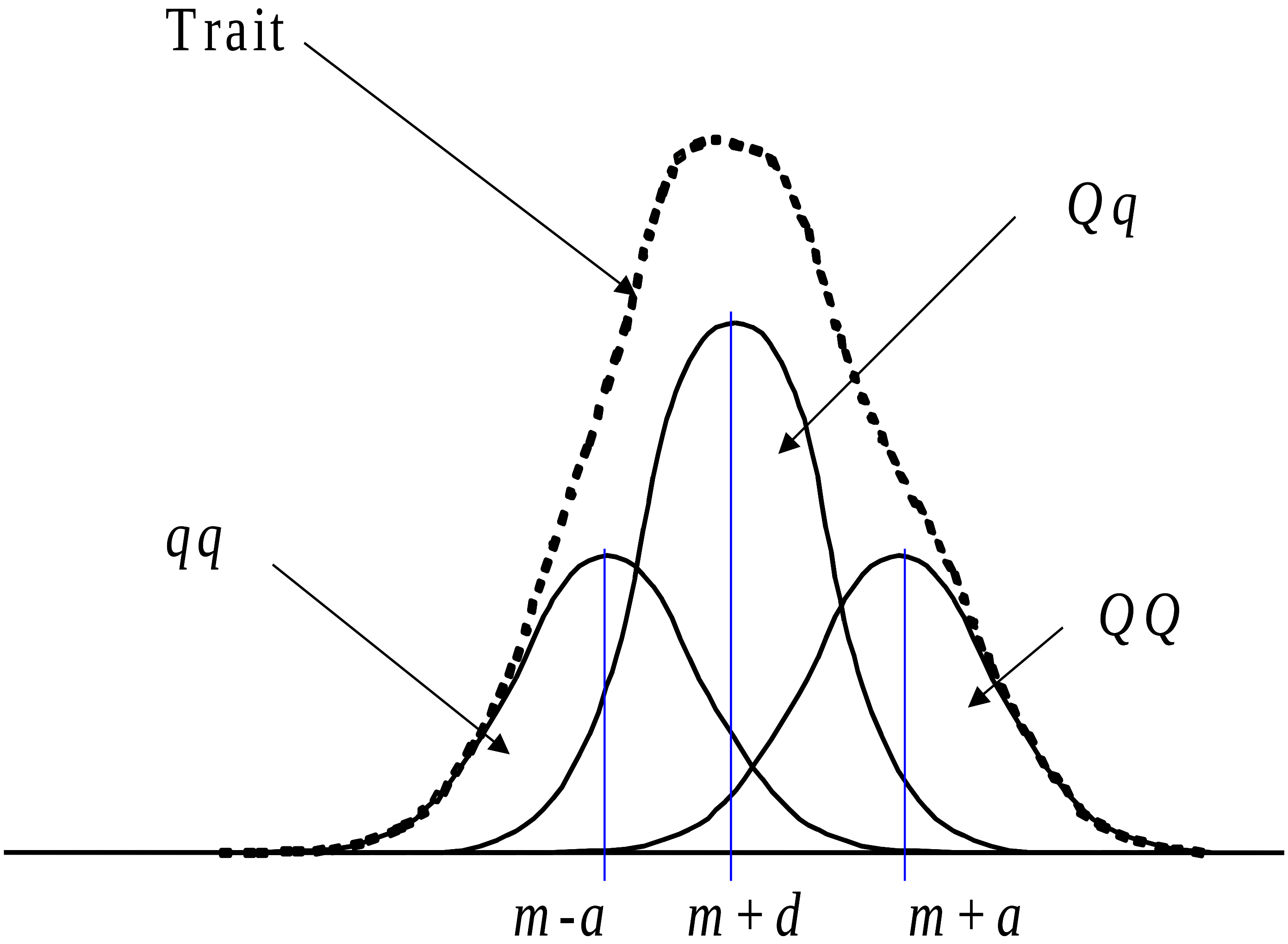
Connect phenotypes with genotypes (genetic factors)

Consider a marker A with two alleles (forms) A_1 and A_2, with allele frequencies p and q.

$$Y = G + E + e$$

Genotype	Value	Frequency
A_1A_1	a	p^2
A_1A_2	$d = 0$	$2pq$
A_2A_2	$-a$	q^2

Population mean: $a(p-q)+2dpq = a(p-q)$


Trait
Qq
qq
QQ
$m - a$
$m + d$
$m + a$

Partition of Genetic Variance

Model:

$$Y = G + E + e$$

Variance:

$$V_Y = V_G + V_E + V_e = V_A + V_E + V_e$$

Heritability (*h*²):

$$h^2 = V_A / V_Y$$

Heritability can be influenced by many factors in addition to the trait locus (loci), e.g. environment, assortative mating, gene-gene interaction etc.

Covariance between Relatives

Resemblance among relatives

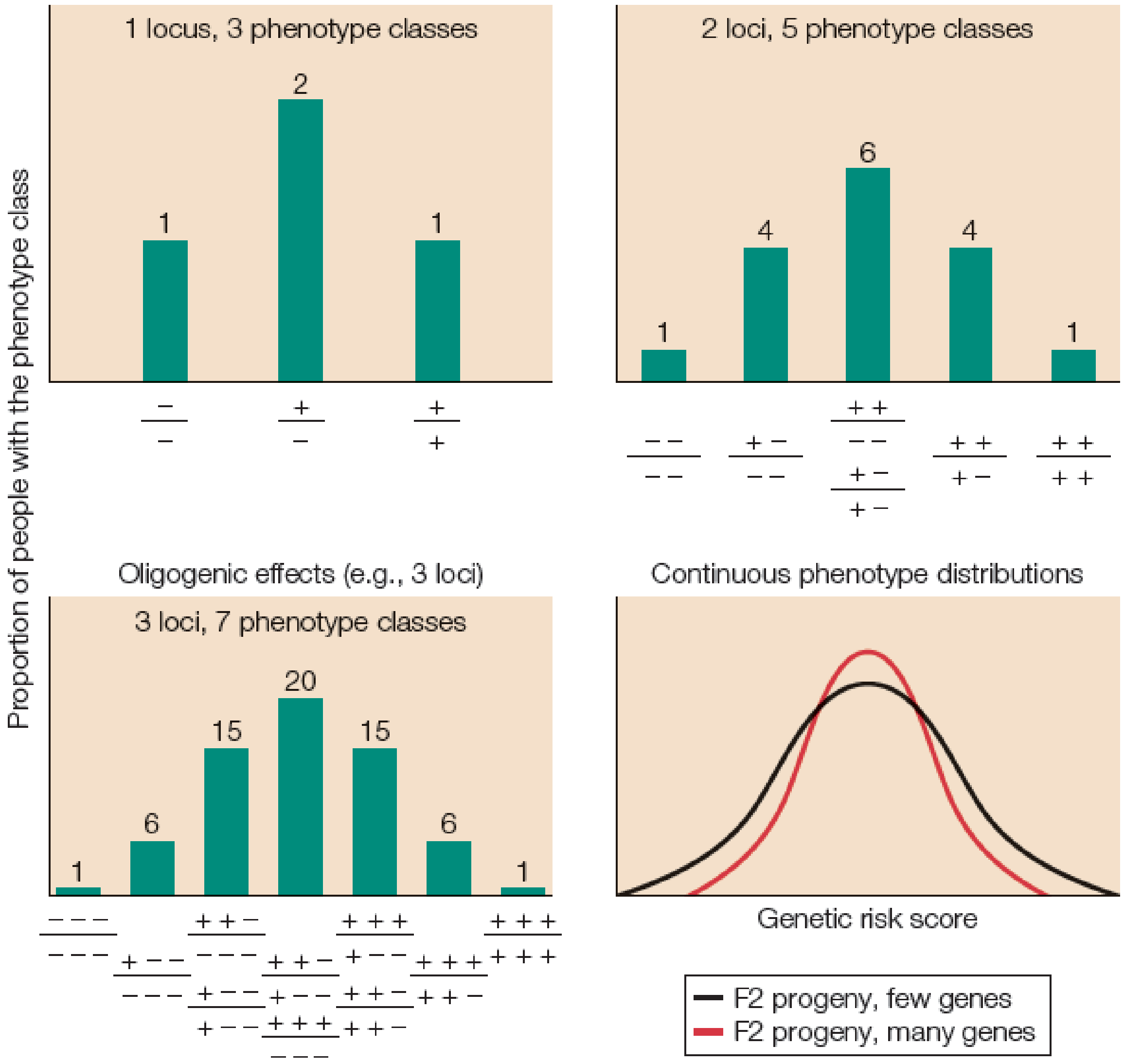
$$\mathbf{Cov(Y_1, Y_2) = c_R V_A}$$

$\mathbf{c_R}$: 2 x kinship coefficient

$$cov(Y_{Parent}, Y_{Offspring}) = ½\ V_A$$

$$cov(Y_{SIB1}, Y_{SIB2}) = ½\ V_A$$

From Mendelian to Quantitative Genetics



G. Gibson

From Mendelian to Quantitative Genetics

$$Y = G + E + e$$

$$\begin{aligned} Y \quad &= G_1 + G_2 + G_3 + \ldots. + G_m + E + e \\ &= \mathbf{G} + E + e \end{aligned}$$

$$V_Y = V_G + V_E + V_e$$

$$Cov(Y_1, Y_2) = c_R V_A$$

Heritability, V_A/V_Y, can be estimated by collecting and analyzing data from relatives

Some Heritability Estimates

Yale

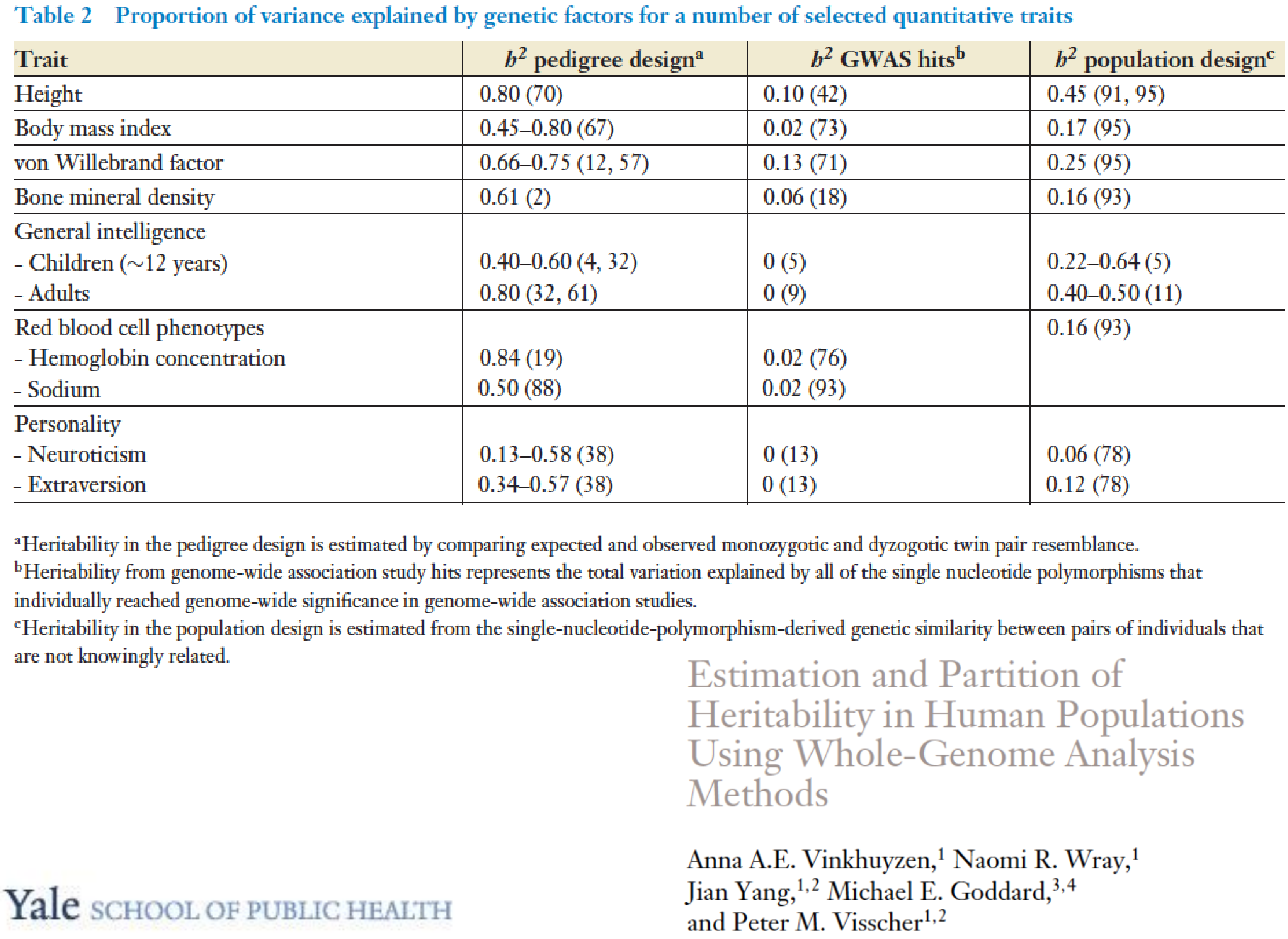


Table 2 Proportion of variance explained by genetic factors for a number of selected quantitative traits

Trait	h^2 pedigree design[a]	h^2 GWAS hits[b]	h^2 population design[c]
Height	0.80 (70)	0.10 (42)	0.45 (91, 95)
Body mass index	0.45–0.80 (67)	0.02 (73)	0.17 (95)
von Willebrand factor	0.66–0.75 (12, 57)	0.13 (71)	0.25 (95)
Bone mineral density	0.61 (2)	0.06 (18)	0.16 (93)
General intelligence			
- Children (~12 years)	0.40–0.60 (4, 32)	0 (5)	0.22–0.64 (5)
- Adults	0.80 (32, 61)	0 (9)	0.40–0.50 (11)
Red blood cell phenotypes			0.16 (93)
- Hemoglobin concentration	0.84 (19)	0.02 (76)	
- Sodium	0.50 (88)	0.02 (93)	
Personality			
- Neuroticism	0.13–0.58 (38)	0 (13)	0.06 (78)
- Extraversion	0.34–0.57 (38)	0 (13)	0.12 (78)

[a] Heritability in the pedigree design is estimated by comparing expected and observed monozygotic and dyzogotic twin pair resemblance.

[b] Heritability from genome-wide association study hits represents the total variation explained by all of the single nucleotide polymorphisms that individually reached genome-wide significance in genome-wide association studies.

[c] Heritability in the population design is estimated from the single-nucleotide-polymorphism-derived genetic similarity between pairs of individuals that are not knowingly related.

Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods

Anna A.E. Vinkhuyzen,[1] Naomi R. Wray,[1]
Jian Yang,[1,2] Michael E. Goddard,[3,4]
and Peter M. Visscher[1,2]

Note that we can estimate heritability

without knowing which genes contribute to traits of interest

If we assume the genetic model is correct (Fisher’s model)

1918

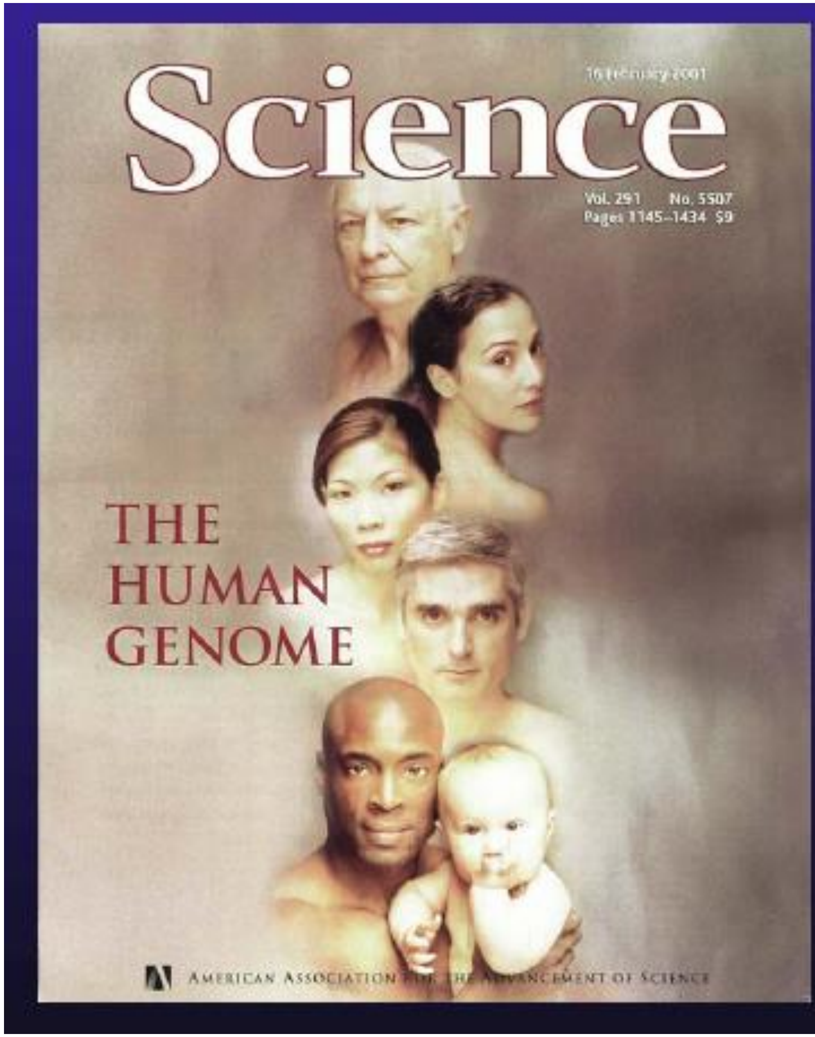
..

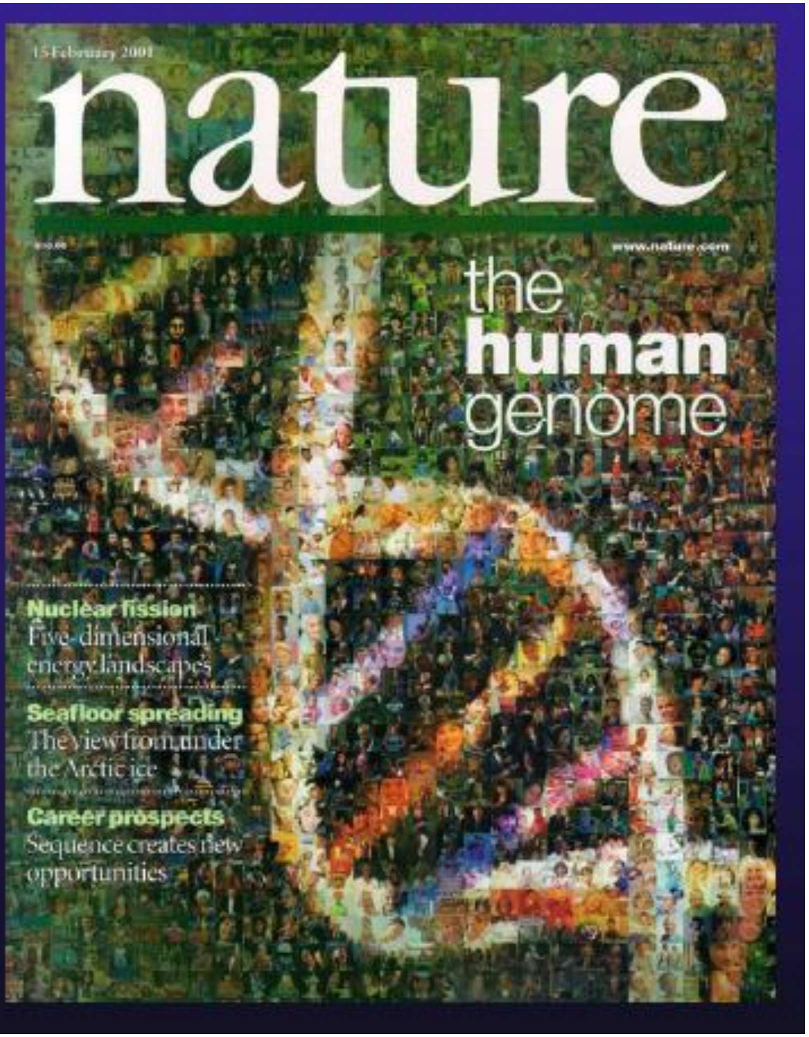
1953: DNA

..

