Estimation of quantitative genetic parameters from distant relatives using marker data

Peter M. Visscher
peter.visscher@uq.edu.au

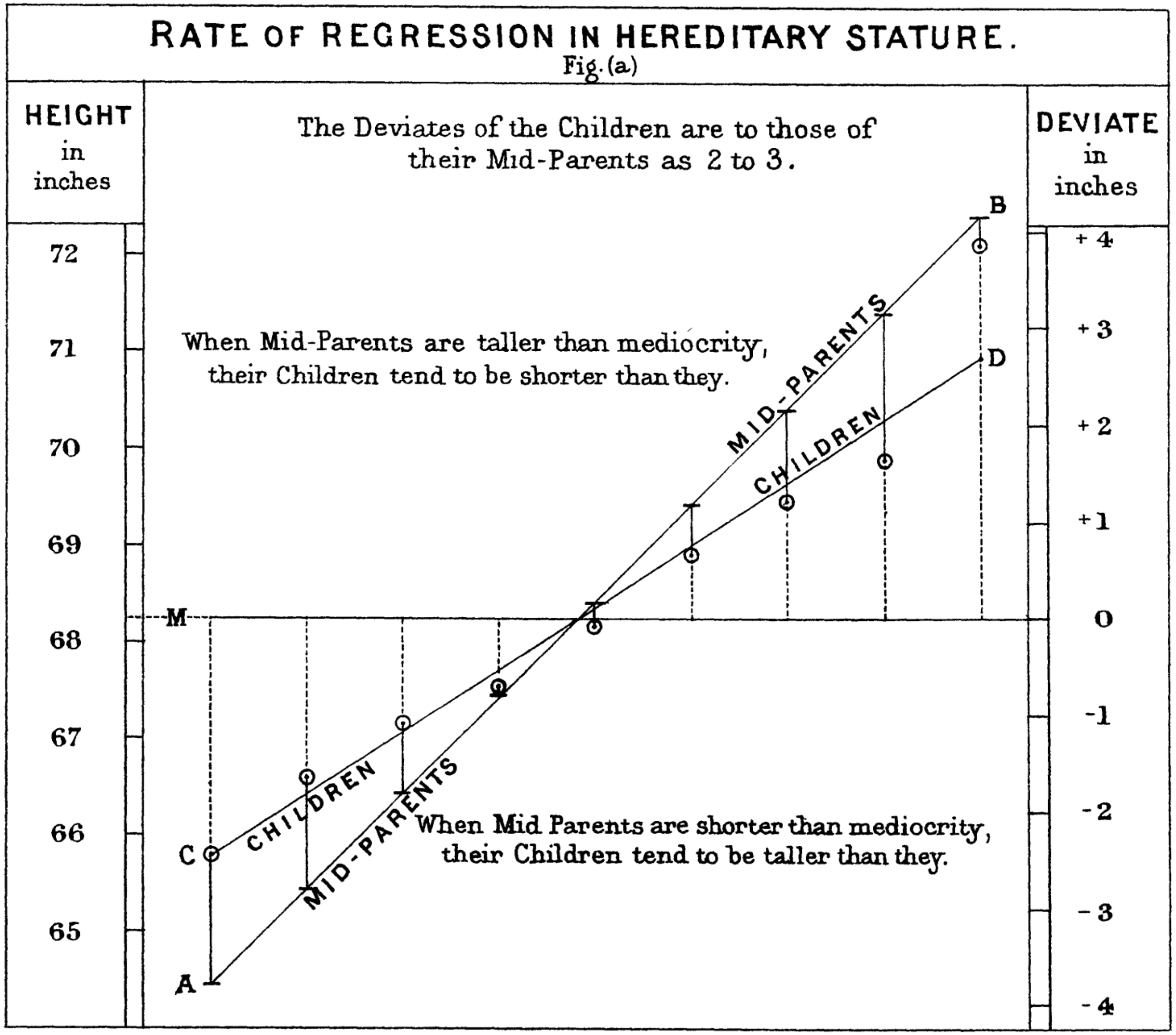
Key concepts

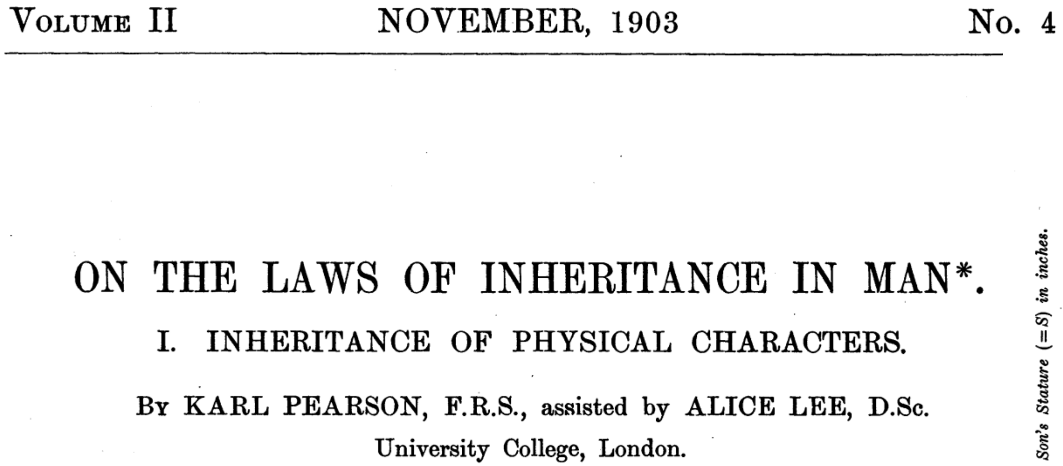
- Dense SNP panels allow the estimation of the expected genetic covariance between distant relatives (‘unrelateds’)
- A model based upon estimated relationships from SNPs is equivalent to a model fitting all SNPs simultaneously
- The total genetic variance due to LD between common SNPs and (unknown) causal variants can be estimated
- Genetic variance captured by common SNPs can be assigned to chromosomes and chromosome segments

ANTHROPOLOGICAL MISCELLANEA.

1886

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.





VOLUME II NOVEMBER, 1903 No. 4

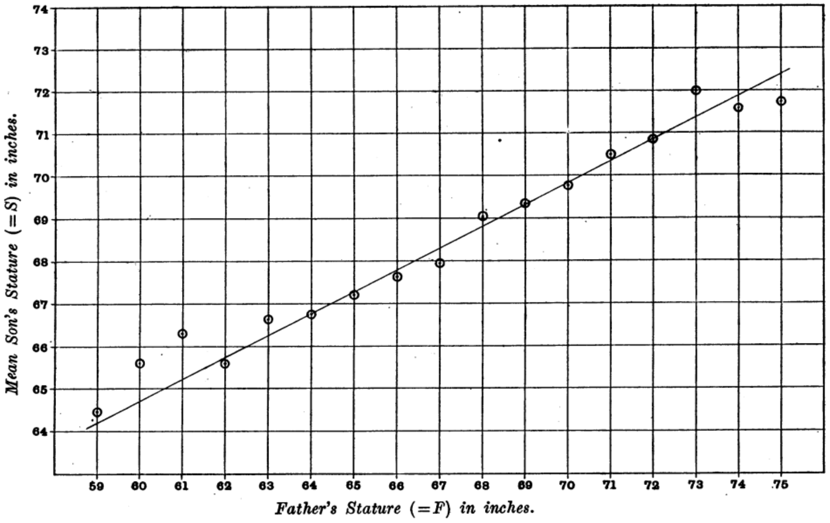
ON THE LAWS OF INHERITANCE IN MAN*.

I. INHERITANCE OF PHYSICAL CHARACTERS.

BY KARL PEARSON, F.R.S., assisted by ALICE LEE, D.Sc.
University College, London.

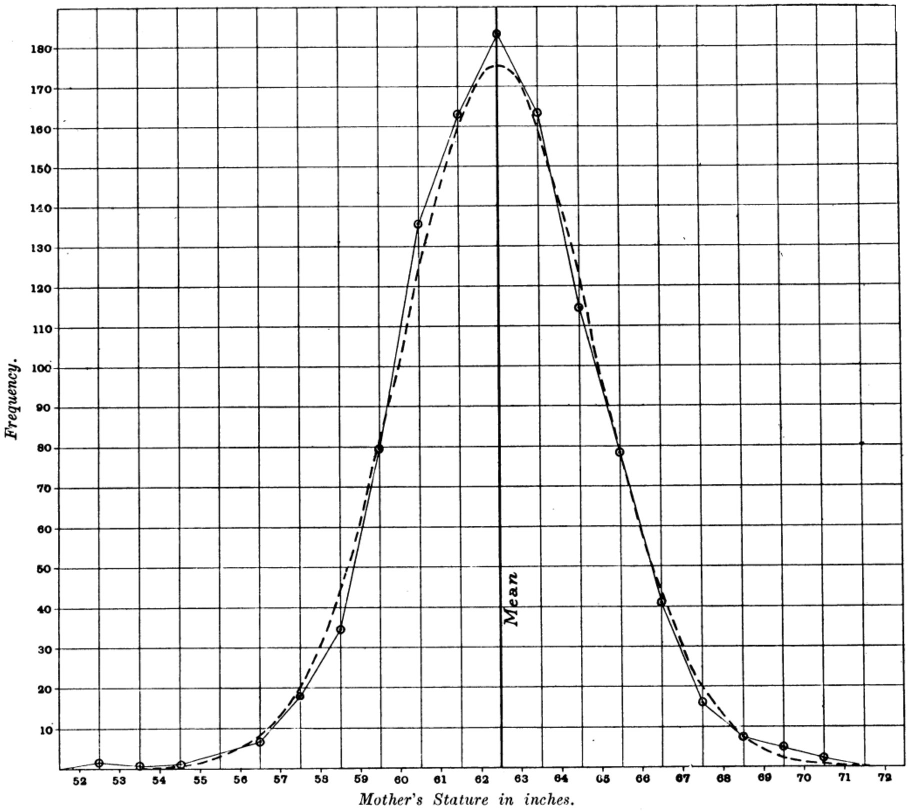
DIAGRAM I. *Probable Stature of Son for given Father's Stature.*

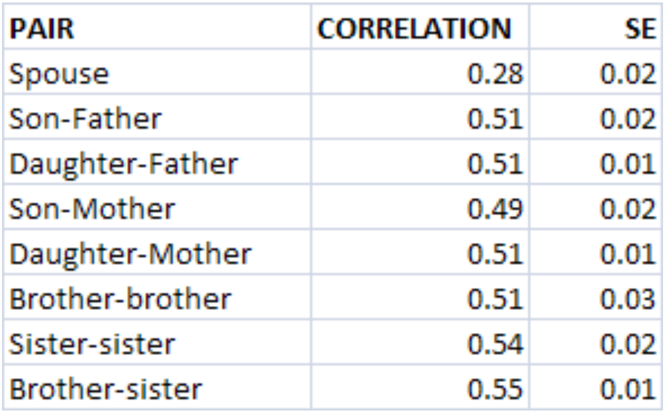
Regression Line: $S = 33{\cdot}73 + {\cdot}516\,F$. 1078 Cases.



