Resemblance between relatives

Mike Goddard

What do we mean by resemble?

Similar values of quantitative traits

Measure by correlation

$$= \text{Covariance}(y_i, y_j) / \text{variance}(y)$$

Why do relatives resemble each other?

Why do relatives resemble each other?

Similar

- Genes
- Family environment
- Country
- School

Model phenotype

Phenotype = genetic effect

+ country

+ year of birth

+ family environment

Fixed effects

Country, year of birth

Random effects

Genetic effect, family environment

We need a model of the covariances between terms

Model phenotype

Phenotype = genetic effect
+ country
+ year of birth
+ family environment
+ individual environment

V(phenotype) = V(genetic effects) + V(family environment)
+ V(individual environment)

Cov(phenotype_i, phenotype_j) = Cov(genetic effects)
+ Cov(family environments)

Model phenotype

Random effects

Genetic effect, family environment

We need a model of the covariances between terms

C(family environments) = 0 if different families

1 * V_{CE} if same family

Covariance between genetic effects of relatives

Model with 1 gene, 2 alleles and additive gene action

We need genetic variances and covariances

Genotype	BB	Bb	bb	
Effect	a	0	-a	
Frequency	p^2	2pq	q^2	(p+q=1)

Mean = $a*p^2 + 0*2pq - a*q^2 = (p-q)*a$

Variance (genetic effect) = genetic variance = V_G

$= E(effect^2) - E(effect)^2$

$V_G = a^2*p^2 + 0*2pq + a^2*q^2 - [(p-q)*a]^2 = 2pqa^2$

Model with 1 gene, 2 alleles and additive gene action

Covariance between parent and offspring

Parent			Offspring			
Genotype		frequency	BB	Bb	bb	mean
BB	a	p^2	p	q		pa
Bb	0	2pq	0.5p	0.5	0.5q	0.5(p-q)a
bb	-a	q^2		p	q	-qa

Cov(parent genetic value, offspring genetic value)

$= p^2 * a * pa + q^2 * (-a) * (-qa) - [(p-q)a] * [(p-q)a] = pqa^2 = 0.5\ V_G$

Model with 1 gene, 2 alleles and additive gene action

Covariance between parent and offspring (another way)

Model genetic value as sum of gametic effects from mother and father

$g = x_m + x_f$

$V(g) = V(x_m) + V(x_f) = 2V(x)$

$C(g_p, g_o) = C(x_{mp} + x_{fp}, x_{mo} + x_{fo})$

$= C(x_{mp}, x_{mo}) + C(x_{mp}, x_{fo}) + C(x_{fp}, x_{mo}) + C(x_{fp}, x_{fo})$

$= 0 + ? + 0 + ?$

$C(x_{mp}, x_{fo})$ $= V(x)$ if x_{mp} is ibd to x_{fo}

$=0$ otherwise

$C(x_{mp}, x_{fo}) = C(x_{fp}, x_{fo}) = 0.5\ V(x)$

$C(g_p, g_o) = 0 + 0.5V(x) + 0 + 0.5V(x) = V(x) = 0.5\ V_G$

Probability that relatives share alleles IBD

Covariance between relatives depends on probability that their alleles are IBD

This probability can be calculated from pedigrees

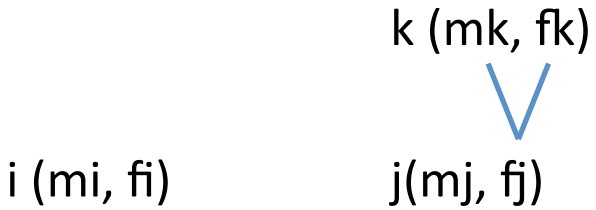
Assume that base individuals at the top of the pedigree (ie those without a pedigree) have unrelated alleles ie the individuals are unrelated

Recurrence formulae for P(IBD)

if i and j are base individuals, $P(x_{.i} \equiv x_{.j}) = 0$

Otherwise, $P(x_{.i} \equiv x_{fj}) = 0.5\ [P(x_{.i} \equiv x_{fk}) + P(x_{.i} \equiv x_{mk})\ ]$ where k is the father of j

Probability that relatives share alleles IBD



Relationships between individuals

P(gametes are IBD) can be stored in a gametic relationship matrix

G(wi,zj) = P(wi≡zj)

But usually we analyse measurements on diploid individuals

$C(g_i, g_j) = A(i,j)\, V_G = [G(mi, mj) + G(mi,fj) + G(fi,mj) + G(fi,fj)]\, V(x)$

$= [G(mi, mj) + G(mi,fj) + G(fi,mj) + G(fi,fj)]\, V_G/2$

A(i,j) = [G(mi, mj) + G(mi,fj) + G(fi,mj) + G(fi,fj)]/2

where A is the numerator relationship matrix

Relationships between individuals

Example: Relationship of individual with herself

Gametic relationship matrix

	mi	fi
mi	1	0
fi	0	1

Numerator relationship A(i,i) = [1+0+0+1]/2 = 1

Relationships between individuals

Example: Relationship of sisters

Gametic relationship matrix

	mi	fi
mj	0.5	0
fj	0	0.5

Numerator relationship A(i,j) = [0.5+0+0+0.5]/2 = 0.5

Relationships between individuals

i = (im, if) and j = (jm,jf)

Co-ancestry of i and j

= Inbreeding co-efficient of an offspring of i and j

= Prob(offspring gets two alleles that are IBD)

= (P(im≡jm) + P(im≡jf) + P(if≡jm) + P(if≡jf))/4

= A(i,j) /2

Additive relationship (NRM) = 2 * co-ancestry

= 2 * kinship

Estimating genetic variance

Data on phenotypes (y) of related subjects

y = fixed effects + g + e

$V(g) = A V_G$

$V(e) = I V_E$

Use ML or REML to estimate variances

Estimating genetic variance

Use ML or REML to estimate variances

- ML finds the value of V_G that maximises the probability of observing the data
- ML estimates all parameters together
  - = estimates variances assuming that fixed effects have been estimated without error
- REML allows for loss of df in estimating fixed effects
- ML $\sigma^2 = \Sigma(y\text{-mean})^2/N$
- REML $\sigma^2 = \Sigma(y\text{-mean})^2/(N-1)$
- Little difference unless many fixed effects
- Use REML computer programs such as ASREML

Estimating genetic variance

Example: Data on phenotypes (y) of full sibs

y = fixed effects = g + e

$Cov(g_i, g_j) = A(i,j)\ V_G = 0.5\ V_G$ if i and j are sibs

Therefore estimate V_G by 2cov(full-sibs)

h^2 by 2 correlation between full-sibs

What is the covariance between twins?

Model with dominance

Covariance between genetic effects of relatives

Model with 1 gene, 2 alleles and additive and dominant gene action

We need genetic variances and covariances

Genotype	BB	Bb	bb	
Effect	a	d	-a	
Frequency	p^2	2pq	q^2	(p+q=1)

Mean = $a*p^2 + d*2pq - a*q^2 = (p-q)*a + 2pqd$

Variance (genetic effect) = genetic variance = V_G

$= E(effect^2) - E(effect)^2$

$V_G = a^2*p^2 + d^2*2pq + a^2*q^2 - [(p-q)*a + 2pqd]^2 = 2pq\alpha^2 + (2pqd)^2$

where $\alpha = a + (q-p)d$

Covariance between genetic effects of relatives

Model with 1 gene, 2 alleles and additive and dominant gene action

but the covariance between relatives doesn't depend directly on VG. We need to decompose VG into an additive and dominance variance.

Parameterise the genetic value as

g = mean + additive effect + dominance deviation

g = mean + paternal allele effect + maternal allele effect + interaction of alleles

Genotype	BB	Bb	bb	
Effect	a	d	-a	
Frequency	p^2	2pq	q^2	(p+q=1)
mean	(p-q)a + 2pqd	(p-q)a + 2pqd	(p-q)a + 2pqd	
additive	$2q\alpha$	$(q-p)\alpha$	$-2p\alpha$	$\alpha=a+(q-p)d$
dominance dev.	$-q^2d$	2pqd	$-p^2d$	

Mean(additive effect) =0, mean(dominance deviation)=0, cov(additive effect, dominance dev) =0

Genetic variance = V_G = $2pq\alpha^2$ + $(2pqd)^2$

= V_A + V_D

Covariance between genetic effects of relatives

Model with 1 gene, 2 alleles and additive and dominant gene action

$$Cov(g_i, g_j) = Cov(a_i+d_i, a_j+d_j) = Cov(a_i, a_j) + cov(d_i, d_j)$$
$$= A(i,j)\, V_A + D(i,j)\, V_D$$

D(i,j) = prob(i and j inherit the same genotype IBD)

Eg

D(i,j) = 1 for MZ twins, 0.25 for full-sibs, 0 for parent and offspring

Covariance between genetic effects of relatives

Model with 1 gene, 2 alleles and additive and dominant gene action

Relationships	MZ twins	full-sibs	1/2sibs	P-O
A	1	0.5	0.25	0.5
D	1	0.25	0	0

Therefore can estimate both V_A and V_D by using multiple relationships

Covariance between environmental effects of relatives

y = mean + genetic effect + common environment effect + individual environment effect

$y = mean + g + e_c + e$

Model with a common environmental effect within the same family

$Cov(e_{ci}, e_{cj}) = V_C$ if i and j in same family, zero otherwise

Relationships	MZ twins	full-sibs	1/2sibs	P-O
A	1	0.5	0.25	0.5
D	1	0.25	0	0
E common	1	1	?	?

Covariance between relatives

Estimating V_A, V_D and V_C

Difficult!

Assume V_D =0

VA = 2(cov(MZ twins) – cov(full-sibs))

Relationships	MZ twins	full-sibs	1/2sibs	P-O
A	1	0.5	0.25	0.5
D	1	0.25	0	0
E common	1	1	?	?

Covariance between relatives

Can add epistatic interactions to model

g = mean + additive + dominace + epistasis

eg g = mean + a + d + aa

$Cov(g_i, g_j) = A(i,j)\, V_A + D(i,j)\, V_D + A(i,j)^2\, V_{AA}$

Relationships	MZ twins	full-sibs	1/2sibs	P-O
A	1	0.5	0.25	0.5
D	1	0.25	0	0
AxA	1	0.25	0.0625	0.25

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 IV. *Distribution of Stature.*

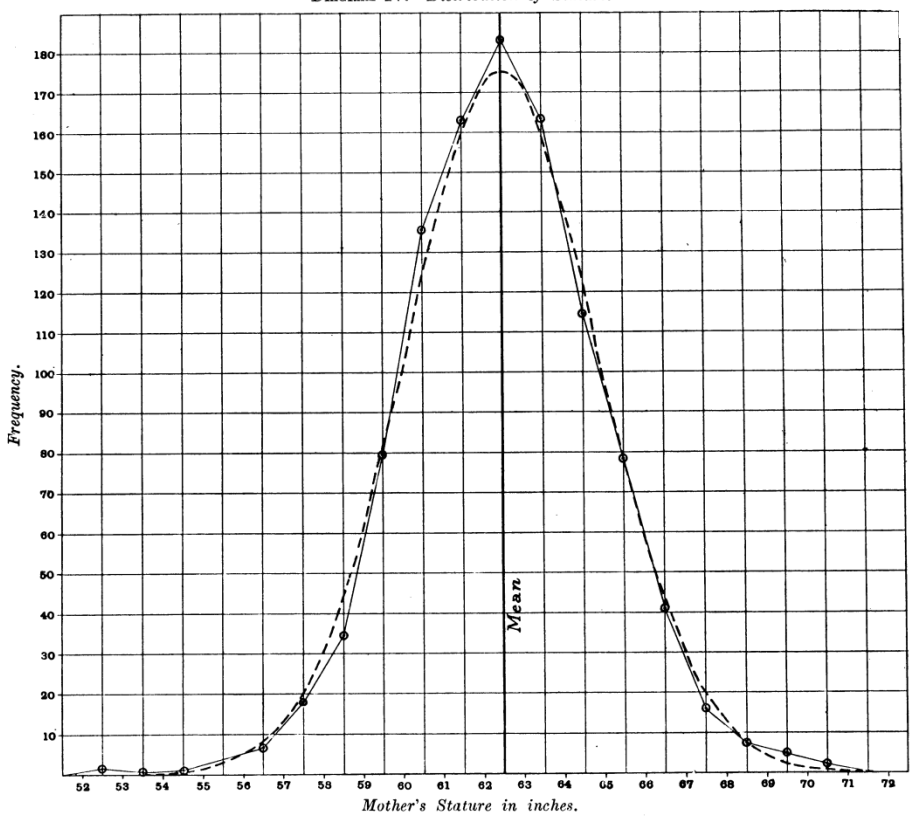
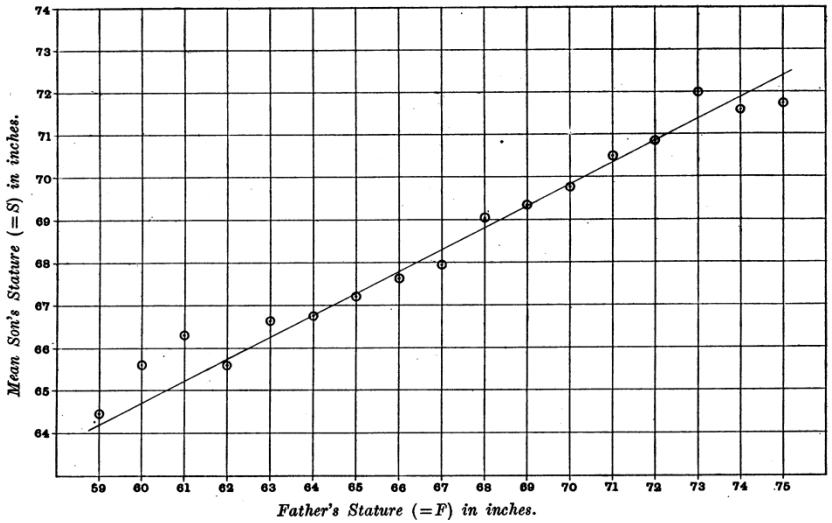


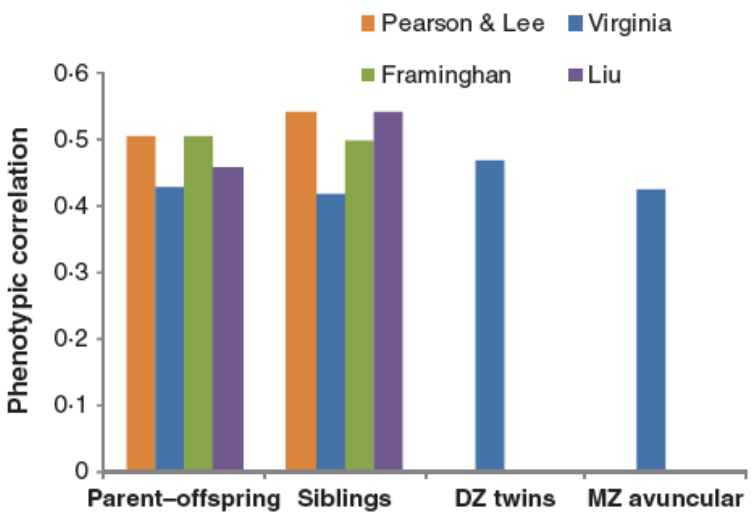
DIAGRAM I. *Probable Stature of Son for given Father's Stature.*

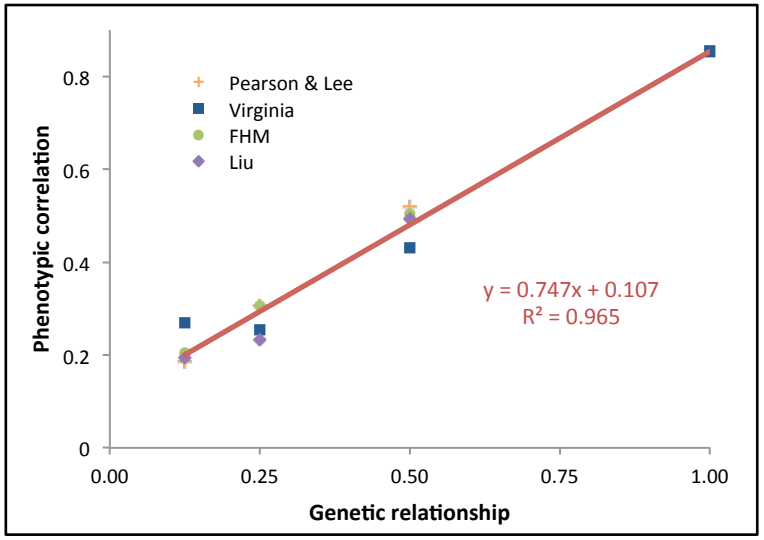
Regression Line: $S = 33{\cdot}73 + {\cdot}516\,F$. 1078 Cases.

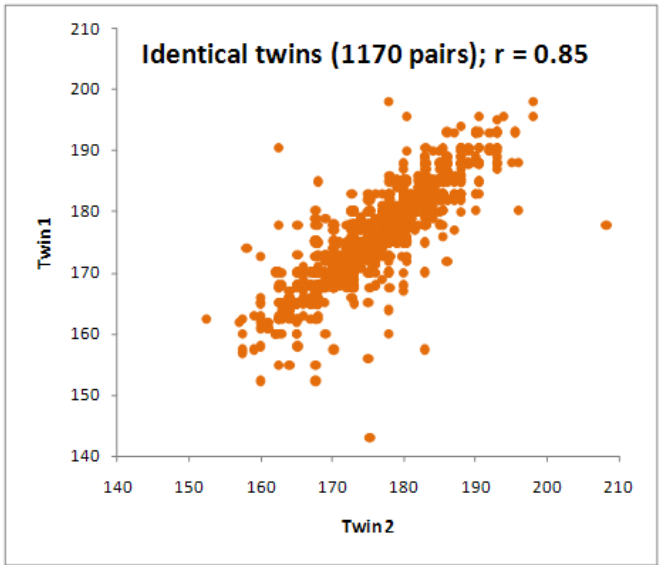


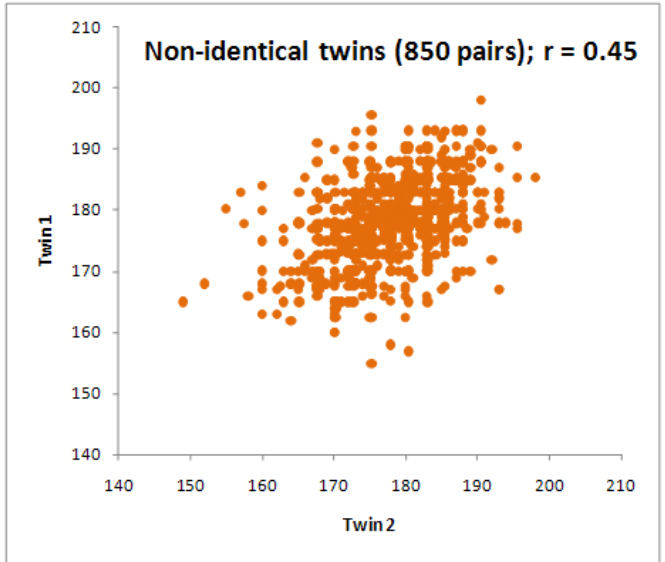
PAIR	CORRELATION	SE
Spouse	0.28	0.02
Son-Father	0.51	0.02
Daughter-Father	0.51	0.01
Son-Mother	0.49	0.02
Daughter-Mother	0.51	0.01
Brother-brother	0.51	0.03
Sister-sister	0.54	0.02
Brother-sister	0.55	0.01

Resemblance between relatives (height)



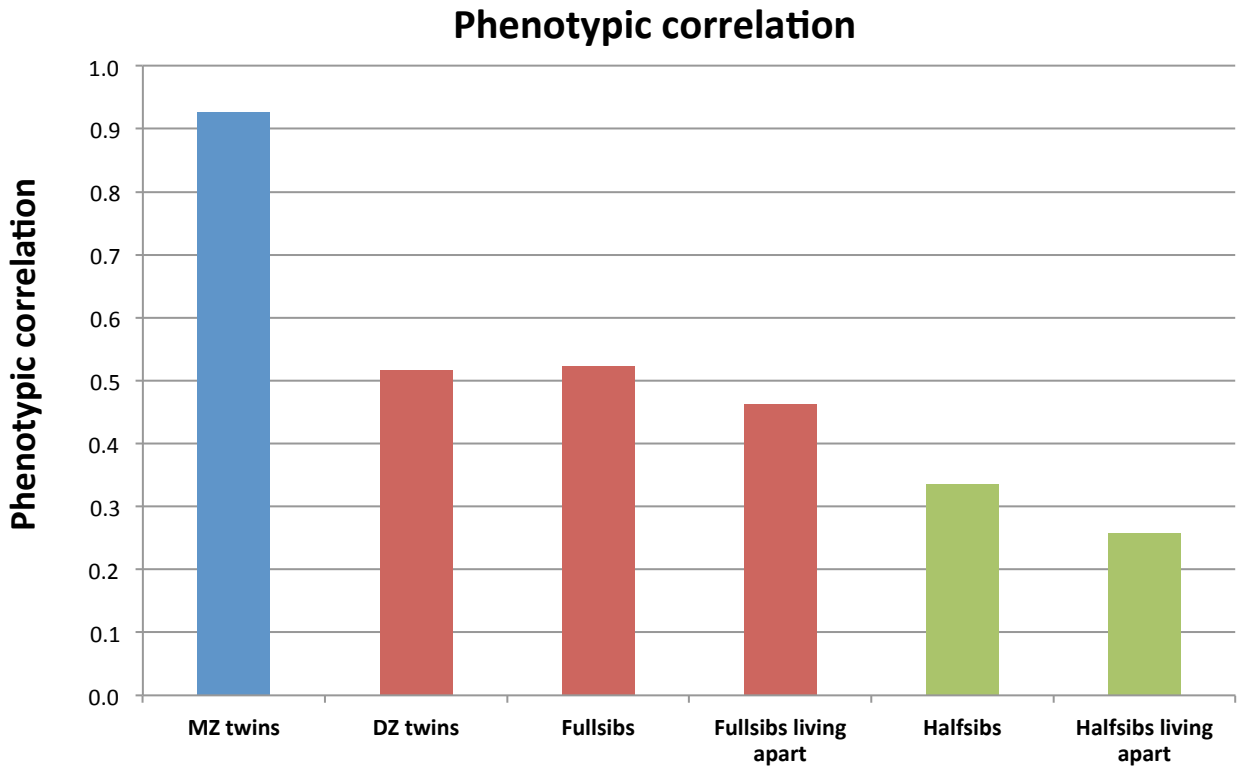






More data on height

Data from ~172,000 18-year old brother pairs



[Magnus Johannesson, David Cesarini]

Sex Differences in Heritability of BMI: A Comparative Study of Results from Twin Studies in Eight Countries

Twin Research October 2003

Karoline Schousboe[1], Gonneke Willemsen[2], Kirsten O. Kyvik[1], Jakob Mortensen[1], Dorret I. Boomsma[2], Belinda K. Cornes[3], Chayna J. Davis[4], Corrado Fagnani[5], Jacob Hjelmborg[1], Jaakko Kaprio[6], Marlies de Lange[7], Michelle Luciano[3], Nicholas G. Martin[3], Nancy Pedersen[4], Kirsi H. Pietiläinen[6,8], Aila Rissanen[8], Suoma Saarni[6], Thorkild I.A. Sørensen[9], G. Caroline M. van Baal[2], and Jennifer R. Harris[10]

Table 5a

Twin Correlations (R) for BMI and Number of Pairs (N) Assessed by Zygosity and Sex for Twins Aged 20–29 years

	Australia R (*N*)	Denmark R (*N*)	Finland R (*N*)	Italy R (*N*)	Netherlands R (*N*)	Norway R (*N*)	Sweden R (*N*)	UK R (*N*)
MZm	0.67 (390)	0.77 (824)	0.74 (247)	0.83 (66)	0.65 (299)	0.69 (563)	0.77 (887)	n.a.
DZm	0.32 (260)	0.35 (897)	0.32 (304)	0.52 (43)	0.31 (222)	0.41 (479)	0.35 (1346)	n.a.
MZf	0.72 (768)	0.73 (1161)	0.78 (411)	0.83 (129)	0.79 (518)	0.74 (738)	0.73 (1054)	0.74 (89)
DZf	0.33 (486)	0.35 (1046)	0.37 (358)	0.58 (76)	0.41 (336)	0.35 (643)	0.36 (1472)	0.52 (75)
DZOS	0.18 (596)	0.30 (1620)	0.22 (668)	0.12 (96)	0.36 (473)	0.18 (968)	n.a.	n.a.

Average correlations	
MZ	0.74
DZ (same sex)	0.36
DZ (opposite sex)	0.25

frontiers in
ENDOCRINOLOGY

ORIGINAL RESEARCH ARTICLE
published: 28 February 2012
doi: 10.3389/fendo.2012.00029

Variability in the heritability of body mass index: a systematic review and meta-regression

*Cathy E. Elks[1], Marcel den Hoed[1], Jing Hua Zhao[1], Stephen J. Sharp[1], Nicholas J. Wareham[1], Ruth J. F. Loos[1] and Ken K. Ong[1,2]**

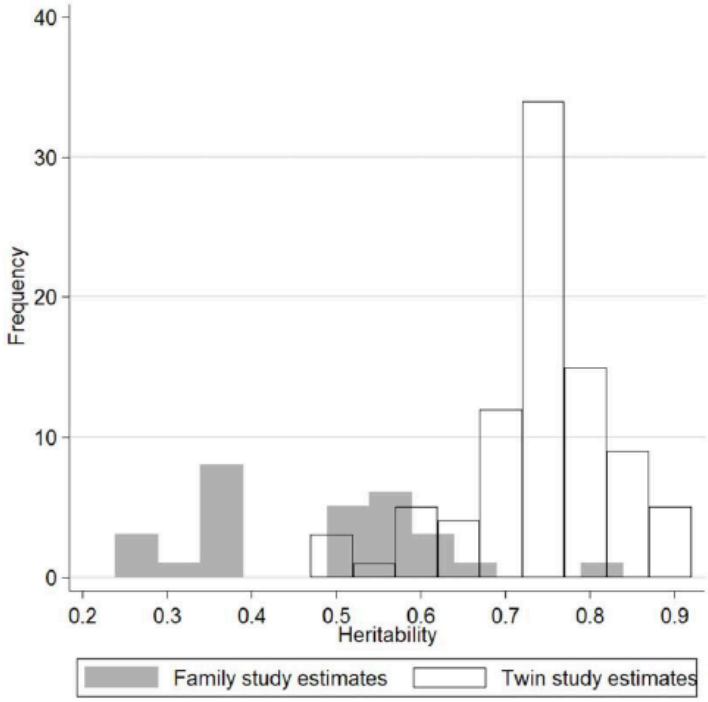
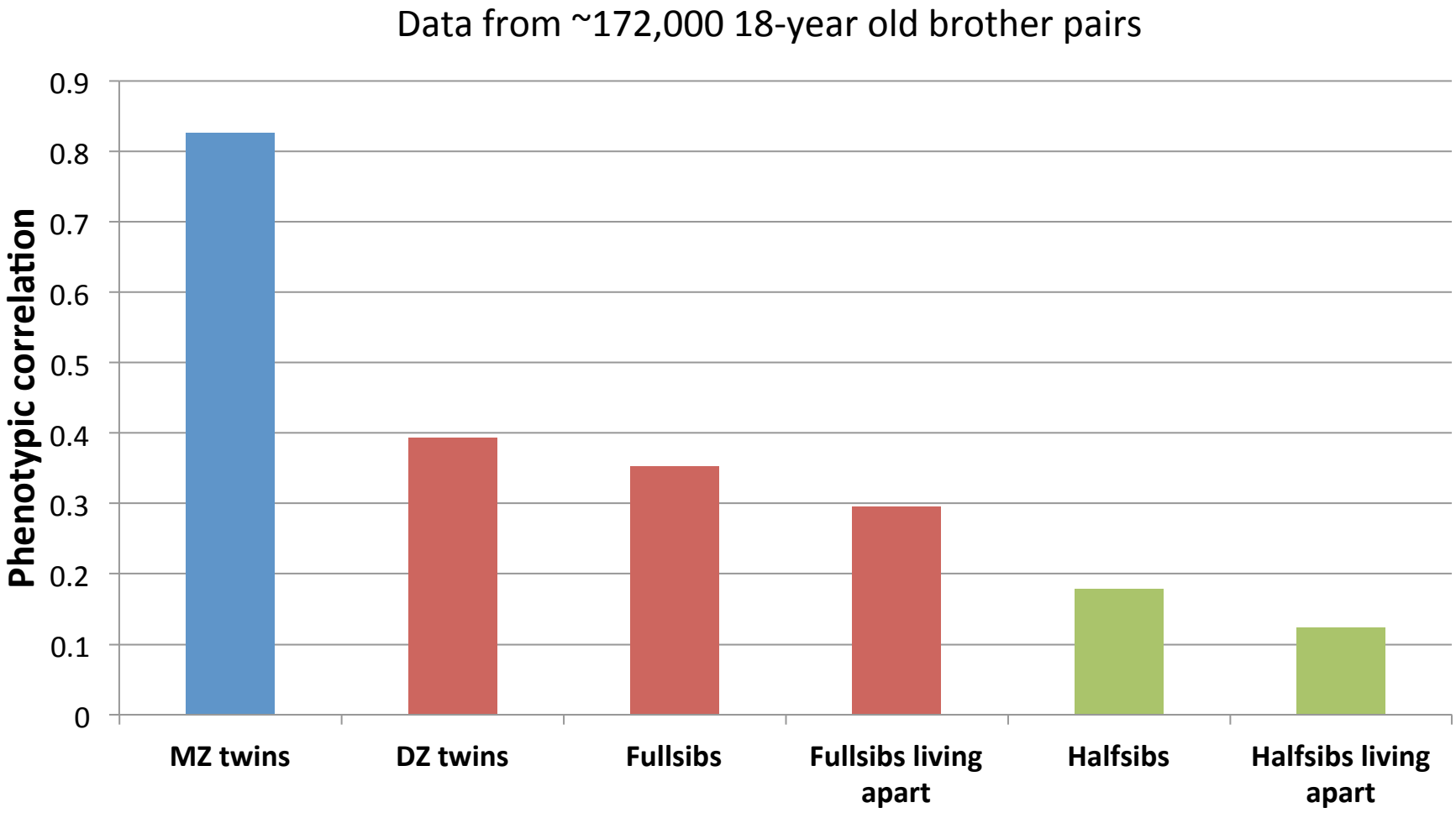


FIGURE 1 | Histogram showing the wide distribution of reported estimates of BMI heritability from twin studies (white bars) and family studies (gray bars).