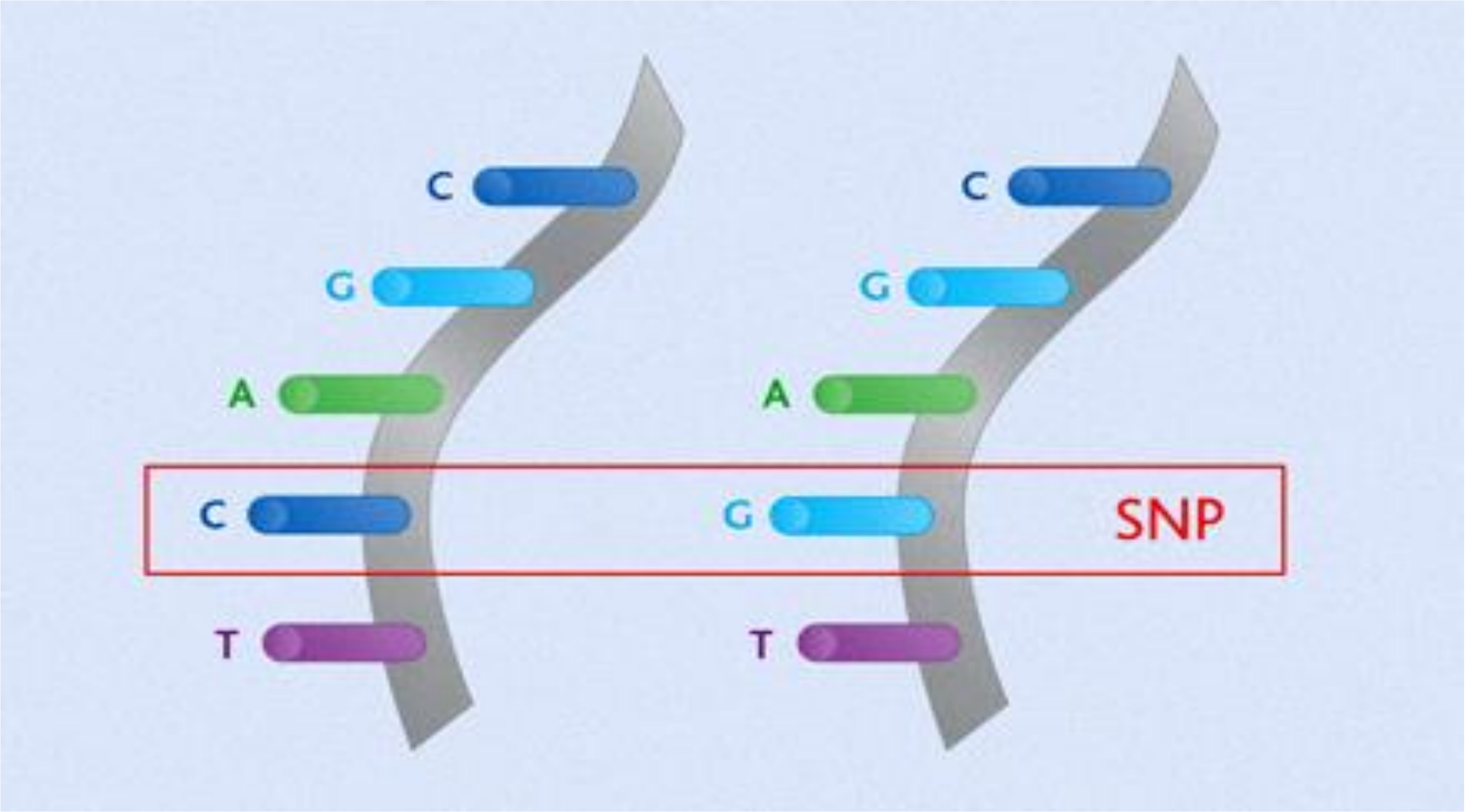
..

2001

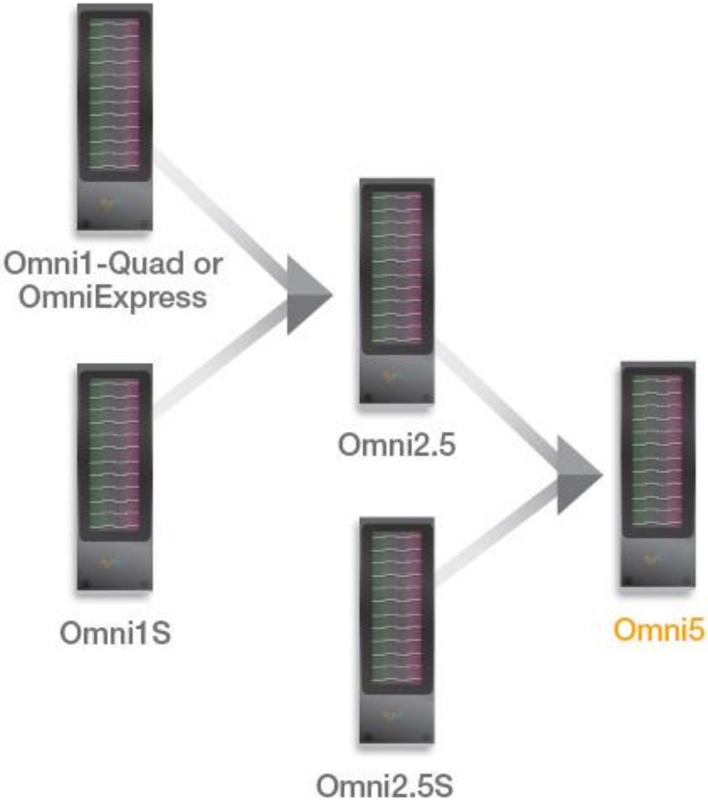

16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE


15 February 2001
nature
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities

Single Nucleotide Polymorphisms (SNPs)



Genotyping Arrays


Omni1-Quad or OmniExpress

Omni1S

Omni2.5

Omni2.5S

Omni5

GWAS General strategy

- Thousands up to hundreds of thousands of subjects
- Genome wide markers, up to several millions
- Correlate SNP with phenotypes to identify markers **significantly** associated with the phenotypes

Large Samples are Available

UK Biobank: **500,000 people**

US Million Veteran Program: 1,000,000 people (goal)

All of US: 1,000,000 people (goal)

China Kadoorie Biobank: 500,000 people

BioBank Japan Project: 200,000 people

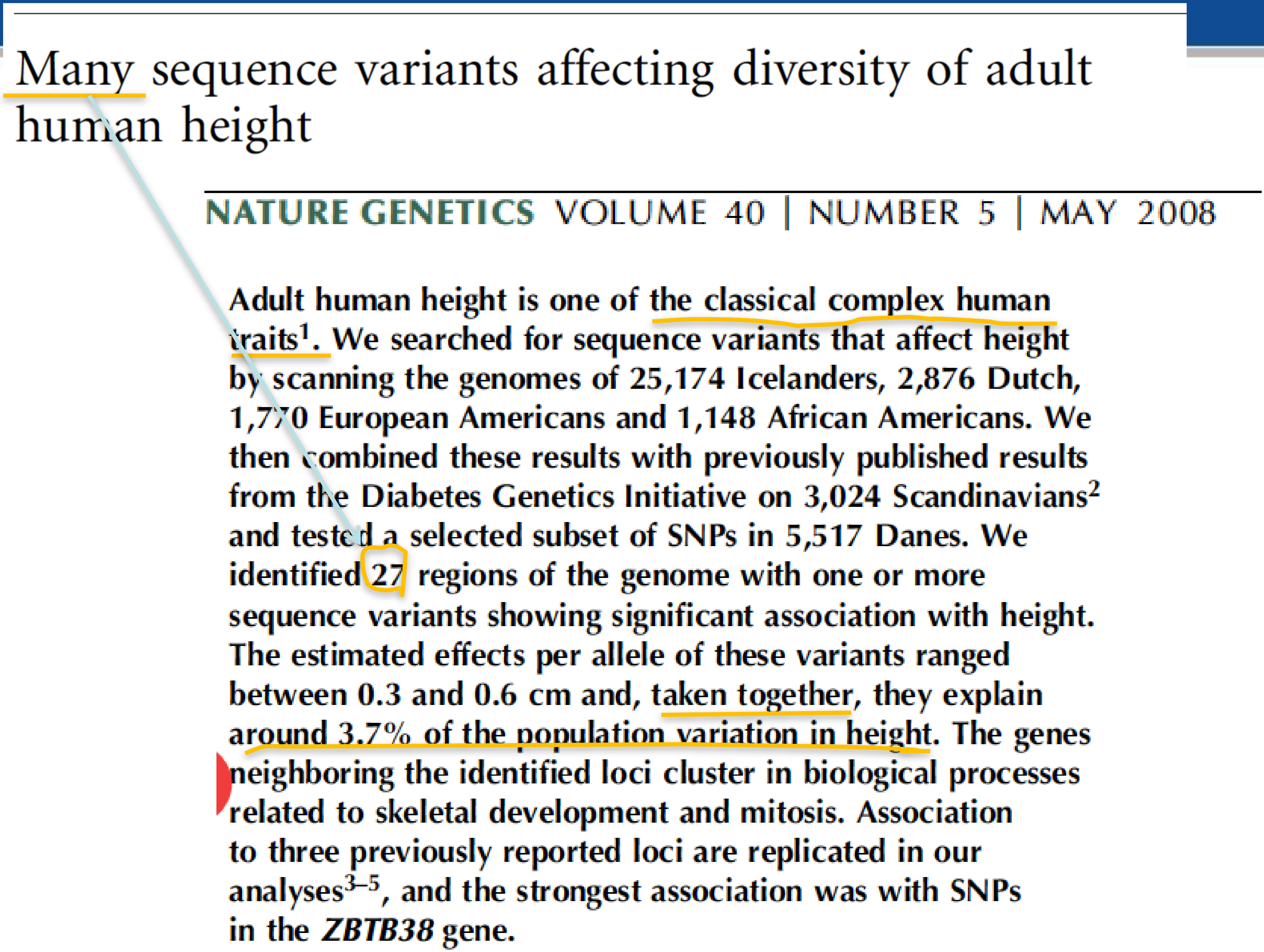
Observed Data

Trait	M_1	M_2 ….	M_m	sex	age	…
1	AA	AB ….	BB	M	23	…
1	AB	?? ….	AA	F	51	…
………………………………………………						
………………………………………………						
0	BB	AA ….	BB	F	64	…

One Question to Ask

Prediction:

How much phenotypic variation can be explained by **the identified variants and their estimated effect sizes**? **i.e. how well can we predict trait values or disease risk based on GWAS results?**



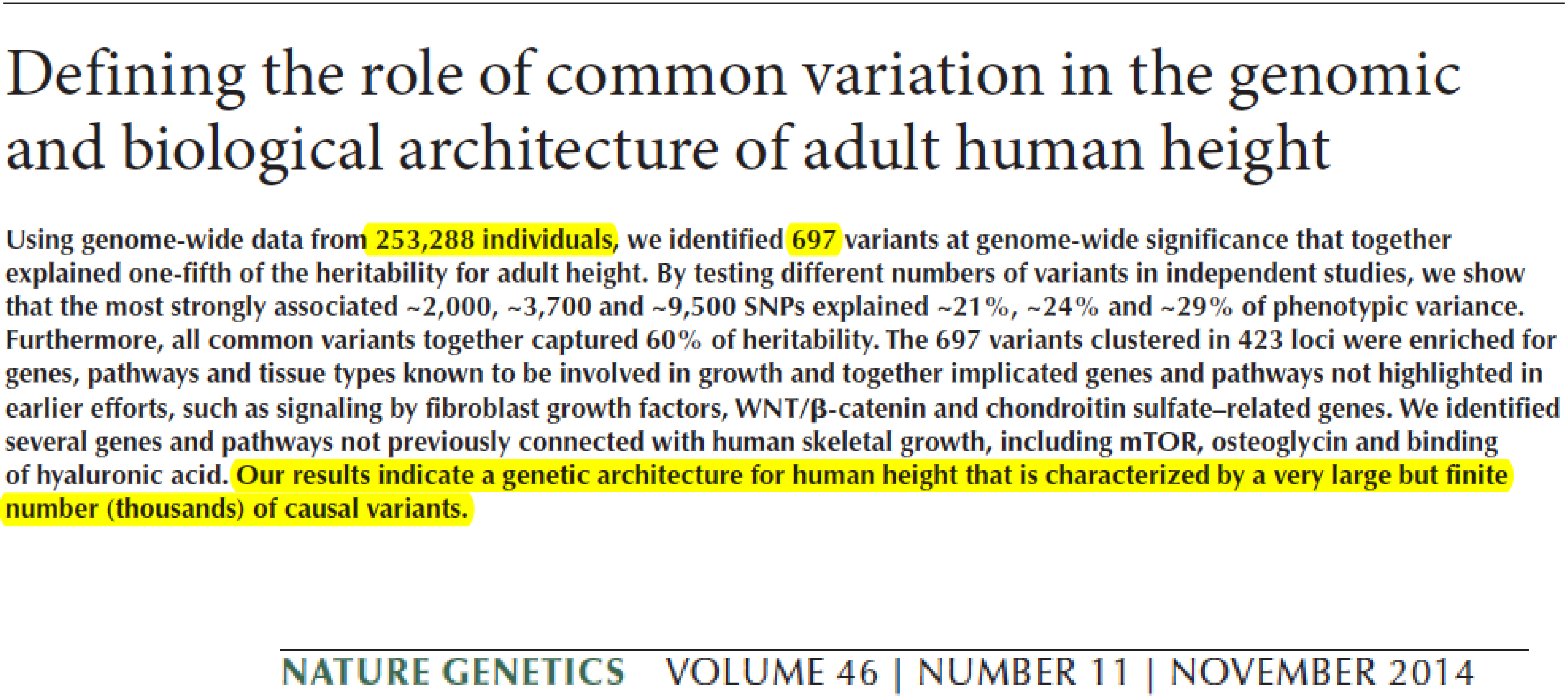
Many sequence variants affecting diversity of adult human height

NATURE GENETICS VOLUME 40 | NUMBER 5 | MAY 2008

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.

~80% heritability for height

Height GWAS



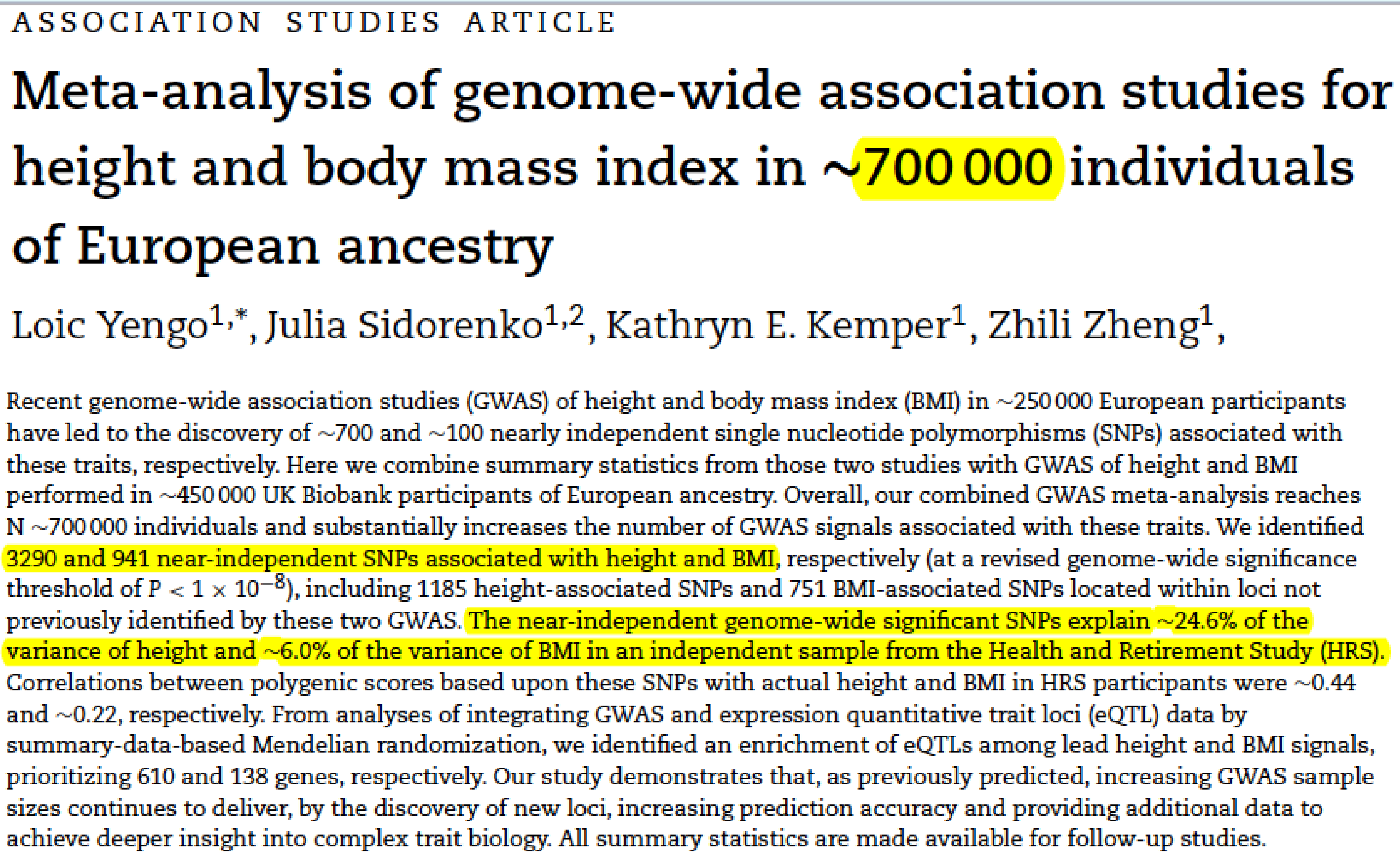
Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

NATURE GENETICS VOLUME 46 | NUMBER 11 | NOVEMBER 2014

~80% heritability for height

Height GWAS



ASSOCIATION STUDIES ARTICLE

Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry

Loic Yengo[1,*], Julia Sidorenko[1,2], Kathryn E. Kemper[1], Zhili Zheng[1],

Recent genome-wide association studies (GWAS) of height and body mass index (BMI) in ~250 000 European participants have led to the discovery of ~700 and ~100 nearly independent single nucleotide polymorphisms (SNPs) associated with these traits, respectively. Here we combine summary statistics from those two studies with GWAS of height and BMI performed in ~450 000 UK Biobank participants of European ancestry. Overall, our combined GWAS meta-analysis reaches N ~700 000 individuals and substantially increases the number of GWAS signals associated with these traits. We identified 3290 and 941 near-independent SNPs associated with height and BMI, respectively (at a revised genome-wide significance threshold of $P < 1 \times 10^{-8}$), including 1185 height-associated SNPs and 751 BMI-associated SNPs located within loci not previously identified by these two GWAS. The near-independent genome-wide significant SNPs explain ~24.6% of the variance of height and ~6.0% of the variance of BMI in an independent sample from the Health and Retirement Study (HRS). Correlations between polygenic scores based upon these SNPs with actual height and BMI in HRS participants were ~0.44 and ~0.22, respectively. From analyses of integrating GWAS and expression quantitative trait loci (eQTL) data by summary-data-based Mendelian randomization, we identified an enrichment of eQTLs among lead height and BMI signals, prioritizing 610 and 138 genes, respectively. Our study demonstrates that, as previously predicted, increasing GWAS sample sizes continues to deliver, by the discovery of new loci, increasing prediction accuracy and providing additional data to achieve deeper insight into complex trait biology. All summary statistics are made available for follow-up studies.