364 *On the Laws of Inheritance in Man*

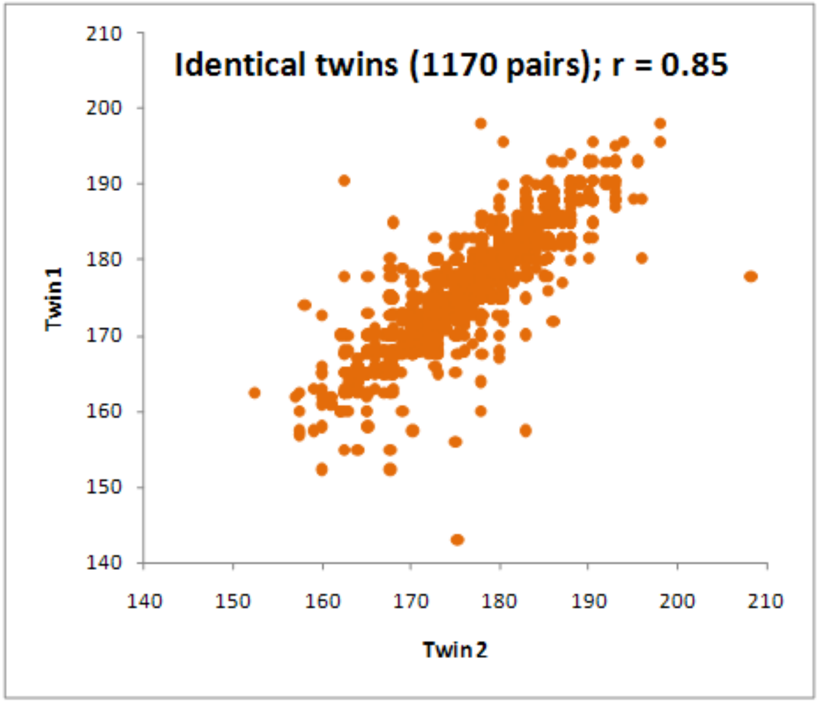
DIAGRAM IV. *Distribution of Stature.*

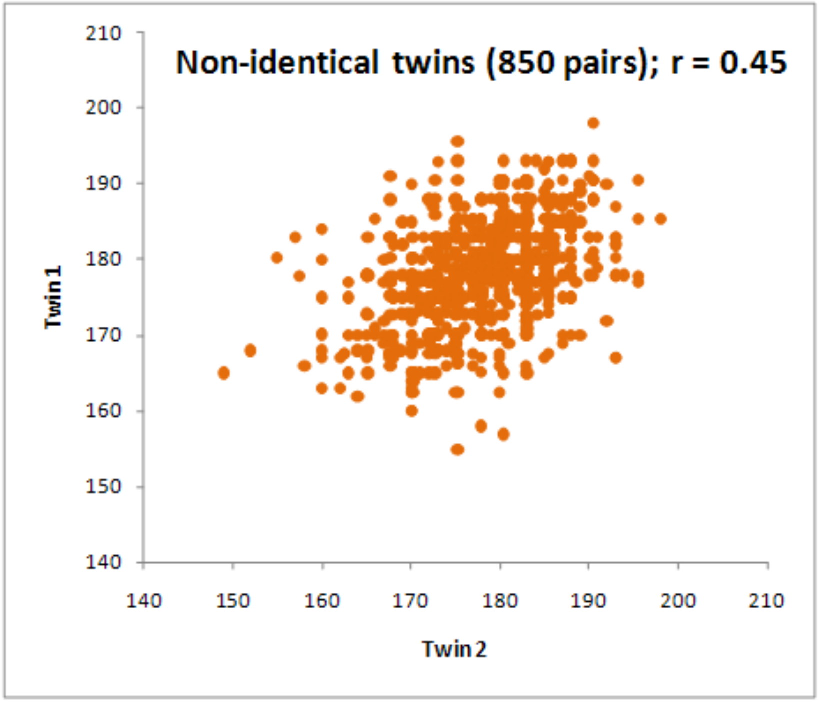




PAIR	CORRELATION	SE
Spouse	0.28	0.02
Son-Father	0.51	0.02
Daughter-Father	0.51	0.01
Son-Mother	0.49	0.02
Daughter-Mother	0.51	0.01
Brother-brother	0.51	0.03
Sister-sister	0.54	0.02
Brother-sister	0.55	0.01

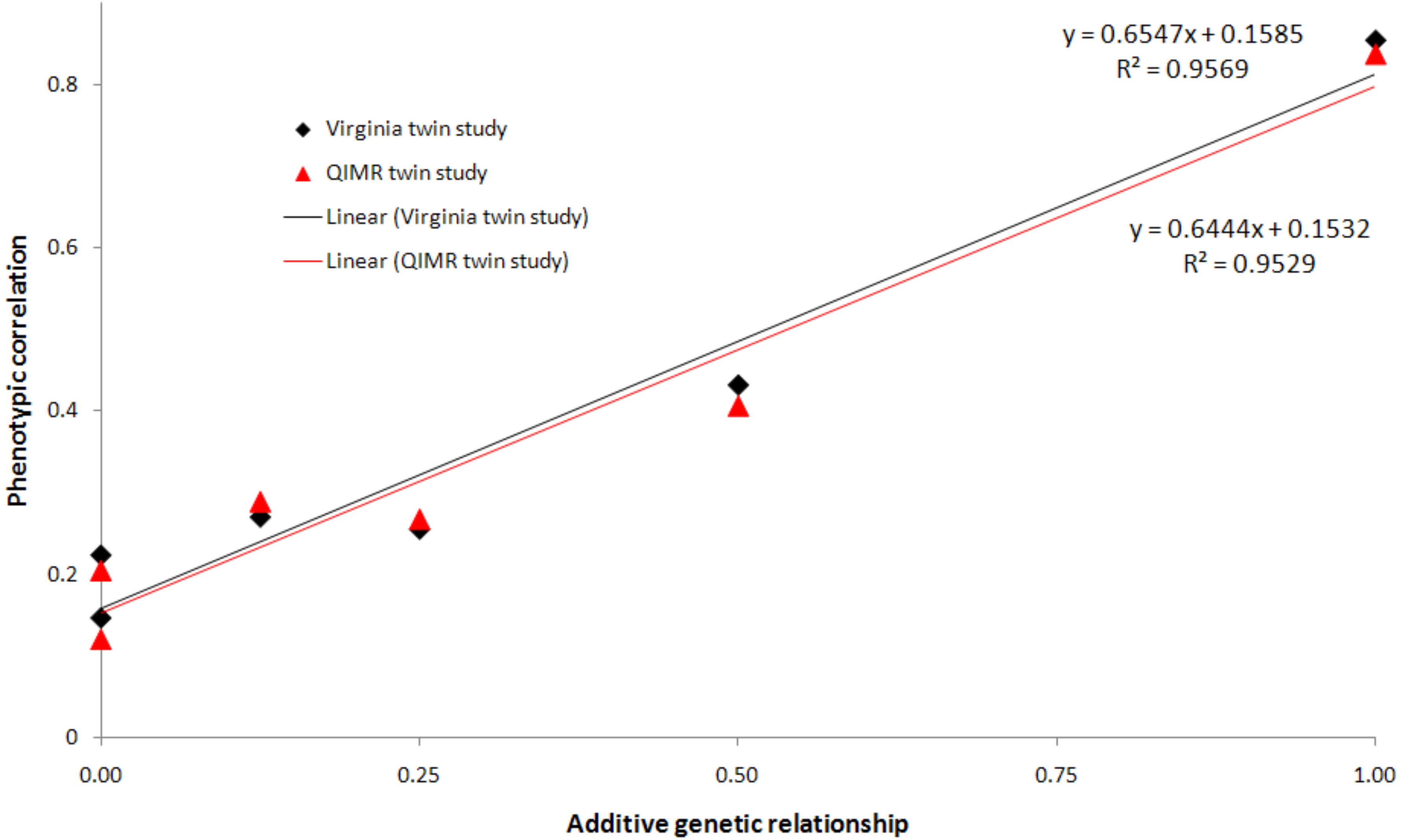
100 years later
Heritability of human height