BMI



[Magnus Johannesson, David Cesarini]

Summary
Resemblance between relatives

Model phenotypes by fixed effects and random effects including genetic value (additive, dominance, epistatic)

Model covariance of genetic effects by relationship estimated from pedigree (or SNP genotypes)

Estimate genetic variance by REML

Estimating genetic variation within families

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

Key concepts

1. There is variation in realised relationships given the expected value from the pedigree;
2. Variation in realised relationships can be captured with genetic markers;
3. Variation in realised relationships can be exploited to estimate genetic variation

Genetic covariance between relatives

$cov_G(y_i,y_j) = a_{ij}\sigma_A^2 + d_{ij}\sigma_D^2$

a = additive coefficient of relationship

= $2 * \theta$ (= $E(\pi_a)$)

d = coefficient of fraternity

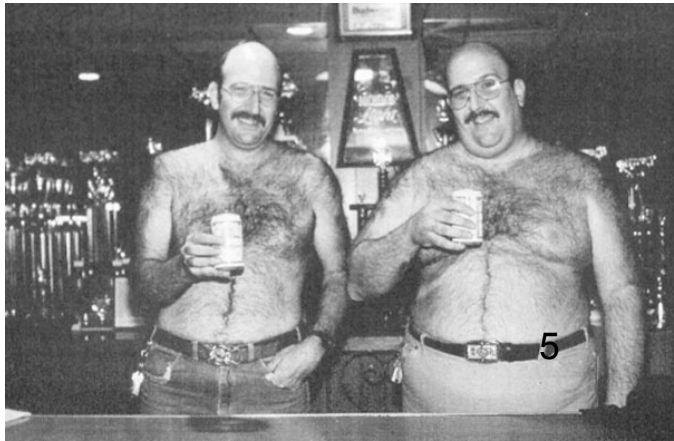
= Prob(2 alleles are IBD) = $\Delta = E(\pi_d)$

Examples (no inbreeding)

Relatives	**a**	**d**
MZ twins	1	1
Parent-offspring	½	0
Fullsibs	½	¼
Double first cousins	¼	$^1/_{16}$

Controversy/confounding: nature vs nurture

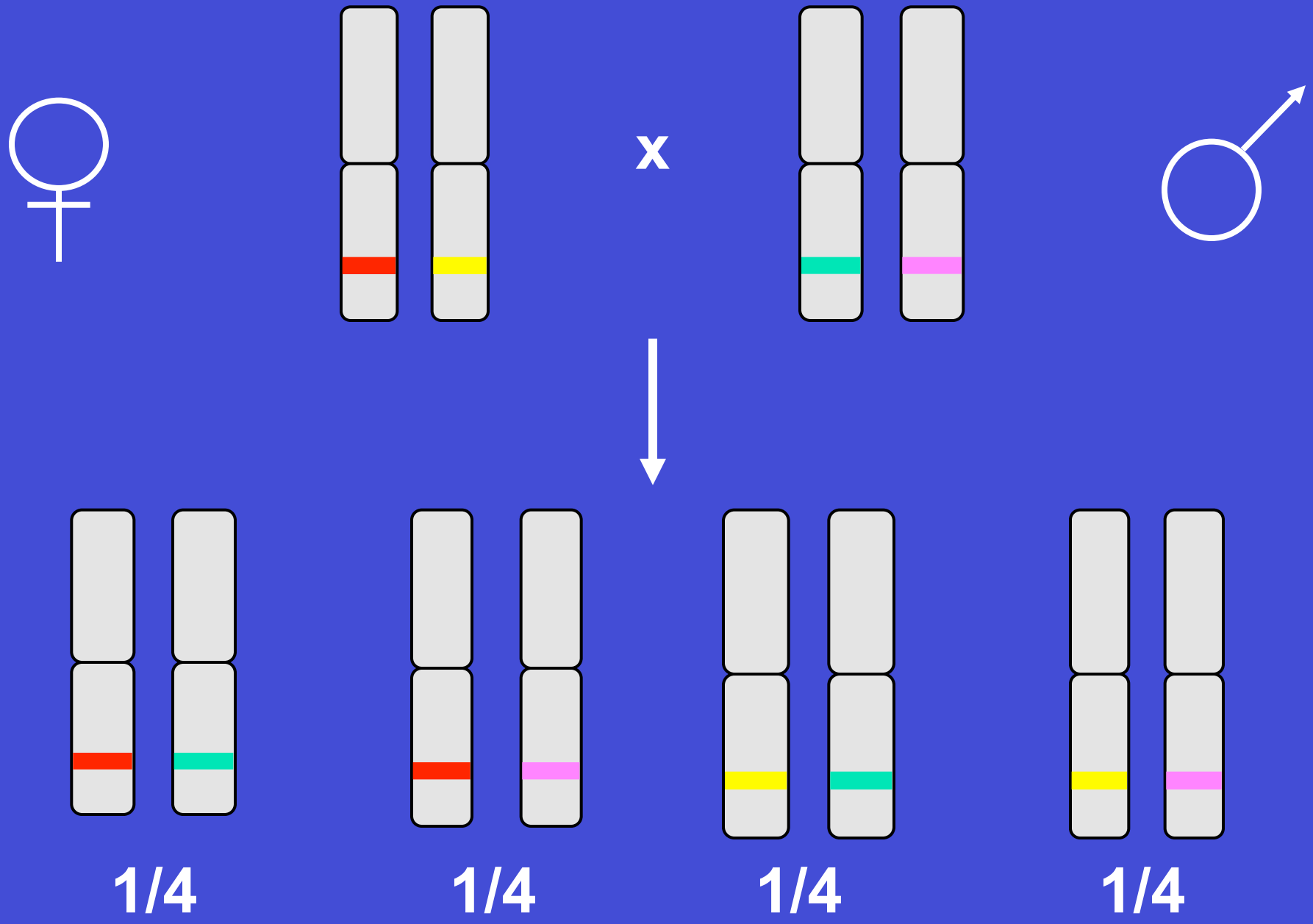
- Is observed resemblance between relatives genetic or environmental?
  - MZ & DZ twins (shared environment)
  - Fullsibs (dominance & shared environment)
- Estimation and statistical inference
  - Different models with many parameters may fit data equally well



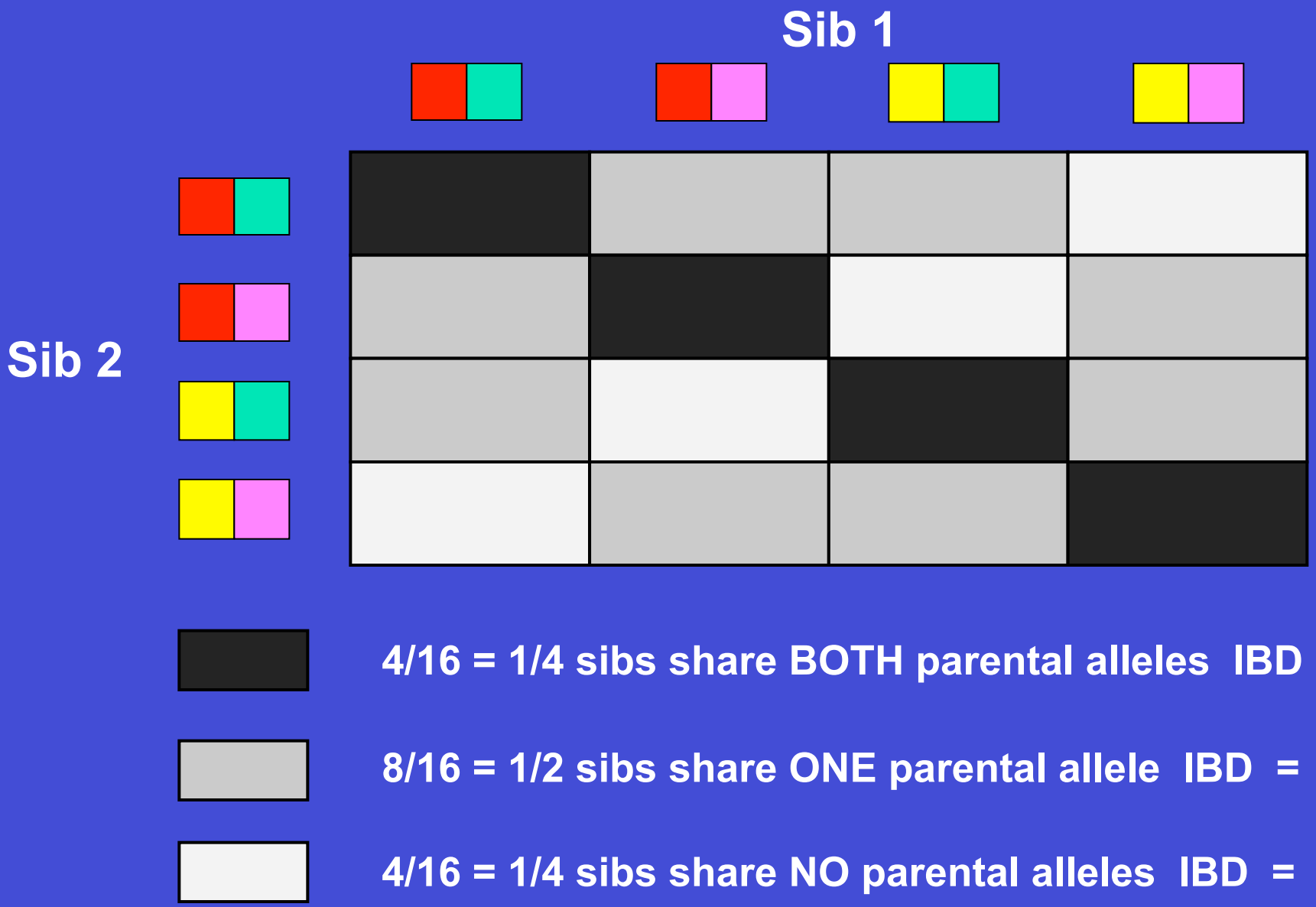
Actual or *realised* genetic relationship

= proportion of genome shared IBD (π_a)

- Varies around the expectation
  - Apart from parent-offspring and MZ twins
- Can be estimated using marker data


x
1/4
1/4
1/4
1/4

IDENTITY BY DESCENT



Single locus

Relatives	$E(\pi_a)$	$var(\pi_a)$
Fullsibs	½	$^1/_8$
Halfsibs	¼	$^1/_{16}$
Double 1st cousins	¼	$^3/_{32}$

Several notations

IBD	Probability	Actual
IBD0	k_0	0 or 1
IBD1	k_1	0 or 1
IBD2	k_2	0 or 1
	$\Sigma=1$	$\Sigma=1$

Realisations

k_0	k_1	k_2
1	0	0
0	1	0
0	0	1

$\pi_a = ½k_1 + k_2 = R = 2\theta$

$\pi_d = k_2 = \Delta_{xy}$

[e.g., LW Chapter 7; Weir and Hill 2011, Genetics Research]

n multiple unlinked loci

Relatives	$E(\pi_a)$	$var(\pi_a)$
Fullsibs	½	$1/_{8n}$
Halfsibs	¼	$1/_{16n}$
Double 1st cousins	¼	$3/_{32n}$

Loci are on chromosomes

- Segregation of large chromosome segments within families
  - increasing variance of IBD sharing
- Independent segregation of chromosomes
  - decreasing variance of IBD sharing

Theoretical SD of π_a

Relatives	1 chrom (1 M)	genome (35 M)
Fullsibs	0.217	0.038
Halfsibs	0.154	0.027
Double 1st cousins	0.173	0.030

[Stam 1980; Hill 1993; Guo 1996; Hill & Weir 2011]

Fullsibs: genome-wide (Total length L Morgan)

$var(\pi_a) \approx 1/(16L) - 1/(3L^2)$ [Stam 1980; Hill 1993; Guo 1996]

$var(\pi_d) \approx 5/(64L) - 1/(3L^2)$

$var(\pi_d)/\ var(\pi_a) \approx 1.3$ if $L = 35$

Genome-wide variance depends more on total genome length than on the number of chromosomes

Fullsibs: Correlation additive and dominance relationships

$$r(\pi_a, \pi_d) = \sigma(\pi_a) / \sigma(\pi_d) \approx [1/(16L) / (5/(64L))]^{0.5} = 0.89.$$

Using $\beta(\pi_a \text{ on } \pi_d) = 1$

Difficult but not impossible to disentangle additive and dominance variance

NB Practical

Summary

Additive and dominance (fullsibs)

	$SD(\pi_a)$	$SD(\pi_d)$
Single locus	0.354	0.433
One chromsome (1M)	0.217	0.247
Whole genome (35M)	0.038	0.043

Predicted correlation 0.89

(genome-wide π_a and π_d)

Estimating IBD from marker data

- Elston-Stewart algorithm

  Handles large pedigrees, but small nr of loci, exact IBD distributions (Elston and Stewart, 1971)
- Lander-Green algorithm

  Handles small pedigrees, but large nr of loci, exact IBD distributions (Lander and Green, 1987). Software: Merlin
- MCMC methods

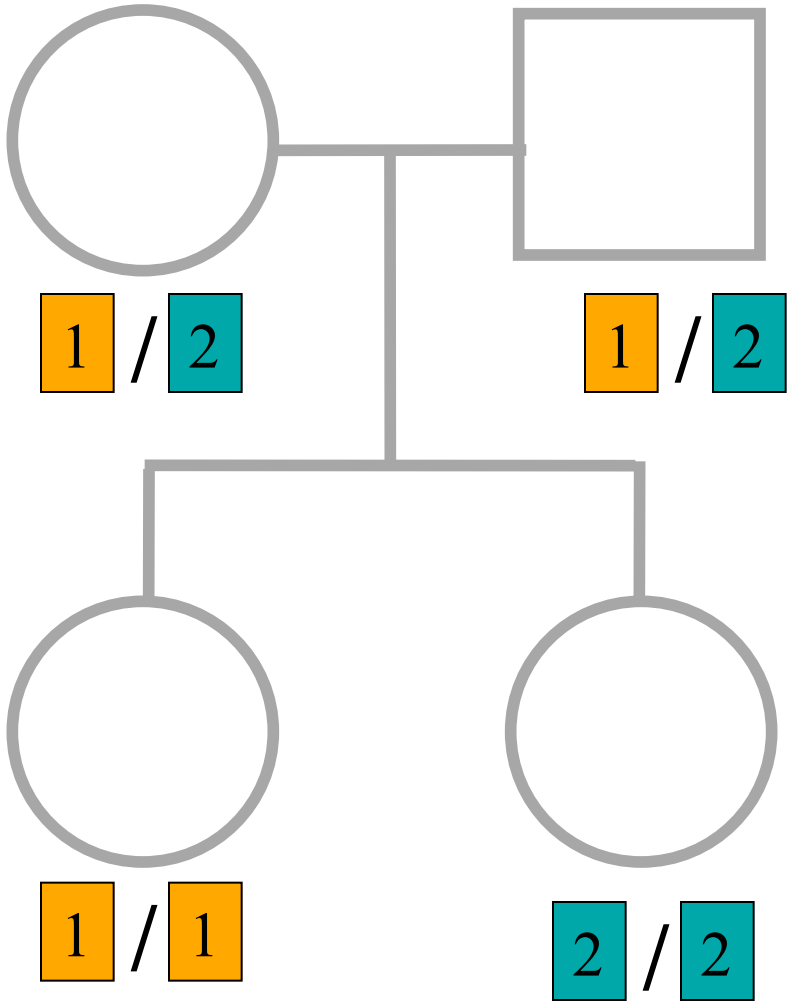
  Calculates approximate IBD distributions (Heath, 1997). Software: Loki
- Average sharing methods.

  Calculates approximate IBD distributions (Fulker et al., 1995; Almasy and Blangero, 1998). Software: SOLAR

Estimating π when marker is not fully informative

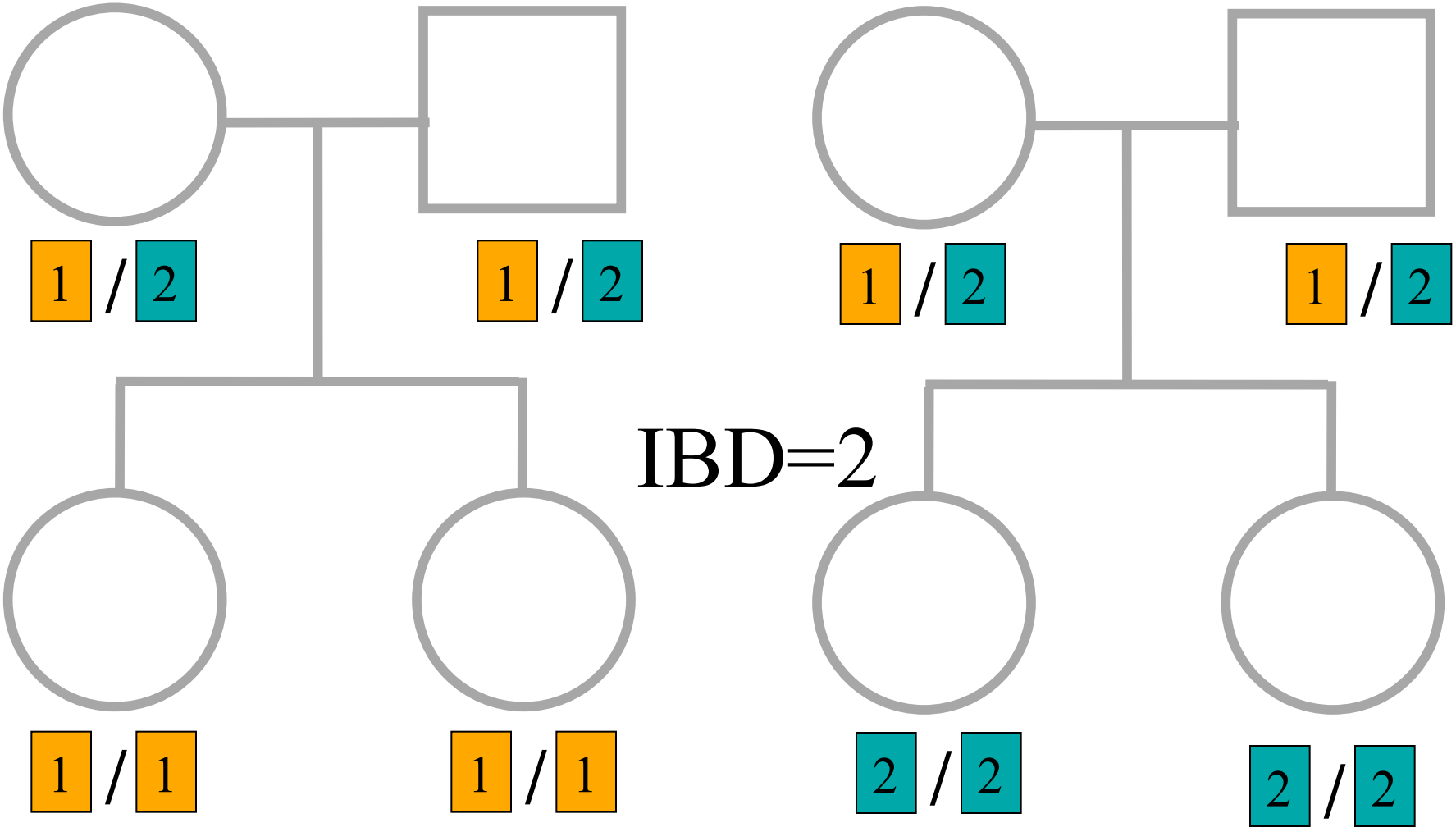
- Using:
  - Mendelian segregation rules
  - Marker allele frequencies in the population

IBD can be trivial…

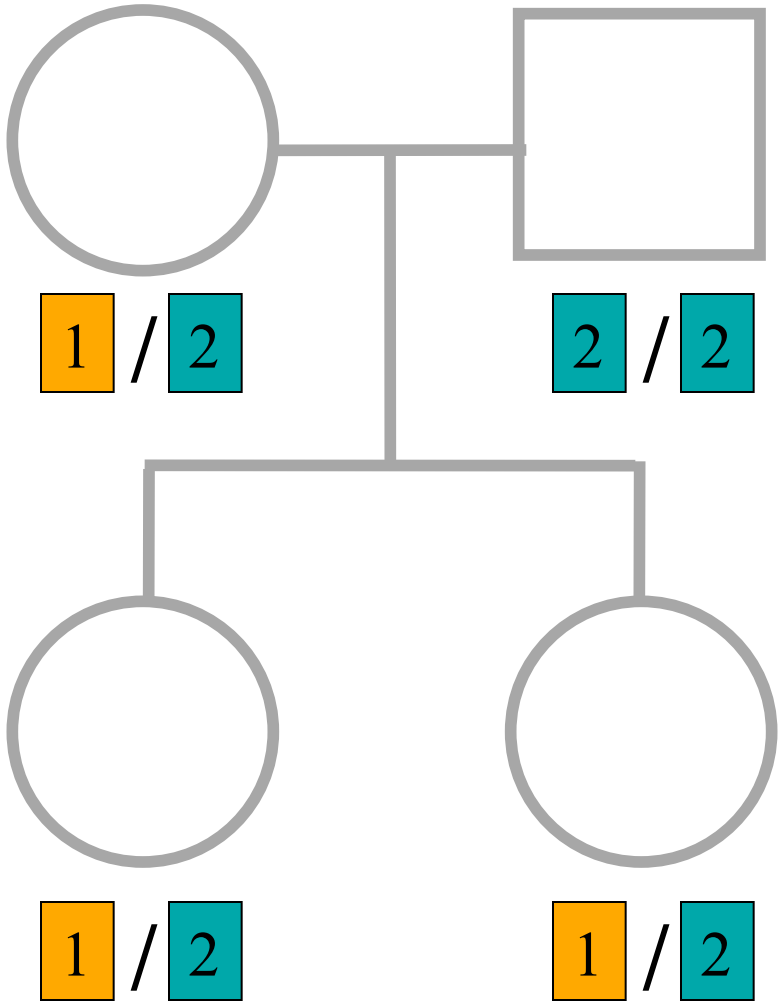


IBD=0

Two Other Simple Cases…



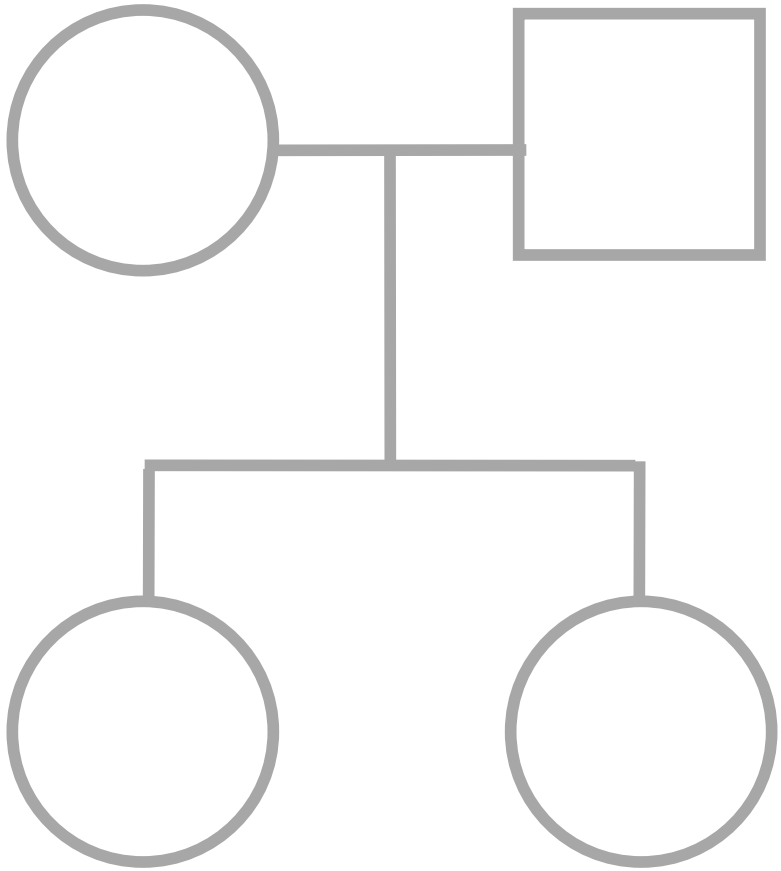
A little more complicated…



IBD=1
(50% chance)

IBD=2
(50% chance)

And even more complicated…



IBD=?

1 / 1

1 / 1

Bayes Theorem for IBD Probabilities

posterior

$$P(IBD = i \mid G) = \frac{\mathrm{P(IBD} = i, G)}{P(G)}$$

prior

$$= \frac{P(IBD = i)P(G \mid IBD = i)}{P(G)}$$

Prob(data)

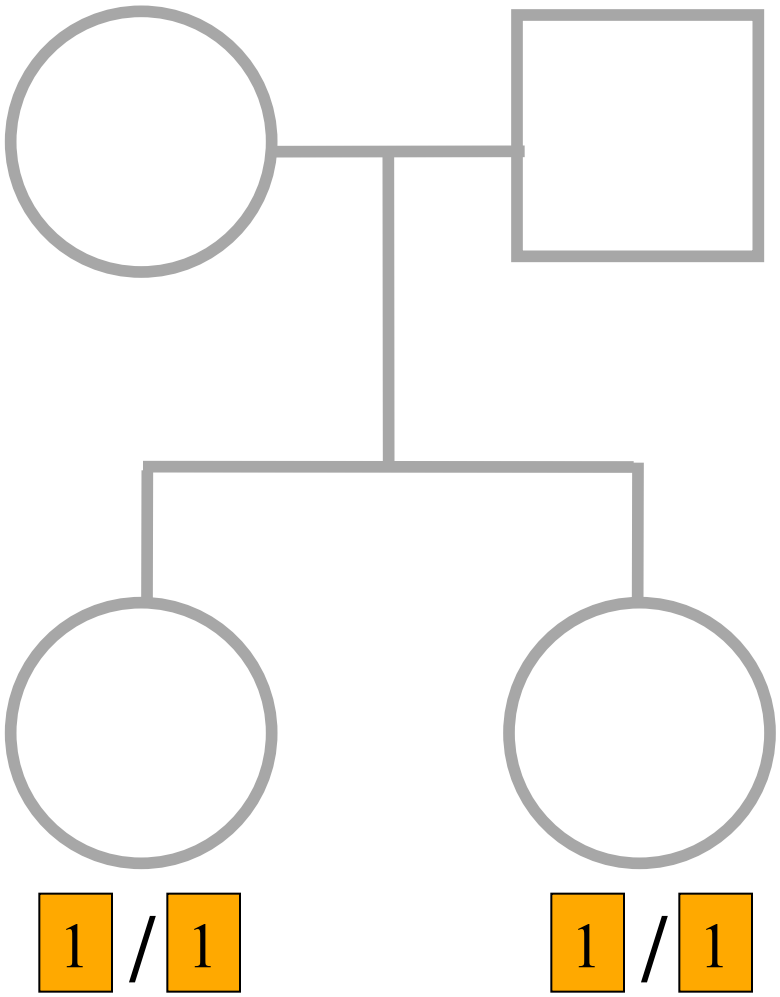
$$= \frac{P(IBD = i)P(G \mid IBD = i)}{\sum_j P(IBD = j)P(G \mid IBD = j)}$$

P(Marker Genotype|IBD State)

		IBD		
Sib	**CoSib**	0	1	2
(a,b)	(c,d)	$p_a p_b p_c p_d$	0	0
(a,a)	(b,c)	$p_a^2 p_b p_c$	0	0
(a,a)	(b,b)	$p_a^2 p_b^2$	0	0
(a,b)	(a,c)	$p_a^2 p_b p_c$	$p_a p_b p_c$	0
(a,a)	(a,b)	$p_a^3 p_b$	$p_a^2 p_b$	0
(a,b)	(a,b)	$p_a^2 p_b^2$	$p_a p_b^2 + p_a^2 p_b$	$p_a p_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

[Assumes Hardy-Weinberg proportions of genotypes in the population]

Worked Example



$$p_1 = 0.5$$

$$P(G \mid IBD = 0) = p_1^4 = 1/16$$

$$P(G \mid IBD = 1) = p_1^3 = 1/8$$

$$P(G \mid IBD = 2) = p_1^2 = 1/4$$

$$P(G) = 1/4\, p_1^4 + 1/2\, p_1^3 + 1/4\, p_1^2 = 9/64$$

$$P(IBD = 0 \mid G) = \frac{1/4\, p_1^4}{P(G)} = 1/9$$

$$P(IBD = 1 \mid G) = \frac{1/2\, p_1^3}{P(G)} = 4/9$$

$$\hat{\pi} = 2/3$$

$$P(IBD = 2 \mid G) = \frac{1/4\, p_1^2}{P(G)} = 4/9$$

Application (1)

Aim: estimate genetic variance from actual relationships between fullsib pairs

- Two cohorts of Australian twin families

	Adolescent	*Adult*
Families	500	1512
Individuals	1201	3804
Sibpairs with genotypes	950	3451
Markers per individual	211-791	201-1717
Average marker spacing	6 cM	5 cM

Application (1)