~80% heritability for height

Gap between 80% and (4%, 10%, 24%):

Reasons:

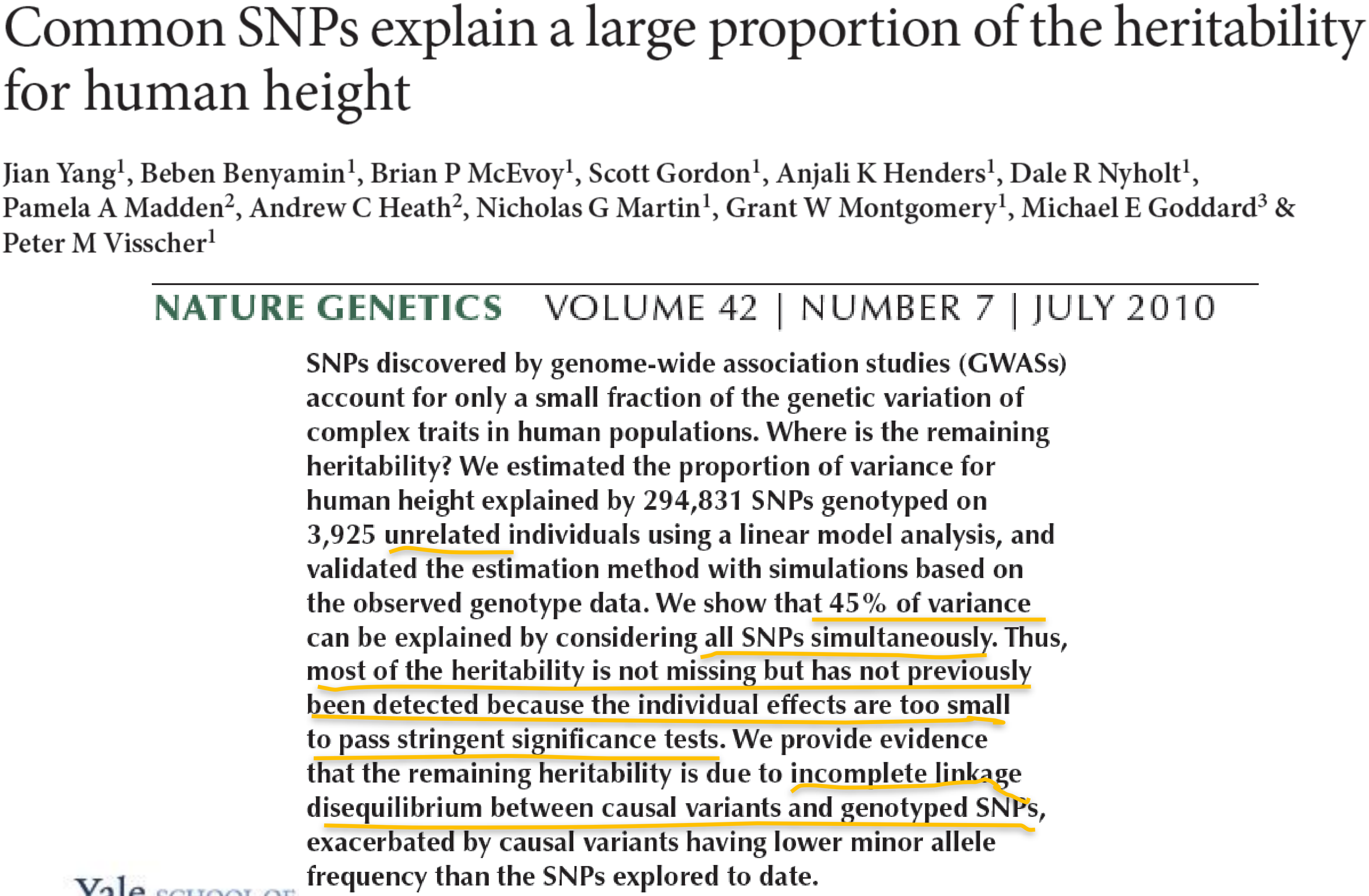
[1] Platform: **beyond common SNPs?**

[2] Models: **interactions and more complex models?**

A Different Question to Ask

Heritability (sort of):

How much heritability can be **potentially** explained by all the studied SNPs (and possibly their true effect sizes), i.e. **all the SNPs on the microarray**?



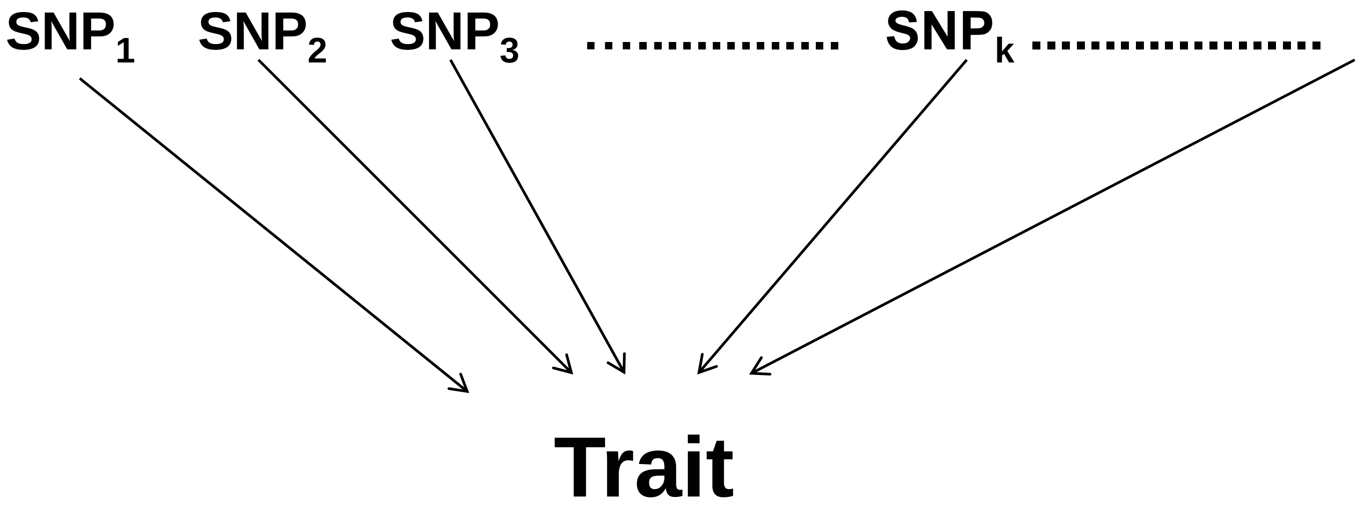
Common SNPs explain a large proportion of the heritability for human height

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]

NATURE GENETICS VOLUME 42 | NUMBER 7 | JULY 2010

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.

Fisher’s model



$$y \quad = \sum_{i=1}^{K} G_i + \epsilon$$

Abstract, no observed genotypes only phenotype data from relatives

GWAS Setting

$$
\begin{aligned}
Y &= G_1 + G_2 + G_3 + \dots + G_m + E + e \\
  &= \mathbf{z_1 u_1 + z_2 u_2 + z_3 u_3 + \dots + z_m u_m} + E + e
\end{aligned}
$$

We can observe genotypes with genotyping arrays
Mostly unrelated individuals

z_i: genotype score
u_i: effect size of the ith SNP

Challenging to infer u_i

Random Effects Model

$$\begin{aligned} Y &= G_1 + G_2 + G_3 + \ldots. + G_m + E + e \\ &= z_1u_1 + z_2u_2 + z_3u_3 + \ldots. + z_mu_m + E + e \end{aligned}$$

z_i: genotype score, u_i: effect size of the ith SNP

Assumption: each SNP contributes equally to the trait variance

$$u_i \sim N(0, \frac{\sigma_g^2}{m})$$

Using **normalized** genotype scores z_i

Observed: {A_1A_1, A_1A_2, A_2A_2} coded as {0, 1, 2}
Frequencies: {p^2, $2pq$, q^2}, $p+q=1$
Normalized: *{-2q/sqrt(2pq), (1-2q)/sqrt(2pq), 2p/sqrt(2pq)}*

$$Y = z_1u_1 + z_2u_2 + z_3u_3 + \ldots. + z_mu_m + e$$

$$Cov(Y) = \frac{ZZ'\sigma_g^2}{m} + I\sigma_e^2$$

When ***causal SNPs are known***, *ZZ'* can be calculated and reflects the genetic relatedness based on causal SNPs

Apply **RE**stricted (**RE**sidual) **M**aximum **L**ikelihood (REML) to estimate σ_g^2

When ***causal SNPs are unknown***, using genome wide SNPs to approximate

Fisher’s model:

Allows estimate of heritability from **family data** without the need to identify genes/variants associated with the traits

Random effects model:

[1] Motivated by the need to understand the contributions of (common) genetic variants to the traits (missing heritability)

[2] Using overall genetic similarity as a surrogate to the similarity across the genes/variants truly having effects on the traits

[3] Assuming that each SNP contributes equally to trait variance

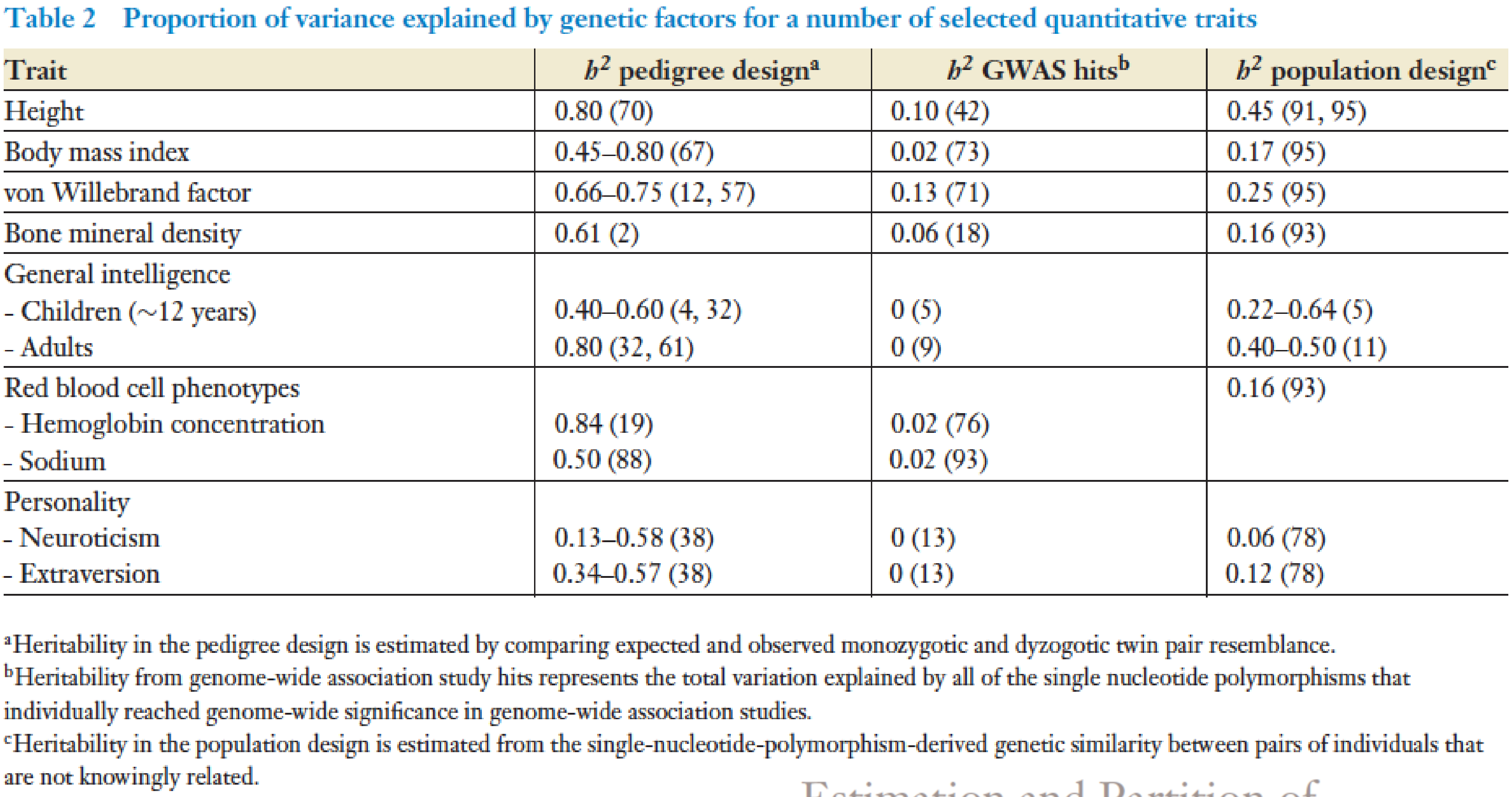


Table 2 Proportion of variance explained by genetic factors for a number of selected quantitative traits

Trait	h^2 pedigree design[a]	h^2 GWAS hits[b]	h^2 population design[c]
Height	0.80 (70)	0.10 (42)	0.45 (91, 95)
Body mass index	0.45–0.80 (67)	0.02 (73)	0.17 (95)
von Willebrand factor	0.66–0.75 (12, 57)	0.13 (71)	0.25 (95)
Bone mineral density	0.61 (2)	0.06 (18)	0.16 (93)
General intelligence			
- Children (~12 years)	0.40–0.60 (4, 32)	0 (5)	0.22–0.64 (5)
- Adults	0.80 (32, 61)	0 (9)	0.40–0.50 (11)
Red blood cell phenotypes			0.16 (93)
- Hemoglobin concentration	0.84 (19)	0.02 (76)	
- Sodium	0.50 (88)	0.02 (93)	
Personality			
- Neuroticism	0.13–0.58 (38)	0 (13)	0.06 (78)
- Extraversion	0.34–0.57 (38)	0 (13)	0.12 (78)

[a] Heritability in the pedigree design is estimated by comparing expected and observed monozygotic and dyzogotic twin pair resemblance.

[b] Heritability from genome-wide association study hits represents the total variation explained by all of the single nucleotide polymorphisms that individually reached genome-wide significance in genome-wide association studies.

[c] Heritability in the population design is estimated from the single-nucleotide-polymorphism-derived genetic similarity between pairs of individuals that are not knowingly related.

Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods

Anna A.E. Vinkhuyzen,[1] Naomi R. Wray,[1]
Jian Yang,[1,2] Michael E. Goddard,[3,4]
and Peter M. Visscher[1,2]

The Random Effects model is clearly **mis-specified** as we do not expect all the SNPs to be associated with a trait

$$y \quad = \sum_{i=1}^{K} Z_i u_i + \epsilon$$

A more appropriate model would assume that the **majority of SNPs** have **no effects** on the trait with **a small proportion** following **certain distribution**, but we have no knowledge on which SNP have or do not have effects on traits.