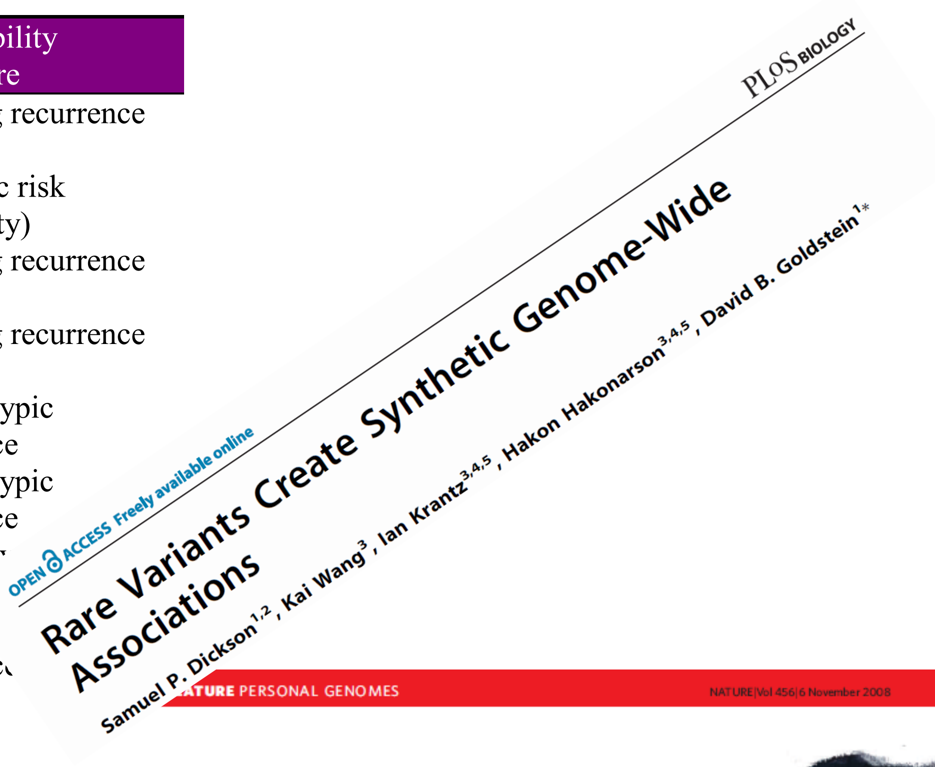


h^2 ~ 80%

Based upon 1000s of twin families



Disease	Number of loci	Percent of Heritability Measure Explained	Heritability Measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn’s disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenot vari
Fasting glucose	4	1.5%	Phenc varianc



PLoS BIOLOGY

OPEN ACCESS Freely available online

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

NATURE PERSONAL GENOMES

NATURE|Vol 456|6 November 2008

Where is the Dark Matter?

Vol 461|8 October 2009|doi:10.1038/nature08494

nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

The case of the missing heritability

Hypothesis testing vs. Estimation

- GWAS = hypothesis testing
  - Stringent p-value threshold
  - Estimates of effects biased ("Winner's Curse")
    - E(bhat| test(bhat) > T) > b {b fixed}
    - var(bhat) = var(b) + var(bhat|b) {b random}

- Can we estimate the total proportion of variation accounted for by all SNPs?

Basic idea

- Estimates of additive genetic variance from known pedigree is unbiased
  - If model is correct
  - Despite variation in identity given the pedigree
  - Pedigree gives correct expected IBD
- Unknown pedigree: estimate genome-wide IBD from marker data
  - Estimate additive genetic variance given this estimate of relatedness
- Idea is not new
  - (Evolutionary) genetics literature (Ritland, Lynch, Hill, others)

Close vs distant relatives

- Detection of close relatives (fullsibs, parent-offspring, halfsibs) from marker data is relatively straightforward
- But close relatives may share environmental factors
  - Biased estimates of genetic variance
- Solution: use only (very) distant relatives

A model for a single causal variant

	AA	AB	BB
frequency	$(1-p)^2$	$2p(1-p)$	p^2
x	0	1	2
effect	0	b	2b
$z = [x-E(x)]/\sigma_x$	$-2p/\sqrt{2p(1-p)}$	$(1-p)/ \sqrt{2p(1-p)}$	$2(1-p)/ \sqrt{2p(1-p)}$

$y_j = \mu' + x_{ij}b_i + e_j$ x = 0, 1, 2 {standard association model}

$y_j = \mu + z_{ij}u_j + e_j$ $u = b\sigma_x$; $\mu = \mu' + b\sigma_x$

Multiple (m) causal variants

$$y_j = \quad \mu + \Sigma z_{ij} u_j + e_j$$

$$= \quad \mu + g_j + e_j$$

$$\mathbf{y} = \quad \mu\mathbf{1} + \mathbf{g} + \mathbf{e}$$

$$\mathbf{=} \quad \mu\mathbf{1} + \mathbf{Zu} + \mathbf{e}$$

Equivalence

Let u be a random variable, u ~ N(0, σ_u^2)

Then $\sigma_g^2 = m\sigma_u^2$ and

$$\begin{aligned} \mathrm{var}(\mathbf{y}) &= \mathbf{ZZ'}\sigma_u^2 + \mathbf{I}\sigma_e^2 \\ &= \mathbf{ZZ'}(\sigma_g^2/m) + \mathbf{I}\sigma_e^2 \\ &= \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2 \end{aligned}$$

Model with individual genome-wide additive values using relationships (**G**) at the causal variants is equivalent to a model fitting all causal variants

We can estimate genetic variance just as if we would do using pedigree relationships

But we don’t have the causal variants

If we estimate **G** from SNPs:

- lose information due to imperfect LD between SNPs and causal variants
- how much we lose depends on
  - density of SNPs
  - allele frequency spectrum of SNPs vs. causal variants
- estimate of variance → missing heritability

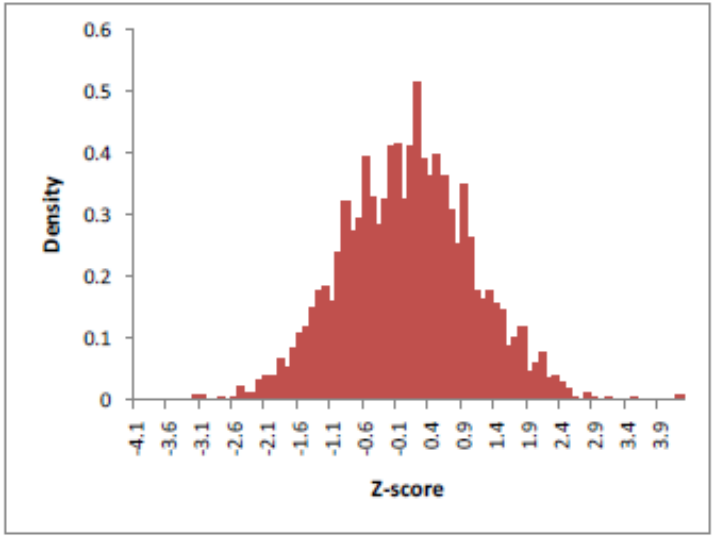
Let **A** be the estimate of **G** from N SNPs:

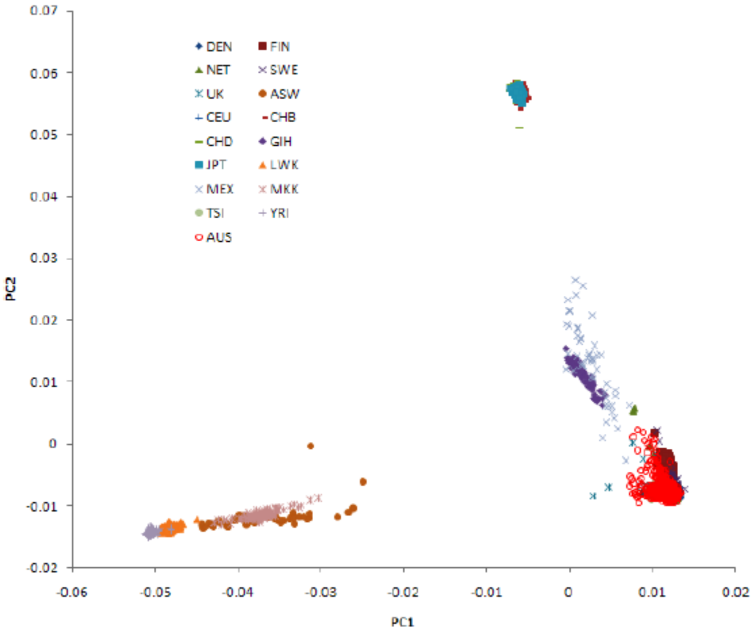
$$A_{jk} = (1/N) \Sigma \{ x_{ij} – 2p_i)(x_{ik} – 2p_i) / \{2p_i(1-p_i)\}$$