- Phenotype = height

Number of sibpairs with phenotypes and genotypes

Adolescent cohort	931
Adult cohort	2444
Combined	3375

Mean IBD sharing across the genome for the *j*th sib pair was based on IBD estimated every centimorgan and averaged over 3500 points (L = 35)

additive

$$\overline{\hat{\pi}}_{a(j)} = \sum_{i=1}^{3500} \hat{\pi}_{a(ij)} / 3500$$

dominance

$$\overline{\hat{\pi}}_{d(j)} = \sum_{i=1}^{3500} p_{2(ij)} / 3500$$

And for the c^{th} chromosome of length l_c cM

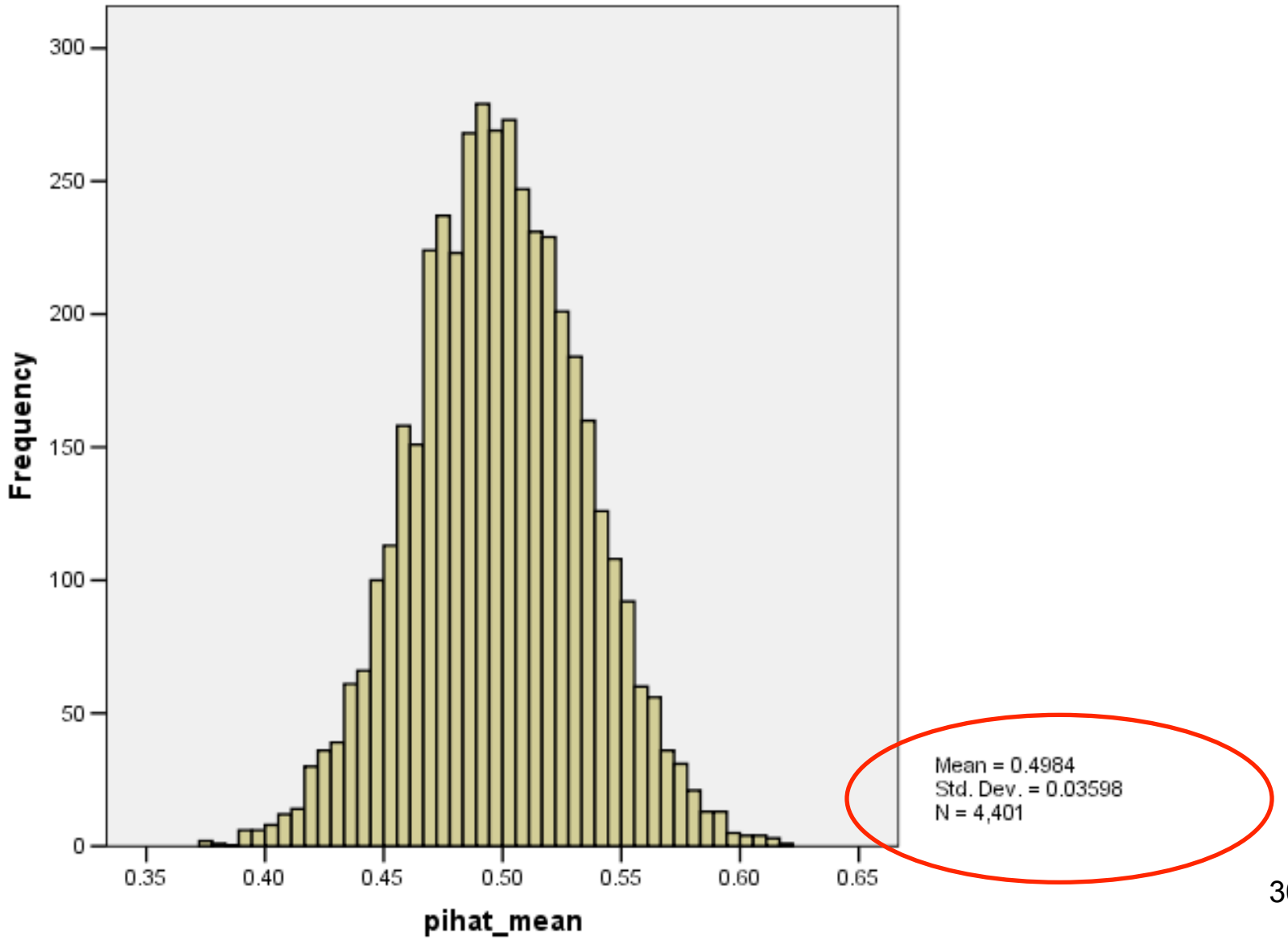
additive

$$\overline{\hat{\pi}}^{c}_{a(j)} = \sum_{i=1}^{l_c} \hat{\pi}^{c}_{a(ij)} / l_c$$

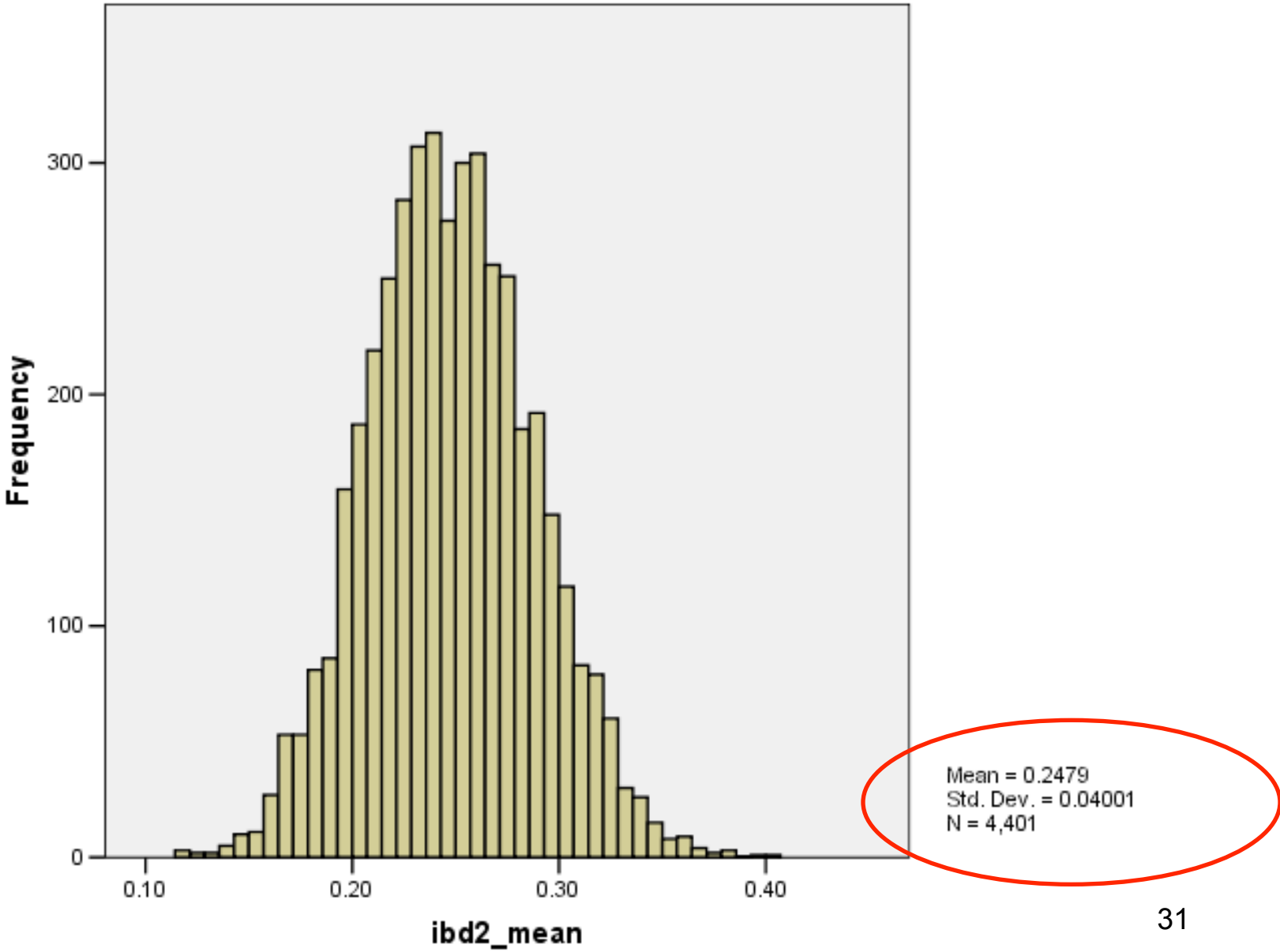
dominance

$$\overline{\hat{\pi}}^{c}_{d(j)} = \sum_{i=1}^{l_c} p_{2(ij)} / l_c$$

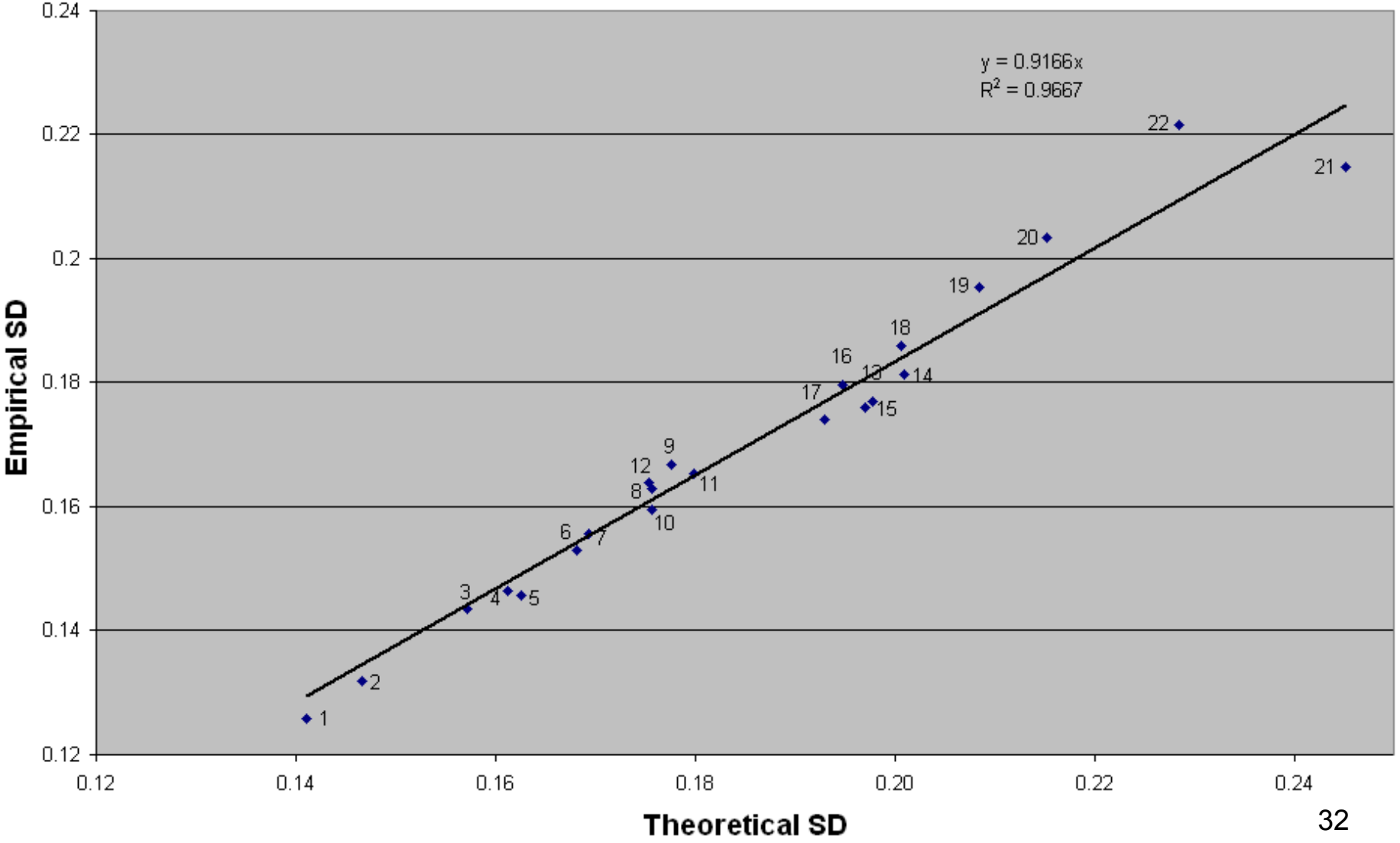
Mean and SD of genome-wide additive relationships



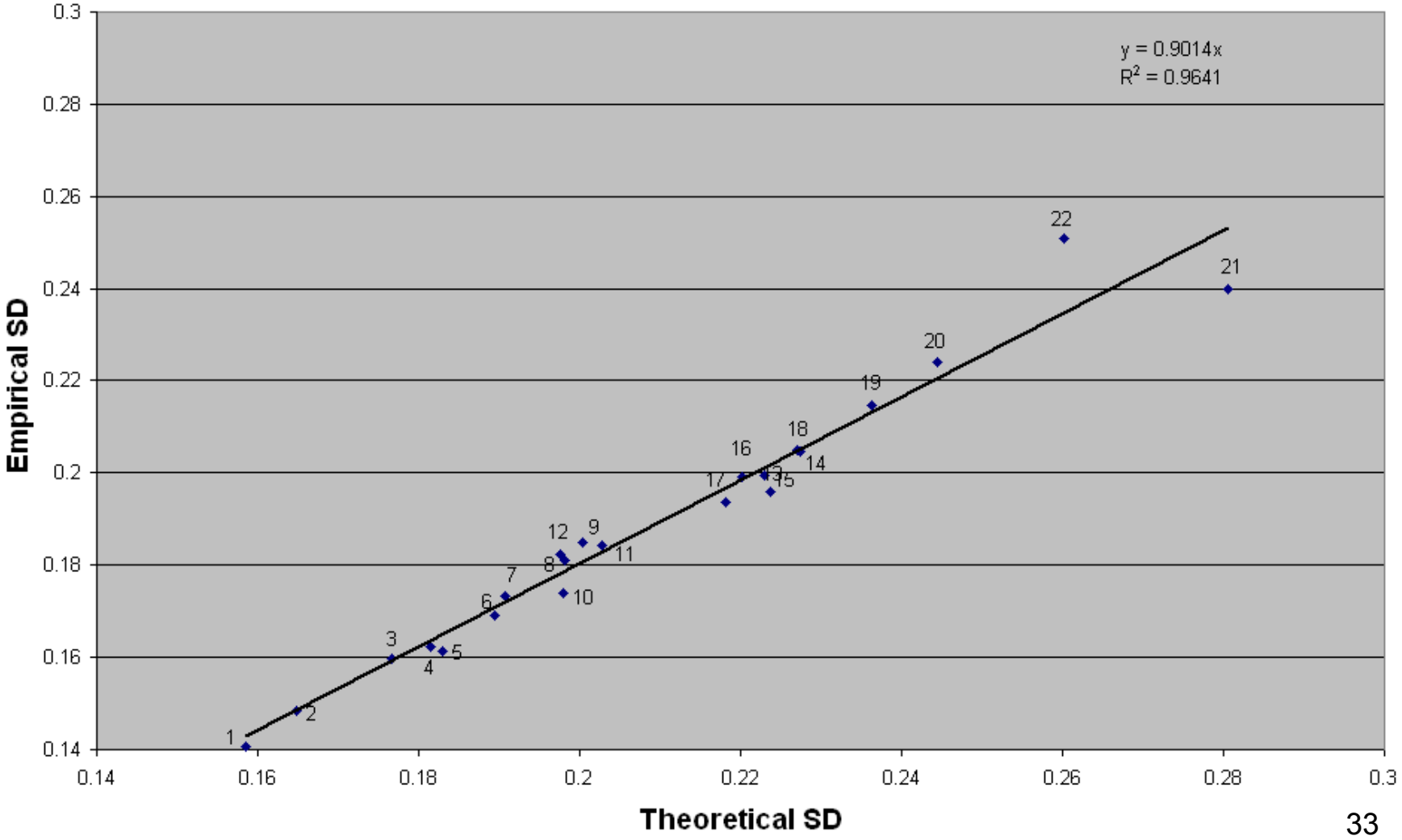
Mean and SD of genome-wide dominance relationships



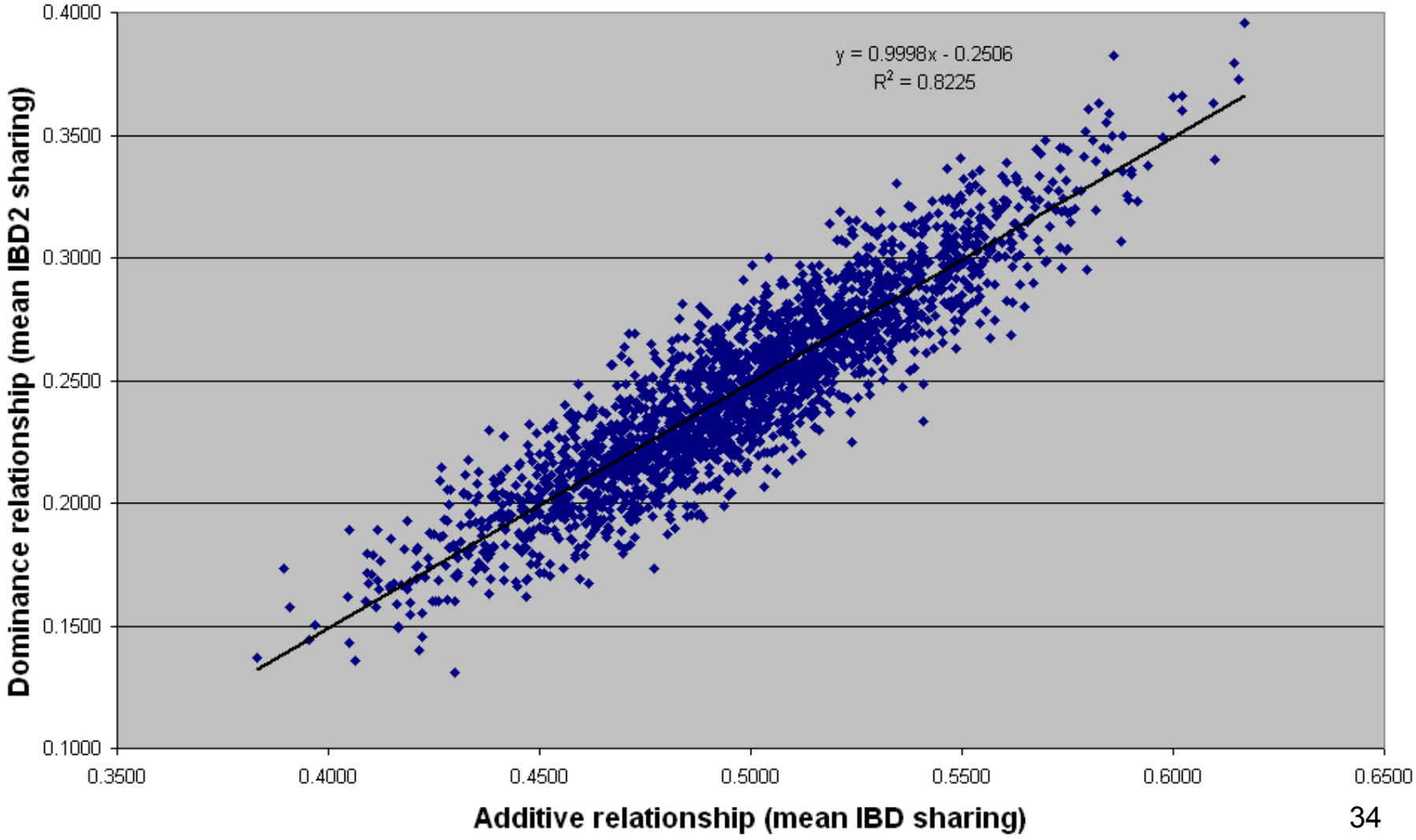
Empirical and theoretical SD of additive relationships correlation = 0.98 (*n* = 4401)



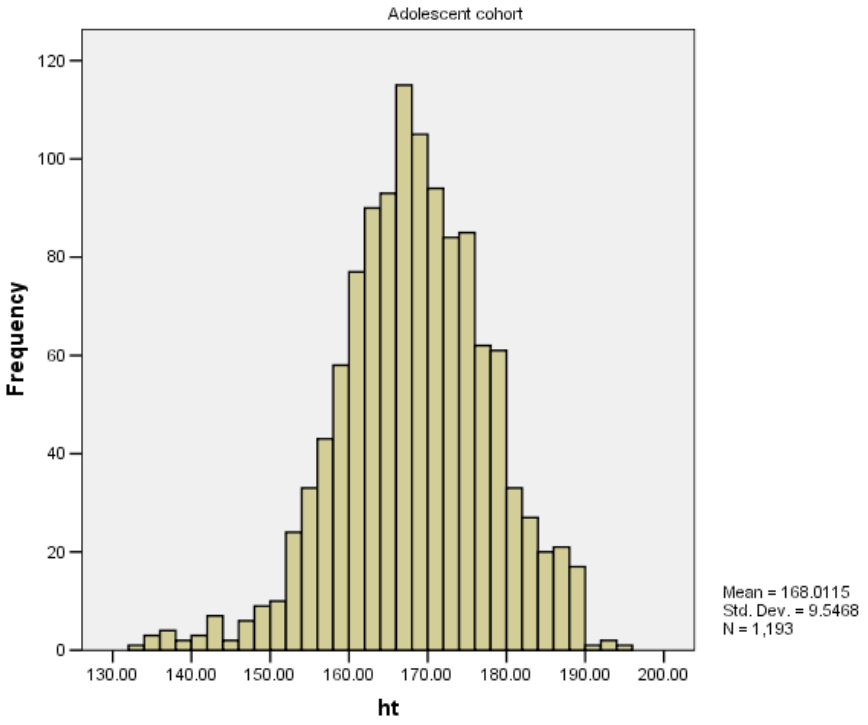
Empirical and theoretical SD of dominance relationships correlation = 0.98 (*n* = 4401)

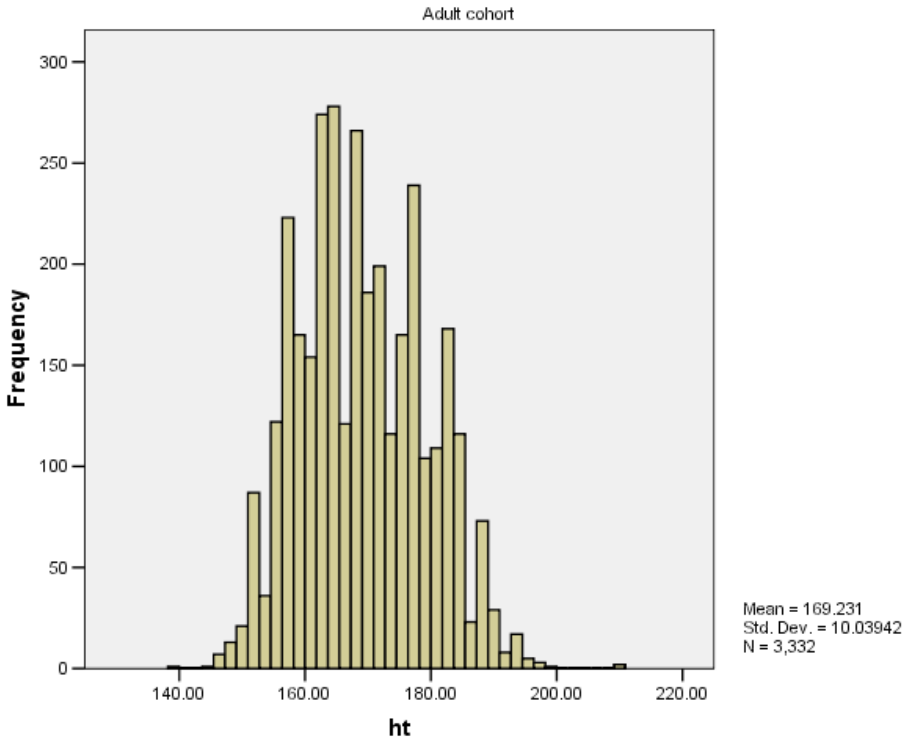


Additive and dominance relationships correlation = 0.91 (*n*= 4401)



Phenotypes





After adjustment for sex and age:

σ_p = 7.7 cm

σ_p = 6.9 cm

Phenotypic correlation between siblings

	Raw	After age & sex
Adolescents	0.33	0.40
Adults	0.24	0.39

Models

$$y_{ij} = \mu + c_i + a_{ij} + e_{ij}$$

$$\mathrm{var}(y) = \sigma_c^2 + \sigma_a^2 + \sigma_e^2$$

$$\mathrm{cov}(y_{ij}, y_{ik}) = \sigma_c^2 + \pi_{a(jk)}\sigma_a^2$$

C = Family effect

A = Genome-wide additive genetic

E = Residual

Full model	C + A + E
Reduced model	C + E

Estimation

- Maximum Likelihood variance components

- Likelihood-ratio-test (LRT) to calculate P-values for hypotheses

  H_0: A = 0

  H_1: A > 0

Estimates: null model (CE)

Cohort	Family effect (C)
Adolescent	0.40 (0.34 – 0.45)
Adult	0.39 (0.36 – 0.43)
Combined	0.39 (0.36 – 0.42)

Estimates: full model (ACE)

Cohort	C	A	*P*
Adolescent	0	0.80	0.0869
Adult	0	0.80	0.0009
Combined	0	0.80	0.0003

- ***All family resemblance due to additive genetic variation***

Sampling variances are large

Cohort	A (95% CI)
Adolescent	0.80 (0.00 – 0.90)
Adult	0.80 (0.43 – 0.86)
Combined	0.80 (0.46 – 0.85)

Power and SE of estimates

- True parameter (*t* = intra-class correlation)
- Sample size (*n* pairs)
- Variance in genome-wide IBD sharing (var(π))

$$\mathrm{var}(\hat{h}^2) \approx (1 - t^2)^2 \,/\left[(1 + t^2)(n\,\mathrm{var}(\pi))\right]$$

$$NCP = nh^4\mathrm{var}(\pi)(1+t^2) \,/\, (1\text{-}t^2)^2$$

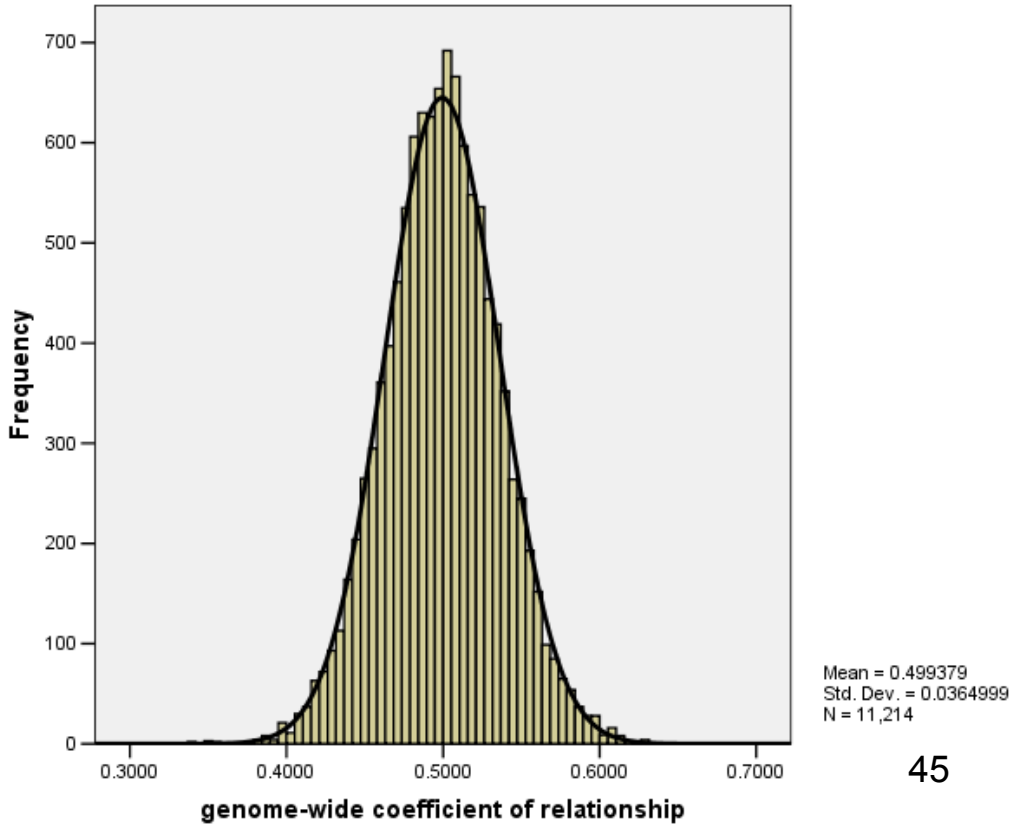
Application (2) Genome partitioning of additive genetic variance for height

- Aims
  - Estimate genetic variance from genome-wide IBD in larger sample
  - Partition genetic variance to individual chromosomes
    - using chromosome-wide coefficients of relationship
  - Test hypotheses about the distribution of genetic variance in the genome

Sample	*# Sibpairs*	*Sib Correlation*
AU	5952	0.43
US	3996	0.50
NL	1266	0.45
Total	11,214	0.46