Robustness of REML

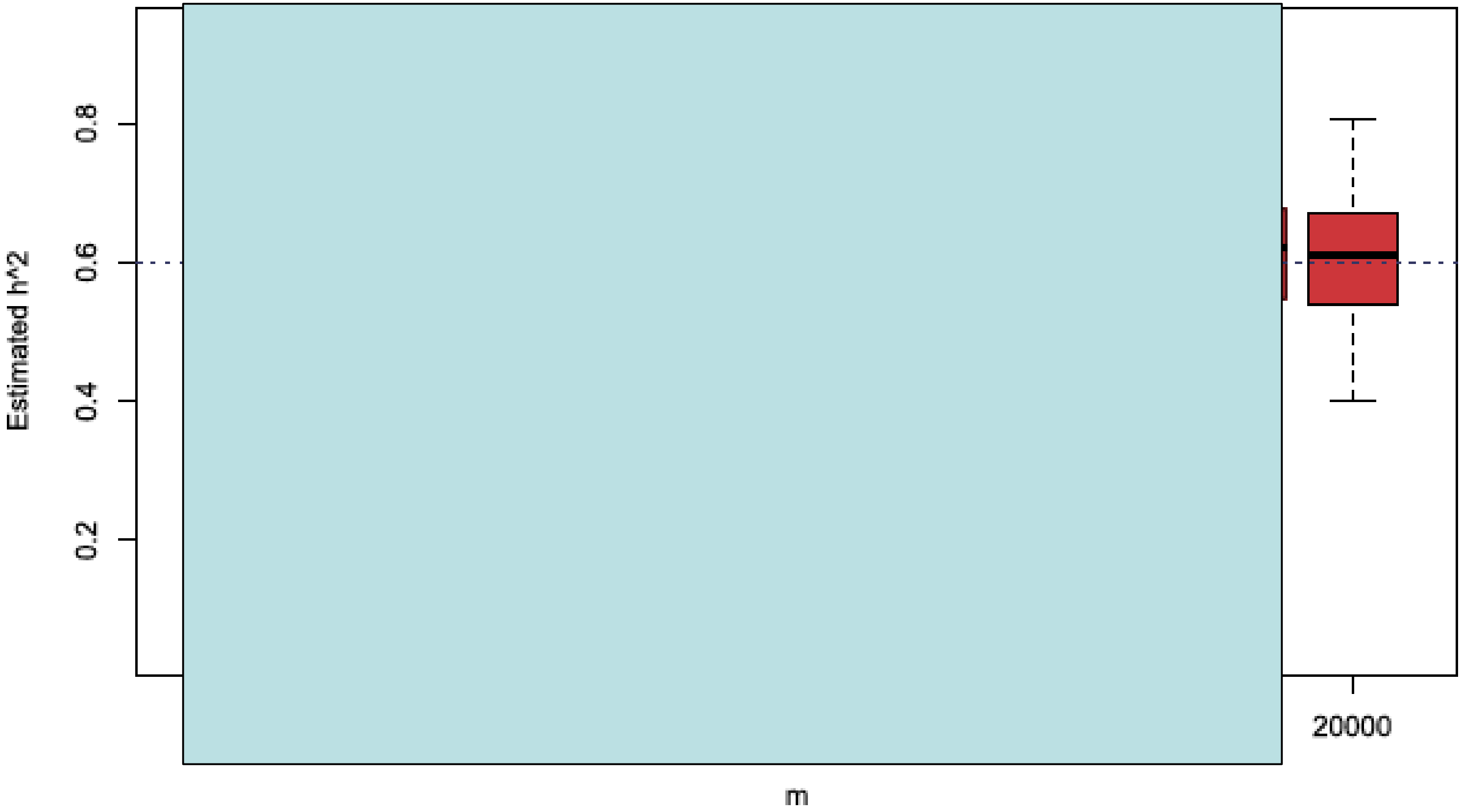


FIG. 1. *Heritability–REML provide right answer despite model misspecification.*

Robustness of REML

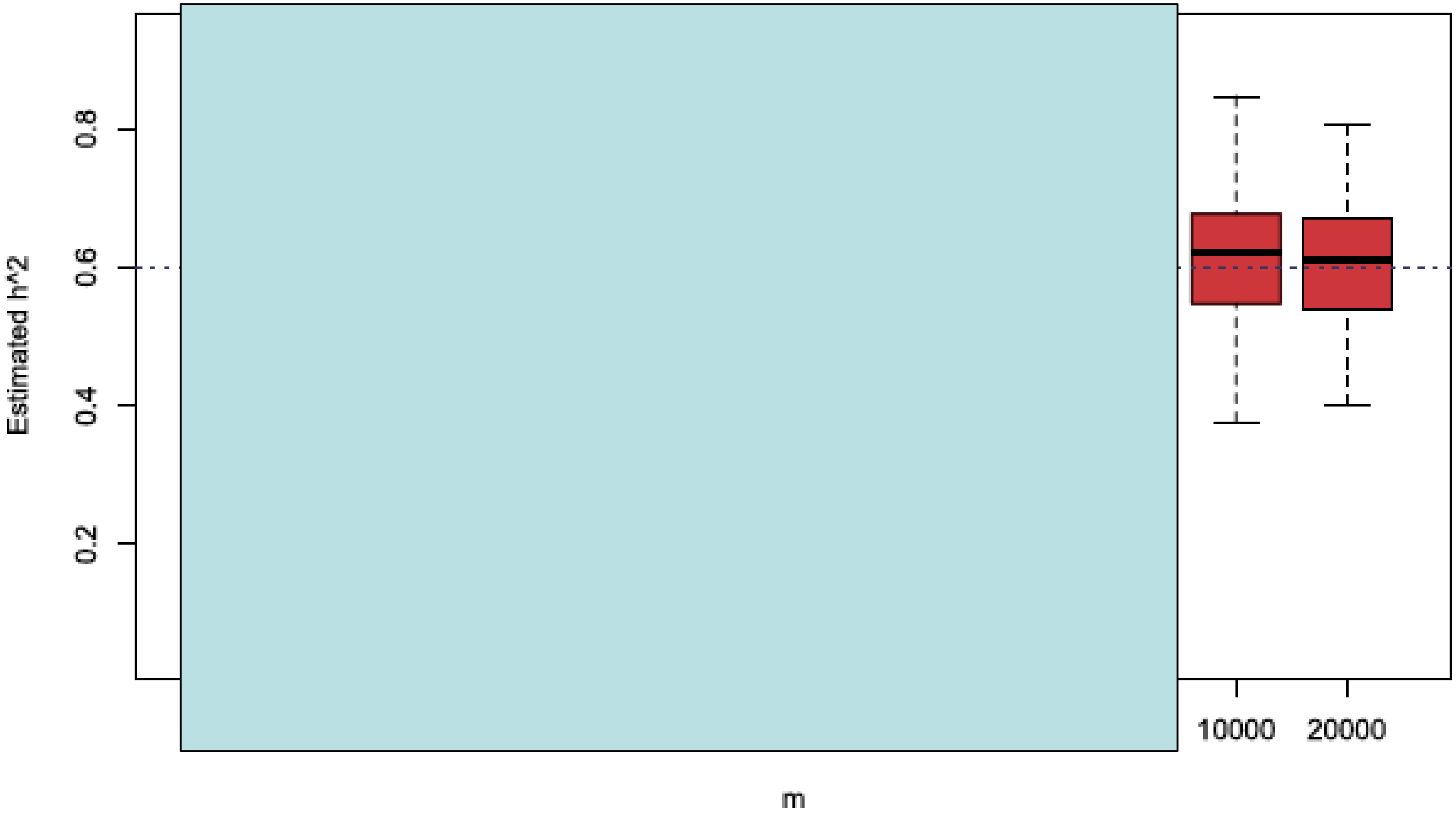


FIG. 1. *Heritability–REML provide right answer despite model misspecification.*

Robustness of REML

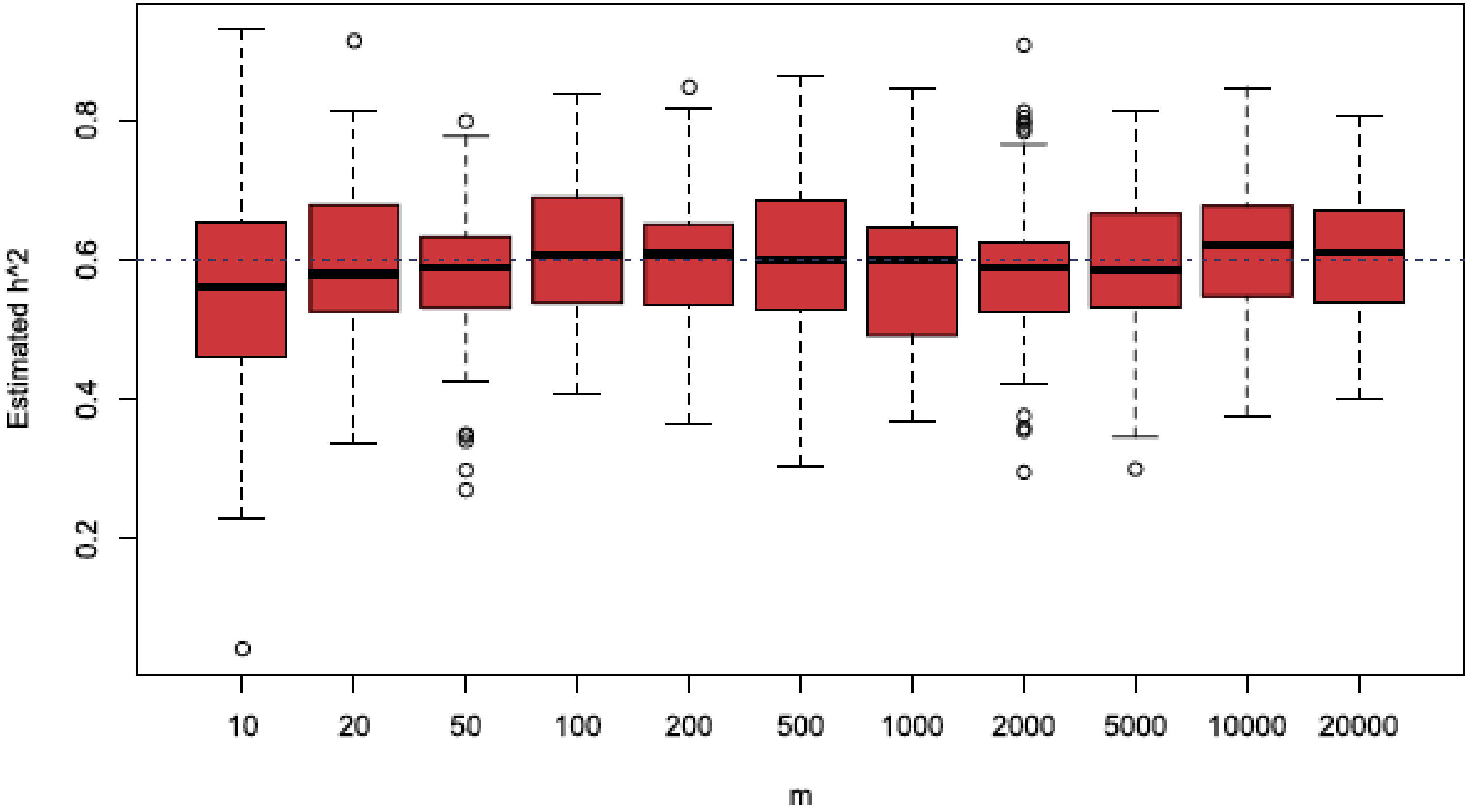
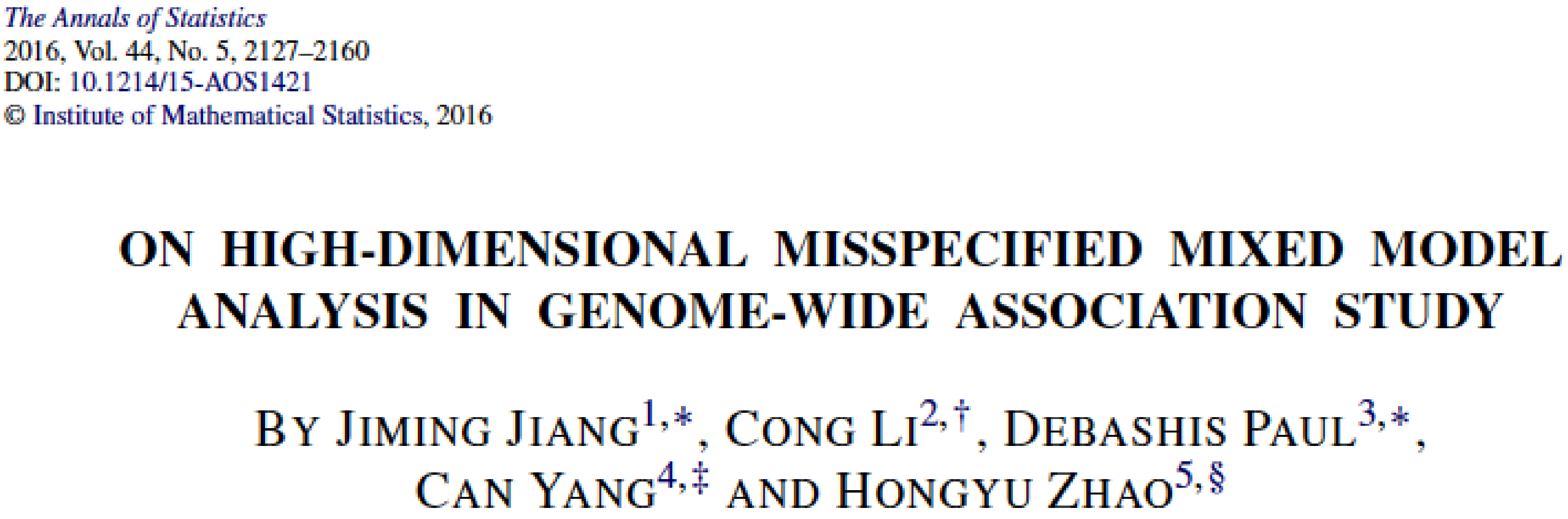


FIG. 1. *Heritability–REML provide right answer despite model misspecification.*



The Annals of Statistics
2016, Vol. 44, No. 5, 2127–2160
DOI: 10.1214/15-AOS1421

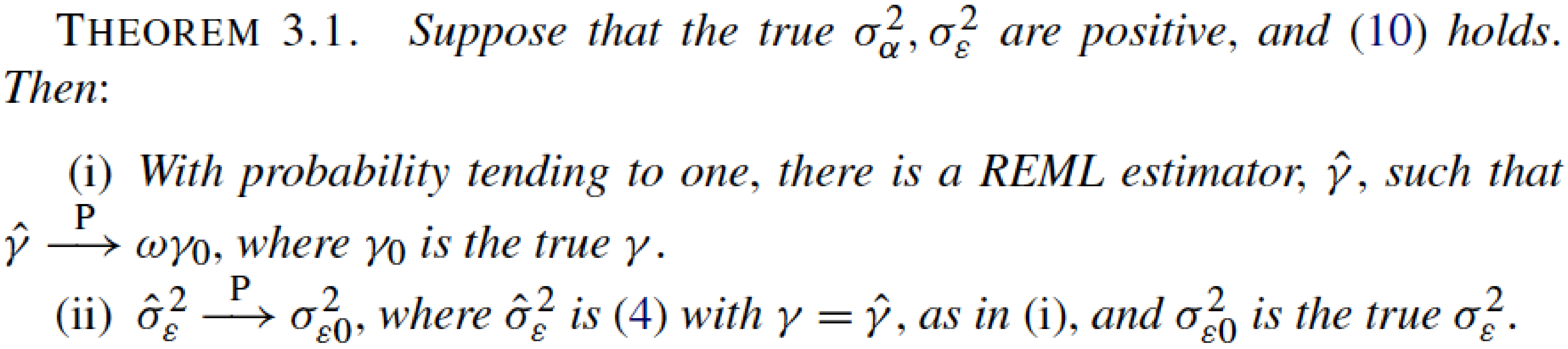
ON HIGH-DIMENSIONAL MISSPECIFIED MIXED MODEL ANALYSIS IN GENOME-WIDE ASSOCIATION STUDY

By Jiming Jiang[1,*], Cong Li[2,†], Debashis Paul[3,*],
Can Yang[4,‡] and Hongyu Zhao[5,§]

n: sample size
p: number of markers
m: number of associated markers

$$\frac{n}{p} \longrightarrow \tau, \qquad \frac{m}{p} \longrightarrow \omega,$$

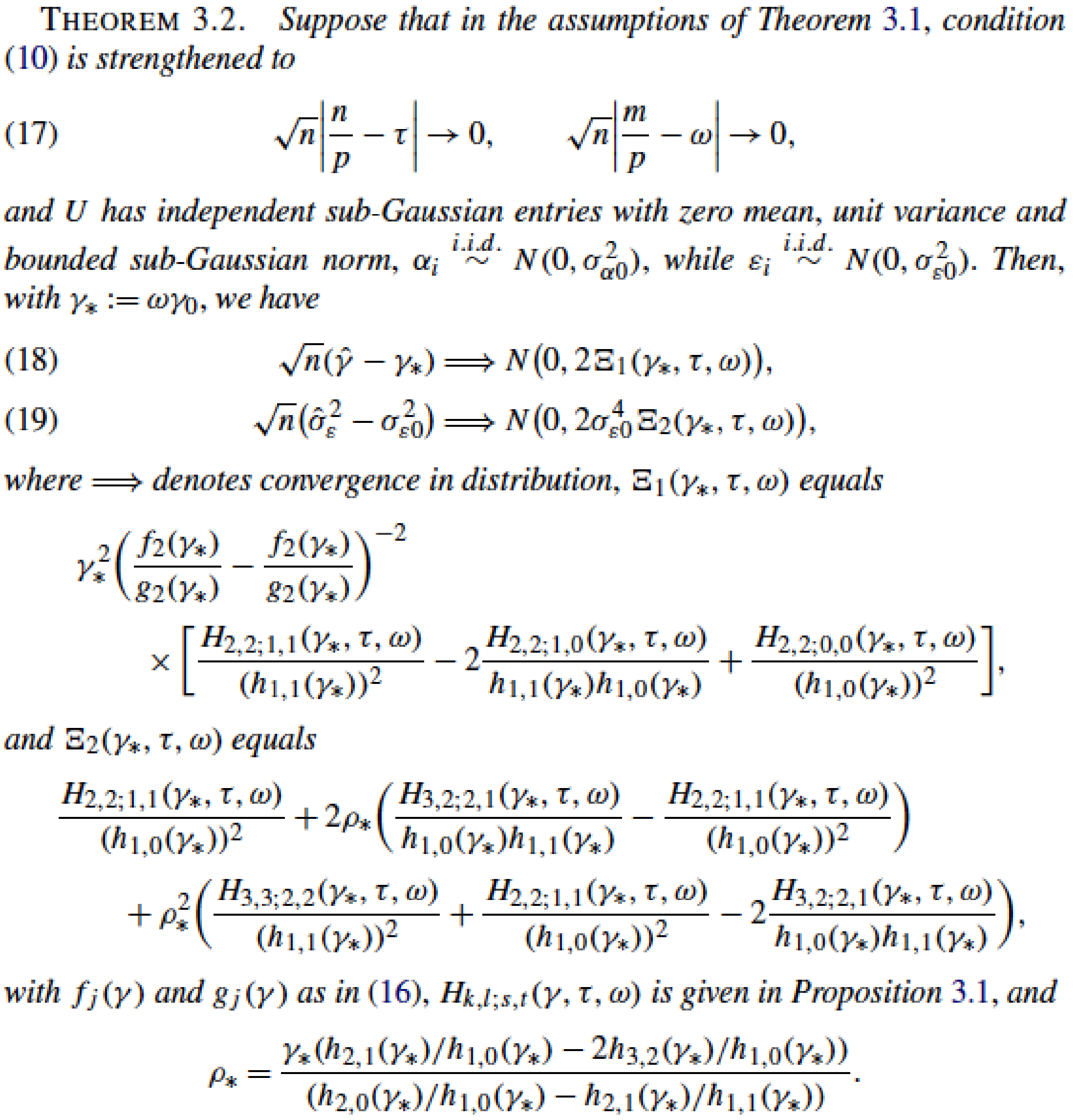
Assuming SNPs are independent



THEOREM 3.1. *Suppose that the true* $\sigma_\alpha^2, \sigma_\varepsilon^2$ *are positive, and* (10) *holds. Then*:

(i) *With probability tending to one, there is a REML estimator*, $\hat{\gamma}$, *such that* $\hat{\gamma} \xrightarrow{\text{P}} \omega\gamma_0$, *where* γ_0 *is the true* γ.

(ii) $\hat{\sigma}_\varepsilon^2 \xrightarrow{\text{P}} \sigma_{\varepsilon 0}^2$, *where* $\hat{\sigma}_\varepsilon^2$ *is* (4) *with* $\gamma = \hat{\gamma}$, *as in* (i), *and* $\sigma_{\varepsilon 0}^2$ *is the true* σ_ε^2.



THEOREM 3.2. *Suppose that in the assumptions of Theorem* 3.1, *condition* (10) *is strengthened to*

$$\sqrt{n}\left|\frac{n}{p}-\tau\right| \to 0, \qquad \sqrt{n}\left|\frac{m}{p}-\omega\right| \to 0, \tag{17}$$

and U has independent sub-Gaussian entries with zero mean, unit variance and bounded sub-Gaussian norm, $\alpha_i \overset{i.i.d.}{\sim} N(0,\sigma_{\alpha 0}^2)$, while $\varepsilon_i \overset{i.i.d.}{\sim} N(0,\sigma_{\varepsilon 0}^2)$. Then, with $\gamma_* := \omega\gamma_0$, we have

$$\sqrt{n}(\hat{\gamma}-\gamma_*) \Longrightarrow N\bigl(0, 2\Xi_1(\gamma_*,\tau,\omega)\bigr), \tag{18}$$

$$\sqrt{n}(\hat{\sigma}_\varepsilon^2-\sigma_{\varepsilon 0}^2) \Longrightarrow N\bigl(0, 2\sigma_{\varepsilon 0}^4\Xi_2(\gamma_*,\tau,\omega)\bigr), \tag{19}$$

where $\Longrightarrow$ denotes convergence in distribution, $\Xi_1(\gamma_*,\tau,\omega)$ equals

$$\gamma_*^2\left(\frac{f_2(\gamma_*)}{g_2(\gamma_*)}-\frac{f_2(\gamma_*)}{g_2(\gamma_*)}\right)^{-2} \times\left[\frac{H_{2,2;1,1}(\gamma_*,\tau,\omega)}{(h_{1,1}(\gamma_*))^2}-2\frac{H_{2,2;1,0}(\gamma_*,\tau,\omega)}{h_{1,1}(\gamma_*)h_{1,0}(\gamma_*)}+\frac{H_{2,2;0,0}(\gamma_*,\tau,\omega)}{(h_{1,0}(\gamma_*))^2}\right],$$

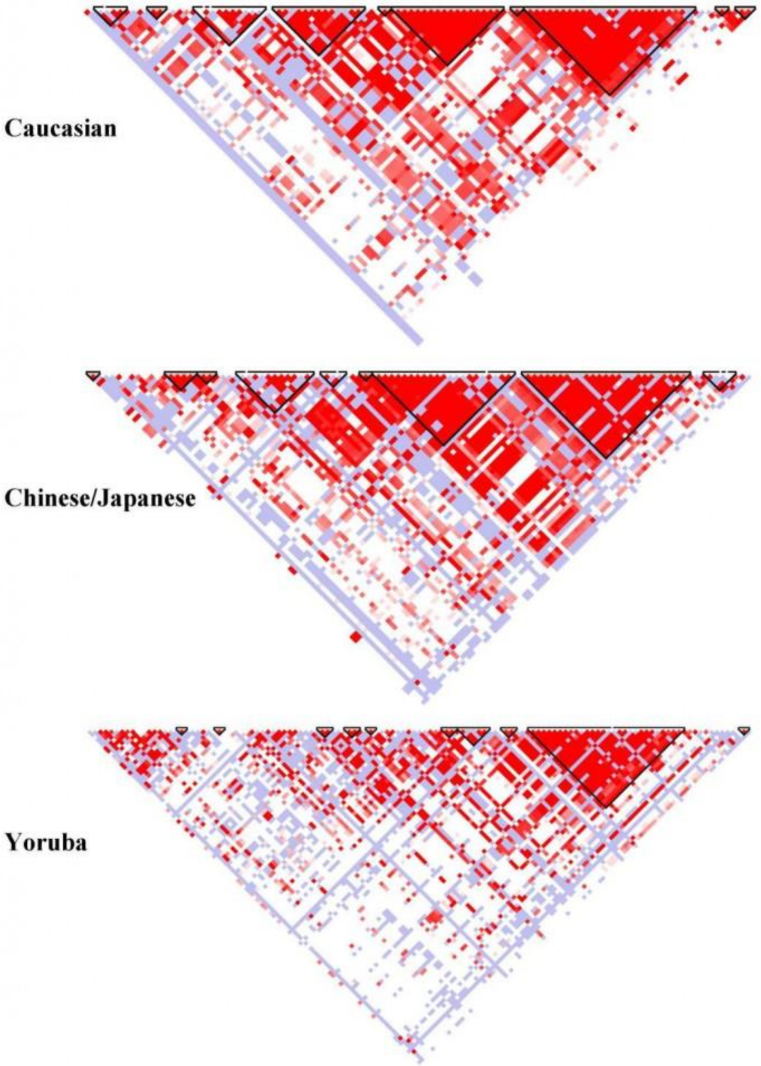
and $\Xi_2(\gamma_*,\tau,\omega)$ equals

$$\frac{H_{2,2;1,1}(\gamma_*,\tau,\omega)}{(h_{1,0}(\gamma_*))^2}+2\rho_*\left(\frac{H_{3,2;2,1}(\gamma_*,\tau,\omega)}{h_{1,0}(\gamma_*)h_{1,1}(\gamma_*)}-\frac{H_{2,2;1,1}(\gamma_*,\tau,\omega)}{(h_{1,0}(\gamma_*))^2}\right) +\rho_*^2\left(\frac{H_{3,3;2,2}(\gamma_*,\tau,\omega)}{(h_{1,1}(\gamma_*))^2}+\frac{H_{2,2;1,1}(\gamma_*,\tau,\omega)}{(h_{1,0}(\gamma_*))^2}-2\frac{H_{3,2;2,1}(\gamma_*,\tau,\omega)}{h_{1,0}(\gamma_*)h_{1,1}(\gamma_*)}\right),$$

with $f_j(\gamma)$ and $g_j(\gamma)$ as in (16), *$H_{k,l;s,t}(\gamma,\tau,\omega)$ is given in Proposition* 3.1, *and*

$$\rho_* = \frac{\gamma_*(h_{2,1}(\gamma_*)/h_{1,0}(\gamma_*)-2h_{3,2}(\gamma_*)/h_{1,0}(\gamma_*))}{(h_{2,0}(\gamma_*)/h_{1,0}(\gamma_*)-h_{2,1}(\gamma_*)/h_{1,1}(\gamma_*))}.$$