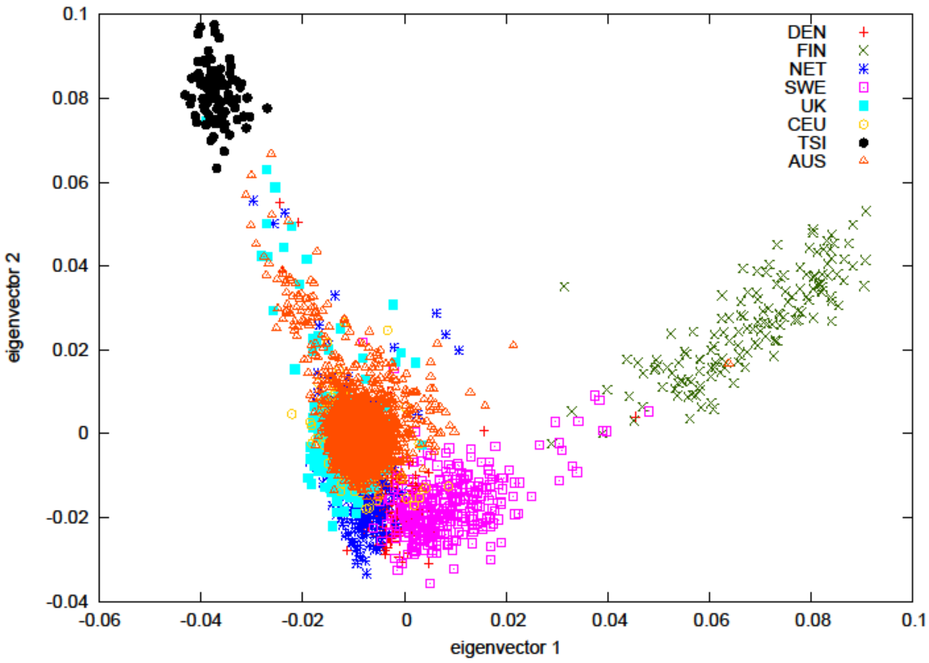
$$= (1/N) \Sigma z_{ij}z_{ik}$$

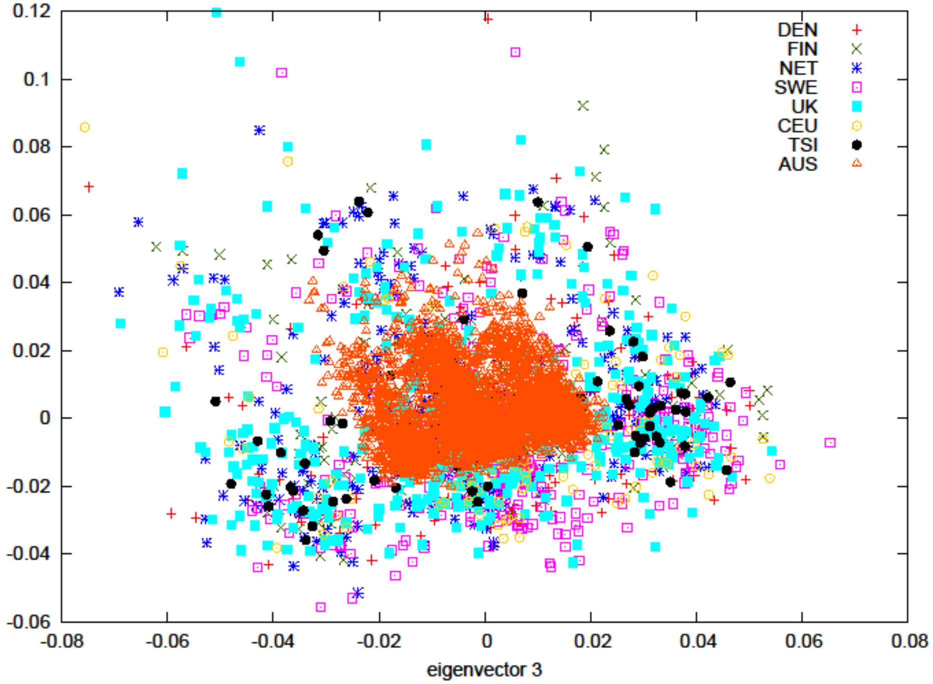
Data

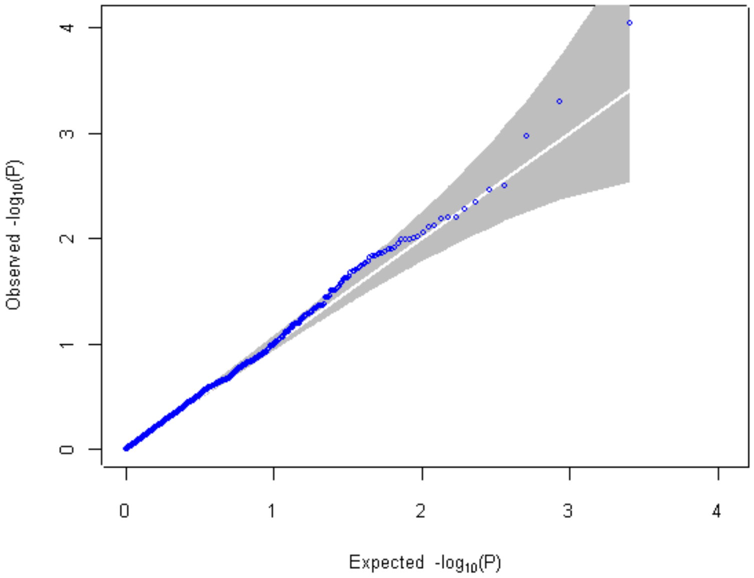
- ~4000 ‘unrelated’ individuals
- Ancestry ~British Isles
- Measurement on height (self-report or clinically measured)
- GWAS on 300k (‘adults’) or 600k (16-year olds) SNPs











Lack of evidence for population stratification within the Australian sample

Methods

- Estimate realised relationship matrix from SNPs $y_i = g_i + e_i$ $\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$
- Estimate additive genetic variance

$$A_{ijk} = \frac{\mathrm{cov}(x_{ij}a_i, x_{ik}a_i)}{\sqrt{\mathrm{var}(x_{ij}a_i)\,\mathrm{var}(x_{ik}a_i)}} = \frac{\mathrm{cov}(x_{ij}, x_{ik})}{2p_i(1-p_i)}$$

Base population = current population

$$A_{jk} = \frac{1}{N}\sum_i A_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, \; j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ij}^2 - (1+2p_i)x_{ij} + 2p_i^2}{2p_i(1-p_i)}, \; j = k \end{cases}$$

Statistical analysis

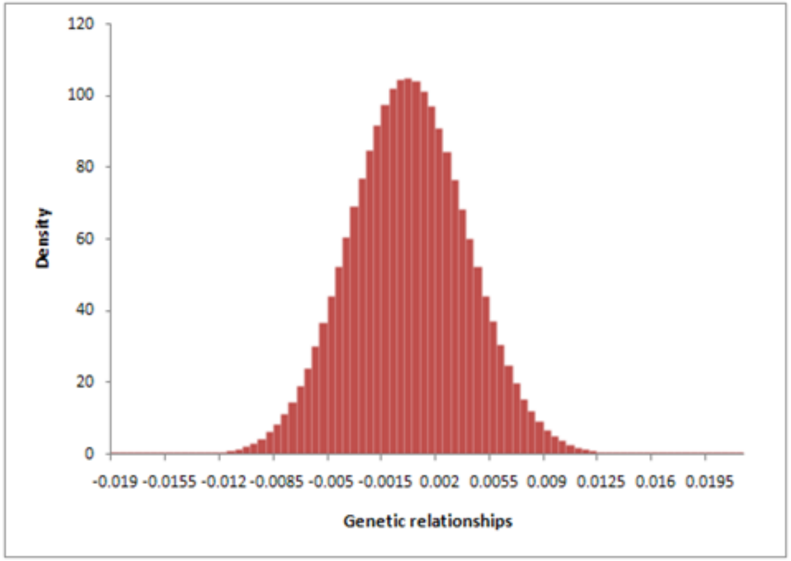
$$\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$$

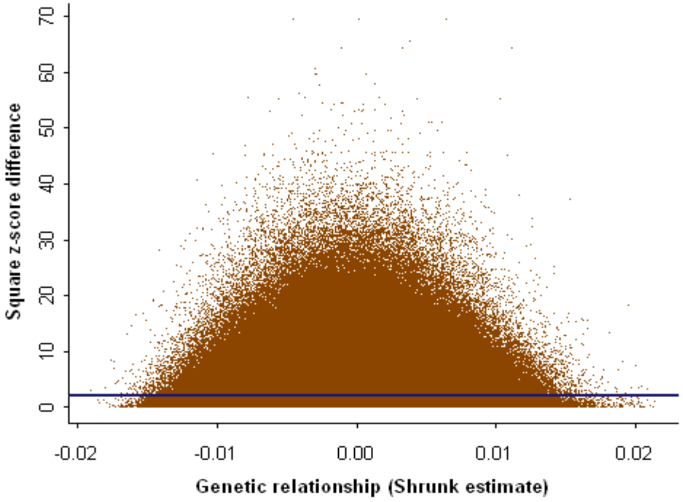
y standardised ~N(0,1)

No fixed effects other than mean

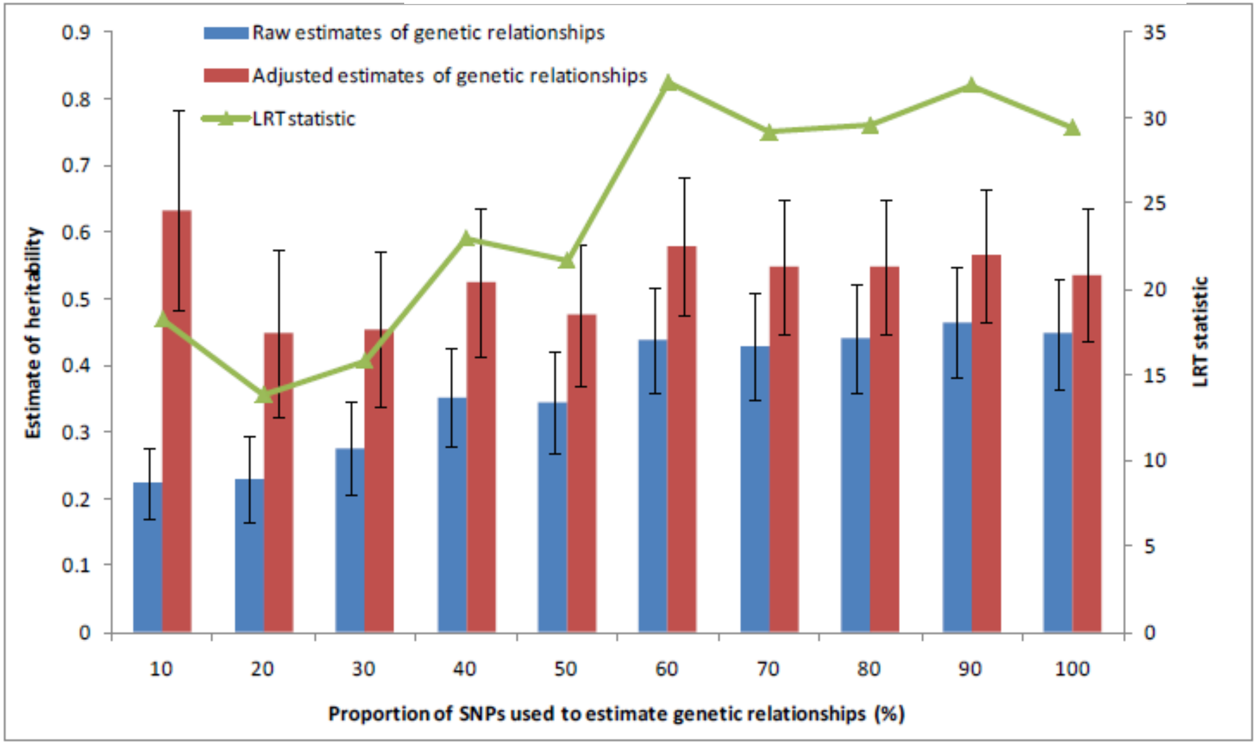
A estimated from SNPs

Residual maximum likelihood (REML)


Density
120
100
80
60
40
20
0
-0.019 -0.0155 -0.012 -0.0085 -0.005 -0.0015 0.002 0.0055 0.009 0.0125 0.016 0.0195
Genetic relationships


Square z-score difference
70
60
50
40
30
20
10
0
-0.02
-0.01
0.00
0.01
0.02
Genetic relationship (Shrunk estimate)

h^2 ~ 0.5 (SE 0.1)