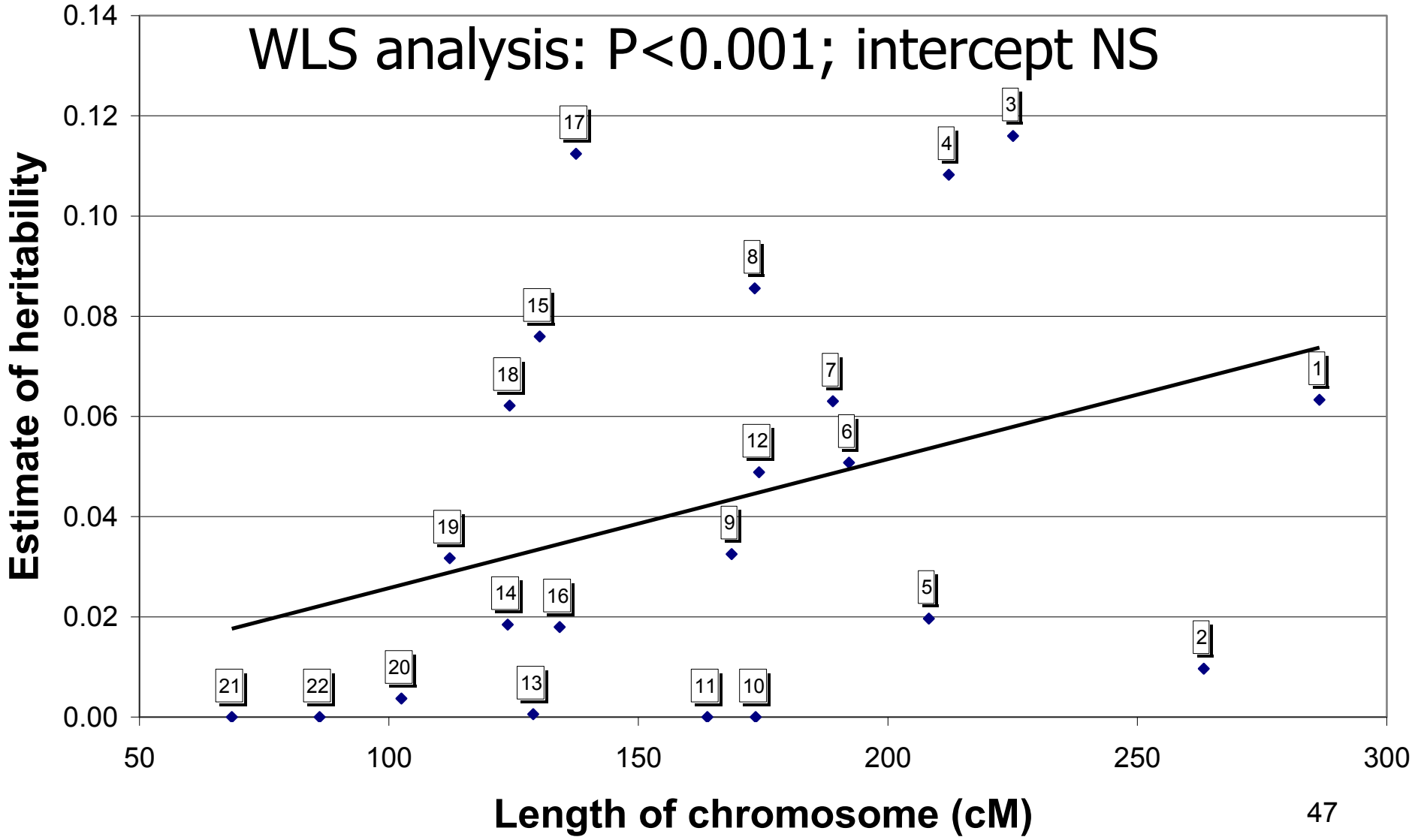
Realised relationships

Mean	0.499
Range	0.31 – 0.64
SD	0.036



Chrom.	Single chromosome analyses					Combined chromosome analysis		
	f^{2} (a)	h_i^{2} (b)	e^{2} (c)	LRT[d]	P-value	h_i^{2}	LRT[e]	P-value
1	0.4285	0.0607	0.5108	1.201	0.137	0.0633	1.418	0.117
2	0.4525	0.0131	0.5344	0.065	0.399	0.0097	0.037	0.424
3	0.4023	0.1134	0.4843	5.704	0.008	0.1160	6.269	0.006
4	0.4036	0.1124	0.4840	5.938	0.007	0.1082	5.705	0.008
5	0.4458	0.0264	0.5278	0.319	0.286	0.0196	0.191	0.500
6	0.4336	0.0506	0.5158	1.294	0.128	0.0508	1.370	0.500
7	0.4284	0.0616	0.5100	2.019	0.078	0.0630	2.230	0.068
8	0.4234	0.0708	0.5058	2.778	0.048	0.0856	4.172	0.021
9	0.4482	0.0216	0.5302	0.277	0.299	0.0325	0.663	0.500
10	0.4590	0.0000	0.5410	0.000	0.500	0.0000	0.000	0.500
11	0.4590	0.0000	0.5410	0.000	0.500	0.0000	0.000	0.500
12	0.4365	0.0451	0.5184	1.121	0.145	0.0489	1.434	0.500
13	0.4545	0.0089	0.5366	0.056	0.406	0.0006	0.000	0.500
14	0.4427	0.0323	0.5250	0.728	0.197	0.0185	0.246	0.500
15	0.4241	0.0703	0.5056	3.353	0.034	0.0760	4.028	0.022
16	0.4556	0.0069	0.5375	0.035	0.426	0.0180	0.251	0.308
17	0.4023	0.1142	0.4834	9.019	0.001	0.1124	8.967	0.001
18	0.4237	0.0703	0.5060	3.753	0.026	0.0622	3.013	0.041
19	0.4437	0.0309	0.5253	0.759	0.192	0.0317	0.840	0.500
20	0.4575	0.0031	0.5395	0.008	0.464	0.0037	0.012	0.456
21	0.4590	0.0000	0.5410	0.000	0.500	0.0000	0.000	0.500
22	0.4590	0.0000	0.5410	0.000	0.500	0.0000	0.000	0.500
SUM		0.9126		38.427		0.9205	40.846	

Longer chromosomes explain more additive genetic variance: ~0.03 per 100 cM



Application (3)

- Using SNP data to estimate IBD
- Data from ~20,000 fullsib pairs
- Height and BMI

Inference of the Genetic Architecture Underlying BMI and Height with the Use of 20,240 Sibling Pairs

Gibran Hemani,[1,2] Jian Yang,[1,2] Anna Vinkhuyzen,[2] Joseph E. Powell,[1,2] Gonneke Willemsen,[3,4] Jouke-Jan Hottenga,[4,5] Abdel Abdellaoui,[3,5] Massimo Mangino,[6] Ana M. Valdes,[6] Sarah E. Medland,[7] Pamela A. Madden,[8] Andrew C. Heath,[8] Anjali K. Henders,[7] Dale R. Nyholt,[7] Eco J.C. de Geus,[3,4,5] Patrik K.E. Magnusson,[9] Erik Ingelsson,[9,10] Grant W. Montgomery,[7] Timothy D. Spector,[6] Dorret I. Boomsma,[3,4,5] Nancy L. Pedersen,[9] Nicholas G. Martin,[7] and Peter M. Visscher[1,2,*]

frontiers in ENDOCRINOLOGY

ORIGINAL RESEARCH ARTICLE
published: 28 February 2012
doi: 10.3389/fendo.2012.00029

Variability in the heritability of body mass index: a systematic review and meta-regression

*Cathy E. Elks[1], Marcel den Hoed[1], Jing Hua Zhao[1], Stephen J. Sharp[1], Nicholas J. Wareham[1], Ruth J. F. Loos[1] and Ken K. Ong[1,2]**

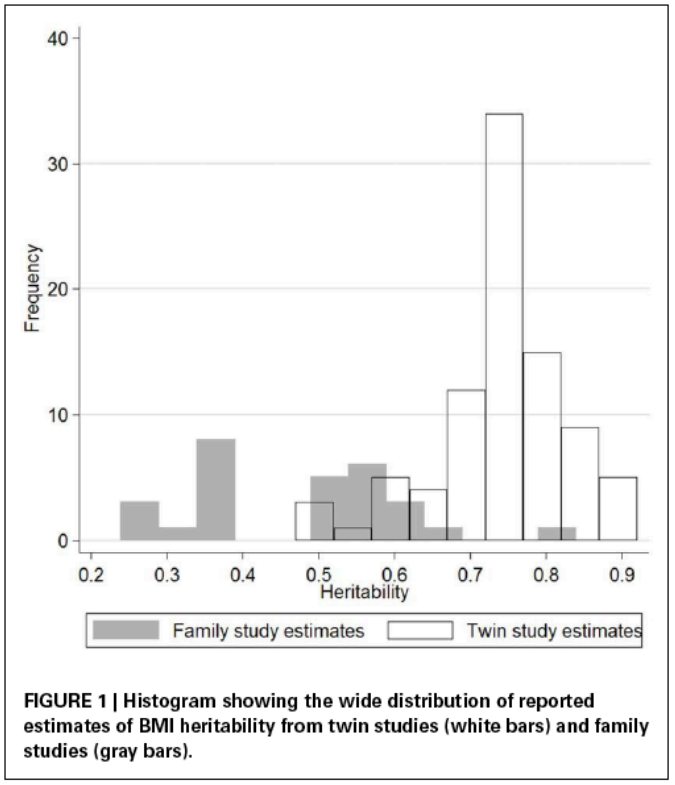
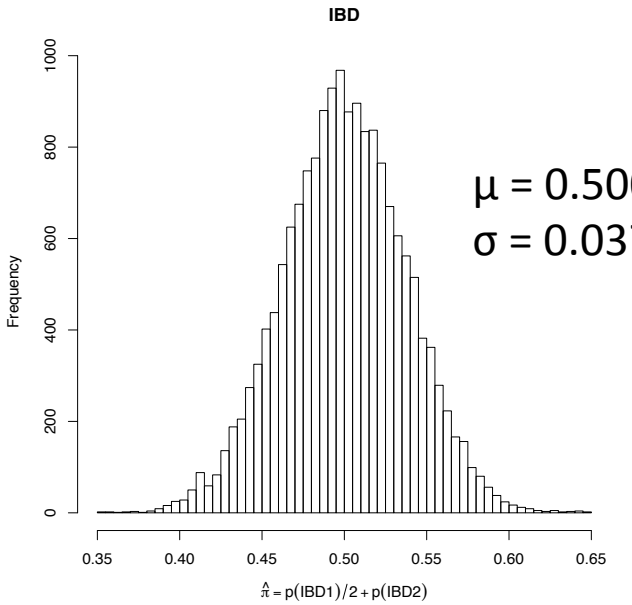


FIGURE 1 | Histogram showing the wide distribution of reported estimates of BMI heritability from twin studies (white bars) and family studies (gray bars).

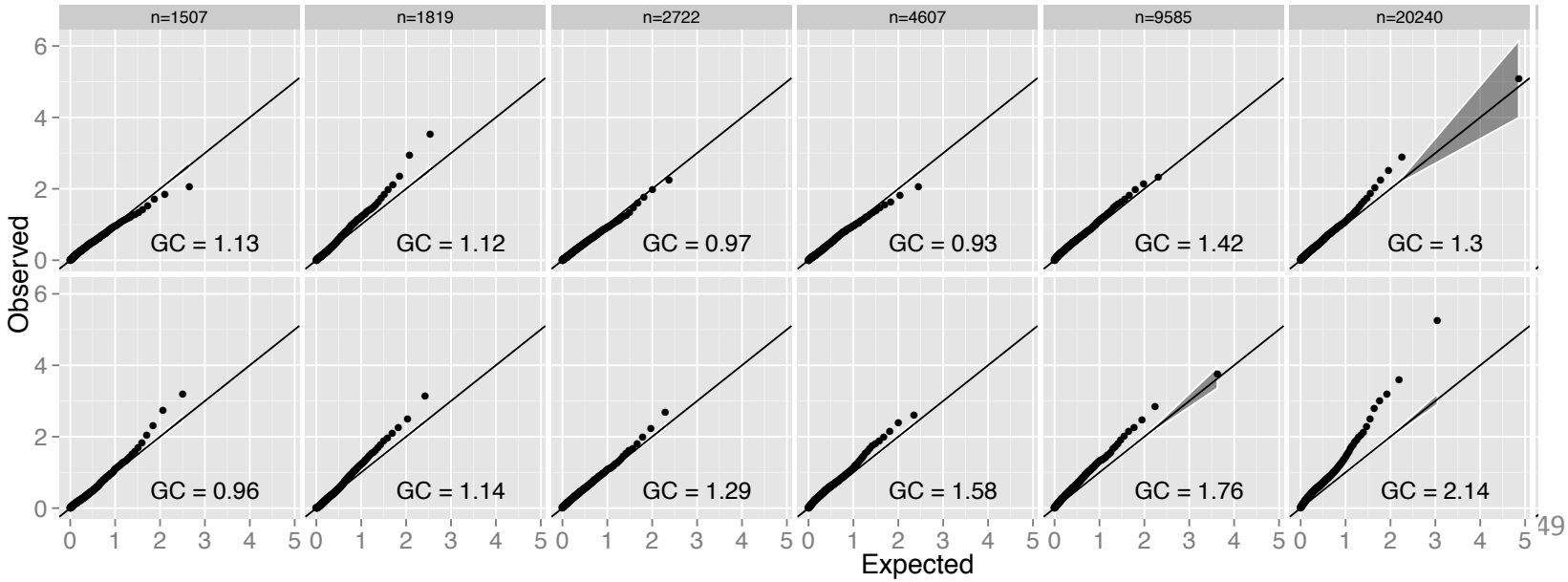
Genetic variation within families using SNP data



$\mu = 0.500$
$\sigma = 0.037$

Heritability estimates from ~20,000 fullsib pairs:

Height	0.7 (SE 0.14)
BMI	0.4 (SE 0.17)



Conclusions

- Empirical variation in genome-wide IBD sharing follows theoretical predictions
- Genetic variance can be estimated from genome-wide IBD within families
  - results for height consistent with estimates from between-relative comparisons
  - no assumptions about nature/nurture causes of family resemblance
- Genetic variance can be partitioned onto chromosomes

Key concepts

1. There is variation in realised relationships given the expected value from the pedigree;
2. Variation in realised relationships can be captured with genetic markers;
3. Variation in realised relationships can be exploited to estimate genetic variation

Estimating relationship from marker genotypes

Mike Goddard

Relationships

We use relationship data

- to estimate genetic variance
- to estimate demographic history
- …

Relationships

Additive genetic relationship G(i, j)

= proportion of the genome in i and j that is IBD

Pedigree relationship A(i,j) = Prob (IBD)

= E(G(i,j))

Actual relationship deviates randomly from this expectation

Relationships

Single locus case, full sibs

Parents A_1A_2 x A_3A_4

offspring A_1A_3
A_1A_4
A_2A_3
A_2A_4

Pairs of sibs share

0 alleles	25% of the time
1 allele	50%
2 alleles	25%

E(G) = A = 0.5 but G varies from 0 to 1

Estimate relationship from markers

G is a more accurate description of relationship than A

G captures unknown pedigree information

pedigree can be incorrect

G captures deviations from A

Therefore, can use G in

Random sample of population (“unrelated individuals”)

Individuals with same pedigree

Estimate relationship from markers

1. Well defined (recent) base
2. No well defined base

3. Well defined, recent base

Eg Data on families of full-sibs and parents of sibs are the base

Estimate relationship from markers

Eg Data on families of full-sibs and parents of sibs are the base

Consider a single SNP

Full sibs can be IBD at either maternal or paternal allele

IBD status		P(IBD status)
Maternal	Paternal	
yes	yes	0.25
yes	no	0.25
no	yes	0.25
no	no	0.25

Estimate relationship from markers

Eg Data on families of full-sibs and parents of sibs are the base

At this SNP, one sib has genotype AA and the other is AB, mother = AB, father = AA

P(IBD status | SNP genotypes)

$$= \frac{P(\text{SNP genotypes} \mid \text{IBD status}) * P(\text{IBD status})}{P(\text{SNP genotypes})}$$

$$= \frac{P(\text{SNP genotypes} \mid \text{IBD status}) * P(\text{IBD status})}{\Sigma\, P(\text{SNP genotypes} \mid \text{IBD status}) * P(\text{IBD status})}$$

Estimate relationship from markers

$$= \frac{P(\text{SNP genotypes} \mid \text{IBD status}) * P(\text{IBD status})}{\Sigma\, P(\text{SNP genotypes} \mid \text{IBD status}) * P(\text{IBD status})}$$

IBD status		P(IBD status)	P(genotypes|IBD status)	P(IBD status | genotypes)	
Maternal	Paternal				G
yes	yes	0.25	0	0	1
yes	no	0.25	0	0	0.5
no	yes	0.25	1	0.5	0.5
no	no	0.25	1	0.5	0

Σ P(SNP genotypes | IBD status)* P(IBD status) = 0.5

E(G) = 0.25 compared with A=0.5

Estimate relationship from markers

1. Well defined, recent base

Eg Data on families of full-sibs and parents of sibs are the base

a) Calculate Bayesian probability of IBD status at each SNP

→ E(G) at each SNP

average over SNPs

b) Use haplotypes ?

Estimate relationship from markers

2. Less well defined, less recent base

Eg Data on current population, base = ancestors 1000 years ago and allele frequencies in base are known (p and q)

Consider haploid gametes of SNP alleles instead of genotypes

What fraction of the gametes are IBD (G)?

At a single SNP, there are 3 possible data sets and their probabilities are

A and A	A and B	B and B
$p^2 + pqG$	$2pq(1-G)$	q^2+pqG

Estimate relationship from markers

SNP genotypes	A and A	A and B	B and B
Probability	p^2 +pqG	2pq(1-G)	q^2+pqG
score (x)	q/p	-1	p/q

Estimate G(i,j) from the mean value of x over SNPs

This is a relationship between gametes. Calculate G for individuals from the 4 gametic relationships.

See Yang et al (2010) and Powell et al (2010) for the diploid formulae.

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W

p(A) = 0.9, q(B) = 0.1

	A	B
	(0.9)	(0.1)
A (0.9)	0.11	-1
B (0.1)	-1	9

Mean G = 0.81 * 0.11 + 0.18 *(-1) + 0.01 *9 = 0

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W

p(A) = 0.9, q(B) = 0.1

AAAAAAAAAAAAAAAAAABB

A and A or A and A B and B

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from same parent

p(A) = 0.9, q(B) = 0.1

Parent	AA	AB	BB
Freq.	0.81	0.18	0.01
	AA (x = 0.11)	AA (0.11)	BB (9)
		AB (-1)	
		BB (9)	

Mean G = 0.81*0.11 + 0.18*(0.25*0.11+0.5*(-1)+0.25*9) + 0.01 *9

= 0.5

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W but after allele frequency has drifted to p(A) = 0.8, q(B) = 0.2

	A	B
	(0.8)	(0.2)
A (0.8)	0.11	-1
B (0.2)	-1	9

Mean G = 0.64 * 0.11 + 0.32 *(-1) + 0.04 *9 = 0.11

Estimate relationship from markers

2. No well defined base

Eg random sample from population but don't know allele frequency in the base.

a) Use the current population as the base

Problem: Some G <0

Cannot interpret as probabilities but still interpret as covariances

If g = genetic value, $V(g) = G V_A$

where G is calculated as above but using allele frequencies in current population.

Estimate relationship from markers

E.g. Score (x) for pairs of gametes from population in H-W but after allele frequency has drifted to p(A) = 0.8, q(B) = 0.2 and using allele frequencies in modern population

	A	B
	(0.8)	(0.2)
A (0.8)	0.25	-1
B (0.2)	-1	4

Mean G = 0.64 * 0.25 + 0.32 *(-1) + 0.04 *4 = 0

Estimate relationship from markers

2. No well defined base

b) Assume SNPs are a random sample of loci as are QTL

$y = mean + g + e$

$y = mean + Zu + e$

Z_{ij} = 0 for AA, 1 for AB or 2 for BB

$u \sim N(0, I\sigma_u^2) \rightarrow g = Zu \sim N(0, ZZ'\sigma_u^2)$, $ZZ'\sigma_u^2 = G\sigma_g^2$, if $\sigma_g^2 = N\sigma_u^2$

where $N = \Sigma 2pq$ across SNPs

Therefore, $G = ZZ'/N$

Estimate relationship from markers

E.g. Score for pairs of gametes from population in H-W

p(A) = 0.8, q(B) = 0.2

		A	B
		(0.8)	(0.2)
	z	0	1
A (0.8)	0	0	0
B (0.2)	1	0	1

Mean G = 0.04 *1 = 0.04

Estimate relationship from markers

E.g. Score for pairs of gametes from population in H-W

p(A) = 0.8, q(B) = 0.2

		A	B
		(0.8)	(0.2)
	z	-0.2	0.8
A (0.8)	-0.2	0.04	-0.16
B (0.2)	0.8	-0.16	0.64

Mean G = 0.64*0.04 + 0.32 * (-0.16) + 0.04 *0.64 = 0

Comparing 2a and 2b

E.g. p(A) = 0.8, q(B) = 0.2

		2b			2a	
		A	B		A	B
		(0.8)	(0.2)			
	z	-0.2	0.8			
A (0.8)	-0.2	0.04	-0.16	A	0.25	-1
B (0.2)	0.8	-0.16	0.64	B	-1	4

Estimate relationship from markers

2a and 2b compared for gametic relationships

SNP data	A and A	A and B	B and B
score (x)	q/p	-1	p/q
weight (w)	pq	pq	pq

2a) G = mean of x

2b) G = weighted mean of x = $\Sigma wx/\Sigma w$

This could be described as using the IBS status of SNPs instead of IBD

Estimate relationship from markers

E.g. Score (x i.e. method 2a) for pairs of gametes p(A) = 0.8, q(B) = 0.2 and weighting by pq = 0.16

	A	B
	(0.8)	(0.2)
A (0.8)	0.25*0.16	-1*0.16
	=0.04	= -0.16
B (0.2)	-1*0.16	4*0.16
	= -0.16	= 0.64

Same as 2b

Estimate relationship from markers

2a) G = mean of x

gives more emphasis to sharing rare alleles

Makes sense because individuals who share rare alleles are more likely to be closely related than individuals who share common alleles.

Gives minimum error variance of relationship under some conditions

Estimate relationship from markers

2. No well defined base

c) Assume SNPs are a random sample of loci as are QTL but effect of SNP decreases as heterozygosity increases

$y = mean + g + e$

$y = mean + Zu + e$

Z_{ij} = 0 for AA, 1 for AB or 2 for BB

$u \sim N(0, D\sigma_u^2) \rightarrow g = Zu \sim N(0, ZDZ'\sigma_u^2)$, $ZDZ'\sigma_u^2 = G\sigma_g^2$, if $\sigma_g^2 = N\sigma_u^2$

where $N = \Sigma(p_i q_i)$

Therefore, $G = ZDZ'/N$

$D_{ii} = 1/(p_i q_i)$

That is, assume the effect of SNPs is proportional to $\sqrt{(p_i q_i)}$

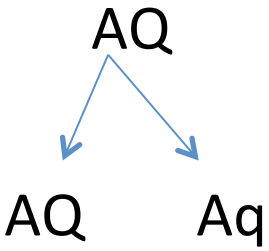
So variance explained by SNPs is not affected by allele frequency

2c = 2a

Estimate relationship from markers

Relationship depends on the markers or QTL

Eg QTL are due to recent mutations



Marker is the same but QTL is different

Rare SNP alleles tend to be a recent mutation

Therefore, treat SNPs differently according to MAF

Estimate relationship from markers

Relationship depends on the markers or QTL

Therefore, treat SNPs differently according to MAF

$y = mean + g1 + g2 + g3 + g4 + g5 + e$

$V(g_i) = (ZZ'/N)\sigma_i^2$ for SNPs in MAF bin i

Estimate relationship from markers

Use haplotypes of markers

New definition of IBD for chromosome segments

Two segments are IBD if they coalesce without recombination

Avoids definition of a base population

Chromosome segment homozygosity (CSH)

= P(2 segments are IBD)

$E(csh) = 1/(1+4N_e c)$

Estimate relationship from markers

Problem: cant observe CSH directly

only observe haplotype homozygosity (HH)

or runs of homozygosity (ROH)

Estimate relationship from markers

Can use HH or ROH in QTL mapping

Additive effects

Calculate P(QTL in position x is IBD) = P(csh for surrounding chr)

Eg P(QTL IBD) = 0.9 if in middle of 10 identical markers

Recessive effects

ROH within individual → homozygous QTL within the run

Estimate relationship from markers

Recessive effects

ROH within individual → homozygous QTL within the run

Detect embryonic lethals by missing ROH

Estimate relationship from markers Summary

1. In families

2. In the general population

    Express relationship relative to current population

        G can be negative

        G is not a probability

        $V(g) = G \sigma_g^2$

    two formulae (2a and 2b)

    Same except 2a gives more weight to rare alleles

(Genome-wide) association analysis

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

Key concepts

- Mapping QTL by association relies on linkage disequilibrium in the population;
- LD can be caused by close linkage between a QTL and marker (= good) or by confounding between a marker and other effects (= usually bad);
- The power of QTL detection by LD depends on the proportion of phenotypic variance explained at a marker;
- Mixed models are good for performing GWAS
- Genetic (co)variance can be estimated from GWAS summary statistics

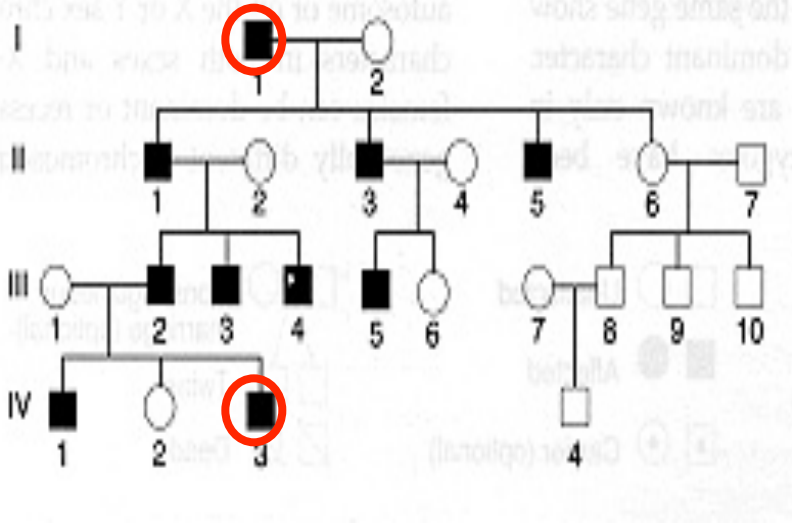
Outline

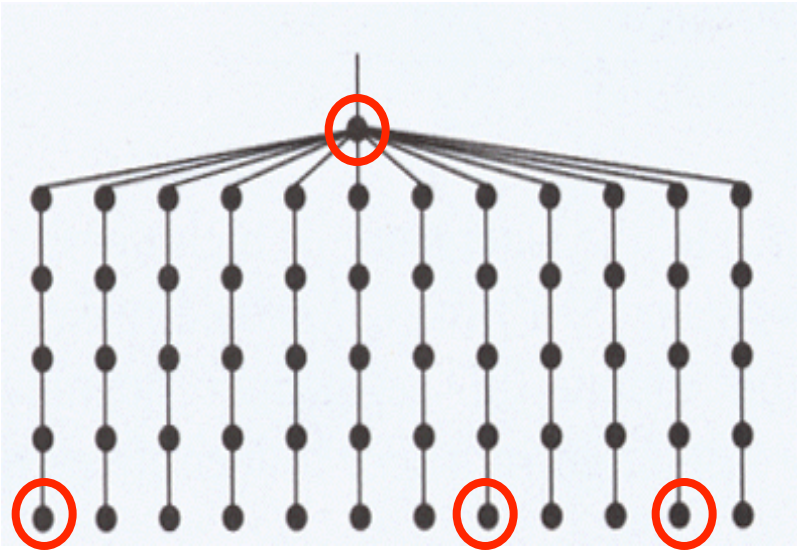
- Association vs linkage
- Linkage disequilibrium
- Analysis: single SNP

- GWAS: design, power
- GWAS: analysis

Linkage

Association

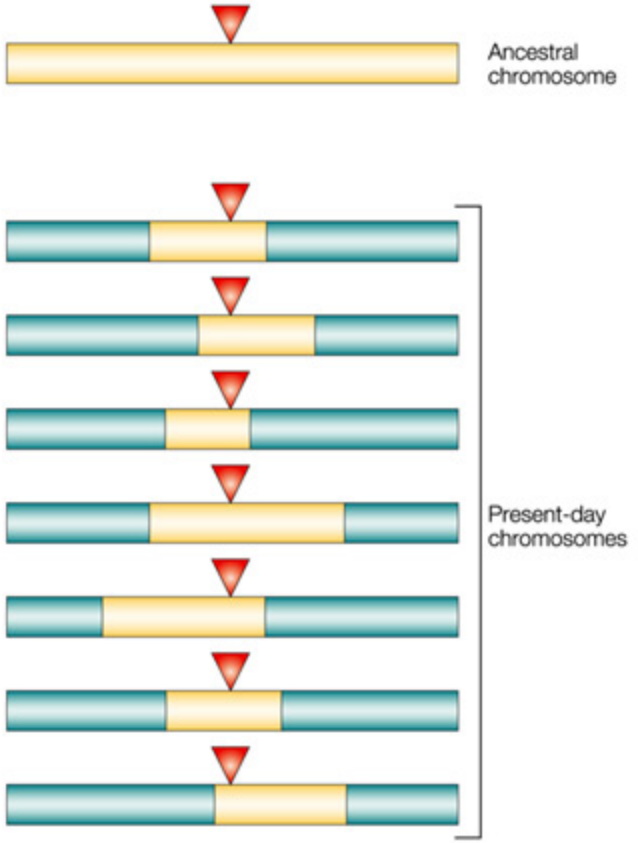




Families

Populations

Linkage disequilibrium around an ancestral mutation



[Ardlie et al. 2002]

LD

- Non-random association between alleles at different loci
- Many possible causes
  - mutation
  - drift / inbreeding / founder effects
  - population stratification
  - selection
- Broken down by recombination

Definition of D

- 2 bi-allelic loci
  - Locus 1, alleles A & a, with freq. p and (1-p)
  - Locus 2, alleles B & b with freq. q and (1-q)
  - Haplotype frequencies p_{AB}, p_{Ab}, p_{aB}, p_{ab}

$D = p_{AB} - pq$

r^2

$r^2 = D^2 / [pq(1-p)(1-q)]$

- Squared correlation between presence and absence of the alleles in the population
- 'Nice' statistical properties

[Hill and Robertson 1968]

Properties of r and r^2

- Population in 'equilibrium'

  $E(r) = 0$

  $E(r^2) = var(r) \approx 1/[1 + 4Nc] + 1/n$

  N = effective population size

  n = sample size (haplotypes)

  c = recombination rate

LD depends on population size and recombination distance

- $nr^2 \sim \chi_{(1)}^2$
- Human population is NOT in equilibrium

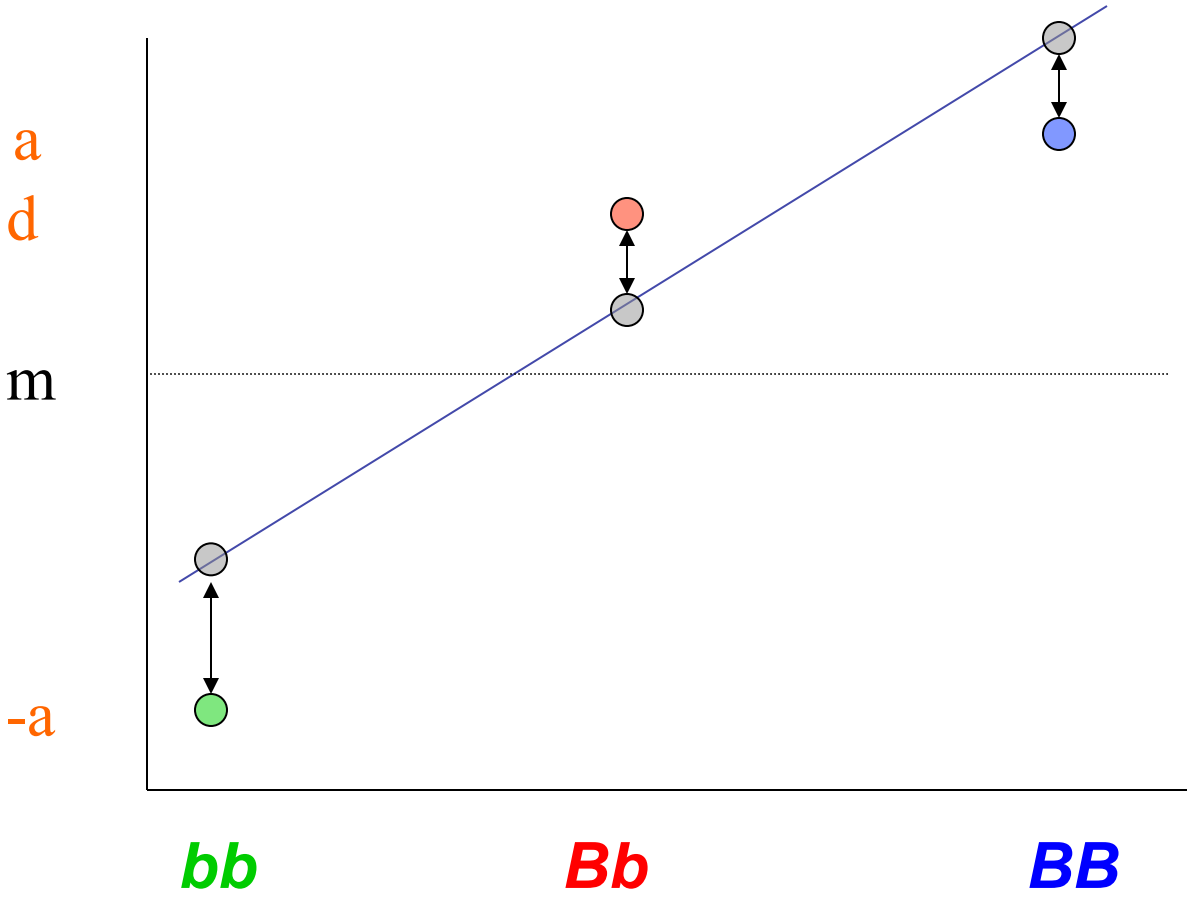
[Sved 1971; Weir and Hill 1980]