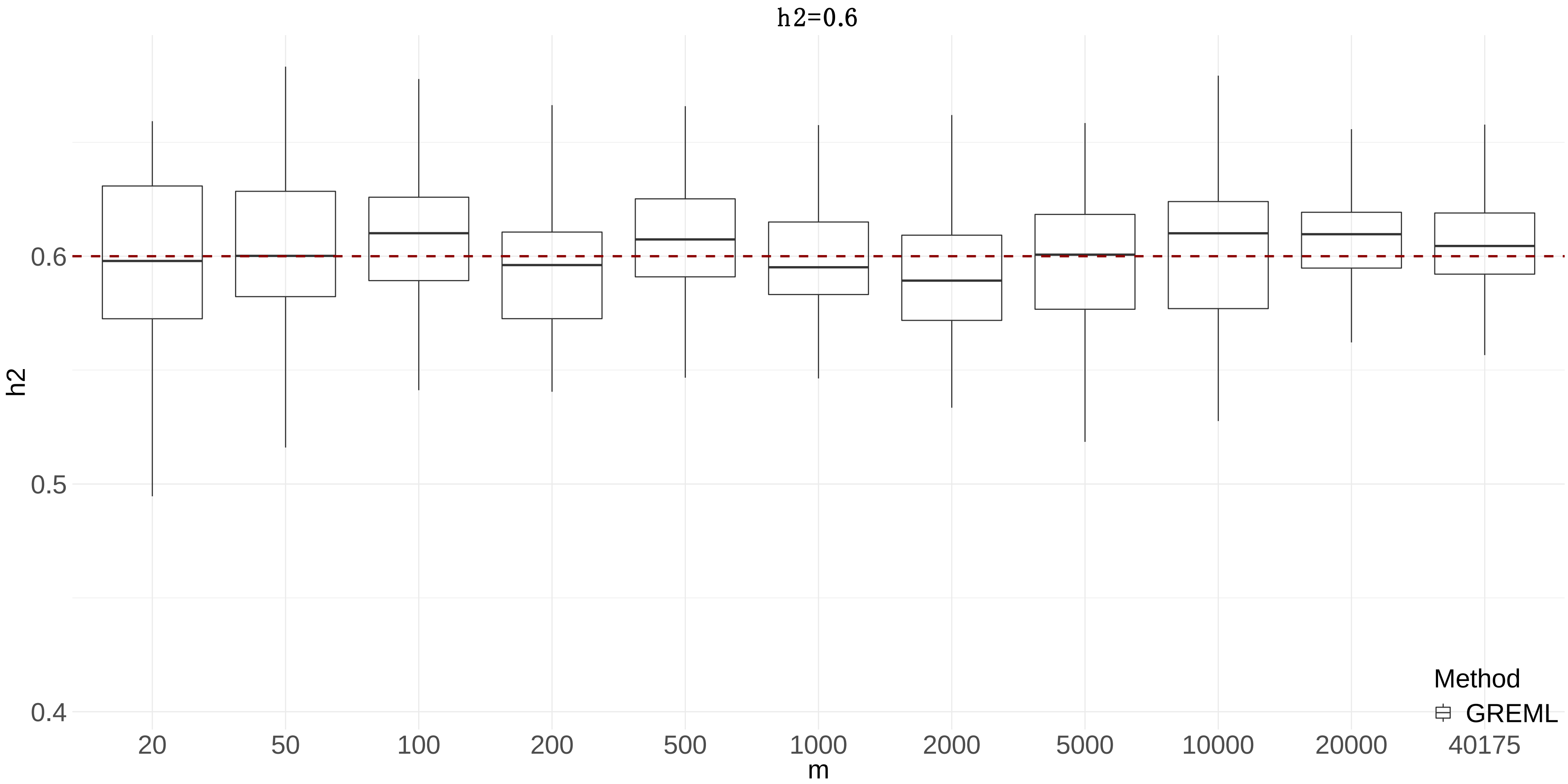
SNPs are NOT independent



http://graphics.cs.wisc.edu/WP/vis10/archives/458-hapmap-linkage-disequilibrium-plot

Robustness of REML with LD

Model dependence after UKBB (UK BioBank)

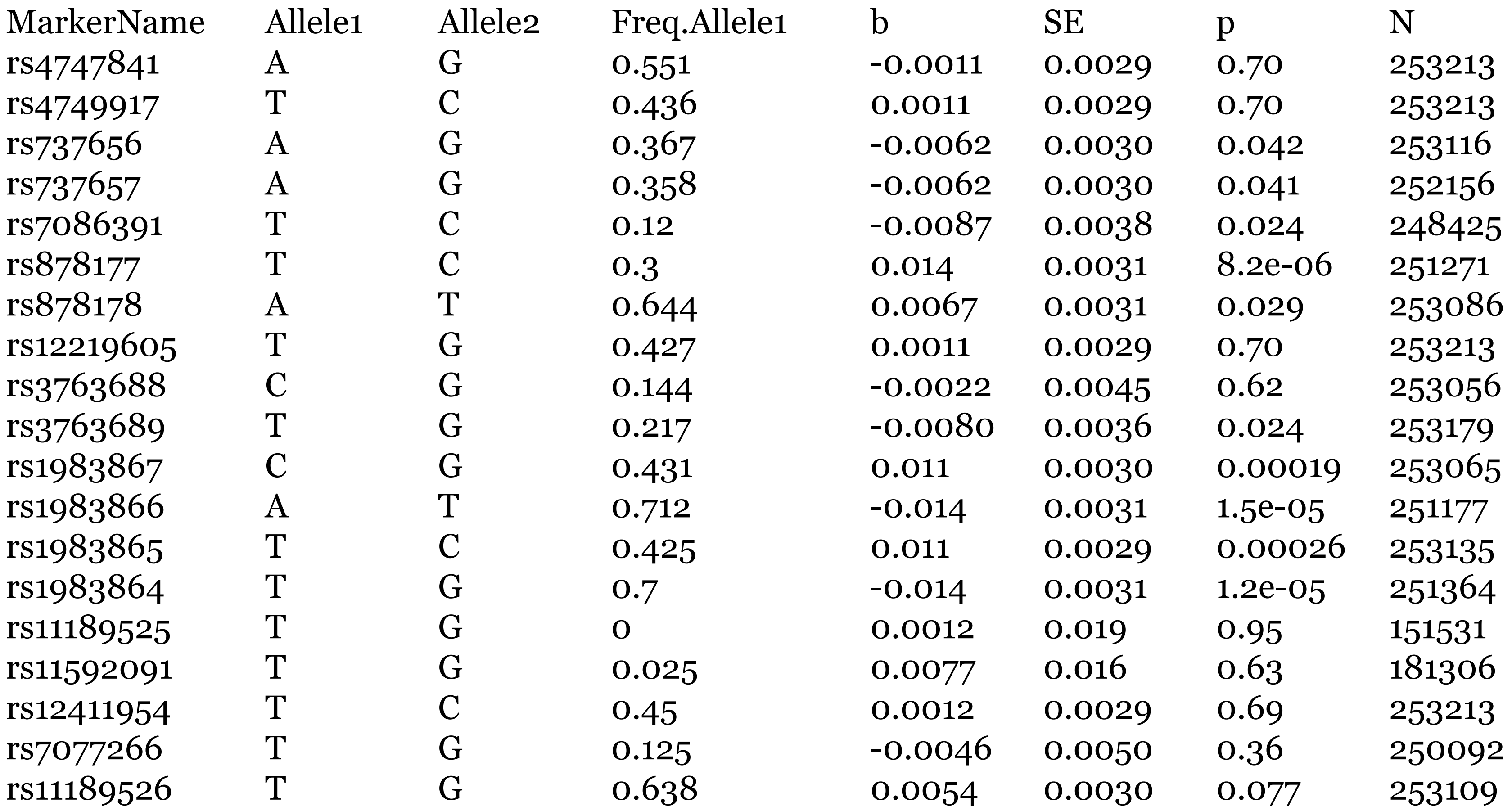


REML requires **individual level data,**
e.g. individual genotypes and phenotypes

However, there are many challenges of accessing individual level GWAS data, e.g. privacy

Much easier to get **summary statistics**

Example of Summary Statistics



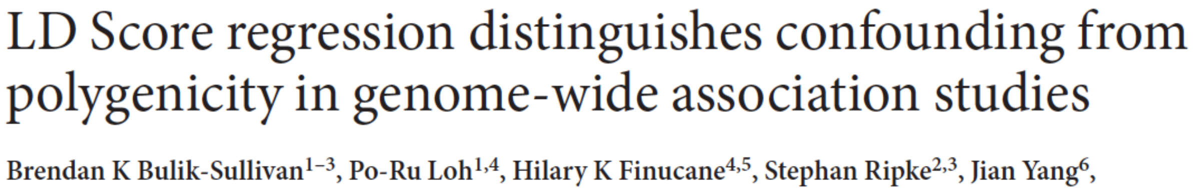
MarkerName	Allele1	Allele2	Freq.Allele1	b	SE	p	N
rs4747841	A	G	0.551	-0.0011	0.0029	0.70	253213
rs4749917	T	C	0.436	0.0011	0.0029	0.70	253213
rs737656	A	G	0.367	-0.0062	0.0030	0.042	253116
rs737657	A	G	0.358	-0.0062	0.0030	0.041	252156
rs7086391	T	C	0.12	-0.0087	0.0038	0.024	248425
rs878177	T	C	0.3	0.014	0.0031	8.2e-06	251271
rs878178	A	T	0.644	0.0067	0.0031	0.029	253086
rs12219605	T	G	0.427	0.0011	0.0029	0.70	253213
rs3763688	C	G	0.144	-0.0022	0.0045	0.62	253056
rs3763689	T	G	0.217	-0.0080	0.0036	0.024	253179
rs1983867	C	G	0.431	0.011	0.0030	0.00019	253065
rs1983866	A	T	0.712	-0.014	0.0031	1.5e-05	251177
rs1983865	T	C	0.425	0.011	0.0029	0.00026	253135
rs1983864	T	G	0.7	-0.014	0.0031	1.2e-05	251364
rs11189525	T	G	0	0.0012	0.019	0.95	151531
rs11592091	T	G	0.025	0.0077	0.016	0.63	181306
rs12411954	T	C	0.45	0.0012	0.0029	0.69	253213
rs7077266	T	G	0.125	-0.0046	0.0050	0.36	250092
rs11189526	T	G	0.638	0.0054	0.0030	0.077	253109

LD Score Regression

$$E[\chi_j^2 | l_j] = \frac{N\sigma^2 l_j}{m} + Na + 1$$

N:	sample size
m:	number of markers
σ^2/m:	average heritability explained per SNP
a:	contribution of confounding bias

$$l_j = \sum_k r_{jk}^2$$

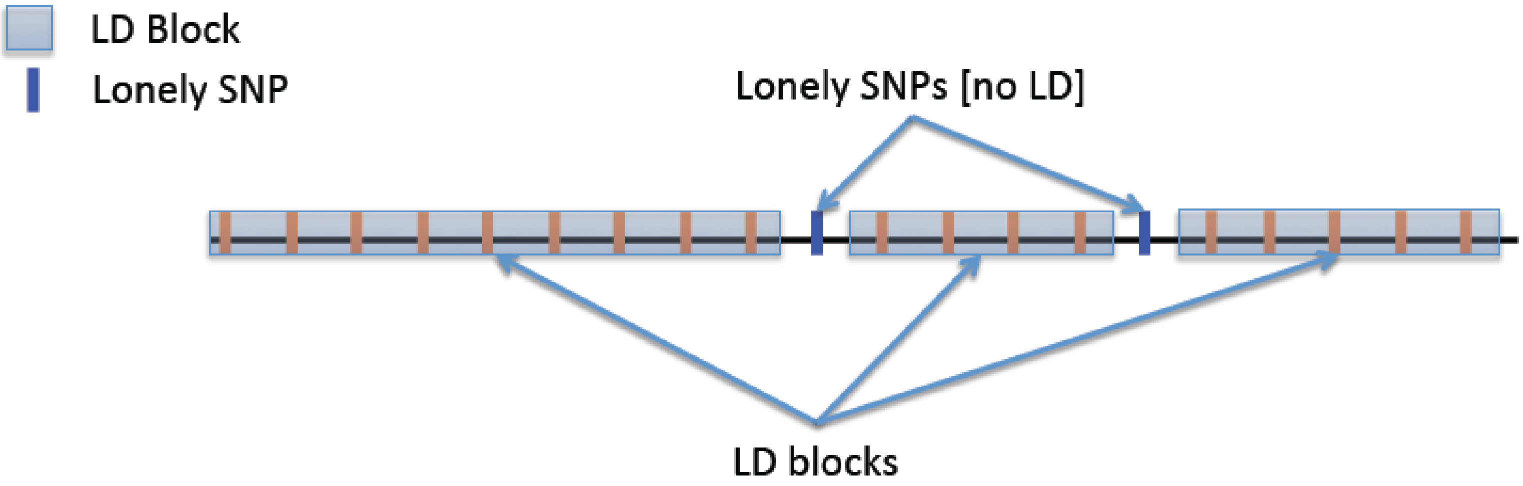


LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6],

LD Score Regression

Toy Illustration of Genome

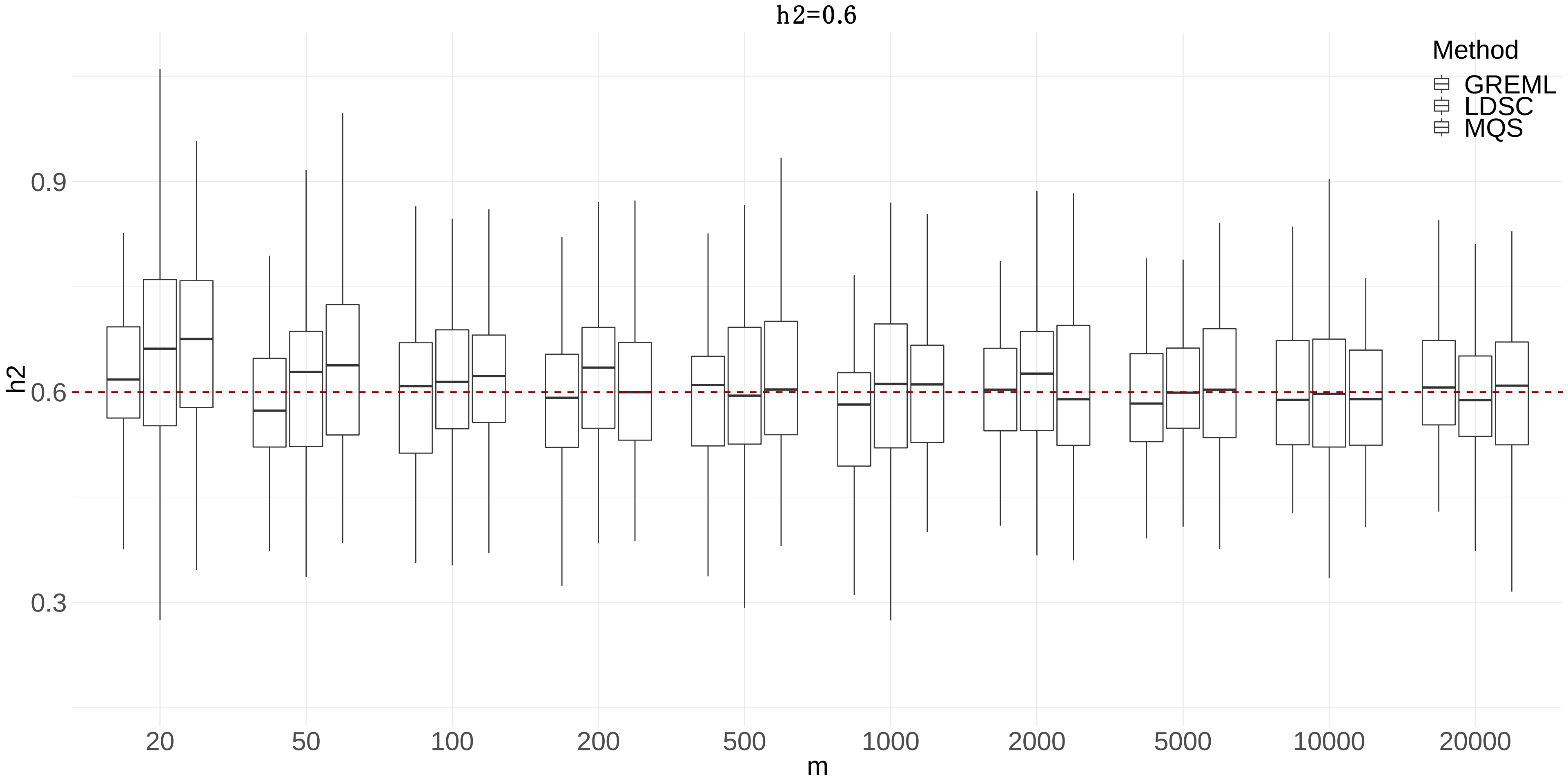


LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6],

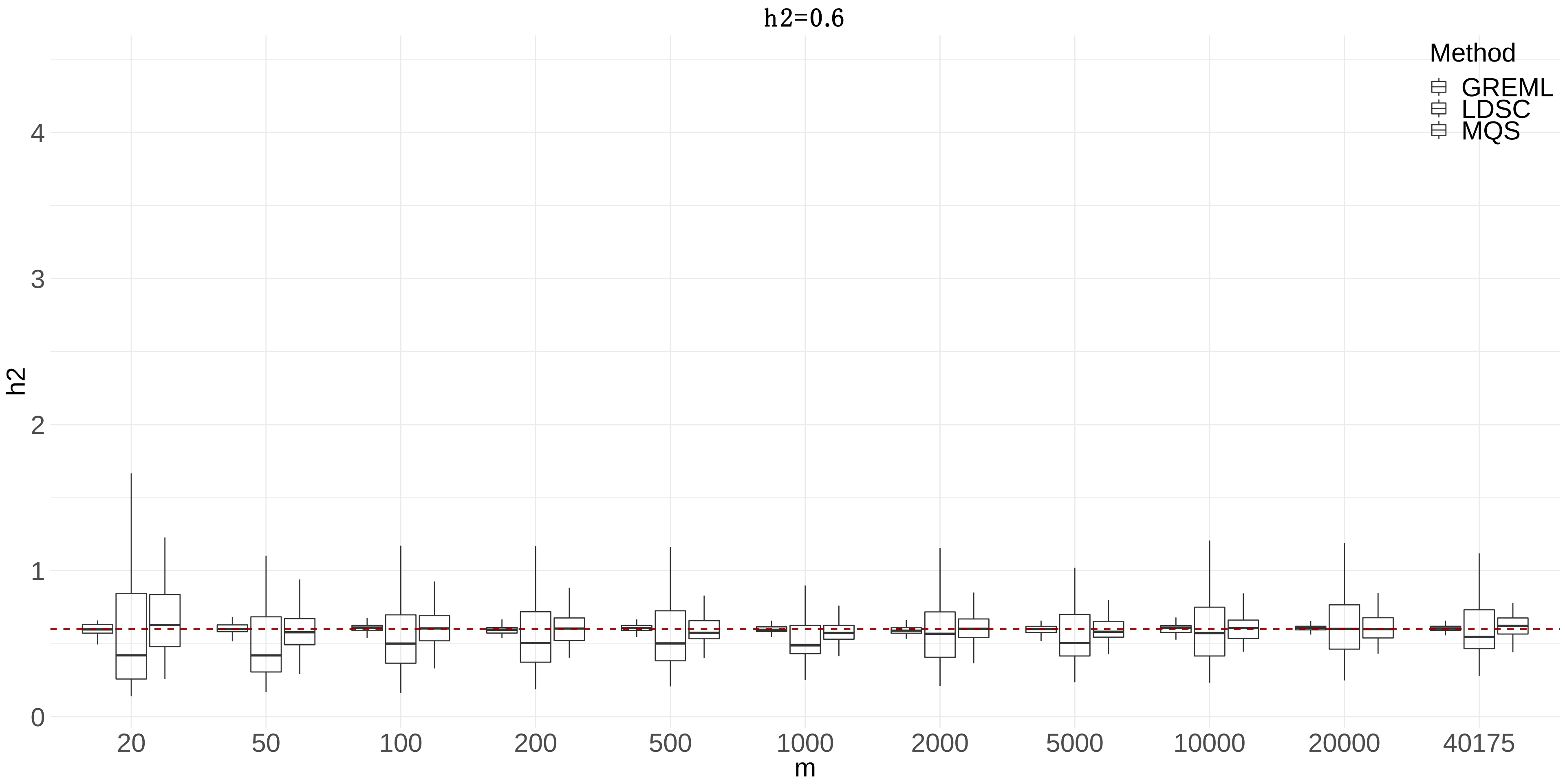
REML versus LDSC

Small variation of LD score (independent)