Raw estimates of genetic relationships
Adjusted estimates of genetic relationships
LRT statistic
Estimate of heritability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
LRT statistic
35
30
25
20
15
10
5
0
10
20
30
40
50
60
70
80
90
100
Proportion of SNPs used to estimate genetic relationships (%)

Partitioning variation

- If we can estimate the variance captured by SNPs genome-wide, we should be able to partition it and attribute variance to regions of the genome
- “Population based linkage analysis”

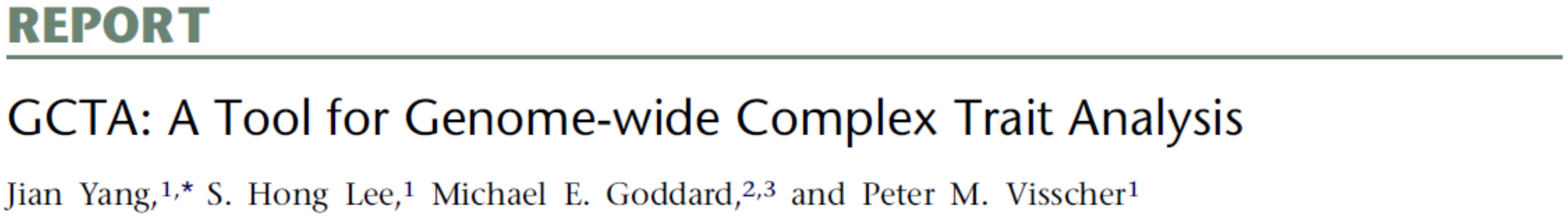
Genome partitioning

- Partition additive genetic variance according to groups of SNPs
  - Chromosomes
  - Chromosome segments
  - MAF bins
  - Genic vs non-genic regions
  - Etc.

- Estimate genetic relationship matrix from SNP groups

- Analyse phenotypes by fitting multiple relationship matrices

- Linear model & REML (restricted maximum likelihood)



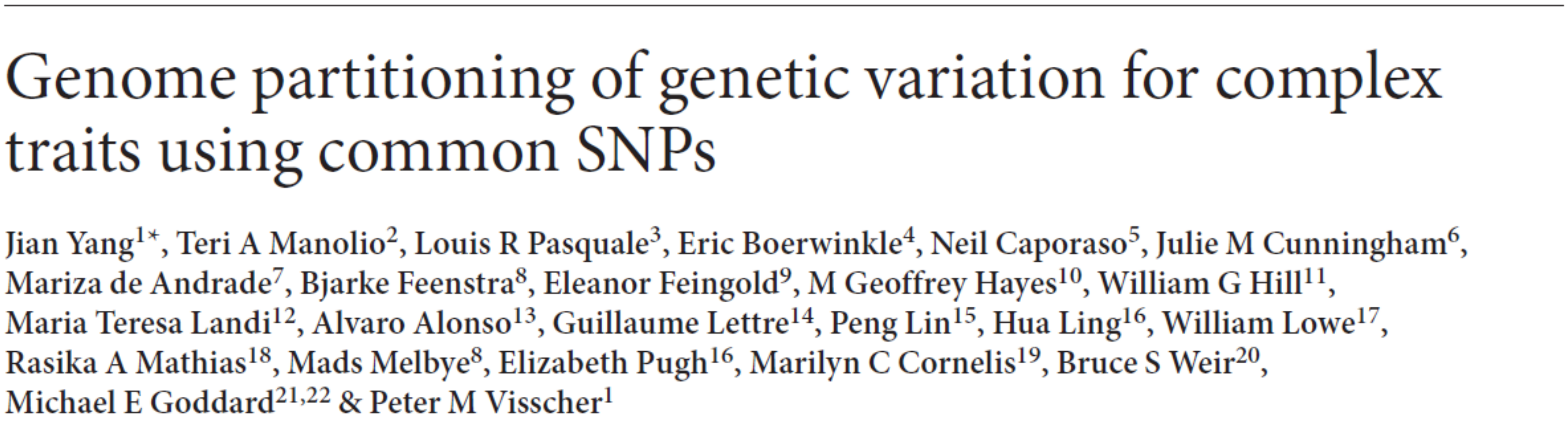
REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

Application: the GENEVA Consortium

- Data
  - ~14,000 European Americans
    - ARIC
    - NHS
    - HPFS
  - Affy 6.0 genotype data
    - ~600,000 after stringent QC
  - Phenotypes on height, BMI, vWF and QT Interval



Genome partitioning of genetic variation for complex traits using common SNPs

Jian Yang[1]*, Teri A Manolio[2], Louis R Pasquale[3], Eric Boerwinkle[4], Neil Caporaso[5], Julie M Cunningham[6], Mariza de Andrade[7], Bjarke Feenstra[8], Eleanor Feingold[9], M Geoffrey Hayes[10], William G Hill[11], Maria Teresa Landi[12], Alvaro Alonso[13], Guillaume Lettre[14], Peng Lin[15], Hua Ling[16], William Lowe[17], Rasika A Mathias[18], Mads Melbye[8], Elizabeth Pugh[16], Marilyn C Cornelis[19], Bruce S Weir[20], Michael E Goddard[21,22] & Peter M Visscher[1]

QC of SNPs

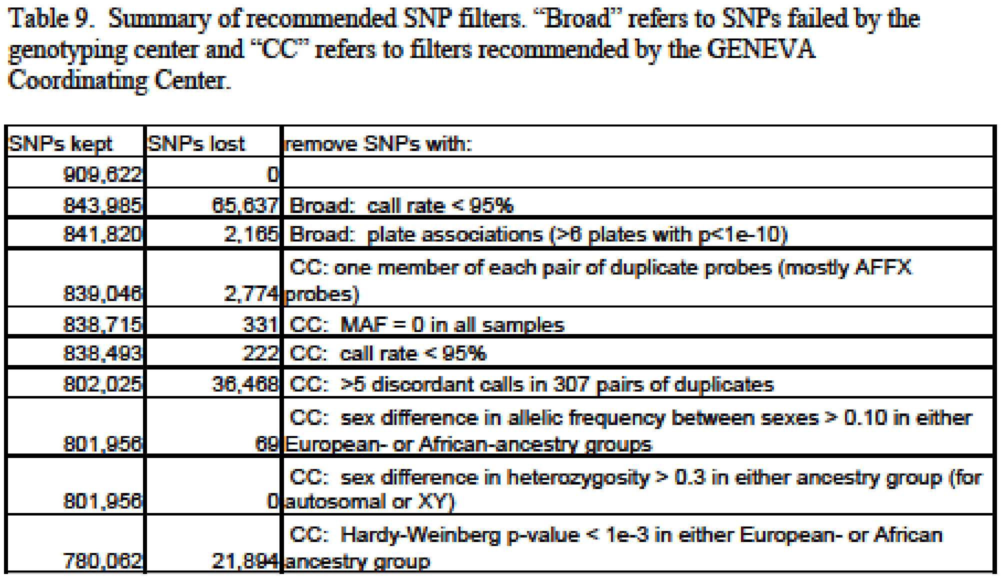


Table 9. Summary of recommended SNP filters. "Broad" refers to SNPs failed by the genotyping center and "CC" refers to filters recommended by the GENEVA Coordinating Center.

SNPs kept	SNPs lost	remove SNPs with:
909,622	0	
843,985	65,637	Broad: call rate < 95%
841,820	2,165	Broad: plate associations (>6 plates with p<1e-10)
839,046	2,774	CC: one member of each pair of duplicate probes (mostly AFFX probes)
838,715	331	CC: MAF = 0 in all samples
838,493	222	CC: call rate < 95%
802,025	36,468	CC: >5 discordant calls in 307 pairs of duplicates
801,956	69	CC: sex difference in allelic frequency between sexes > 0.10 in either European- or African-ancestry groups
801,956	0	CC: sex difference in heterozygosity > 0.3 in either ancestry group (for autosomal or XY)
780,062	21,894	CC: Hardy-Weinberg p-value < 1e-3 in either European- or African ancestry group

- 780,062 SNPs after QC steps listed in the table.
- Exclude 141,772 SNPs with MAF < 0.02 in European-ancestry group.
- Exclude 36,949 SNPs with missingness > 2% in all samples.
- Include autosomal SNPs only.
- End up with 577,778 SNPs.

Results (genome-wide)