Analysis

- Single locus association
- GWAS

- Least squares
- ML
- Bayesian methods

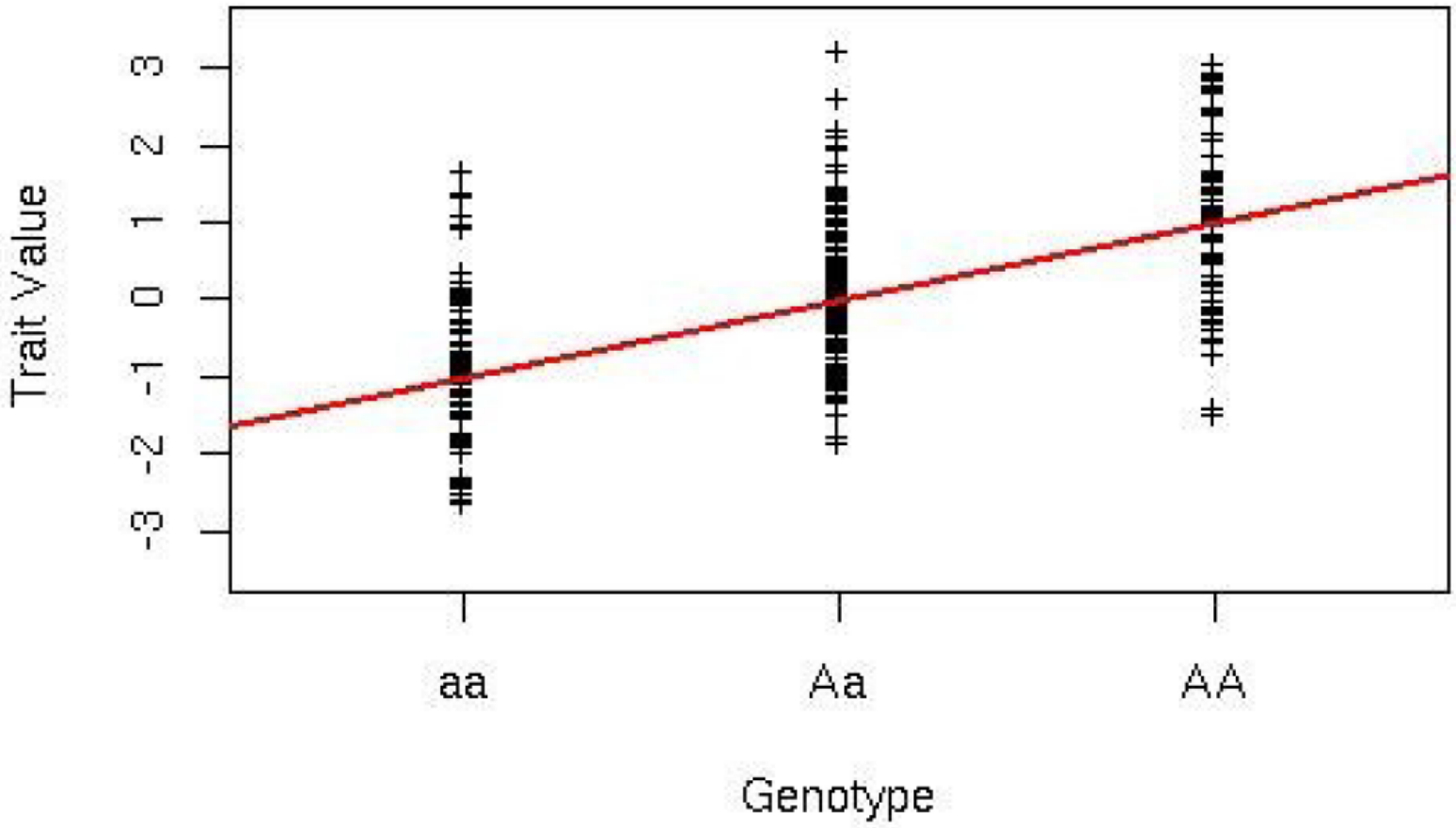
Falconer model for single biallelic QTL



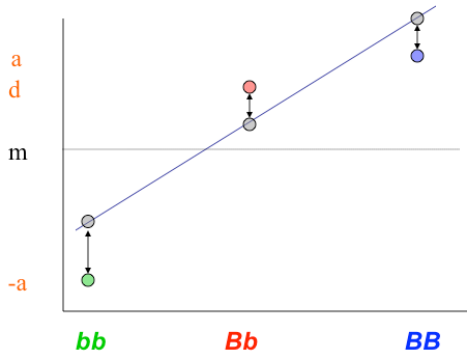
Var (X) = Regression Variance + Residual Variance
= Additive Variance + Dominance Variance

Unrelated Samples

$$\hat{y}_i = \mu + \hat{\beta}\, x_i$$



Statistical power (linear regression)



$y = \mu + \beta * x + e, \quad x = 0, 1, 2$

$\sigma_y^2 = \sigma_q^2 + \sigma_e^2$ regression + residual

$\sigma_x^2 = 2p(1-p)$ p = allele frequency for indicator x

{HWE: note x is usually considered fixed in regression}

$\sigma_q^2 = \beta^2 \sigma_x^2 = [a + d(1-2p)]^2 * 2p(1-p)$

$q^2 = \sigma_q^2 / \sigma_y^2$ {QTL heritability}

Statistical Power

χ^2 test with 1 df:

$E(X^2) = 1 + n R^2 / (1 - R^2)$

$= 1 + nq^2/(1-q^2)$

$= 1 + NCP$

NCP = non-centrality parameter

Power of association proportional to q^2
(Power of linkage proportional to q^4)

Statistical Power (*R*)

```
alpha= 5e-8
threshold= qchisq(1-alpha,1)
q2= 0.005
n= 10000
ncp= n*q2/(1-q2)
power= 1-pchisq(threshold,1,ncp)
threshold
ncp
power
```

```
> alpha= 5e-8
> threshold= qchisq(1-alpha,1)
> q2= 0.005
> n= 10000
> ncp= n*q2/(1-q2)
> power= 1-pchisq(threshold,1,ncp)
> threshold
[1] 29.71679
> ncp
[1] 50.25126
> power
[1] 0.9492371
```

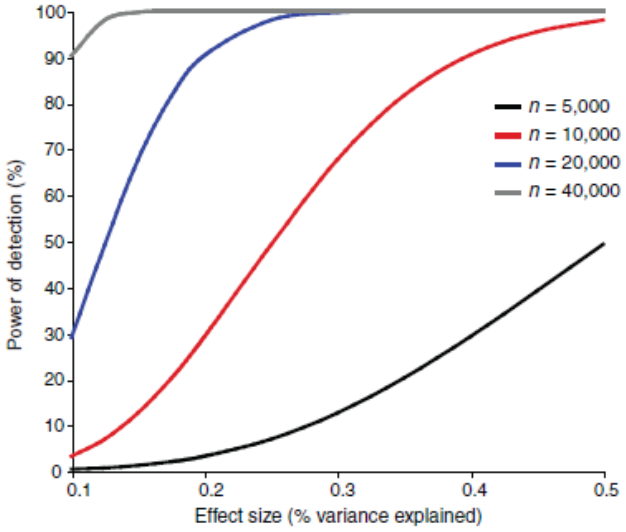


Figure 1 Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5×10^{-7} (calculated using the Genetic Power Calculator[15]). Shown is the power to detect a variant with a given effect size, assuming this type I error rate, which is typical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

Power by association with SNP

(small effect; HWE)

$NCP(SNP) = n\ r^2\ q^2$

$= r^2 * NCP(\text{causal variant})$

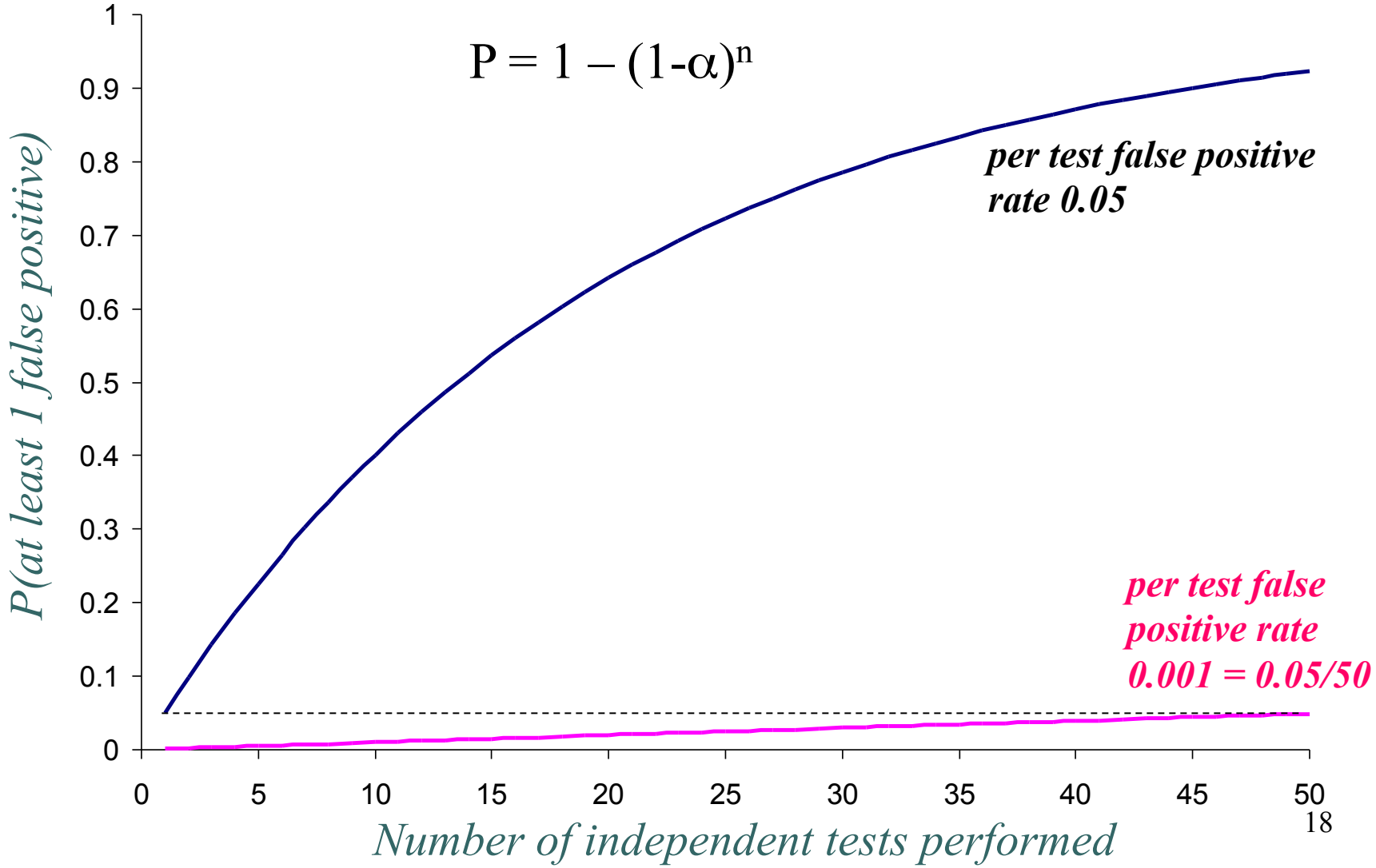
$= n * \{r^2\ q^2\} = n * (\text{variance explained by SNP})$

Power of LD mapping depends on the experimental sample size, variance explained by the causal variant and LD with a genotyped SNP

GWAS

- Same principle as single locus association, but additional information
  - QC
    - Duplications, sample swaps, contamination
  - Power of multi-locus data
    - Unbiased genome-wide association
    - Relatedness
    - Population structure
    - Ancestry
    - More powerful statistical analyses

The multiple testing burden



Population stratification (association unlinked genes)

Both populations are in linkage equilibrium; genes unlinked

	Allele frequency		Haplotype frequency			
	p_{A1}	p_{B1}	p_{A1B1}	p_{A1B2}	p_{A2B1}	p_{A2B2}
Pop. 1	0.9	0.9	0.81	0.09	0.09	0.01
Pop. 2	0.1	0.1	0.01	0.09	0.09	0.81
Average	0.5	0.5	0.41	0.09	0.09	0.41

Combined population: $D = 0.16$ and $r^2 = 0.41$

Population stratification (genes and phenotypes)

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not. His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes.

The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use, enough to account for nearly half of the observed variance. When the experiment was repeated with students from a different university, precisely the same marker lit up. Eureka! The delighted scientist popped a bottle of champagne and quickly submitted an article to *Molecular Psychiatry* heralding the discovery of the 'successful-use-of-selected-hand-instruments gene' (SUSHI).

[Hamer & Sirota 2010 Mol Psych]

Population stratification (genes and phenotypes)

It took another 2 years to discover that SUSHI is a histocompatibility antigen gene that has nothing to do with chopstick use but just happens to have different allele frequencies in Asians and Caucasians, who of course differ in chopstick use for purely cultural rather than biological reasons. Even though the association data were highly significant and readily replicated, they were biologically meaningless.

[Hamer & Sirota 2010 Mol Psych]

Population stratification (genes and phenotypes)

The source of confounding in the chopstick example is better thought of as the environment. The problem arises because different subgroups have different levels of exposure to chopsticks. This type of confounding is extremely familiar to genetic epidemiologists, but it is unimportant in settings where the environment can be experimentally controlled or randomized (as is routinely done in plant breeding, for example).

There is another source of confounding, however, and that is the genetic background. The estimate of the effect of a particular locus can be confounded by the other causal loci in the genome. This genetic background effect will always be present to some extent, even

[Vilhjalmsson & Nordborg 2013 Nature Reviews Genetics]

Demonstrating stratification in a European American population

Catarina D Campbell[1,2], Elizabeth L Ogburn[1], Kathryn L Lunetta[3,8], Helen N Lyon[1,2], Matthew L Freedman[4–6], Leif C Groop[7], David Altshuler[2,4,5], Kristin G Ardlie[3] & Joel N Hirschhorn[1,2,4]

Table 2 No evidence for stratification using standard methods

		χ^2 values[a]		Estimates of stratification parameters[b]		
	SNPs	Median	Mean	λ_{max}	λ	*P*
Random SNPs	111	0.37	0.96	3.21	1	0.61
AIMs	67	0.58	0.95	–	–	0.61
Total	178	0.49	0.95	–	–	0.66

Table 3 A strong association of *LCT* –13910C→T and height is reduced by rematching subjects on the basis of ancestry

			Origin of grandparents[a]			
		All	Four US-born	Southeastern	Northwestern	Combined[b]
N	Total	2,179	1,282	354	543	–
	Tall	1,123	645	127	351	–
	Short	1,056	637	227	192	–
LCT –13910 genotype counts[c]	Total	392:918:869	142:543:596	182:141:31	68:233:243	–
	Tall	161:474:489	66:265:314	54:55:18	41:154:157	–
	Short	231:444:380	76:278:282	128:86:13	27:79:86	–
Hardy-Weinberg *P*	Total	5.6×10^{-7}	0.57	0.89	0.89	–
	Tall	0.03	0.66	0.81	0.92	–
	Short	2.5×10^{-5}	0.86	0.96	0.45	–
Association *P*		3.6×10^{-7}	0.098	0.0016	0.71	0.0074
OR (95% c.i.)[d]		1.37 (1.22–1.54)	1.15 (0.97–1.36)	1.70 (1.22–2.38)	1.05 (0.81–1.37)	1.19 (1.05–1.36)

Table 4 No association of *LCT* –13910C/T and height in other European populations

		Polish	Scandinavian	Combined
Genotypes (CC:CT:TT)	Tall	166:251:86	–	–
	Short	174:235:96	–	–
Transmissions of T allele (T:U)[a]	Tall	–	65:68	–
	Short	–	76:66	–
P		0.92	0.43	0.58
OR (95% c.i.)[b]		0.99 (0.83–1.18)	0.91 (0.72–1.15)	0.96 (0.83–1.11)

Stratification

$y = \Sigma g_i + \Sigma e_i$

$r(y,g_i)$ due to

- causal association with g_i
- correlation g_i and g_j and causal association with g_j (LTC and height)
- correlation g_i and environmental factor e_j (chopsticks)

How to deal with structure?

- Detect and discard ‘outliers’
- Detect, analysis and adjustment
  - E.g. genomic control
- Account for structure during analysis
  - Fit a few principal components as covariates
  - Fit GRM

GWAS using mixed linear models

$y = \mathbf{Xb} + \beta * x + \mathbf{g} + \mathbf{e}$

$var(g) = \mathbf{G}\sigma_g^2$

G = genetic relationship matrix (GRM)

Model conditions on effects of all other variants

Power depends on whether x is included (MLMi) or excluded (MLMe) from the construction of **G**.

[Yang et al. 2014 Nature Genetics]

GWAS using mixed linear models: statistical power

For linear regression (LR), the expected mean of χ^2 association statistics (λ_{mean}) is

$$\lambda_{mean}\,(\mathrm{LR}) = 1 + Nh_g^2 / M \qquad (1)$$

regardless of the genetic architecture of the trait[24].

For MLMi, the λ_{mean} value at markers used to construct the GRM is

$$\lambda_{mean}\,(\mathrm{MLMi}) = 1 \qquad (2)$$

Equation (2) highlights the dangers of using λ_{mean} (or λ_{median}) to assess the presence of population stratification or other artifacts. A researcher who observes lower λ_{mean} (or λ_{median}) values for MLMi than for linear regression might conclude that this difference is due to correction for confounding, but this result is in fact expected, even in the absence of any confounding.

Finally, for MLMe,

$$\lambda_{mean}\,(\mathrm{MLMe}) = 1 + \frac{Nh_g^2 M}{1 - r^2 h_g^2} \qquad (3)$$

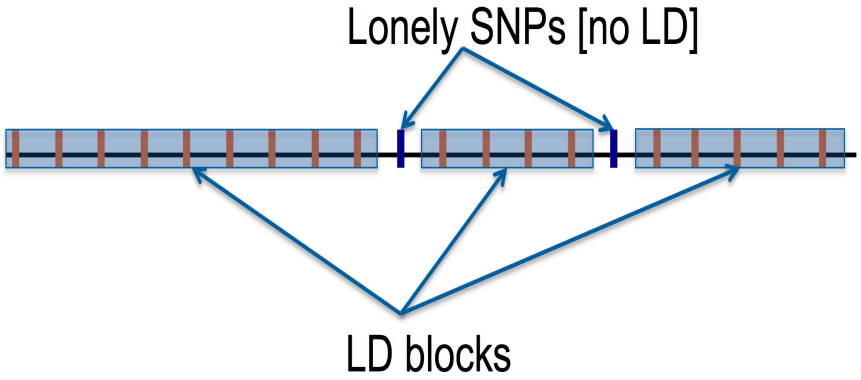
r^2 here is the squared correlation between **g**-hat and **g**

[Yang et al. 2014 Nature Genetics]

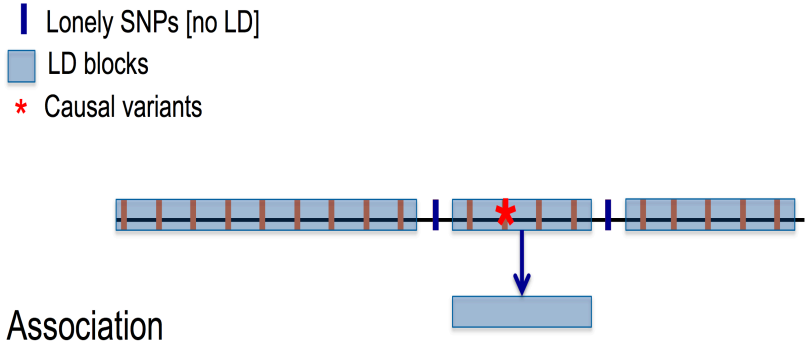
How does LD shape association

A set of markers along a chromosome region:

Superimpose LD between markers



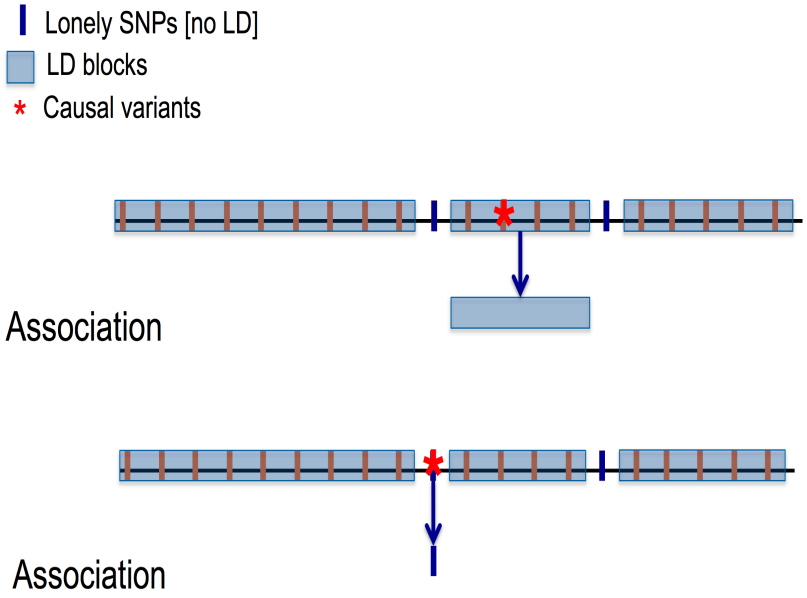
Consider causal SNPs



All markers correlated with a causal variant show association

How does LD shape association

Consider causal SNPs



All markers correlated with a causal variant show association.
Lonely SNPs only show association if they are causal
The more you tag the more likely you are to tag a causal variant

Assuming all SNPs gave an equal probability of association given LD status,
we expect to see more association for SNPs with more LD friends.
This is a reasonable assumption under a polygenic genetic architecture

LD score regression

$$l_j = \sum_{k \neq j} r_{jk}^2$$

Quantifies local LD for SNP j

$$E[\chi^2 \mid \ell_j] = Nh^2\ell_j / M + Na + 1$$

Test statistic is linear in LD score

→ regression of test statistic on LD score provides an estimate of SNP heritability

Use GWAS summary statistics and reference sample for LD score estimation

[Yang 2011 EJHG; Bulik-Sullivan 2015]

Same principle for genetic covariance

$$E[z_{1j}z_{2}\ell_j] = \frac{\sqrt{N_1N_2}\,\varrho_g}{M}\ell_j + \frac{\varrho N_s}{\sqrt{N_1N_2}}$$

N_s is the number of overlapping samples

z = test statistics from GWAS summary statistics
N = sample size
M = number of markers
ρ_g = genetic covariance between traits
ρ = phenotypic correlation between traits

[Bulik-Sullivan 2015 Nature Genetics]

Key concepts

- Mapping QTL by association relies on linkage disequilibrium in the population;
- LD can be caused by close linkage between a QTL and marker (= good) or by confounding between a marker and other effects (= usually bad);
- The power of QTL detection by LD depends on the proportion of phenotypic variance explained at a marker;
- Mixed models are good for performing GWAS
- Genetic (co)variance can be estimated from GWAS summary statistics

Estimation of quantitative genetic parameters from distant relatives using marker data

Peter M. Visscher

peter.visscher@uq.edu.au

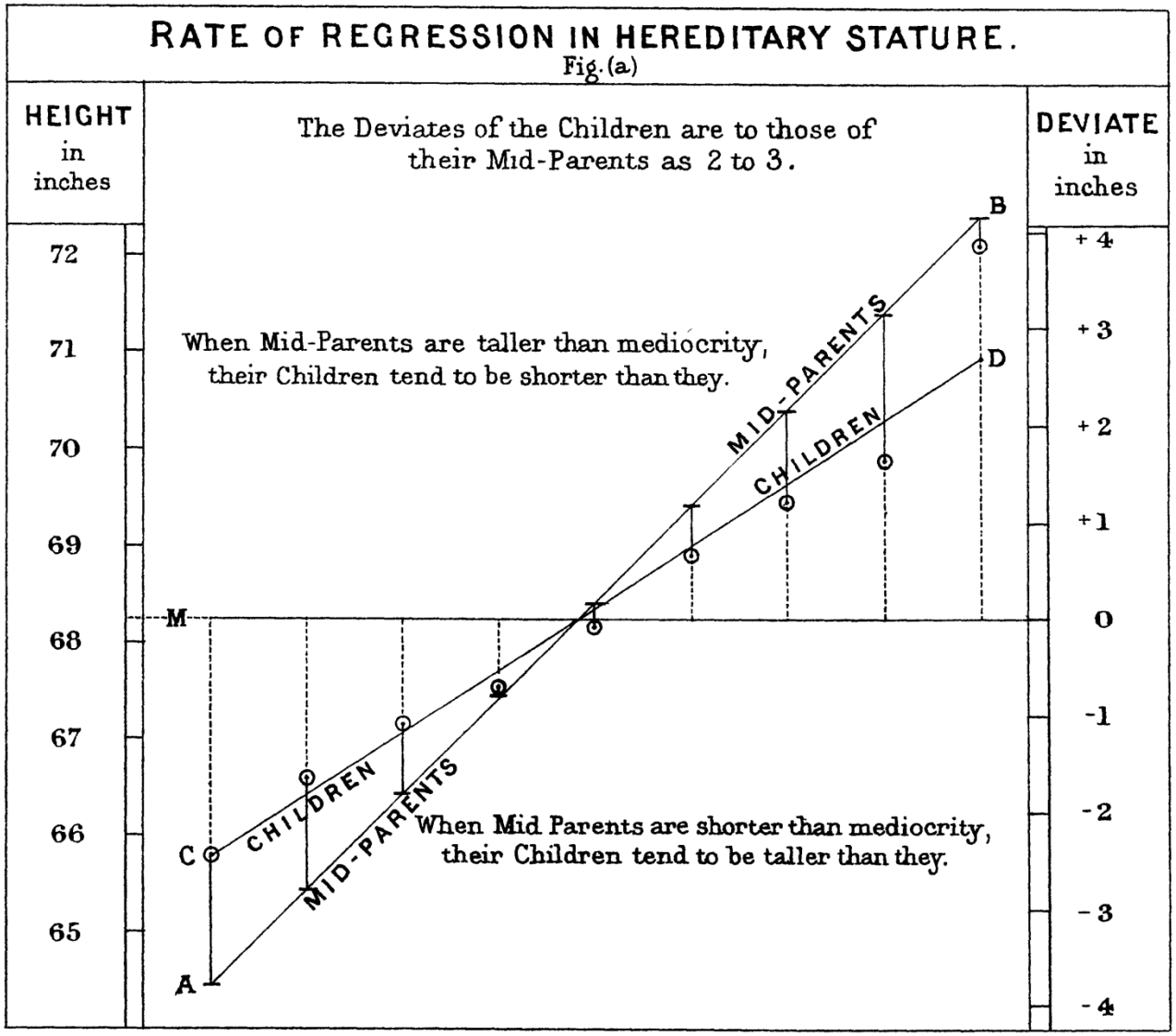
Key concepts

- 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

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 IV. *Distribution of Stature.*

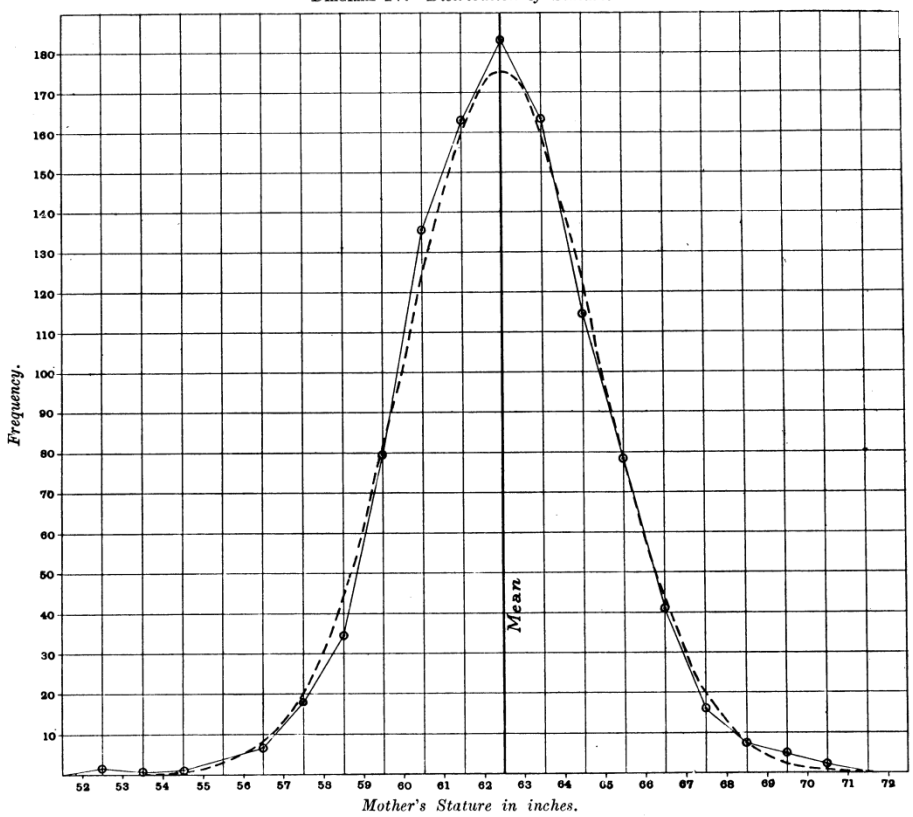
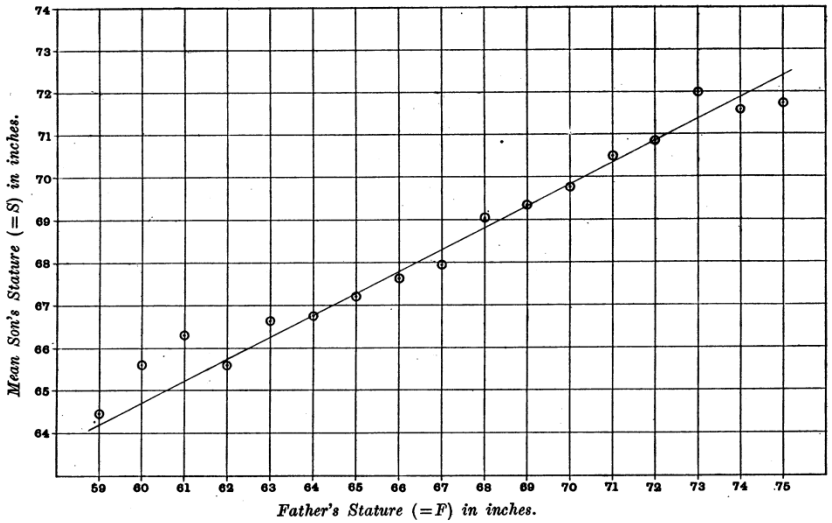