REML versus LDSC

Large variation of LD score (UKBB)



SNPs are NOT equal

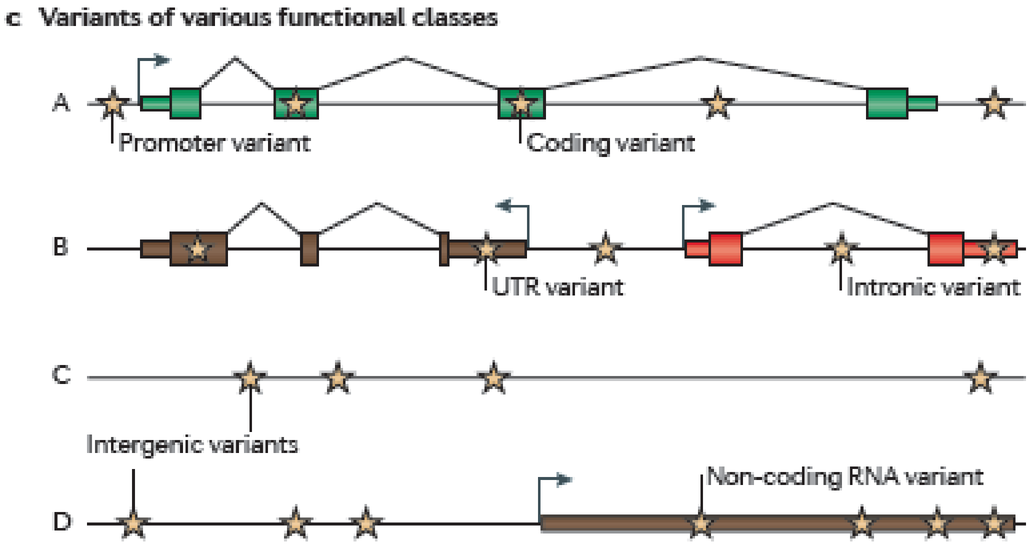
Trait	M1	M2….	Mm	x1	x2 …	xk
1	AA	AB ….	BB	1.1	2.3…	1.5
1	AB	?? ….	AA	1.5	0.5…	2.9
…………………………………………………						
…………………………………………………						
0	BB	AA ….	BB	0.5	?? …	1.8

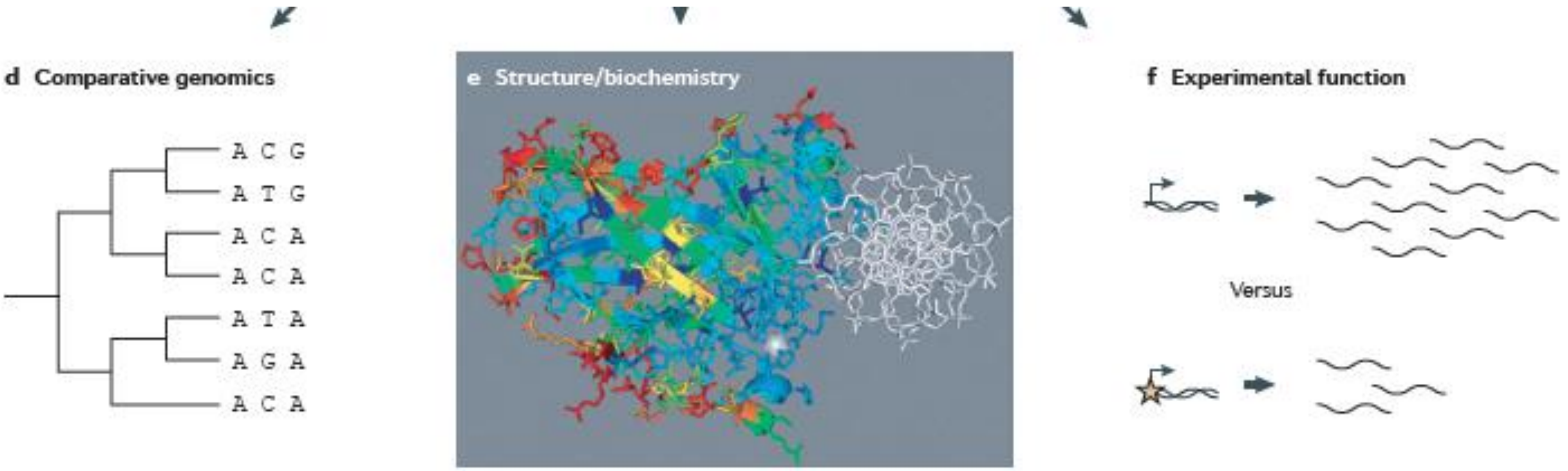
Abundant Annotation Information

Yale

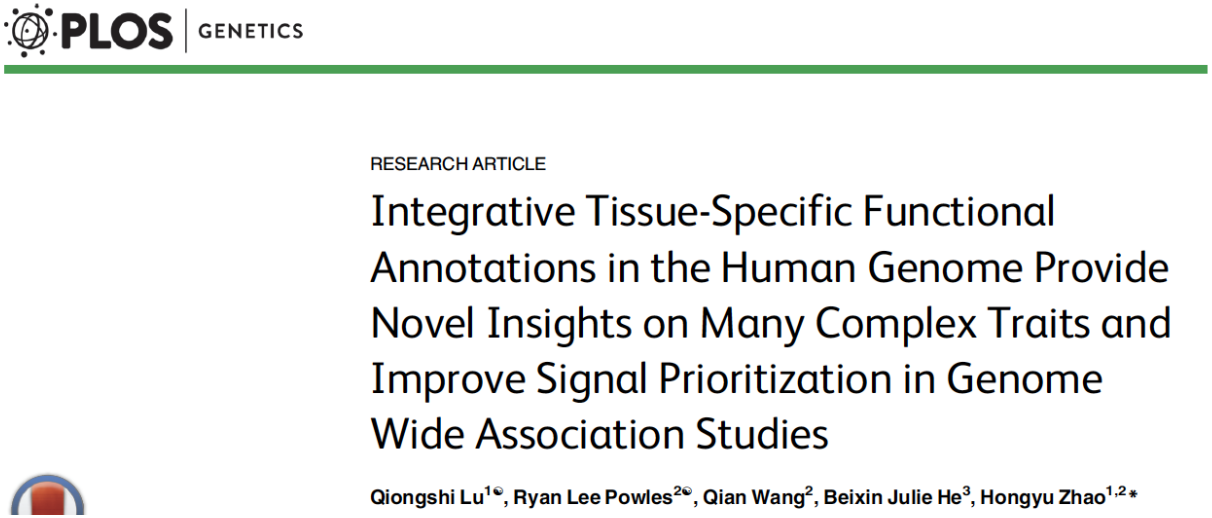
Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data

Gregory M. Cooper and Jay Shendure‡*





Tissue/Cell Type Specific Annotations

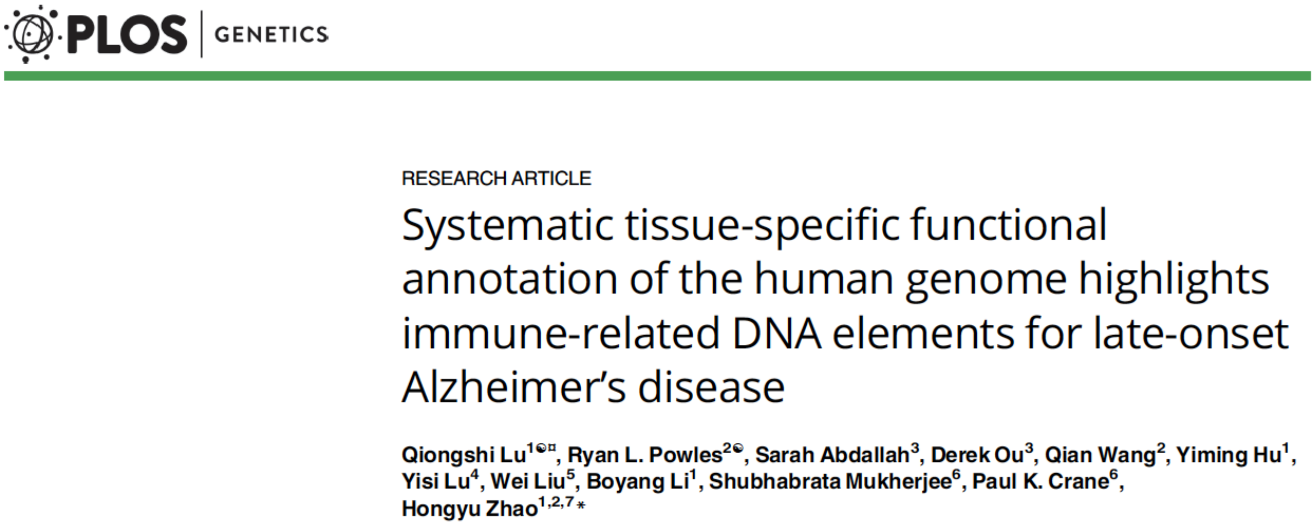


PLOS GENETICS

RESEARCH ARTICLE

Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies

Qiongshi Lu[1☯], Ryan Lee Powles[2☯], Qian Wang[2], Beixin Julie He[3], Hongyu Zhao[1,2]*



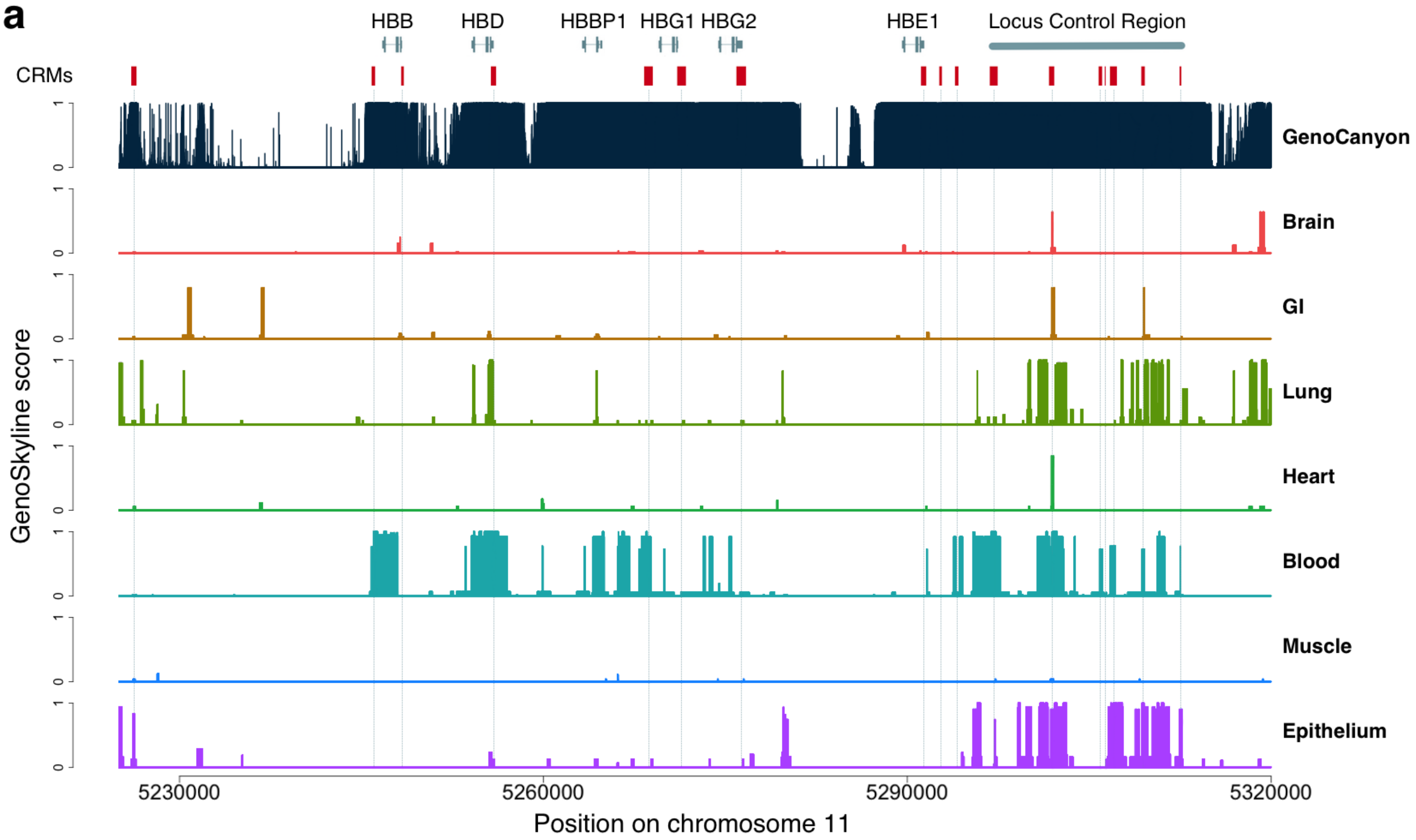
PLOS GENETICS

RESEARCH ARTICLE

Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease

Qiongshi Lu[1☯¤], Ryan L. Powles[2☯], Sarah Abdallah[3], Derek Ou[3], Qian Wang[2], Yiming Hu[1], Yisi Lu[4], Wei Liu[5], Boyang Li[1], Shubhabrata Mukherjee[6], Paul K. Crane[6], Hongyu Zhao[1,2,7]*

Tissue Specific Annotations



Which tissues are relevant to a trait?

Basic idea: Consider genetic contributions of SNPs from the “functional” regions of different tissues, i.e. tissue specific heritability

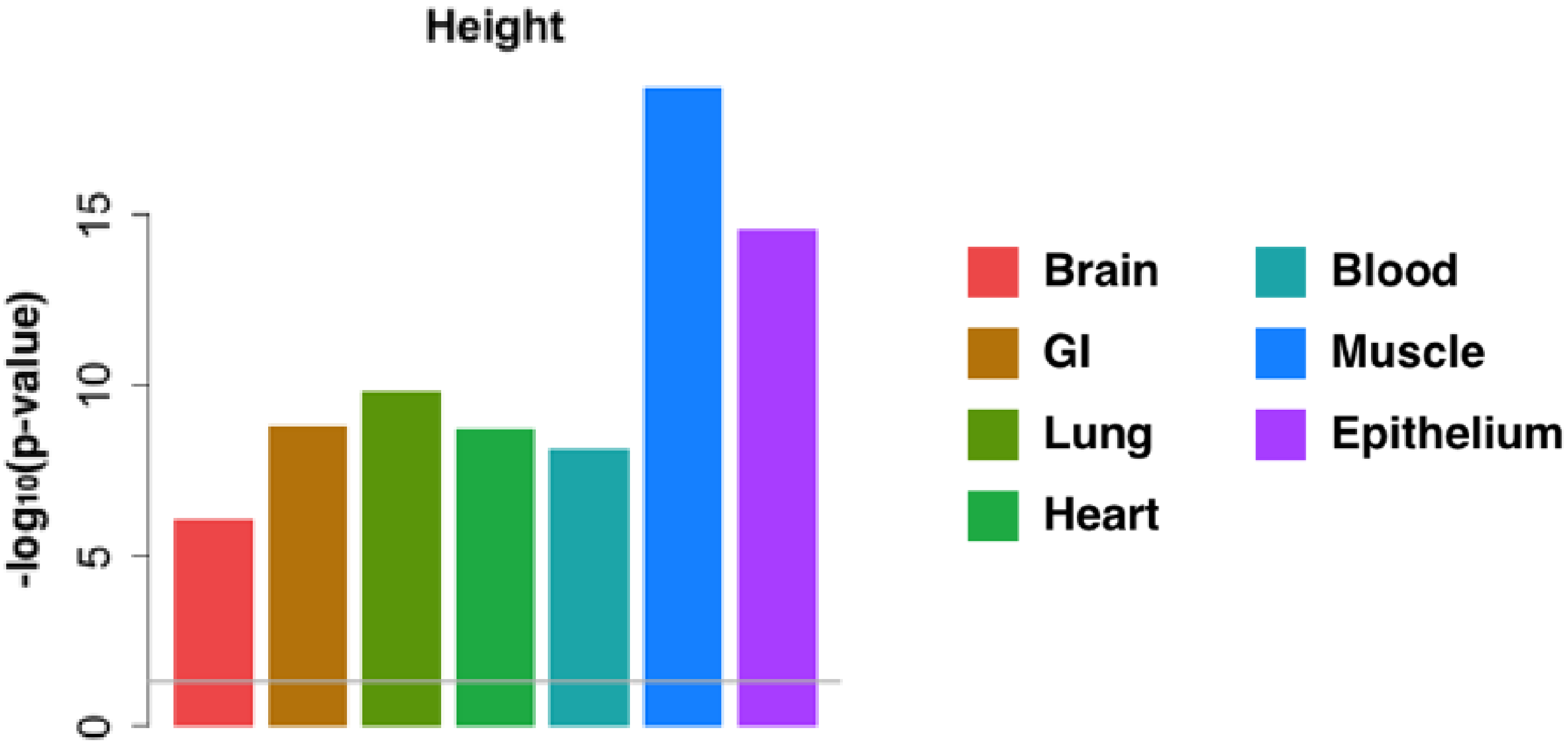
Example:

Suppose **10%** of the genome in a tissue is “functional”