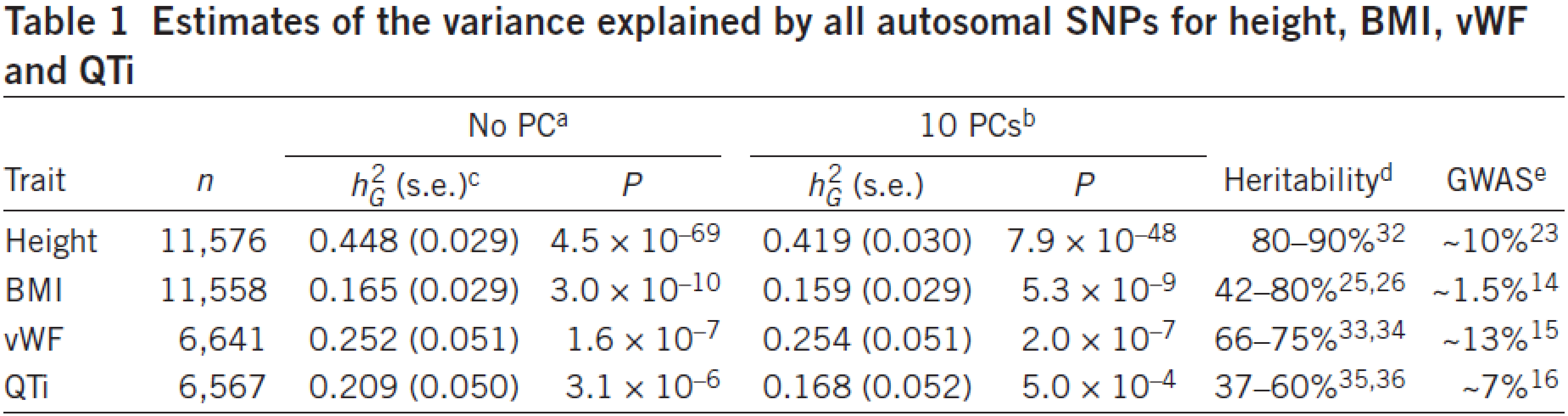
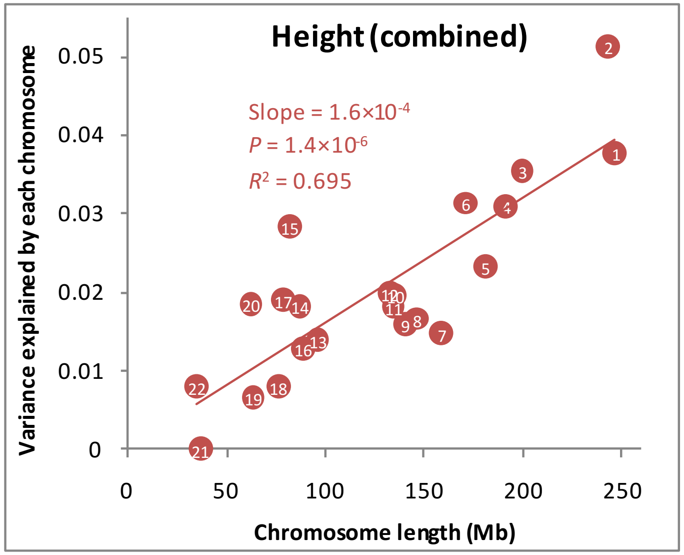
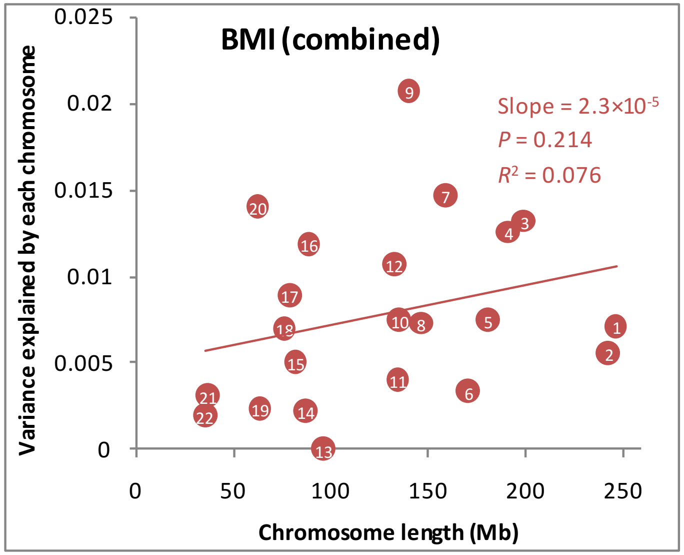


Table 1 Estimates of the variance explained by all autosomal SNPs for height, BMI, vWF and QTi

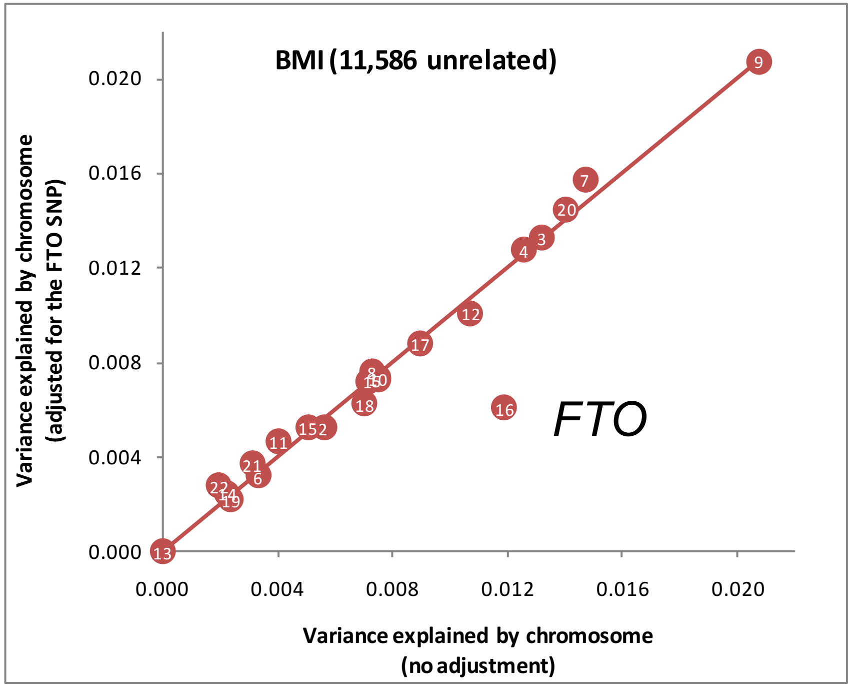
		No PC[a]		10 PCs[b]			
Trait	*n*	h^2_G (s.e.)[c]	*P*	h^2_G (s.e.)	*P*	Heritability[d]	GWAS[e]
Height	11,576	0.448 (0.029)	4.5×10^{-69}	0.419 (0.030)	7.9×10^{-48}	80–90%[32]	~10%[23]
BMI	11,558	0.165 (0.029)	3.0×10^{-10}	0.159 (0.029)	5.3×10^{-9}	42–80%[25,26]	~1.5%[14]
vWF	6,641	0.252 (0.051)	1.6×10^{-7}	0.254 (0.051)	2.0×10^{-7}	66–75%[33,34]	~13%[15]
QTi	6,567	0.209 (0.050)	3.1×10^{-6}	0.168 (0.052)	5.0×10^{-4}	37–60%[35,36]	~7%[16]

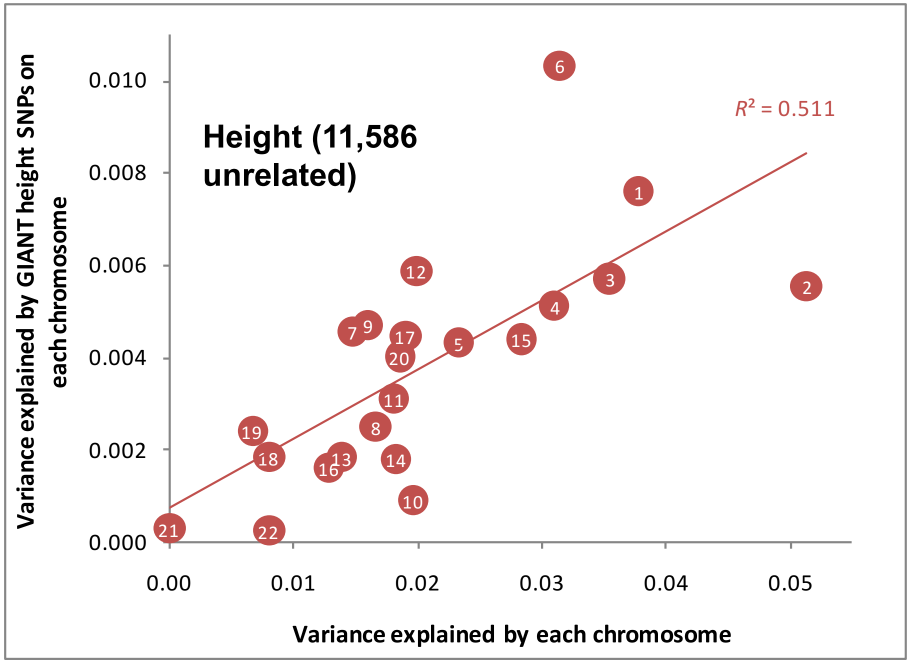
Genome-partitioning: longer chromosomes explain more variation



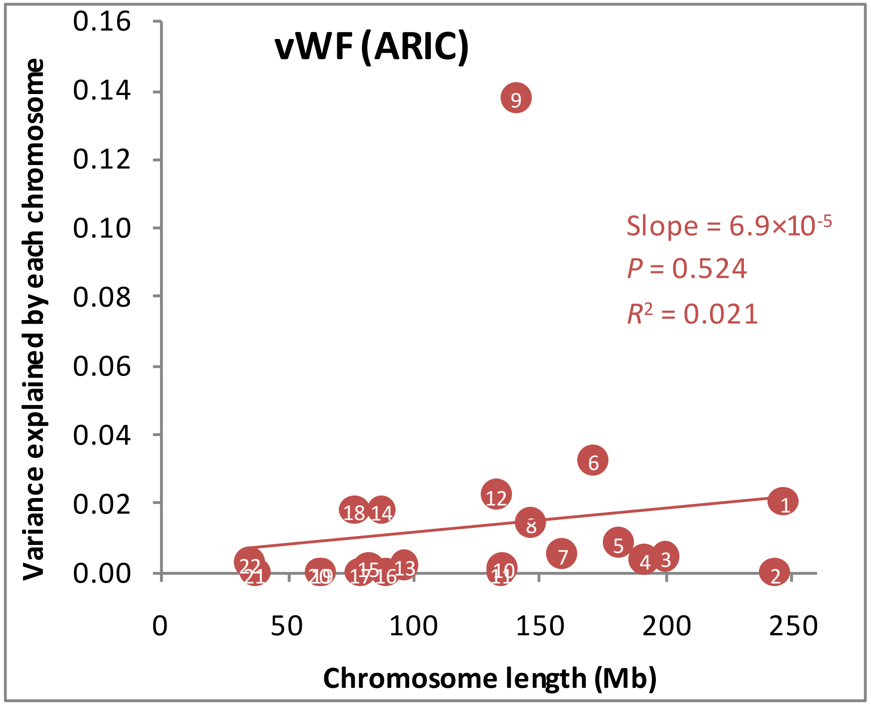


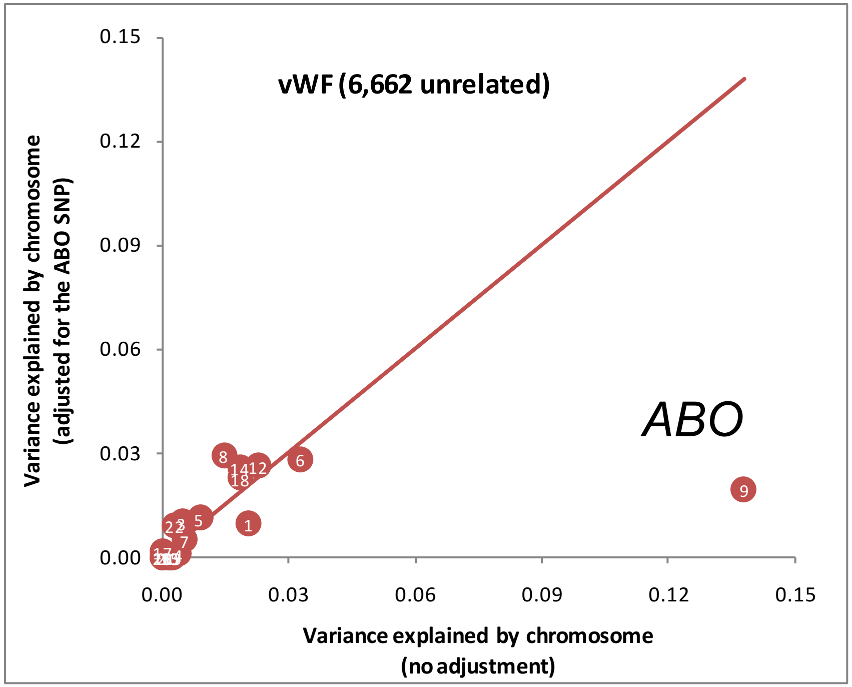
Results are consistent with reported GWAS