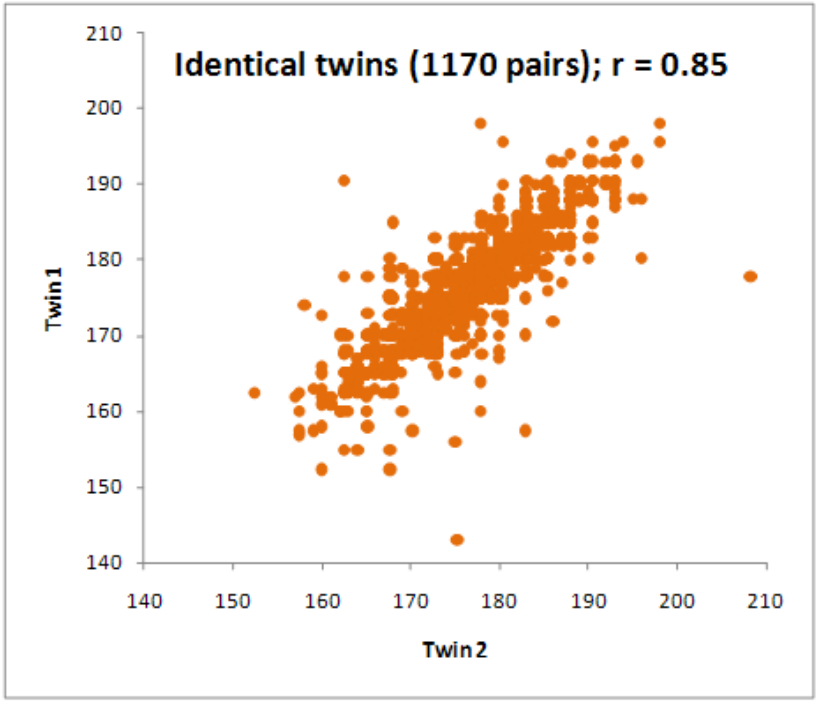
DIAGRAM I. *Probable Stature of Son for given Father's Stature.*

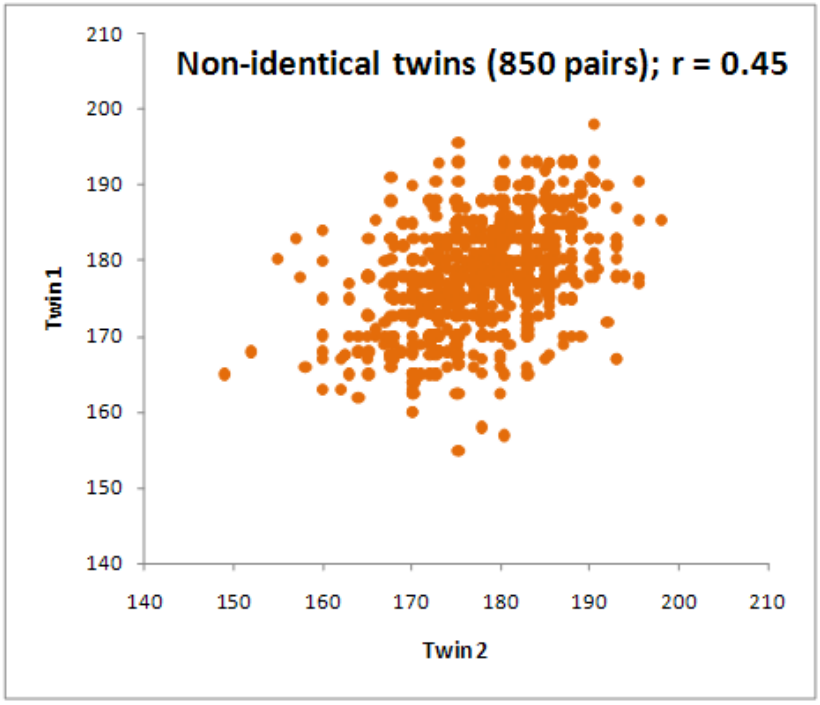
Regression Line: $S = 33{\cdot}73 + {\cdot}516\,F$. 1078 Cases.



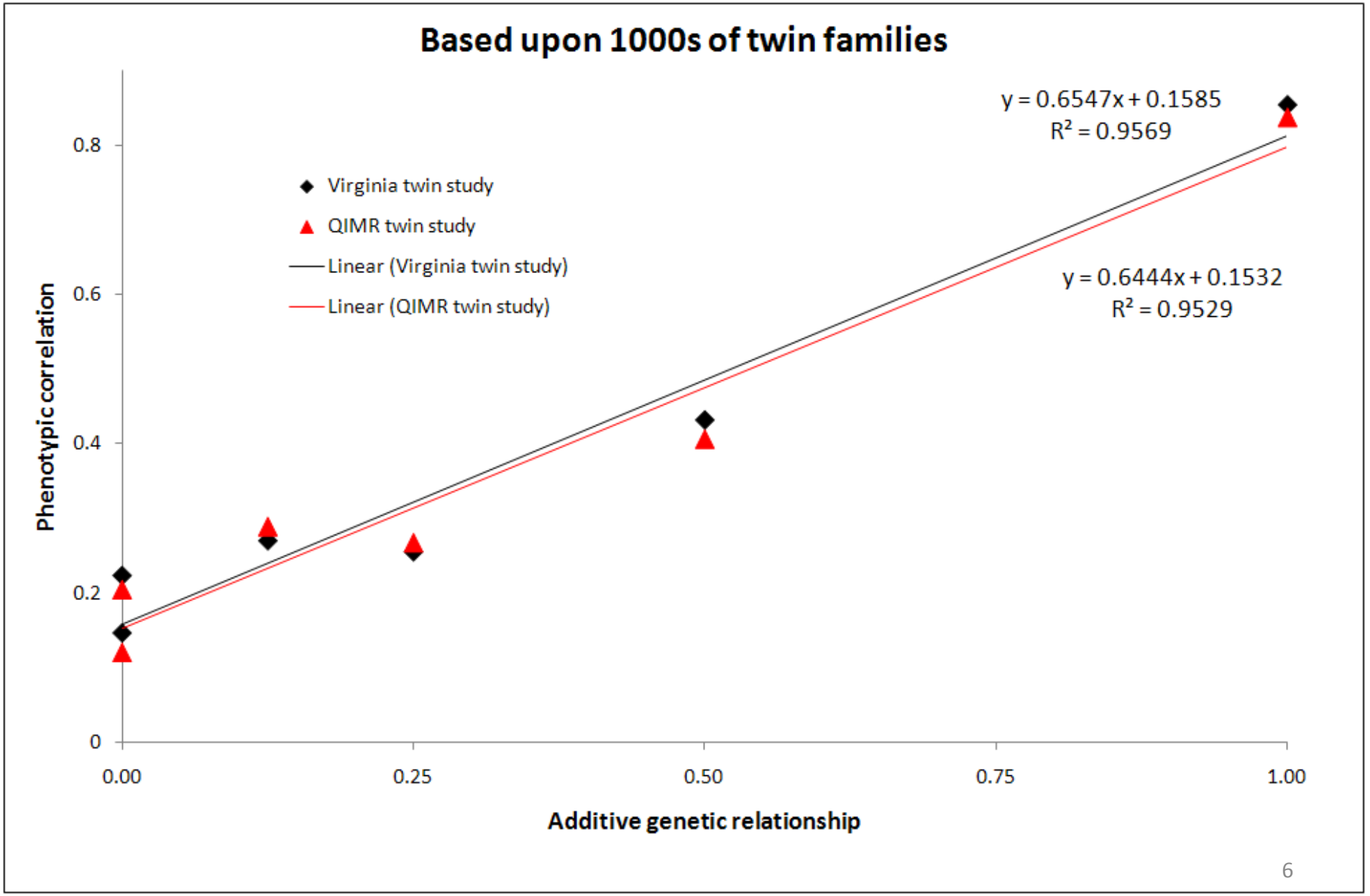
PAIR	CORRELATION	SE
Spouse	0.28	0.02
Son-Father	0.51	0.02
Daughter-Father	0.51	0.01
Son-Mother	0.49	0.02
Daughter-Mother	0.51	0.01
Brother-brother	0.51	0.03
Sister-sister	0.54	0.02
Brother-sister	0.55	0.01

100 years later
Heritability of human height





$h^2 \sim 80\%$


Based upon 1000s of twin families
y = 0.6547x + 0.1585
R² = 0.9569
y = 0.6444x + 0.1532
R² = 0.9529
Virginia twin study
QIMR twin study
Linear (Virginia twin study)
Linear (QIMR twin study)
Phenotypic correlation
0
0.2
0.4
0.6
0.8
0.00
0.25
0.50
0.75
1.00
Additive genetic relationship

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*]

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?

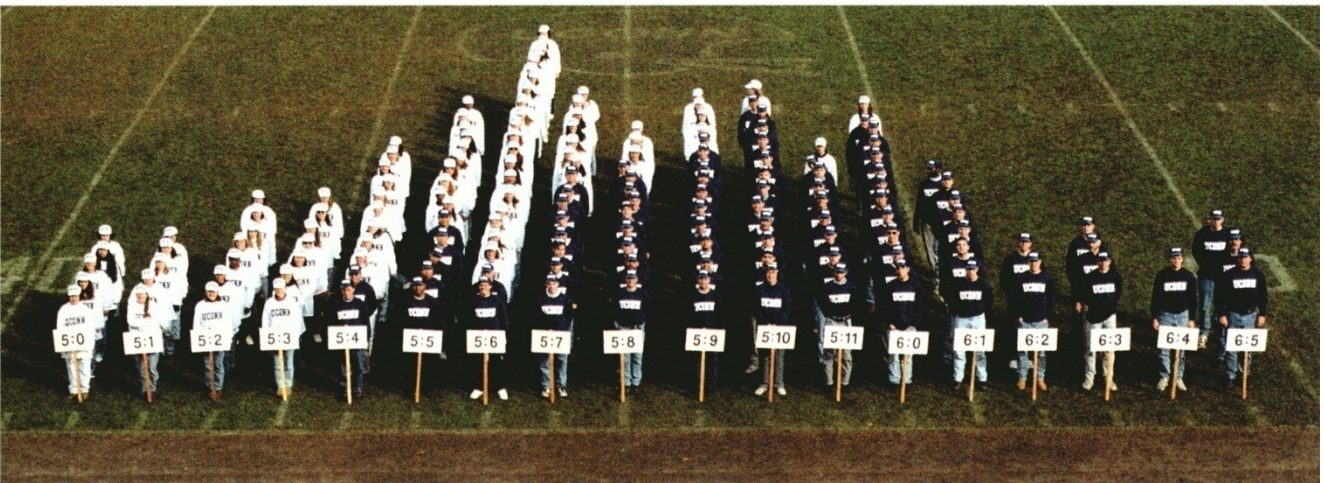
ANALYSIS

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]

Are very distant relatives that share more of their genome by descent phenotypically more similar than those that share less?



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

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

$$= \mu + g_j + e_j$$

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

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

Equivalence

Let u be a random variable, $u \sim N(0, \sigma_u^2)$

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

$$\text{var}(\mathbf{y}) = \mathbf{ZZ'}\,\sigma_u^2 + \mathbf{I}\sigma_e^2$$
$$= \mathbf{ZZ'}\,(\sigma_g^2/m) + \mathbf{I}\sigma_e^2$$
$$= \mathbf{G}\,\sigma_g^2 + \mathbf{I}\sigma_e^2$$

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

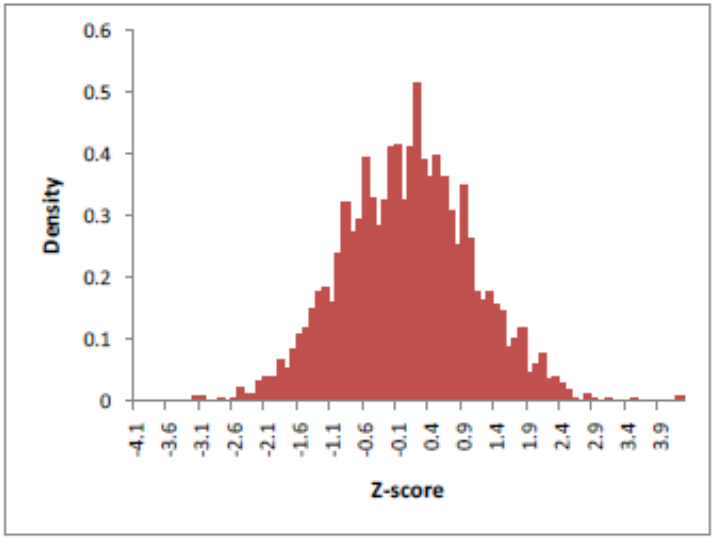
Let **A** be the estimate of **G** from N SNPs:

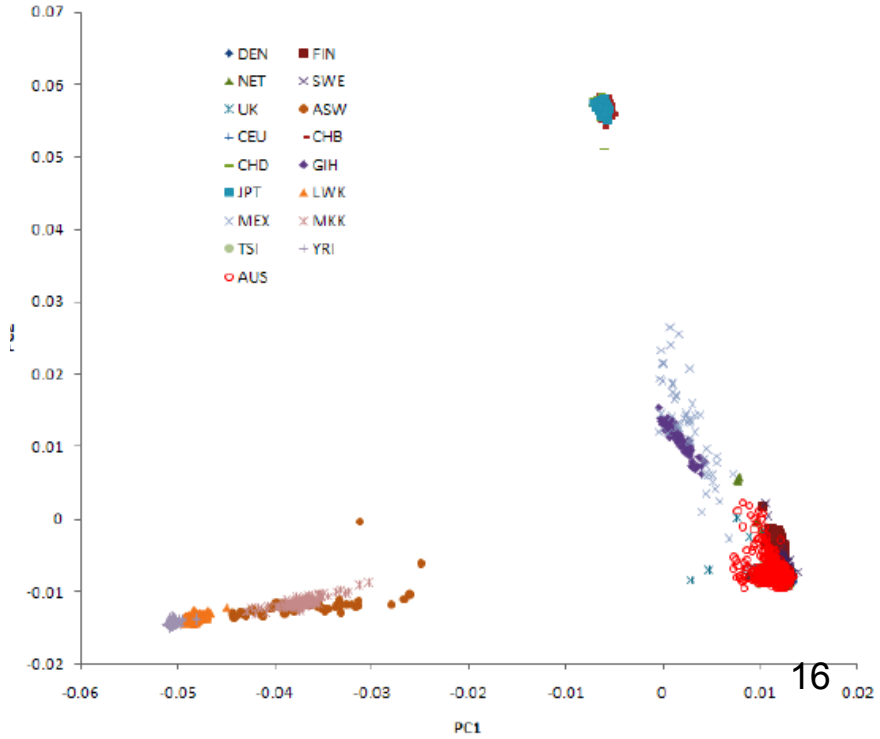
$$A_{jk} = (1/N) \Sigma \{ x_{ij} - 2p_i)(x_{ik} - 2p_i) / \{2p_i(1-p_i)\}$$

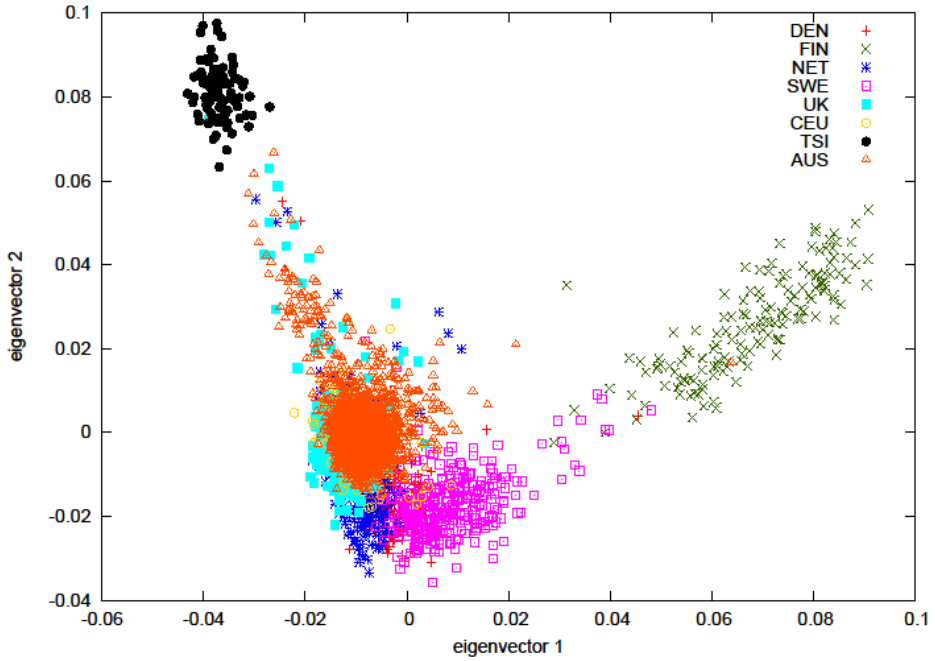
$$= (1/N) \Sigma z_{ij} z_{ik}$$

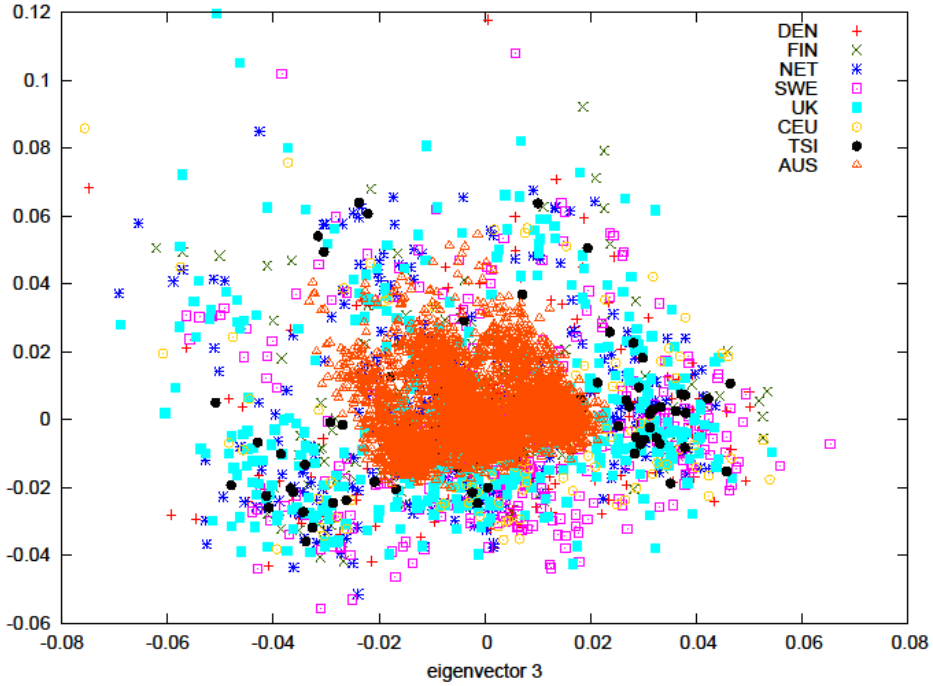
Data

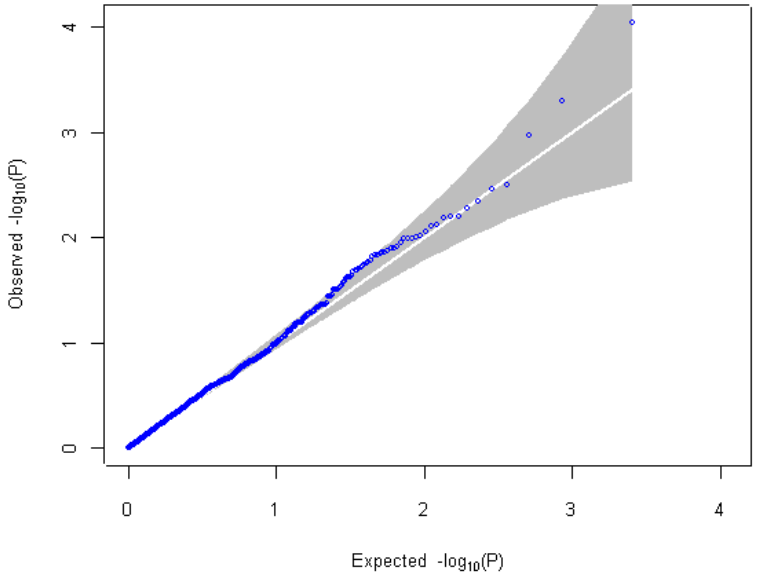
- ~4000 ‘unrelated’ individuals
- Ancestry ~British Isles
- Measurement on height (self-report or clinically measured)
- GWAS on 300k (‘adults’) or 600k (16-year olds) SNPs











Lack of evidence for population stratification within the Australian sample

Methods

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

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

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

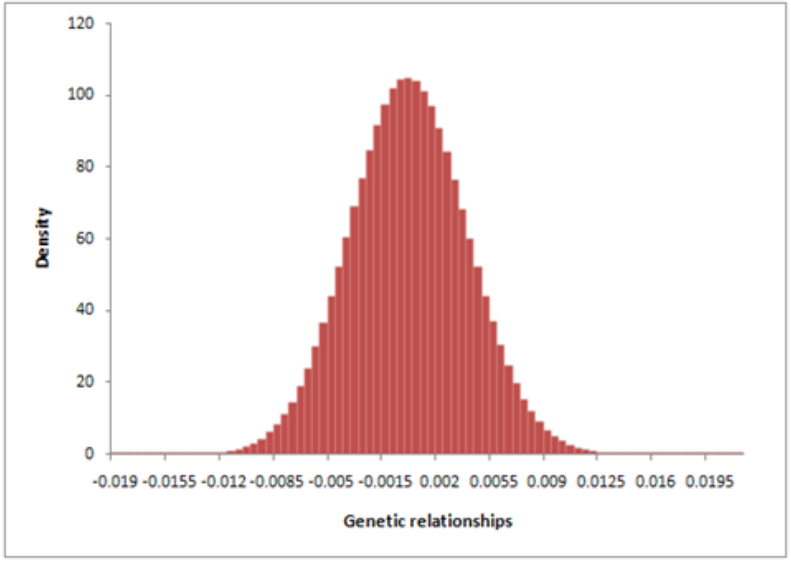
$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$$

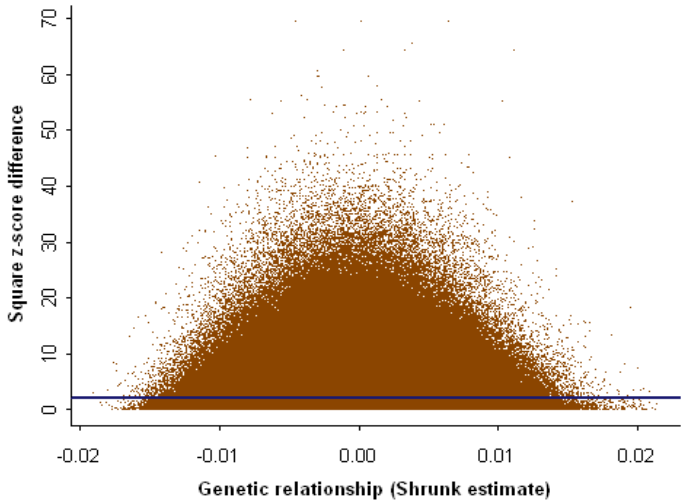
y standardised ~N(0,1)

No fixed effects other than mean

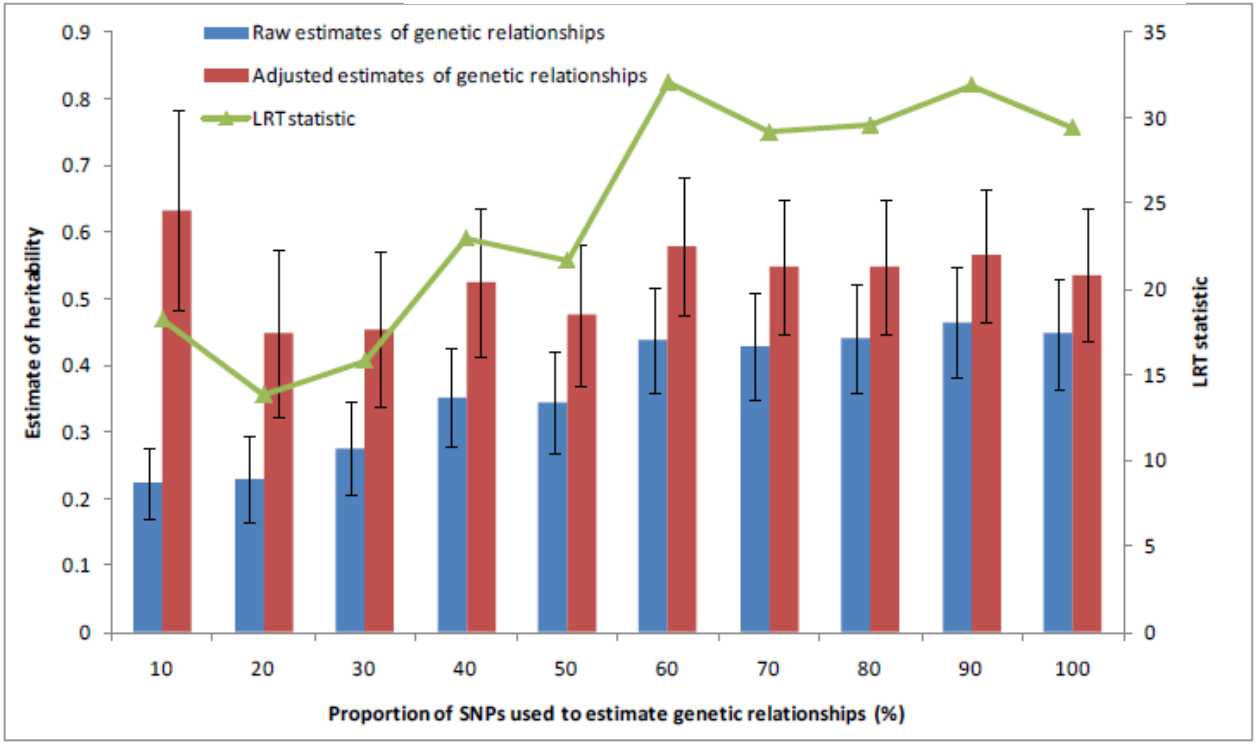
A estimated from SNPs

Residual maximum likelihood (REML)





$h^2 \sim 0.5$ (SE 0.1)



Checking for population structure

Table 1

Estimates of the Variance Explained by the SNPs on Even Chromosomes from 10 Simulation Replicates

Replicate	h^2	SE
1	0.045	0.055
2	0.025	0.057
3	0.0	0.058
4	0.0	0.057
5	0.0	0.059
6	0.0	0.056
7	0.057	0.056
8	0.0	0.062
9	0.0	0.057
10	0.0	0.054

Note: A total of 1,000 causal variants were simulated on the odd chromosomes, with a total heritability of 0.8. Genetic variance was estimated from a relationship matrix constructed from all SNPs on the even chromosomes. The same genotypes were used as in Yang et al. (2010). If there is population structure then estimated relatedness on the even chromosomes is correlated with relatedness on the odd chromosomes (where the causal variants are simulated) and therefore genetic variance will be associated with the even chromosomes.

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]

Data from the GENEVA Consortium

- Investigators: Bruce Weir, Teri Manolio and many others
- 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

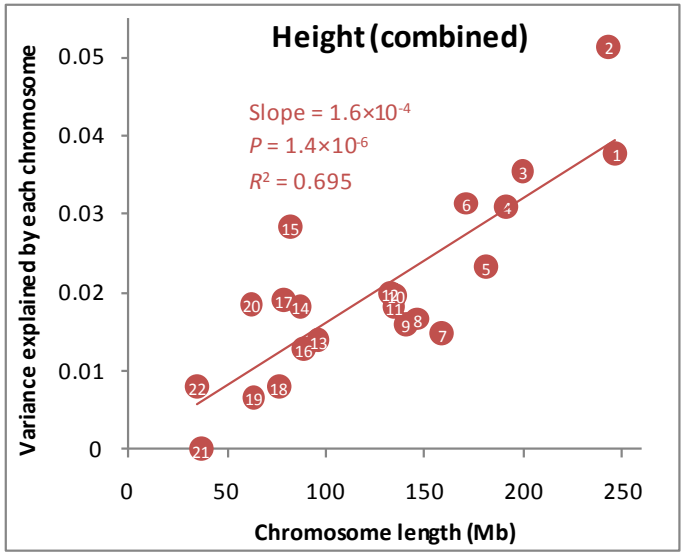
- 780,062 SNPs after QC steps listed in the table.
- Exclude 141,772 SNPs with MAF < 0.02 in European-ancestry group.
- Exclude 36,949 SNPs with missingness > 2% in all samples.
- Include autosomal SNPs only.
- End up with 577,778 SNPs.