SNPs in the “functional” genome explain **30%** of variance

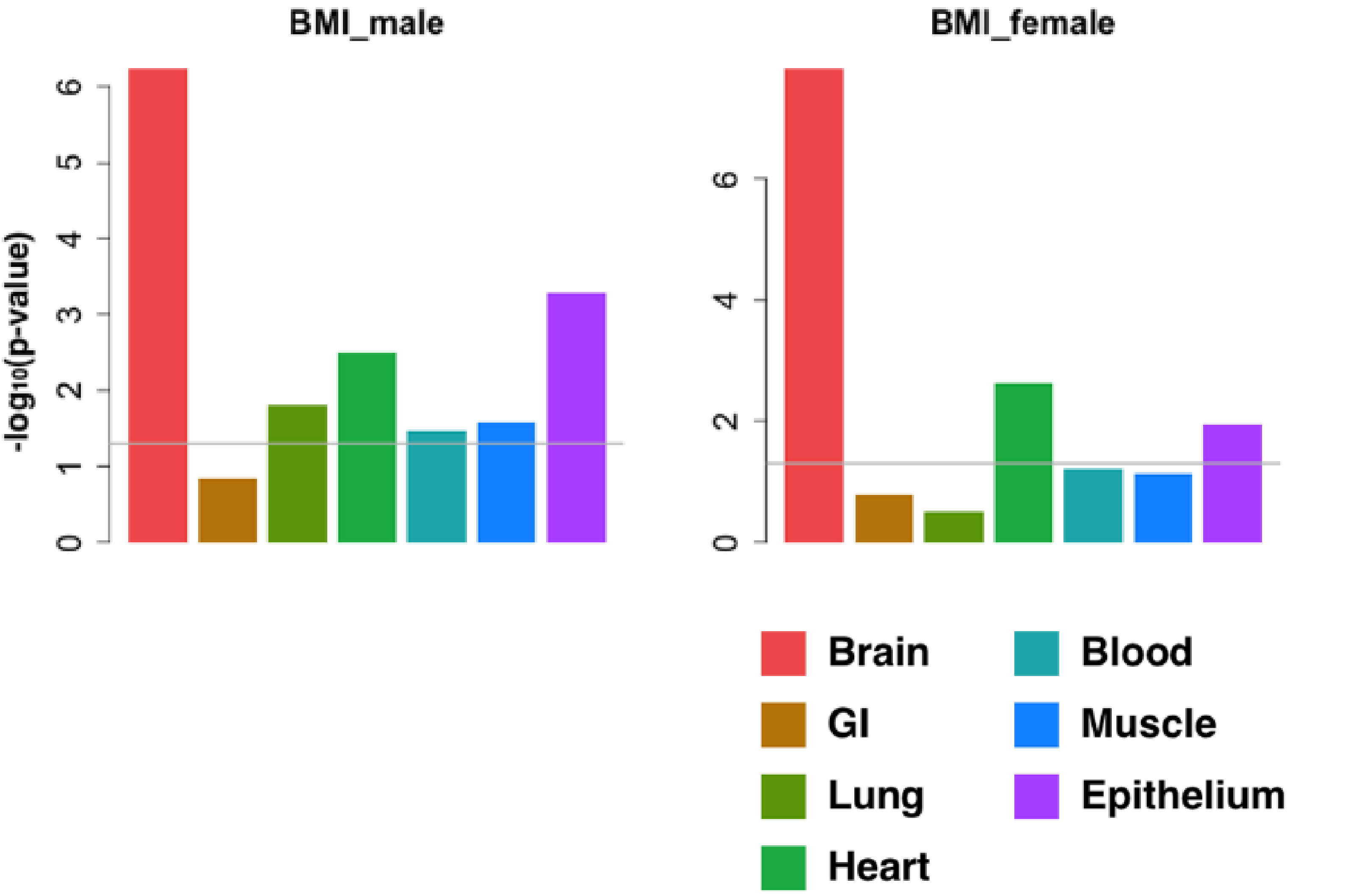
3-fold enrichment

Height GWAS

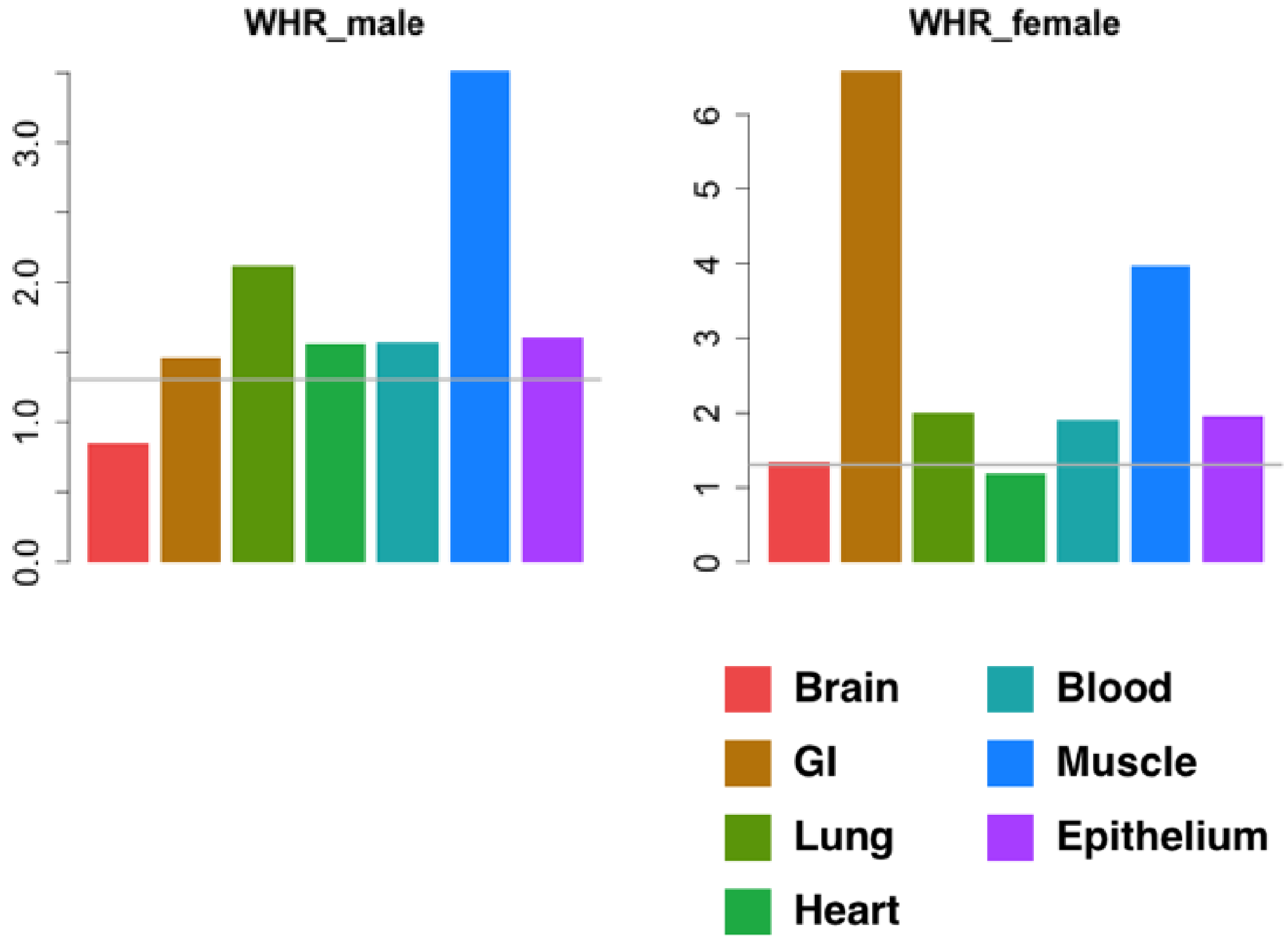

Height
-log10(p-value)
0
5
10
15
Brain
GI
Lung
Heart
Blood
Muscle
Epithelium

Body Mass Index GWAS

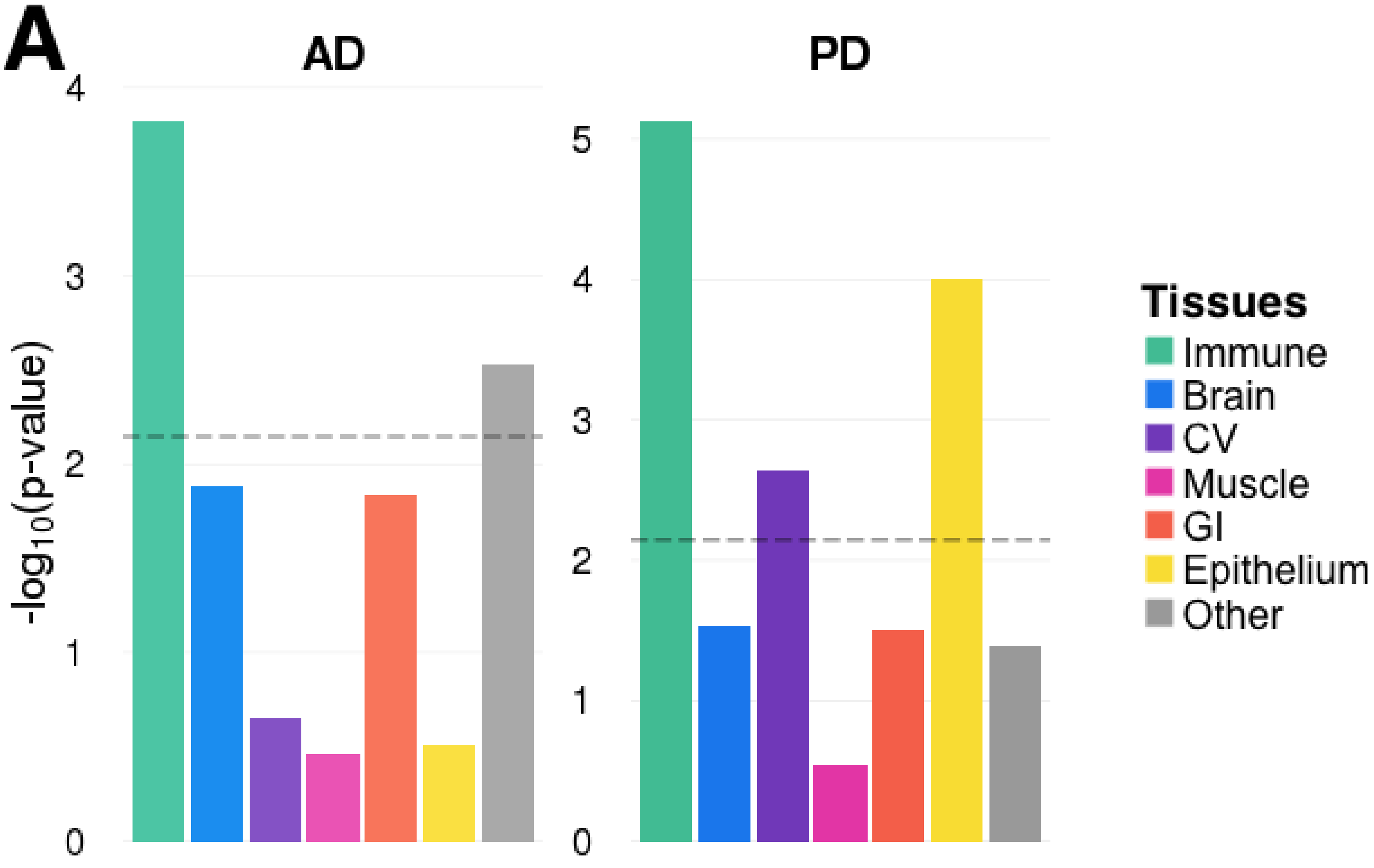
Yale



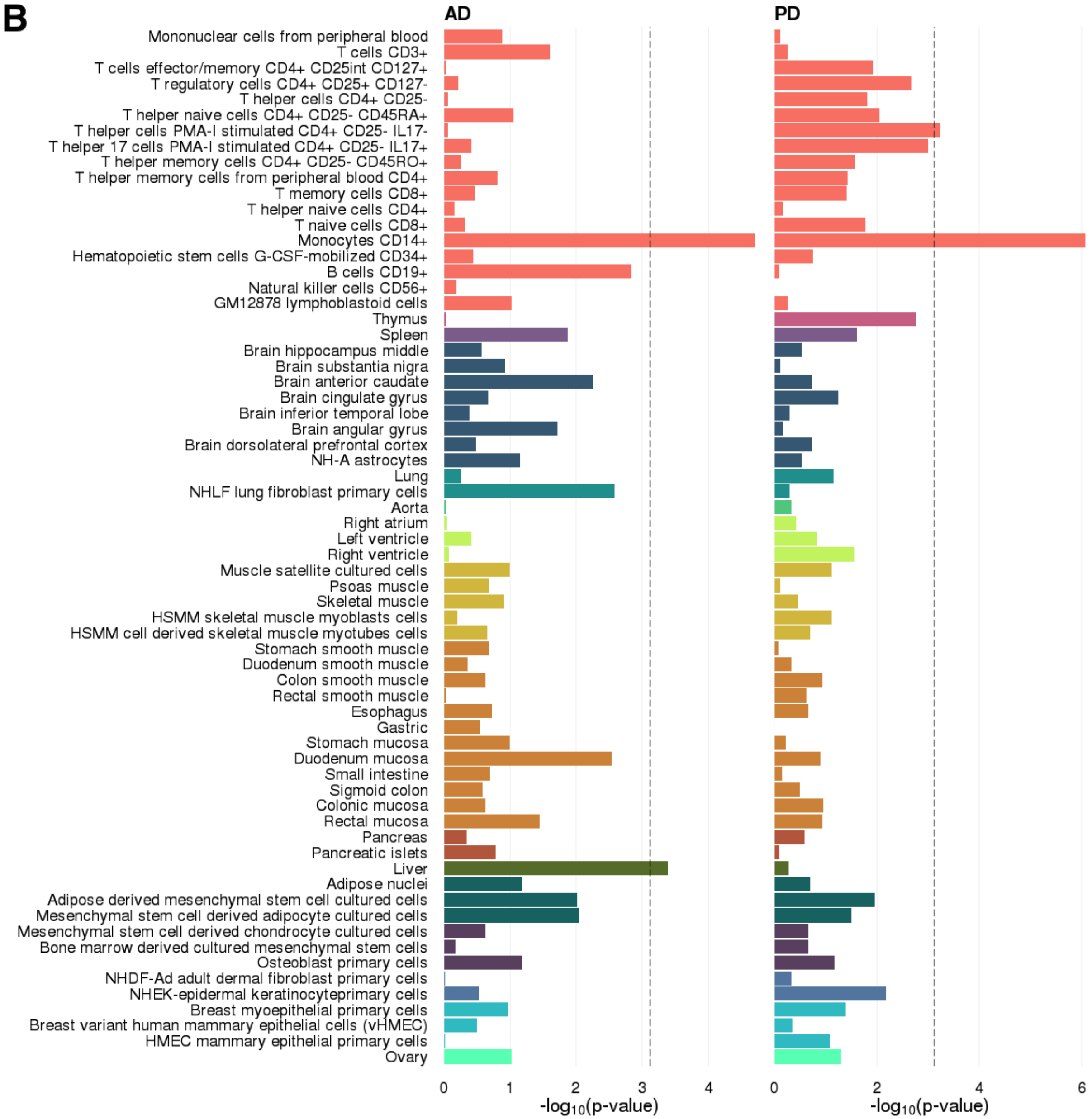
Waist Hip Ratio GWAS



Alzheimer and Parkinson



Alzheimer and Parkinson



Shared Genetics Between Two Traits

$$y_1 = \sum_{i=1}^{K} X_i\beta_i + \epsilon$$

$$y_2 = \sum_{i=1}^{K} Z_i\gamma_i + \delta$$

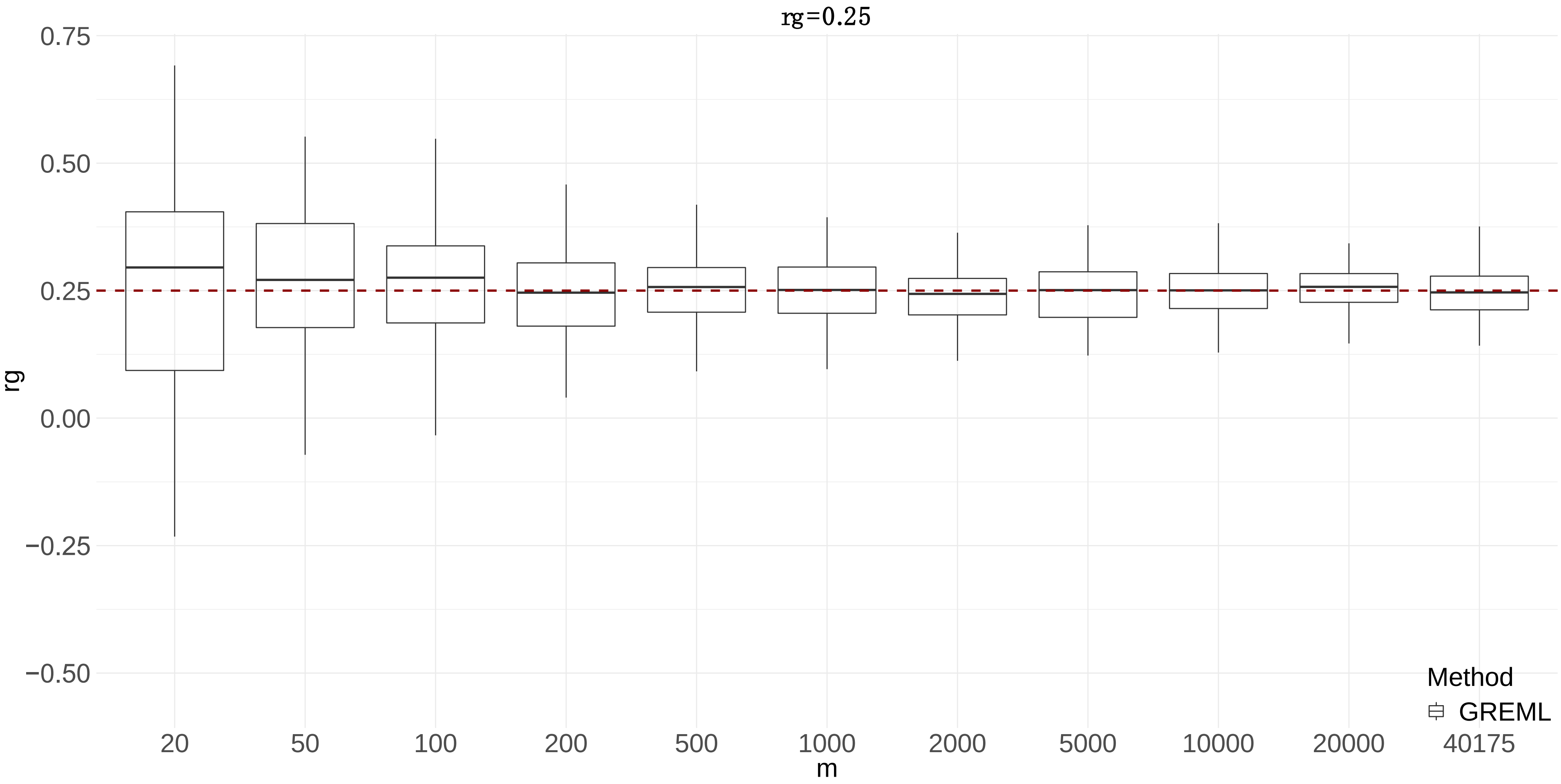
genetic covariance

$$\mathbb{E}(\beta_i) = \mathbb{E}(\gamma_i) = 0 \;\; and \;\; \mathbb{E}\left(\gamma_i\beta_i^T\right) = \frac{\rho_i}{m_i} I, \quad i = 1, \dots, K$$