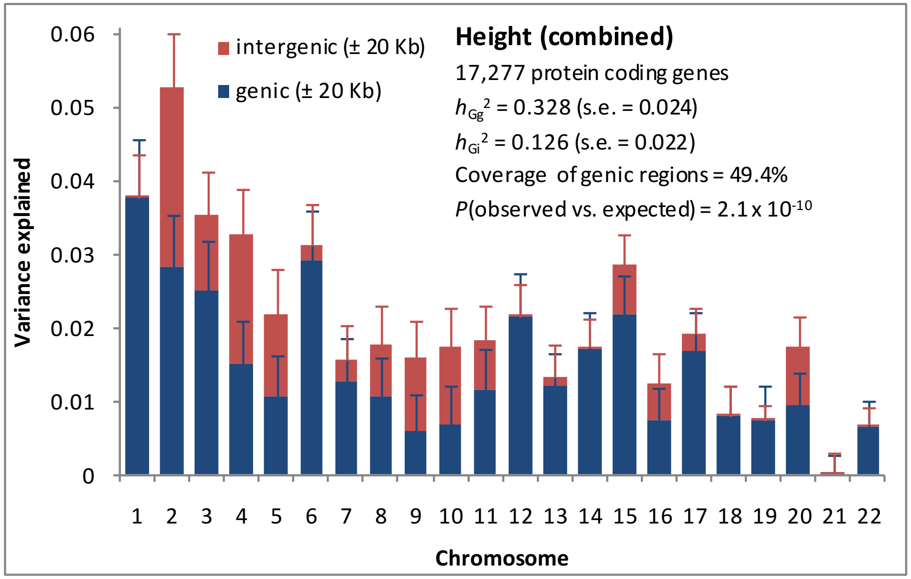


Inference robust with respect to genetic architecture





Genic regions explain variation disproportionately



Height (combined)

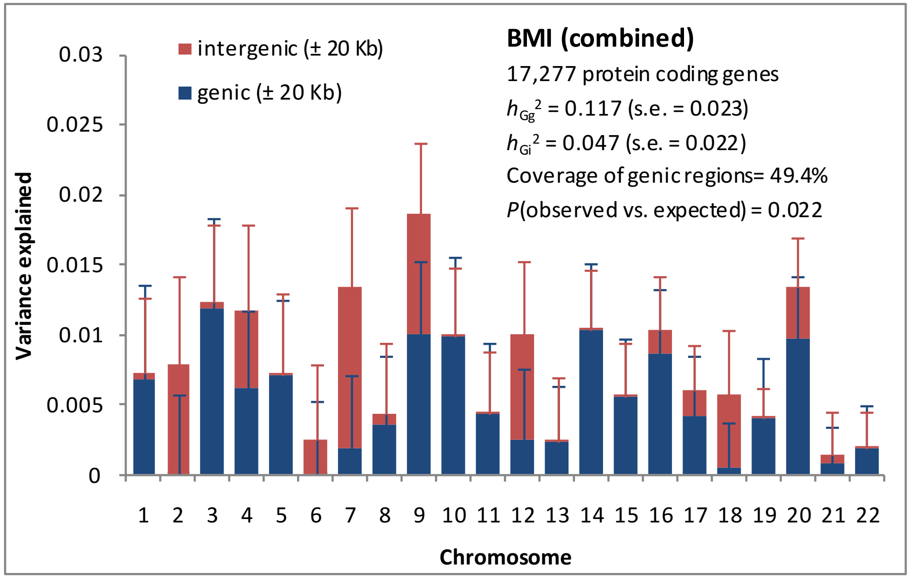
17,277 protein coding genes

h_{Gg}^2 = 0.328 (s.e. = 0.024)

h_{Gi}^2 = 0.126 (s.e. = 0.022)

Coverage of genic regions = 49.4%

P(observed vs. expected) = 2.1 x 10^{-10}



BMI (combined)

17,277 protein coding genes

h_{Gg}^2 = 0.117 (s.e. = 0.023)

h_{Gi}^2 = 0.047 (s.e. = 0.022)

Coverage of genic regions= 49.4%

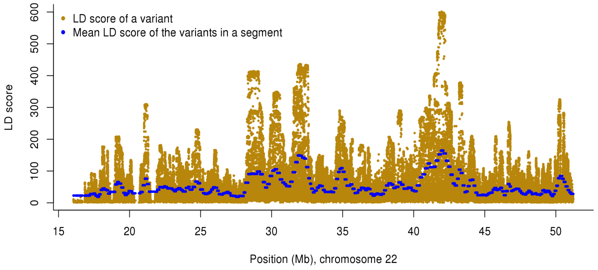
P(observed vs. expected) = 0.022

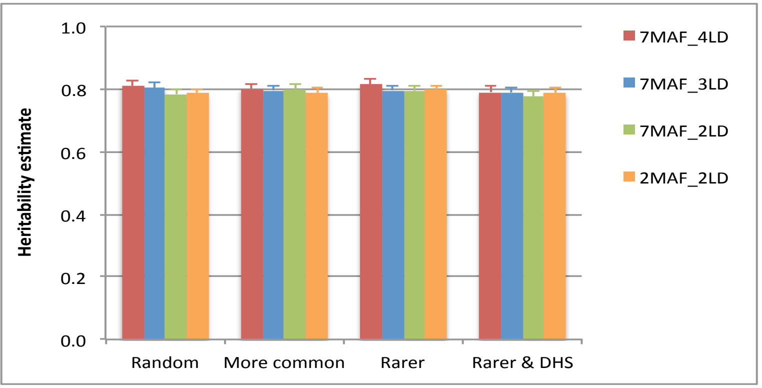
Using imputed sequence data

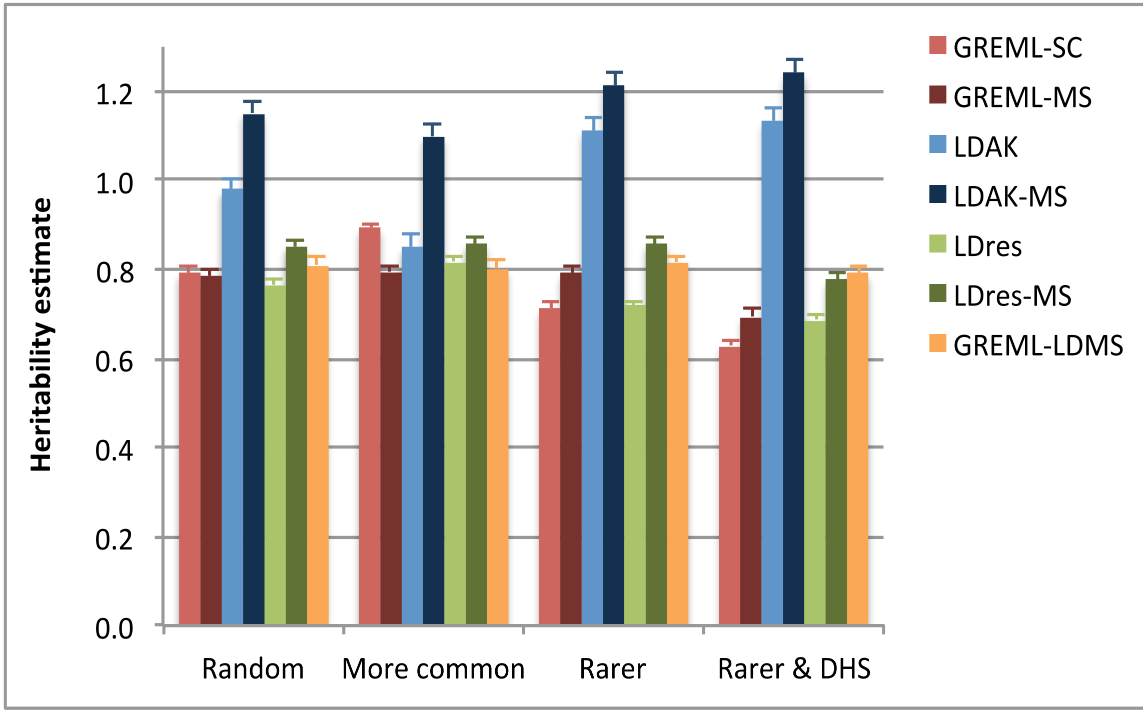
- How much information is gained by using SNP array data imputed to a fully sequenced reference?
- How much is lost relative to whole genome sequencing?