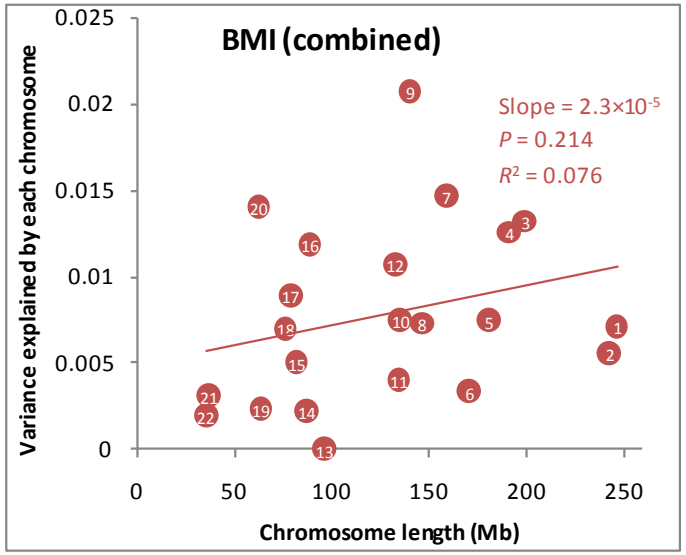
Results (genome-wide)

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

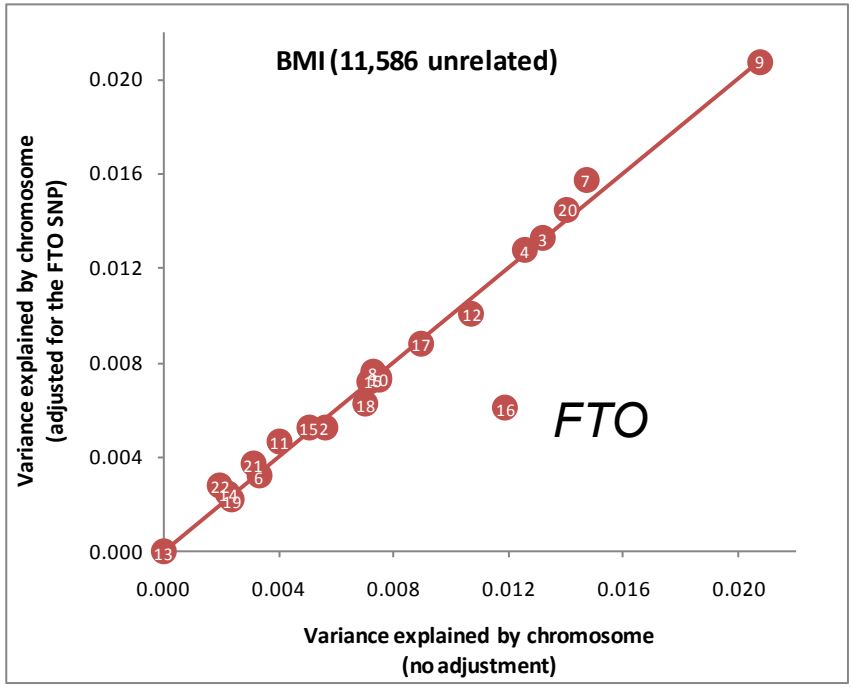
		No PC[a]		10 PCs[b]			
Trait	*n*	h^2_G (s.e.)[c]	*P*	h^2_G (s.e.)	*P*	Heritability[d]	GWAS[e]
Height	11,576	0.448 (0.029)	4.5×10^{-69}	0.419 (0.030)	7.9×10^{-48}	80–90%[32]	~10%[23]
BMI	11,558	0.165 (0.029)	3.0×10^{-10}	0.159 (0.029)	5.3×10^{-9}	42–80%[25,26]	~1.5%[14]
vWF	6,641	0.252 (0.051)	1.6×10^{-7}	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]

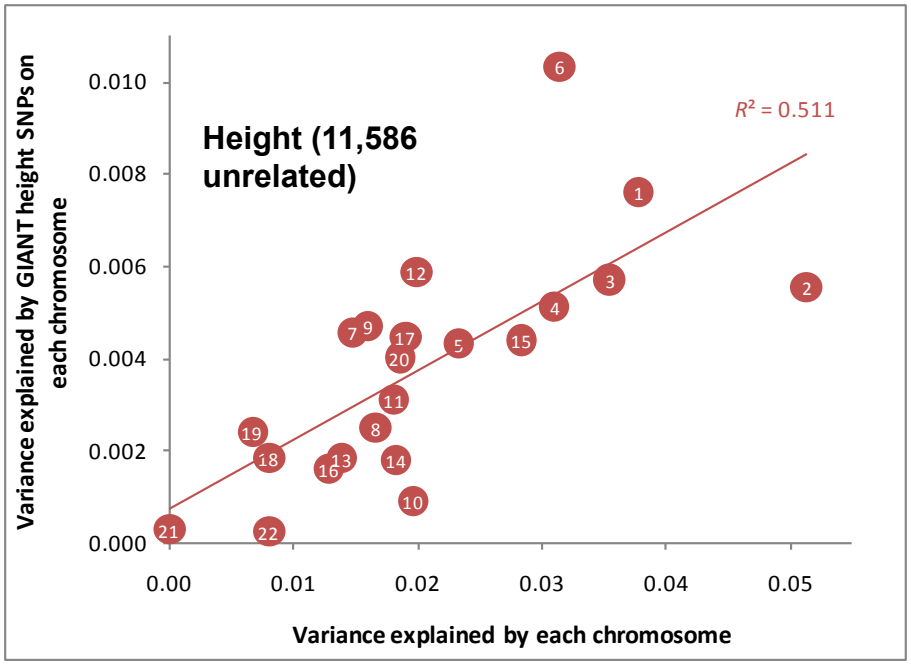
Genome-partitioning: longer chromosomes explain more variation



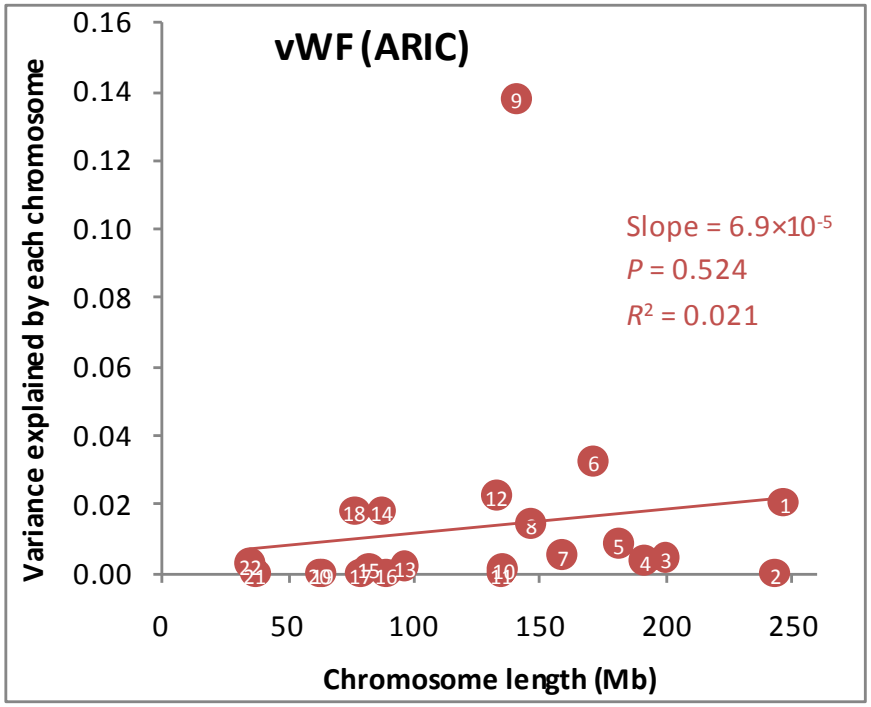


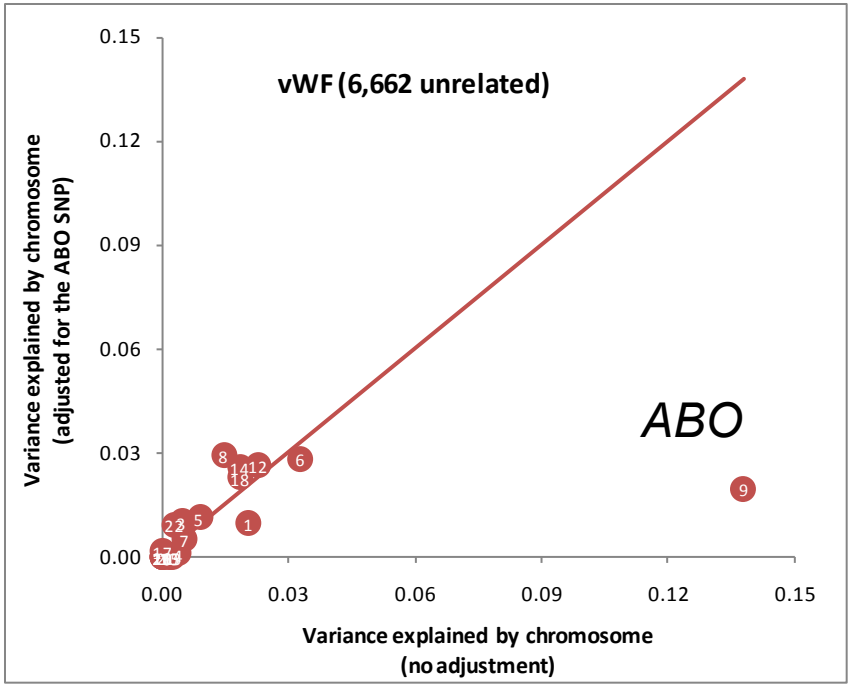
Results are consistent with reported GWAS



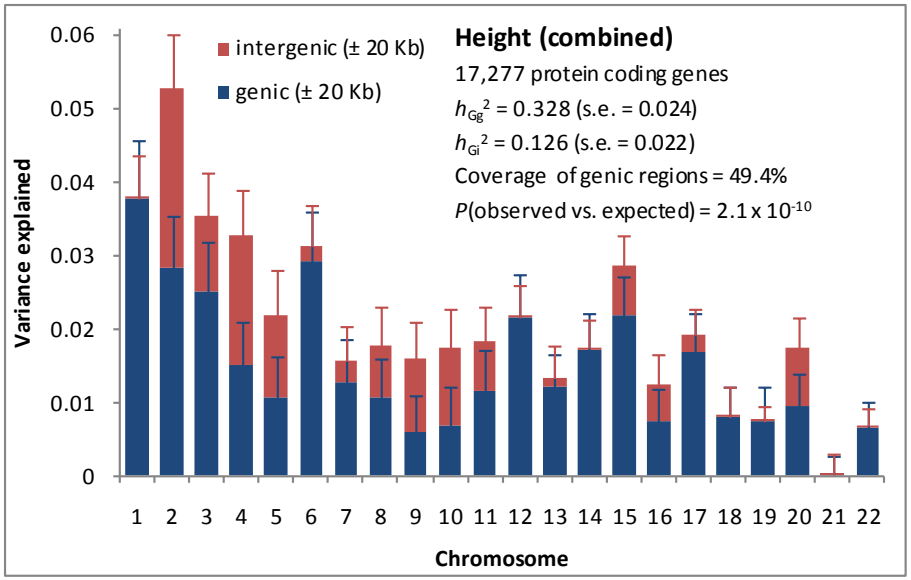


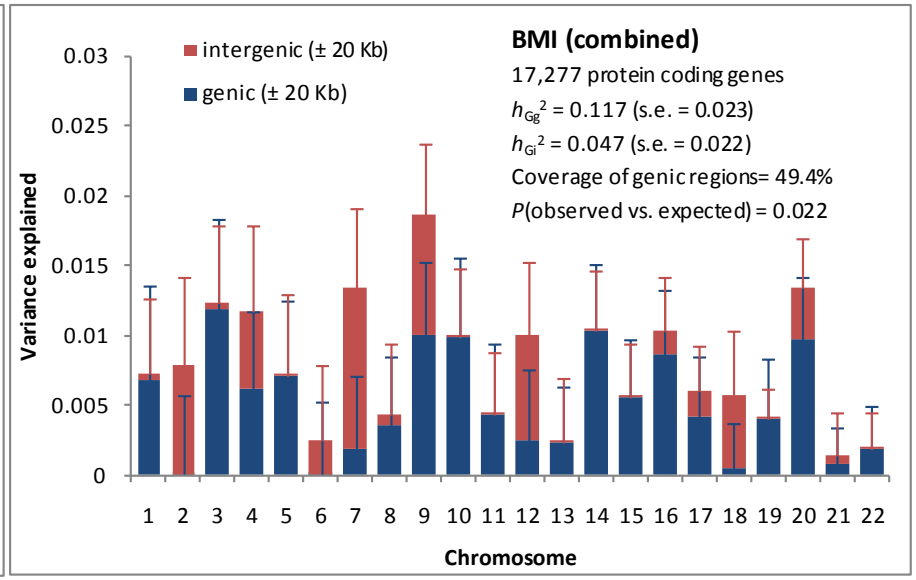
Inference robust with respect to genetic architecture





Genic regions explain variation disproportionately





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

Prediction of quantitative traits using marker data

Peter M. Visscher & Michael E. Goddard

peter.visscher@uq.edu.au

Mike.Goddard@depi.vic.gov.au

Key concepts

- Prediction of phenotypic values is limited by heritability
- Accuracy of prediction depends on
  - how well marker effects are estimated (sample size)
  - how well marker effects are correlated with causal variants (LD)
- Estimation of marker effects and prediction in the same data leads to (severe) bias
  - winner's curse; over-fitting
- Variance explained by a SNP-based predictor is not the same as the variance explained by those SNPs
- Marker data captures both between and within family genetic variation
- Best prediction methods take genetic values as random effects

WORLD WIDE SIRES, LTD.
YOUR FOUNDATION...YOUR FUTURE

7HO10780 UNICORN MILLION ABERLIN-ET *TR *TV
*TL *TY *TD
USA 000066985571
MILLION X GOLDWYN X O MAN
100% Registered Holstein Ancestry

ABERLIN

Unicorn Million Aberlin-ET

DAM RC-LC Goldwyn ATM

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes† and M. E. Goddard†,‡

Production				Management Traits	
TPI	2206	PTA [illegible]	3.32	SCE / Rel.%	7/69
NM$	561	Udder Compo[illegible]	3.53	DCE / Rel.%	7/65
PTA Milk (lbs)	836	Feet & Leg Compo[illegible]	2.53	SSB / Rel.%	7.4/56
PTA Protein (lbs)	29	Body Composite	2.31	DSB / Rel.%	8.1/57
PTA Protein (%)	0.01	Dairy Composite	1.90	SCS	2.65
PTA Fat (lbs)	38	[illegible] Reliability %	72	Productive Life	4.6
PTA Fat (%)	0.02	Dtrs / [illegible]erds	0/0	DPR / Rel.%	0.8/61
Production Reliability %	76	aAa		[illegible]	[illegible]
Dtrs / Herds	0/0				

"Genomic selection" =
individual prediction in a commercial setting

Take-home from animal breeding

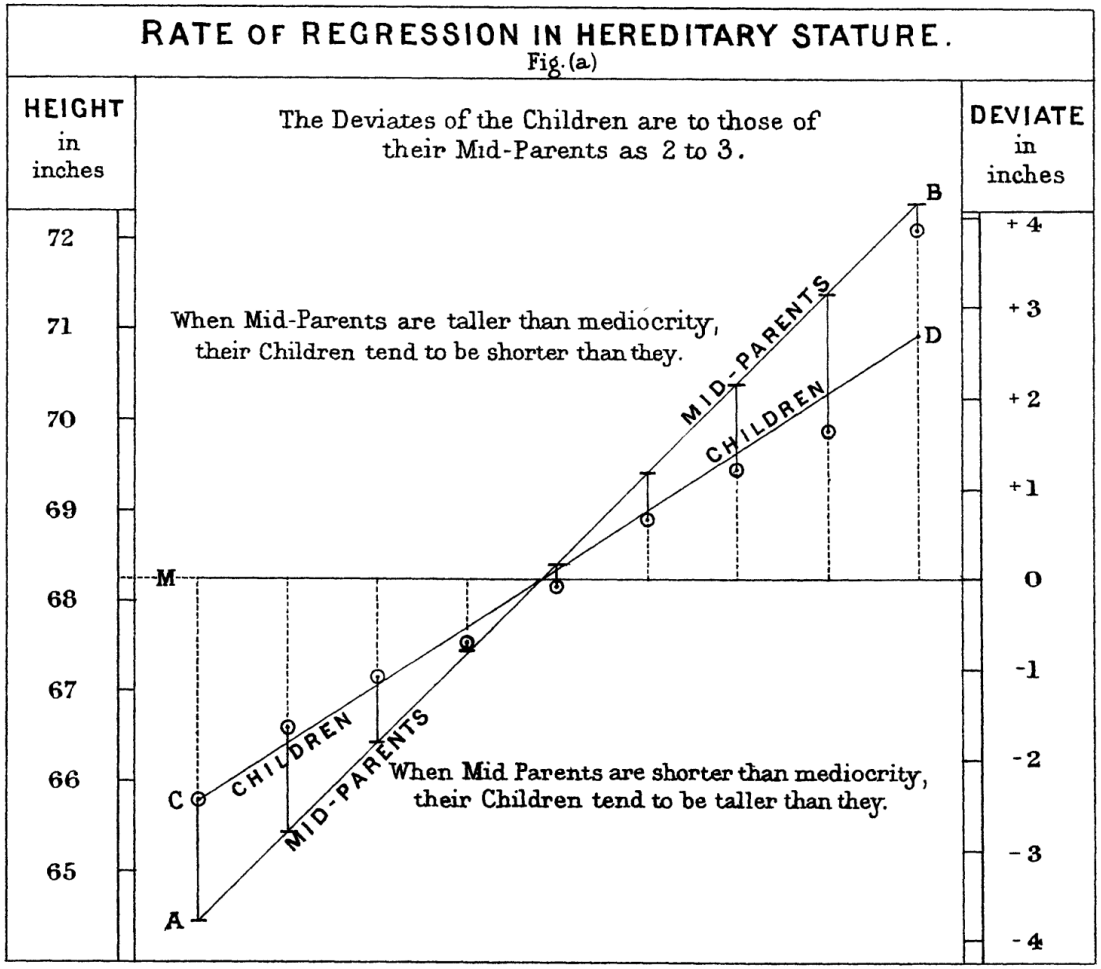
(1) Don’t need genome-wide significant effects

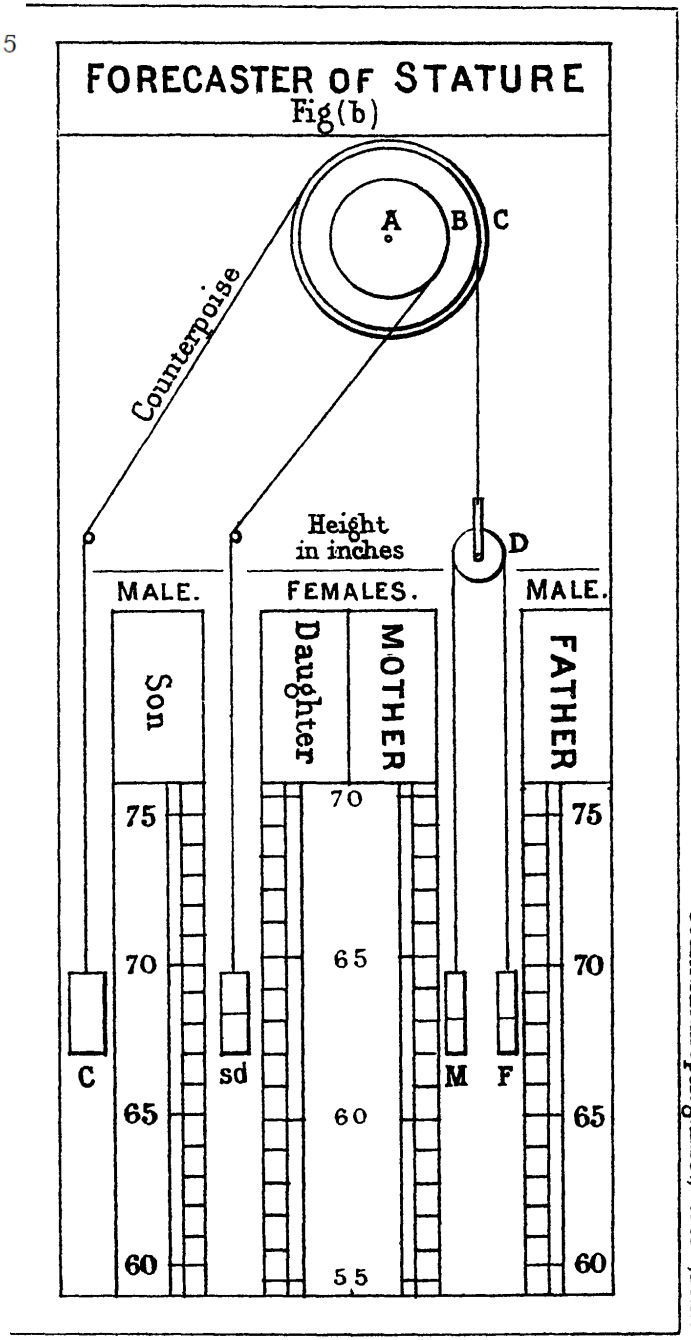
(2) Don’t need to know causal variants

(3) Don’t need to know function

(4) Fit all SNPs simultaneously

Regression Towards Mediocrity in Hereditary Stature.
Author(s): Francis Galton
Source: *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15 (1886), pp. 246-263





Journ. Anthropolog. Inst., Vol. XV, Pl. IX

A quantitative genetics model

y = fixed effects + G + E

G = A + D + I

Possible predictions:

- Predict y from fixed effects and G
- Predict G from A
- Predict y from A
- **Predict y from A using markers**

Prediction using linear regression

$y = \beta * x + e$

- Usually, β and x are considered 'fixed'
- For SNPs, x is random with variance 2p(1-p) assuming HWE
- Later we will consider the case where β is random

Chance association

m markers, sample size *N*

All $\beta = 0$

Multiple linear regression of *y* on *m* markers

$E(R^2) = m/N$ {strictly $m/(N-1)$}

→ Variation "explained" by chance

[Wishart, 1931]

Selection bias

- Select *m* 'best' markers out of *M* in total
- 'Prediction' in same sample (in-sample prediction)

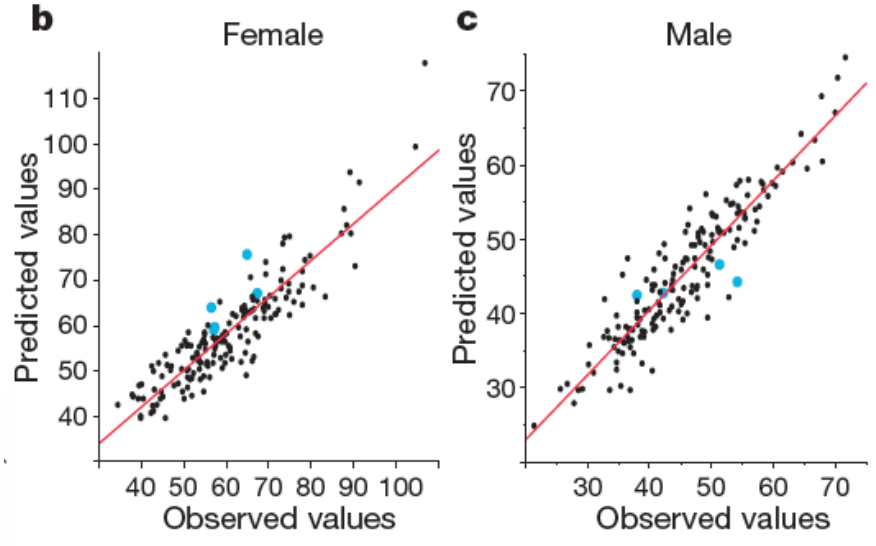
$E(R^2) >> m/N$

→ Lots of variation explained by chance

ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel



~15 best markers selected from 2.5 million markers

Least squares prediction

$$R_m^2 = \mathrm{var}(a) / \mathrm{var}(y) = h^2$$

$$E(\hat{R}_{y,\hat{y}}^2) \approx h^2 / [1 + m / \{Nh^2\}]$$

Even if we knew all *m* causal variants but needed to estimate their effect sizes then the variance explained by the predictor is less than the variance explained by the causal variants in the population.

[Daetwyler et al. 2008, PLoS Genetics; Visscher, Yang, Goddard 2010, Twin Research Human Genetics 2010]

Take-home

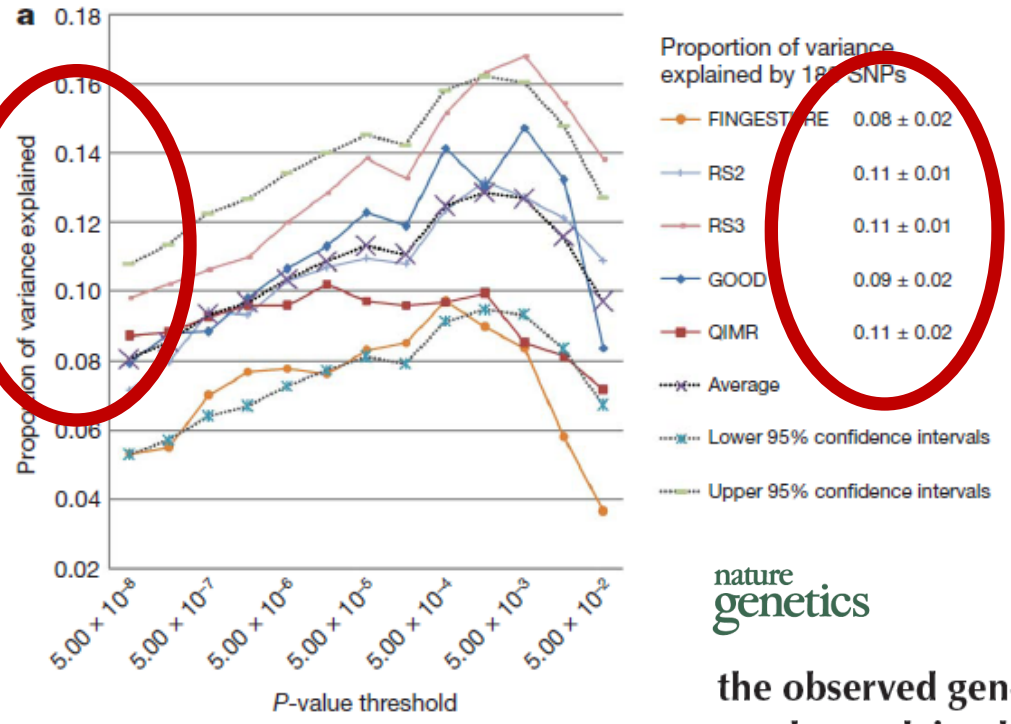
(4) Estimation of variance contributed by (all) loci is not the same as prediction accuracy

unless the effect sizes are estimated without error

LETTER

doi:10.1038/nature09410

Hundreds of variants clustered in genomic loci and biological pathways affect human height



nature genetics

the observed genotype data. We show tha 45% f variance can be explained by considering all SNPs si ultaneously. Thus,

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]

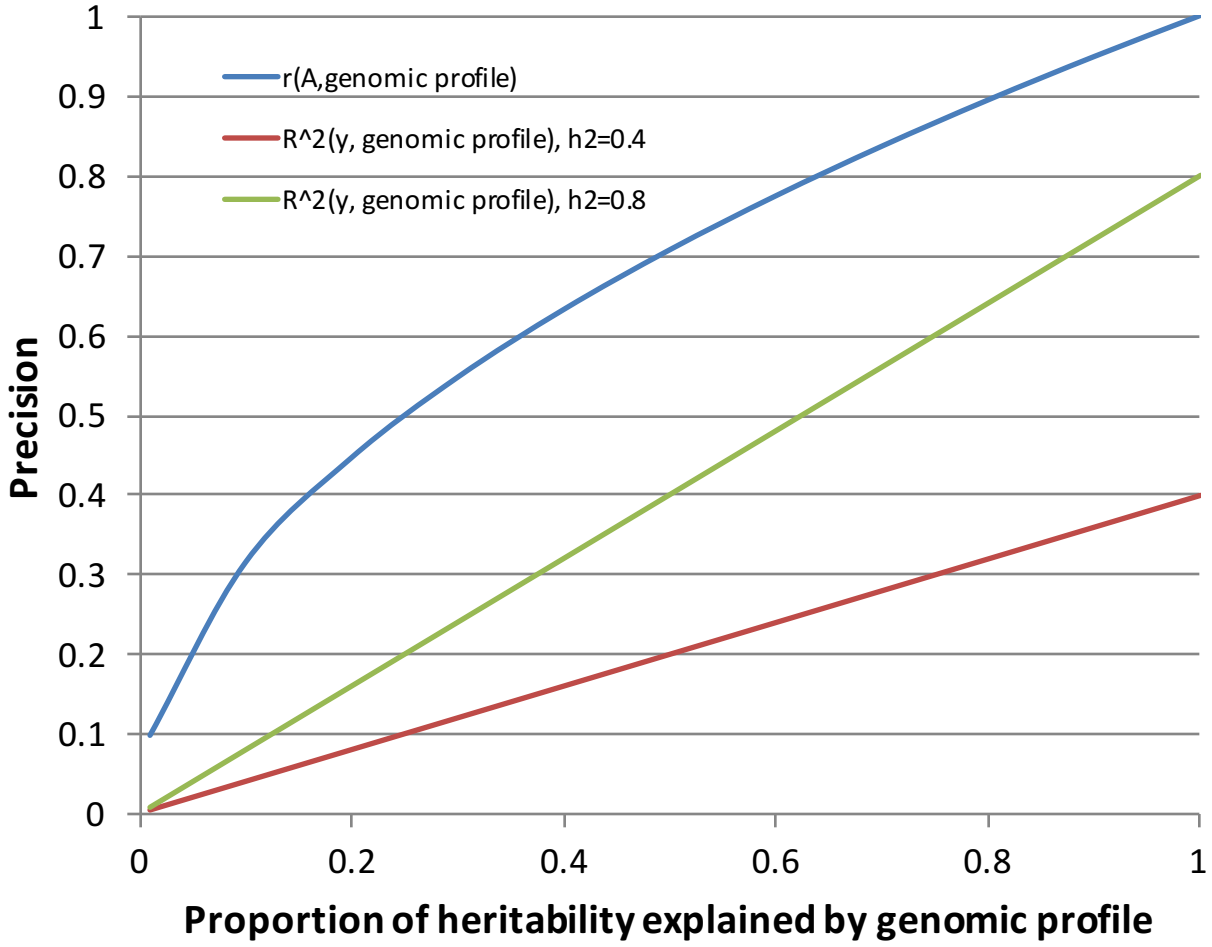
Measures of how well a predictor works

- "Accuracy" (animal breeding)
  - Correlation between true genome-wide genetic value and its predictor
- R^2 from a regression of outcome on predictor (human genetics)
- Area-under-curve from ROC analyses (disease classification)