Pleiotropy: one gene (variant) affects more than one trait

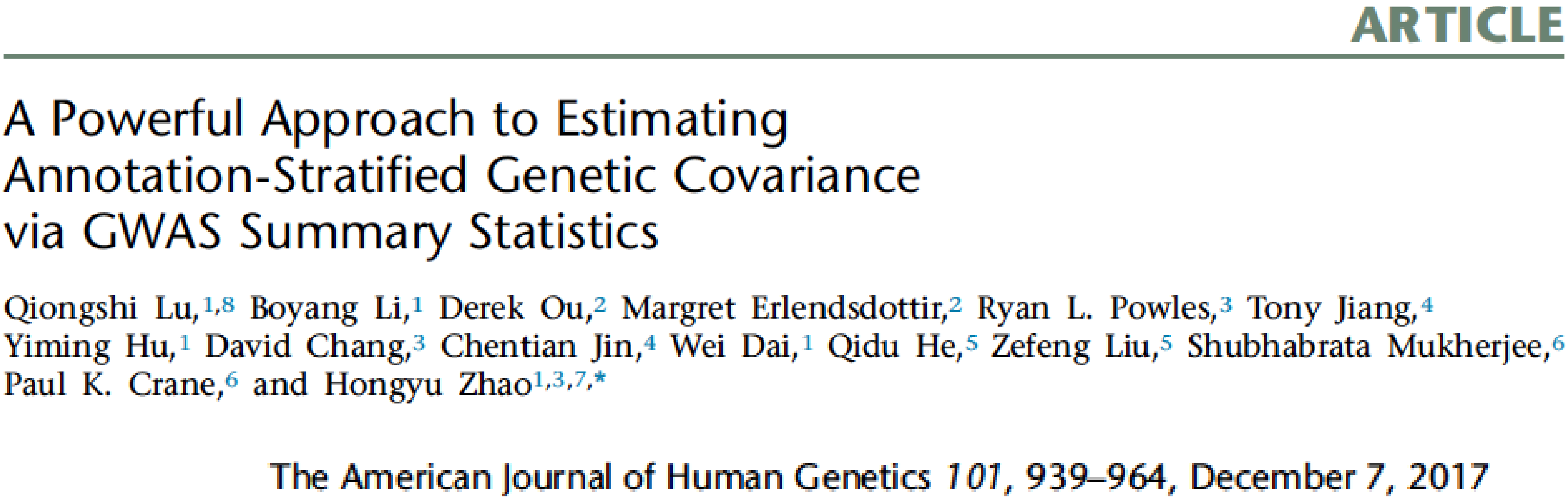
Robustness of REML with LD

UKBB



Genetic correlation can also be estimated using only summary statistics

Shared Genetics Between Two Traits



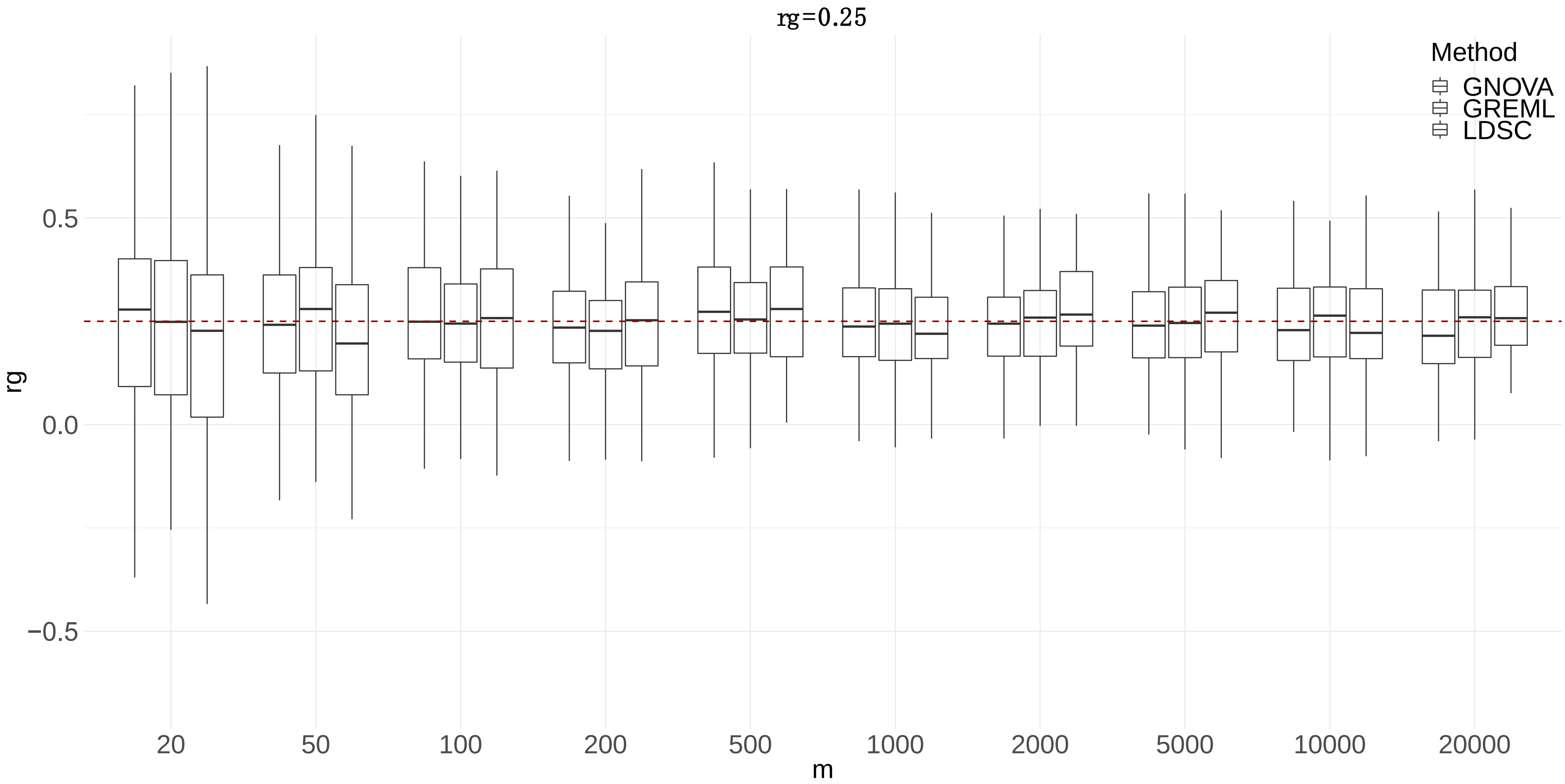
ARTICLE

A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics

Qiongshi Lu,[1,8] Boyang Li,[1] Derek Ou,[2] Margret Erlendsdottir,[2] Ryan L. Powles,[3] Tony Jiang,[4] Yiming Hu,[1] David Chang,[3] Chentian Jin,[4] Wei Dai,[1] Qidu He,[5] Zefeng Liu,[5] Shubhabrata Mukherjee,[6] Paul K. Crane,[6] and Hongyu Zhao[1,3,7,*]

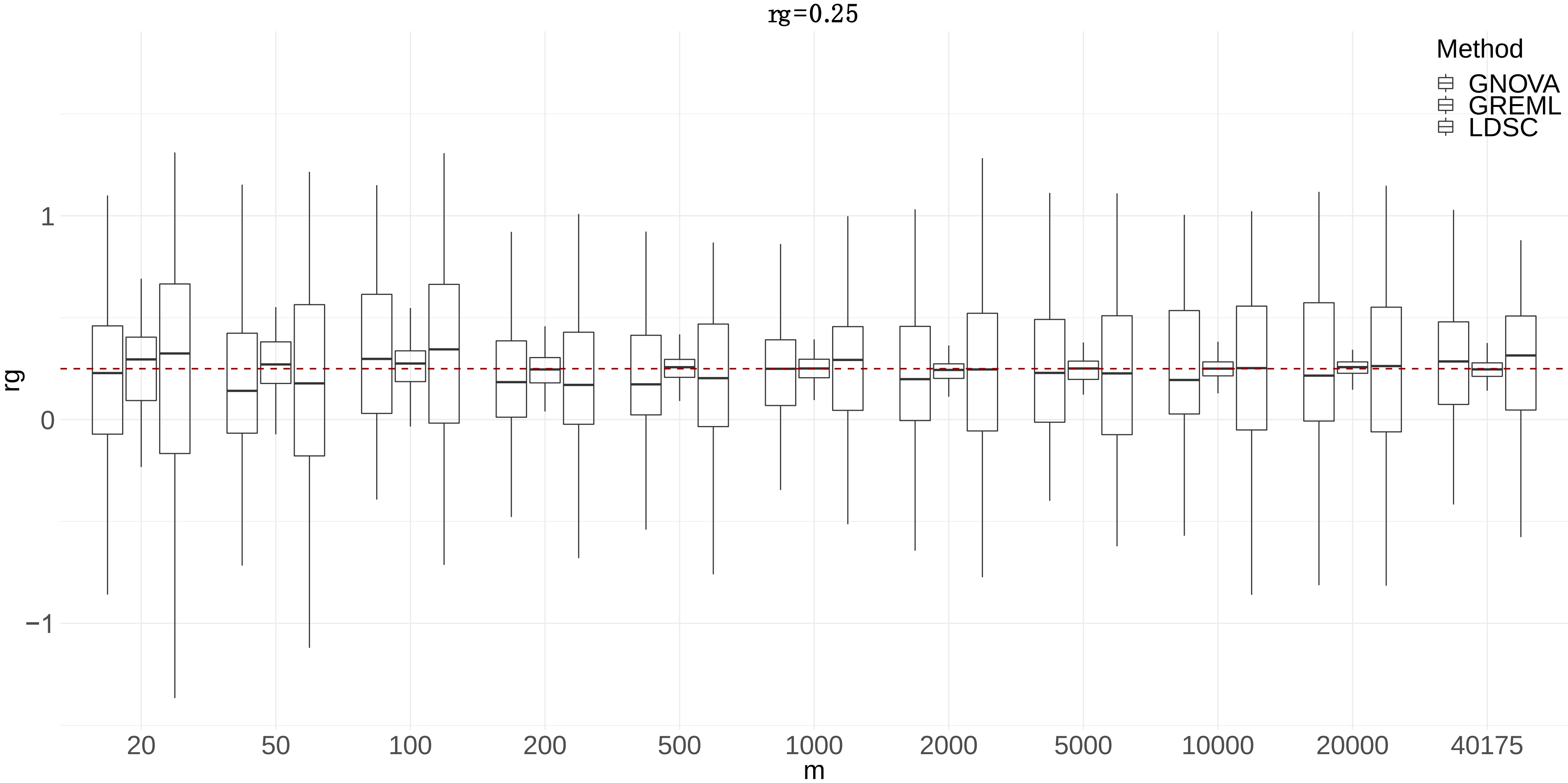
The American Journal of Human Genetics *101*, 939–964, December 7, 2017

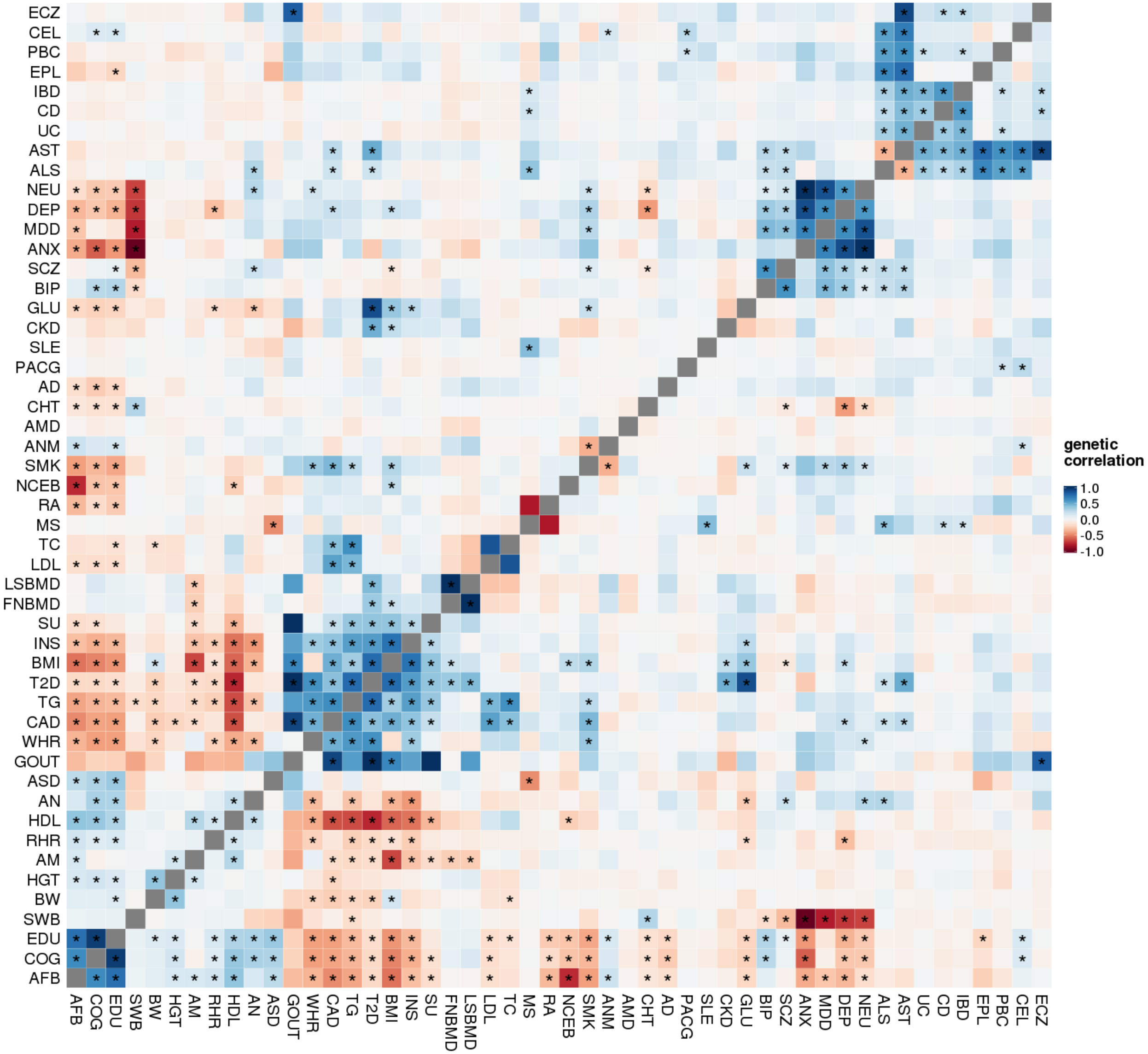
Independent SNPs

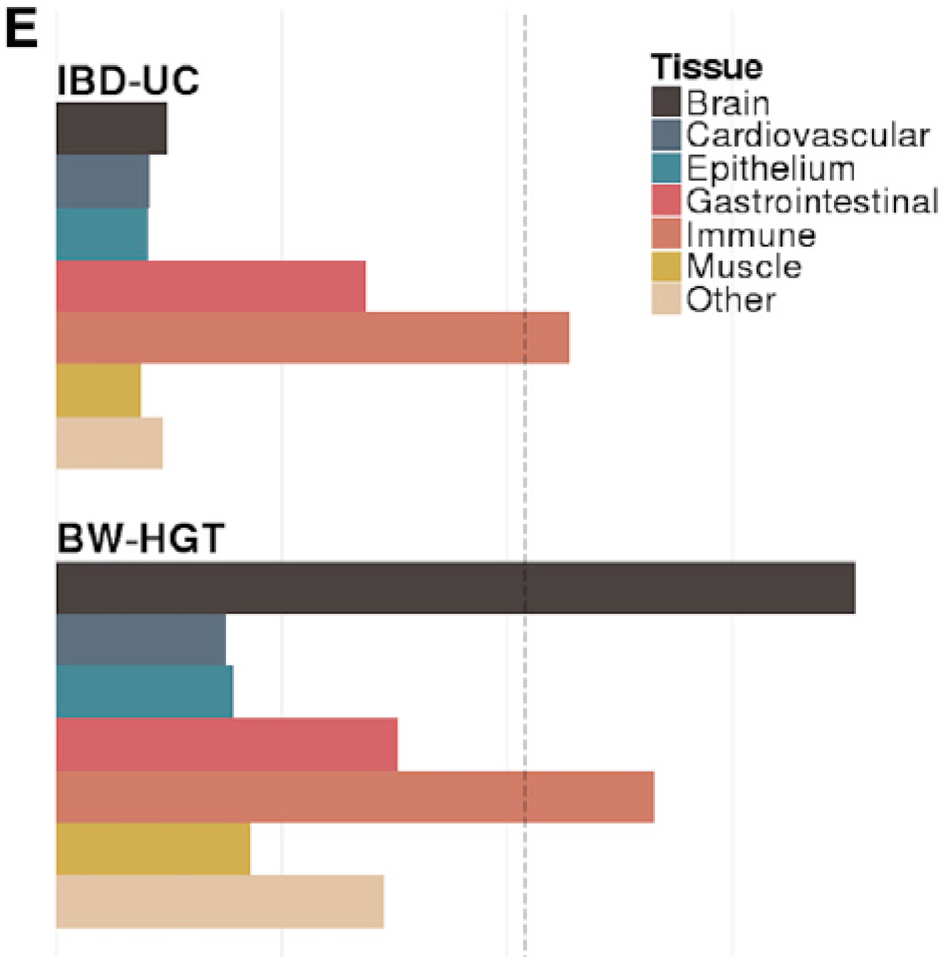


In the Presence of LD

UKBB




ECZ
CEL
PBC
EPL
IBD
CD
UC
AST
ALS
NEU
DEP
MDD
ANX
SCZ
BIP
GLU
CKD
SLE
PACG
AD
CHT
AMD
ANM
SMK
NCEB
RA
MS
TC
LDL
LSBMD
FNBMD
SU
INS
BMI
T2D
TG
CAD
WHR
GOUT
ASD
AN
HDL
RHR
AM
HGT
BW
SWB
EDU
COG
AFB
AFB
COG
EDU
SWB
BW
HGT
AM
RHR
HDL
AN
ASD
GOUT
WHR
CAD
TG
T2D
BMI
INS
SU
FNBMD
LSBMD
LDL
TC
MS
RA
NCEB
SMK
ANM
AMD
CHT
AD
PACG
SLE
CKD
GLU
BIP
SCZ
ANX
MDD
DEP
NEU
ALS
AST
UC
CD
IBD
EPL
PBC
CEL
ECZ
genetic correlation
1.0
0.5
0.0
-0.5
-1.0


E
IBD-UC
BW-HGT
Tissue
Brain
Cardiovascular
Epithelium
Gastrointestinal
Immune
Muscle
Other

A Journey of Fisher's model

Fisher's 1918 model

$$Y = G_1 + G_2 + G_3 + \ldots. + G_m + E + e$$

Random effects model

$$y_1 = \sum_{i=1}^{K} X_i \beta_i + \epsilon$$

$$y_2 = \sum_{i=1}^{K} Z_i \gamma_i + \delta$$

Summary

Fisher’s 1918 model has turned out to provide a rather accurate description connecting phenotypes to genotypes

This framework has enabled the estimation of chip-based heritability for hundreds of human traits/diseases from genome wide association study data, using either individual genotype data or summary statistics

It further leads to insights on the genetic architecture of many traits when other information (e.g. genome annotations and pleiotropy) is considered

Summary

Much more may be learned under this framework and its extensions to

[1] infer causal relationships among traits, e.g. phenotype networks

[2] analyze whole genome sequence data (~3 billion base pairs)

[3] study interplay between genetic and other factors on disease onset and progression

[4] handle longitudinal data, e.g. electronic health records

These will present challenges in

methodology, computation, and theory

1940:	Math Department, PKU
1945-1947:	UC Berkeley, Columbia, UNC Chapel Hill
1947:	Math Department, PKU
1956:	Director of Institute of Probability and Statistics, PKU

Minping Qian, **Zhongguo Zheng**, **Jiading Chen**, and many others

许宝騄

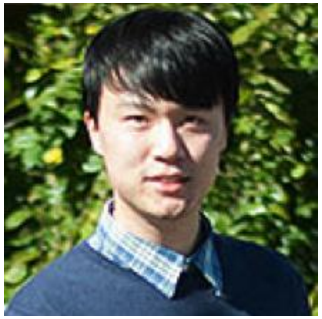
Acknowledgments

Yale

Can Yang
HKUST

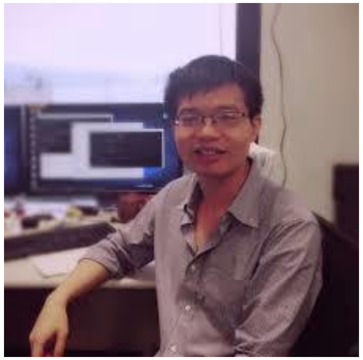
Jiming Jiang
UC Davis

Debashis Paul
UC Davis



Qiongshi Lu
U of Wisconsin Madison

Yiliang Zhang
Yale



Wei Jiang
Yale

Acknowledgments

Yale

Acknowledgments

Yale

ICSA

Award Committee Members

THANK YOU!