Yang et al. 2015 (Nature Genetics)

Accounting for LD and MAF spectrum allows unbiased estimation of genetic variance (GREML-LDMS)

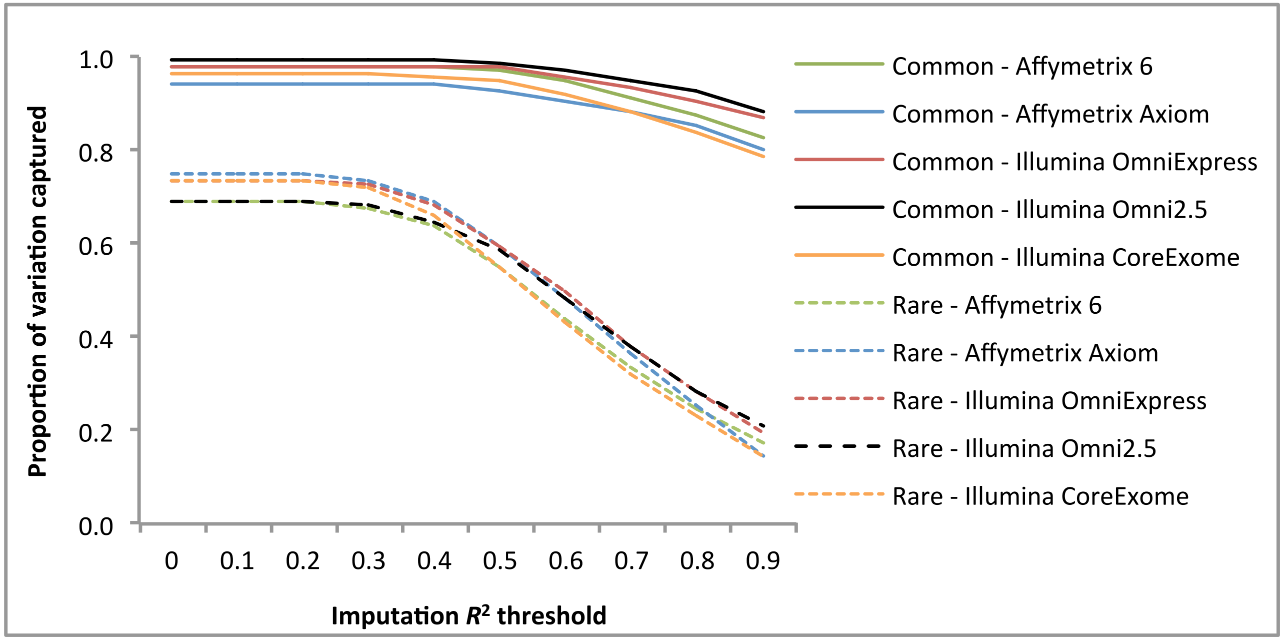






Yang et al. 2015 (Nature Genetics)

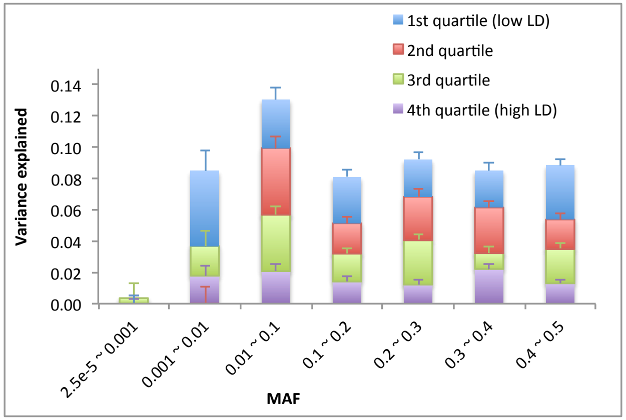
Very little difference in "taggability" between SNP chips

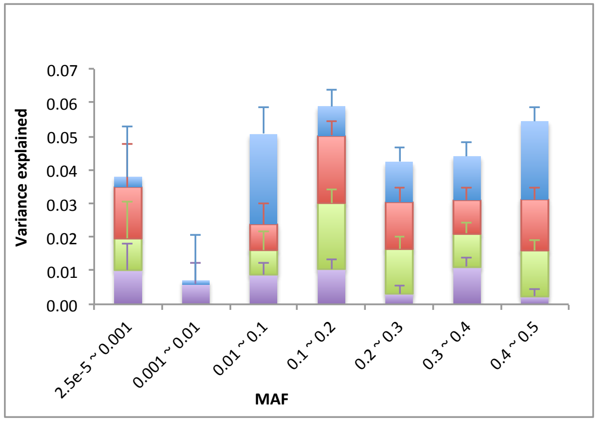


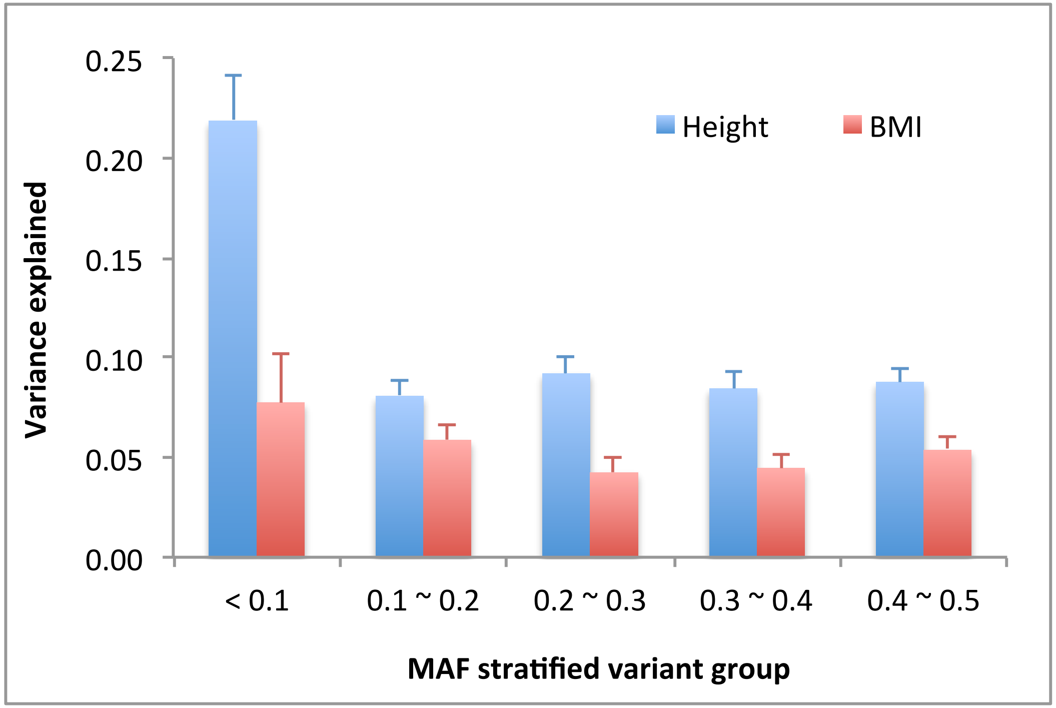
Genetic variation captured after imputation
96% due to common variants
73% due to rare variants

Yang et al. 2015 (Nature Genetics)

n = 45k data on height and BMI



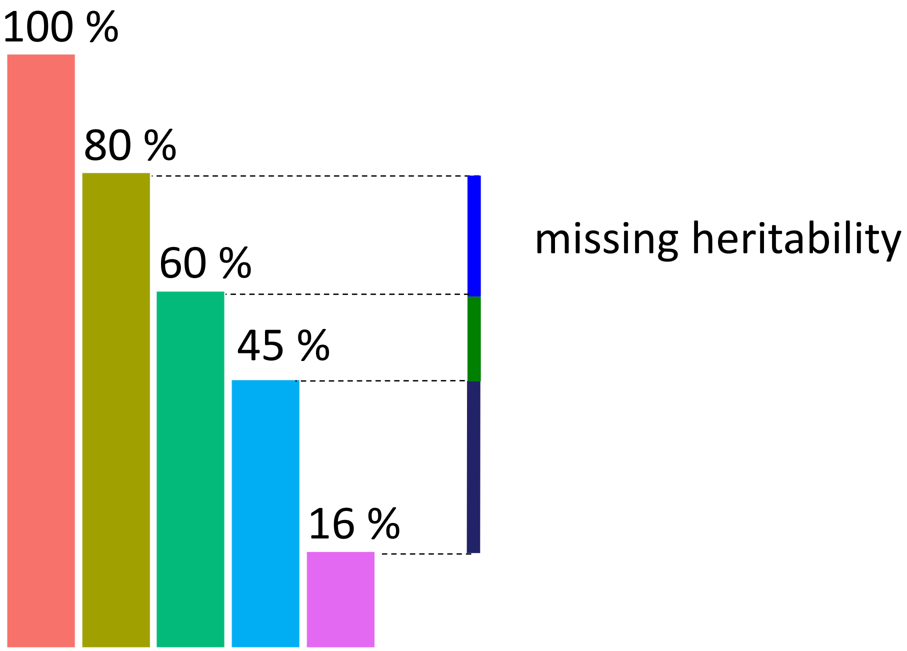




Totals
~60% for height
~30% for BMI

Yang et al. 2015 (Nature Genetics)

Partitioning variance of height



h^2 overestimation?
untagged rare variants?

better tagging of
ungenotyped variants

sample size / power

🟥 Total variance
🟩 Heritability (based on Twin or family studies)
🟩 SNP heritability from imputation to sequenced reference
🟦 SNP-heritability (variance explained by all genotyped SNPs on the Chip)
🟪 Variance explained by genome wide significant SNPs

Slide by Robert Maier

Scaling revisited

$u = b\sigma_x \sim N(0, \sigma_u^2)$ implies

b^2 proportional to $\sigma_u^2/[2p(1-p)]$, so rare variants have larger allelic effect: natural selection

If $b^2 = \sigma_u^2$ then no relationship between frequency and effect size: neutral model

In between: $b^2 = \sigma_u^2\ [2p(1-p)]^{-s}$

Variance explained by SNP: $2p(1-p)\sigma_u^2[2p(1-p)]^{-s}$
$= \sigma_u^2\ [2p(1-p)]^{1-s}$

s = 0: common SNPs explain more variation

s = 1: all SNPs explain the same amount of variation

Multiple methods to estimate additive genetic variance

- Individual-level data
  - GREML
  - Haseman-Elston regression

  $(y_j y_j) = \mu + \beta A_{ij}$
- Summary data

  LDscore regression
- Consideration:
  - data availability
  - model assumptions
  - computation

Key concepts

- Dense SNP panels allow the estimation of the expected genetic covariance between distant relatives (‘unrelateds’)
- A model based upon estimated relationships from SNPs is equivalent to a model fitting all SNPs simultaneously
- The total genetic variance due to LD between common SNPs and (unknown) causal variants can be estimated
- Genetic variance captured by common SNPs can be assigned to chromosomes and chromosome segments