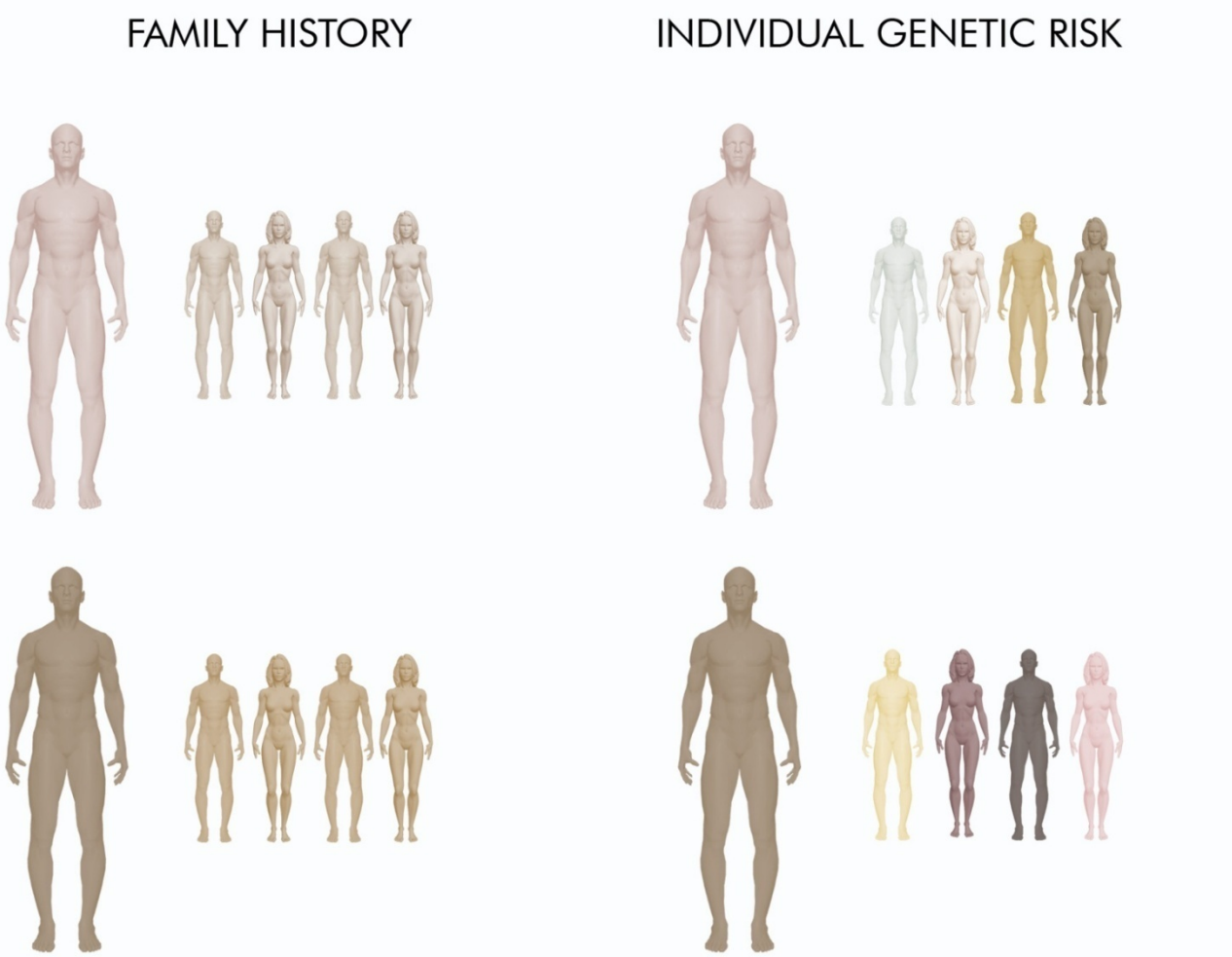
Limits of prediction

- A perfect predictor of A can be a lousy predictor of a phenotype
- The regression R^2 has a maximum that depends on heritability
- The regression R^2 is limited by unknown (eg future) fixed effects and covariates

Predictions from known variants



Prediction using genetic markers: using between and within-family genetic variation



In class demo

- 180 height variants from Lango-Allen et al. 2010
  - Estimation of b from data (N ~ 4000)
    - Note that $E(R^2) = 180/4000 = 0.045$ by chance!
  - Using b from Lango-Allen paper
- Taking the top 180 SNPs from GWAS

Analysis demonstration

- **Data:**
  - Genotype data: 3,924 unrelated individuals and ~2.5M SNPs.
  - Phenotype data: height z-scores (adjusted for age and sex)
  - 180 SNPs identified by the GIANT meta-analysis (MA) of height (n = ~180,000)

- **Analyses:**
  - Estimating effect sizes of the 180 height SNPs in the data.
  - PLINK scoring: 180 GIANT SNPs, using effect sizes estimated from GIANT MA.
  - GWAS analysis in the data, selecting top SNPs at 180 loci and predicting the phenotypes in the same data.

- **Results:**
  - Estimation: $R^2 = 0.134$ ($R^2 = 0.046$ by chance), adjusted $R^2 = 0.093$
  - Prediction: $R^2 = 0.099$
  - Prediction using the top SNPs selected in the same data: $R^2 = 0.429$

Identifying people at high risk: T1D

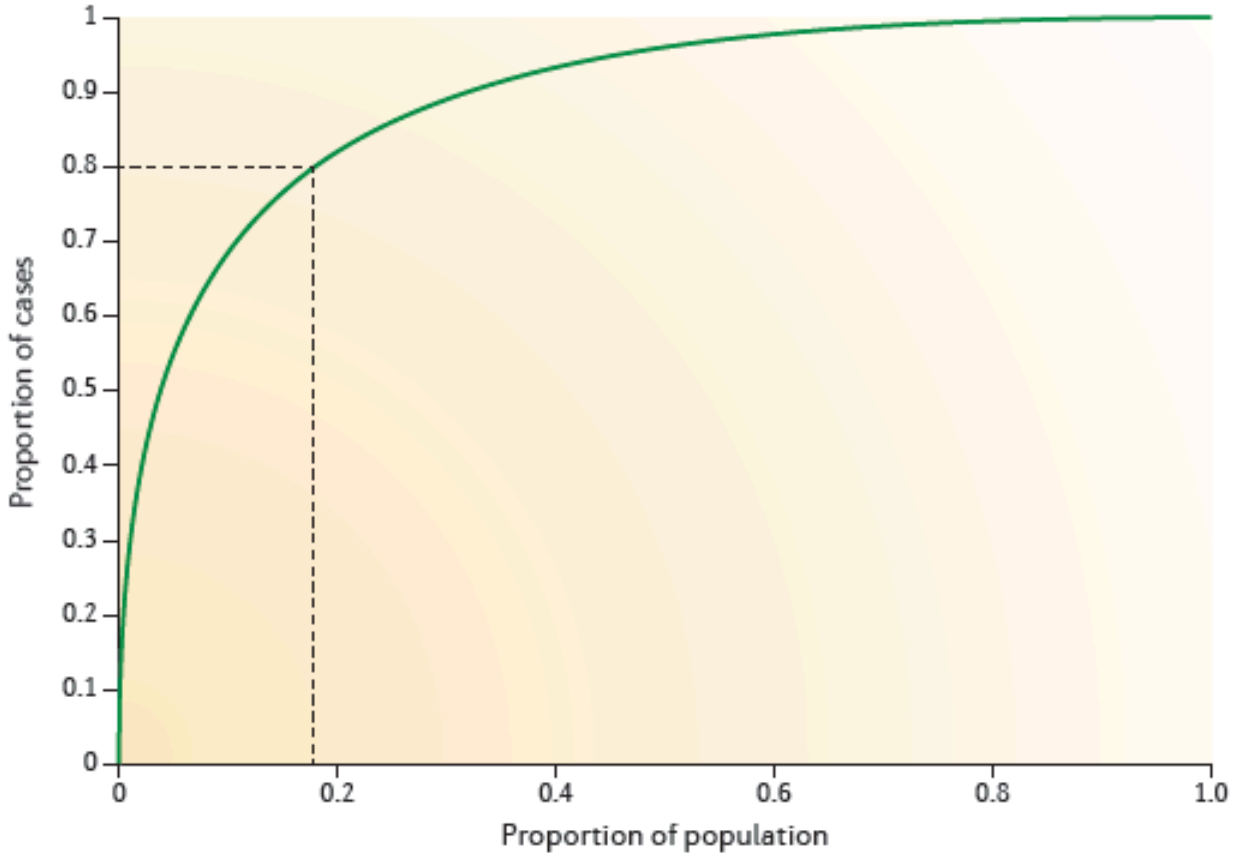


Figure 3 | **The receiver operating characteristic (ROC) curve for the known T1D loci.** The ROC curve plots the sensitivity of genetic type 1 diabetes (T1D) prediction

Per 10,000 people

40 cases
Ratio 1:250

32 cases in 1800 at most risk Ratio 1:56

Most disease is due to people most at risk

Polychronakos & Li NRG 2011
Clayton PLoS Genetics 2009

Prediction of genetic value using better predictors

Model with additive inheritance

$y = g + e$

$V(g) = G\sigma_g^2$, $V(e) = I\sigma_e^2$, $V(y) = V = G\sigma_g^2 + I\sigma_e^2$,

Aim is to predict g for individuals

Eg to predict future risk of a disease

Prediction of genetic value

$y = g + e$

$V(g) = G\sigma_g^2$, $V(e) = I\sigma_e^2$, $V(y) = V = G\sigma_g^2 + I\sigma_e^2$,

Best prediction is

g-hat = $E(g \mid y)$

If y and g are bivariate normal

$E(g \mid y) = b'y = \sigma_g^2 GV^{-1} y$

Prediction of genetic value

Eg Unrelated individuals

$V(g) = Ih^2$, $V(e) = I(1-h^2)$, $V(y) = I$,

Best prediction is

g-hat = $E(g \mid y) = b'y = \sigma_g^2 GV^{-1} y = h^2 y$

Prediction of genetic value

$y = g + e,\ g = Zu$

$V(u) = I\sigma_u^2,\ V(Zu) = ZZ'\sigma_u^2,$

Best prediction is

u-hat = $E(u \mid y)$

If y and u are multivariate normal

$E(u \mid y) = b'y = \sigma_u^2 Z'V^{-1} y$

Prediction of genetic value

$y = g + e,\ g = Zu$

$V(u) = I\sigma_u^2,\ V(Zu) = ZZ'\sigma_u^2,$

$u\text{-hat} = E(u \mid y) = b'y = \sigma_u^2 Z'V^{-1} y$

$g\text{-hat} = Z\ u\text{-hat} = \sigma_u^2 ZZ'V^{-1} y = \sigma_g^2 GV^{-1} y$

Prediction of genetic value

$y = g + e$, $g = Zu$

If y and u are multivariate normal

$E(u \mid y) = b'y = \sigma_u^2 Z'V^{-1} y$

The SNP effects are unlikely to be normally distributed with equal variance

Prediction of genetic value

Best prediction

u-hat = E(u | y)

$= \int u\, P(u \mid y)\, du$

Bayes theorem

$P(u \mid y) = P(y \mid u)\, P(u) / P(data)$

Likelihood prior

Prediction of genetic value

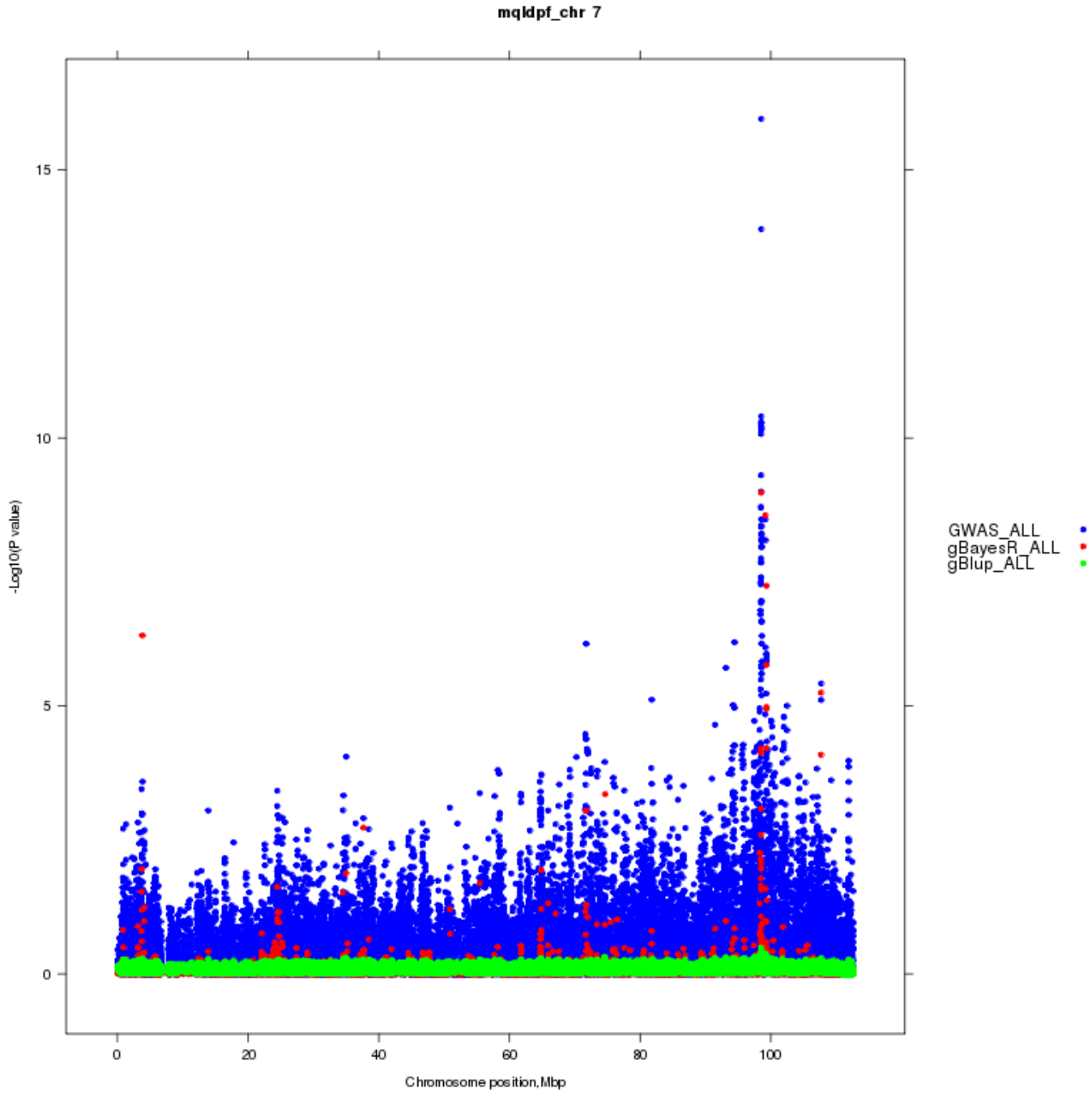
Bayesian estimation

$E(u \mid y) = \int u\, P(y \mid u)\, P(u) / P(y)\, du$

Distribution of SNP effects

Normal → BLUP
t-distribution → Bayes A
Mixture → Bayes B (Meuwissen et al 2001)

Mixture of N → Bayes R (Erbe et al 2012)
$u \sim N(0,\sigma_i^2)$ with probability π_i
$\sigma_i^2 = \{0, 0.0001, 0.001, 0.01\}\ \sigma_g^2$
Accuracy is greatest if assumed distribution matches real distribution.


mqldpf_chr 7
-Log10(P value)
0
5
10
15
0
20
40
60
80
100
Chromosome position, Mbp
GWAS_ALL
gBayesR_ALL
gBlup_ALL

Prediction of genetic value

Other methods of prediction

Estimate effect of each SNP one at a time and add
g-hat = Z u-hat
u-hat estimated from single SNP regression

Biased E(g | g-hat) ≠ g-hat
Less accurate because ignores LD between SNPs
and treats u as fixed effects

Prediction of genetic value

Real data

- 4500 bulls and 12000 cows (Holstein and Jersey)
- 600,000 SNPs genotyped
- Train using bulls born < 2005
- Test using bulls born >= 2005

Correlation of EBV and daughter average

	Protein	Stature	Milk	Fat%
BLUP	0.66	0.52	0.65	0.72
Bayes R	0.66	0.54	0.68	0.82

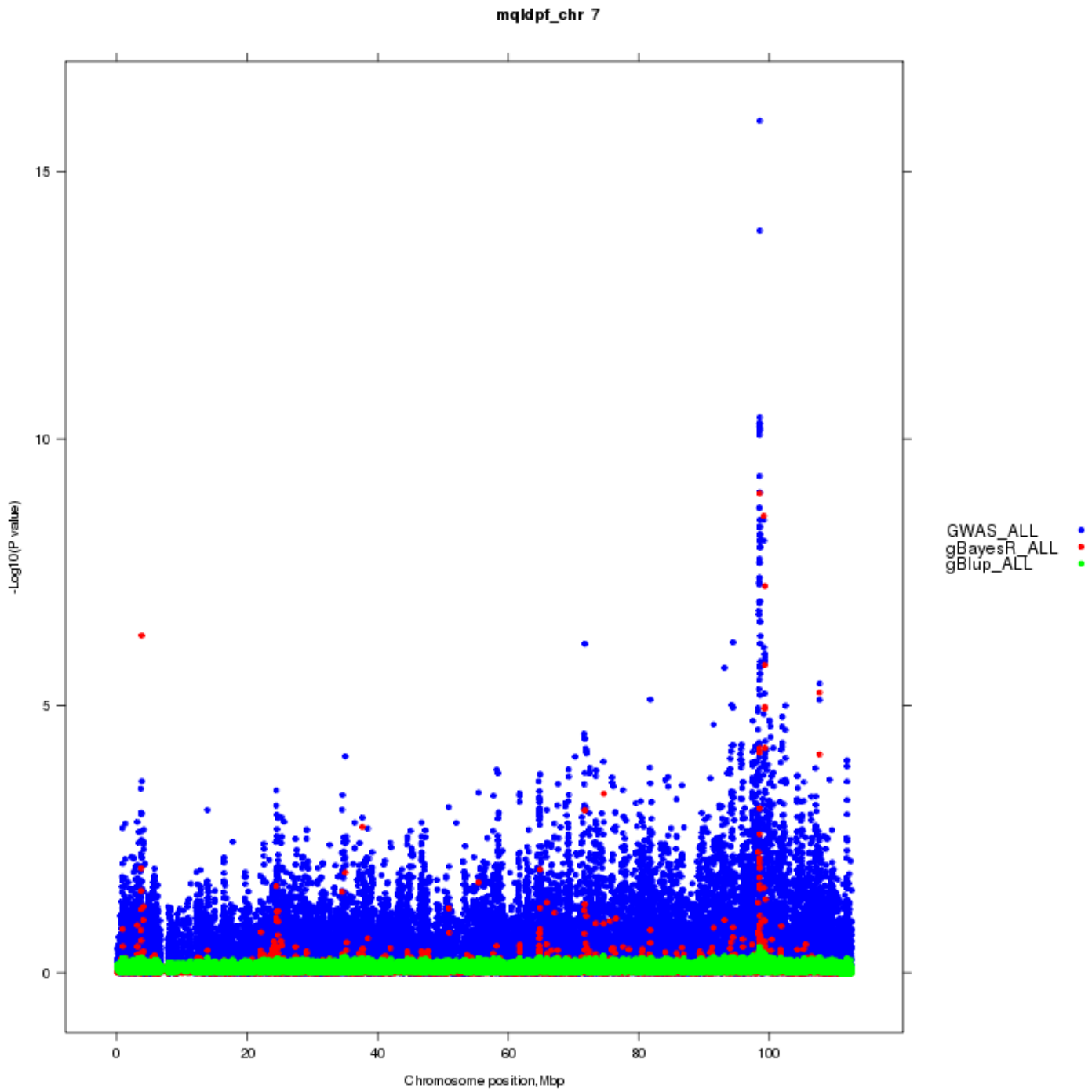
Genetic architecture

Proportion of SNPs from distribution with variance

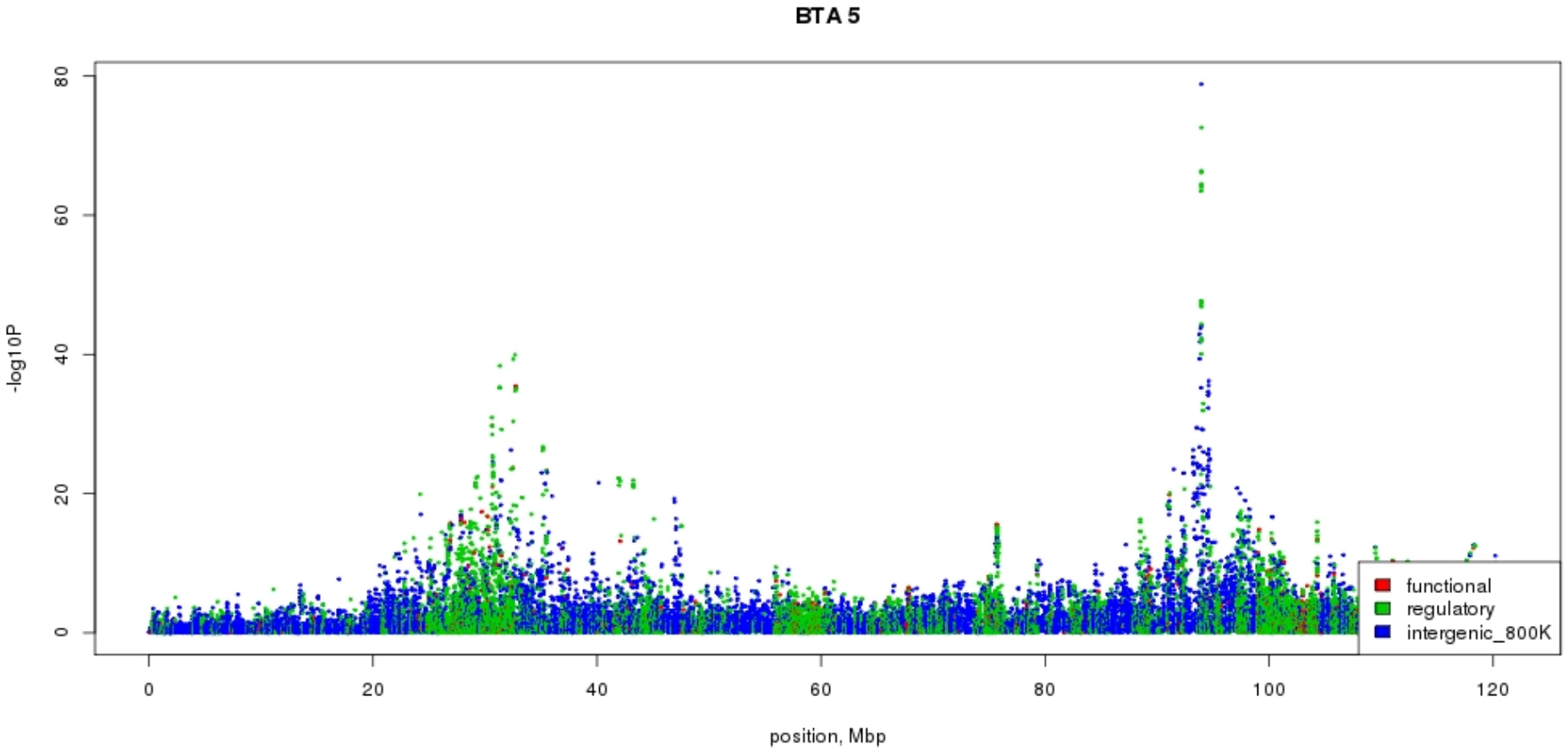
Trait	0.01%	0.1%	1%	polygenic (%)
RFI	7498	296	6	11
LDPF	1419	254	36	27
Mean	4029	271	19	25

Integration of prediction and mapping of causal variants

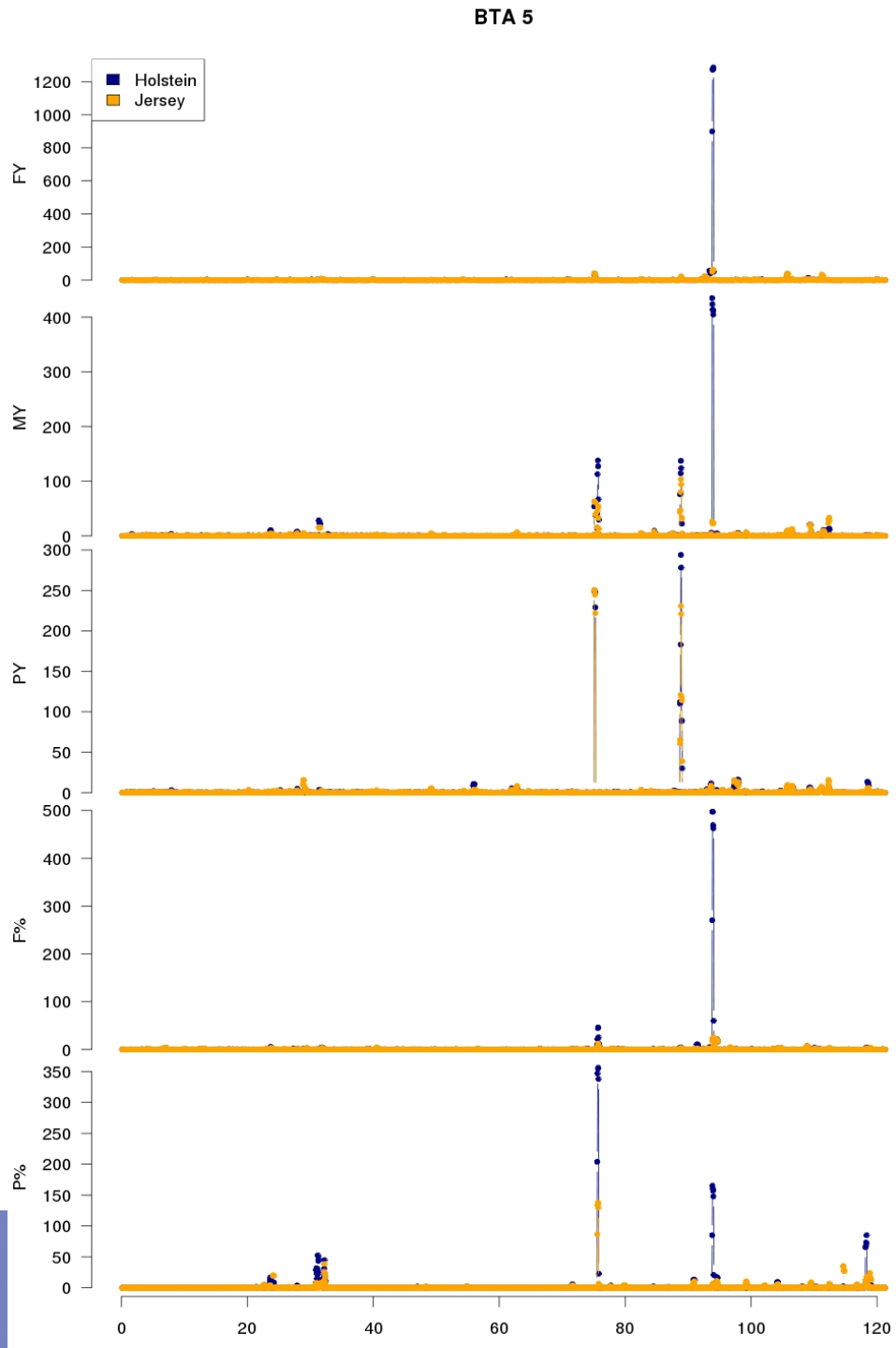
Same Bayesian models as used for prediction can be used for mapping causal variants of complex traits


mqldpf_chr 7
-Log10(P value)
15
10
5
0
0
20
40
60
80
100
Chromosome position, Mbp
GWAS_ALL
gBayesR_ALL
gBlup_ALL

Mapping QTL – Milk on BTA5



Mapping QTL – Milk on BTA5



Application to human disease data (WTCCC)

PLOS | GENETICS

RESEARCH ARTICLE

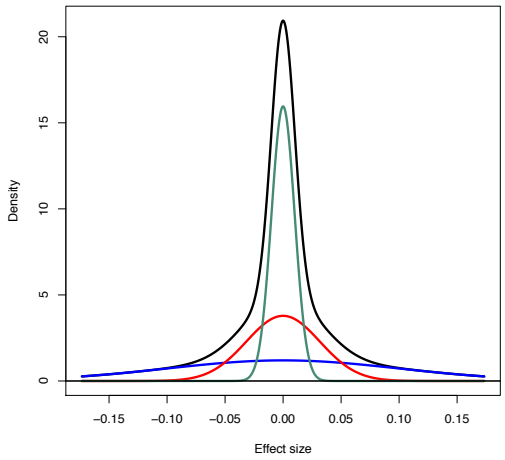
Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model

Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

Model

- Assumes true SNP effects are derived from a series of normal distributions
- Prior assumptions
  - Effects size of SNP *k*

$$\sigma_k^2 = \begin{cases} \pi_1 \times N\left(0, 0 \times \sigma_g^2\right) \\ \pi_2 \times N\left(0, 10^{-4} \times \sigma_g^2\right) \\ \pi_3 \times N\left(0, 10^{-3} \times \sigma_g^2\right) \\ \pi_4 \times N\left(0, 10^{-2} \times \sigma_g^2\right) \end{cases}$$



  - Mixing proportion, **π**
    - *Dirichlet* distribution, $p(\pi_1, \ldots, \pi_4) \sim D(\delta, \ldots, \delta), with\ \delta = 1$
  - Genetic variance
    - hyper-parameter estimated from data, $\sigma_g^2 \sim \chi^{-2}(v_0, S_0^2)$

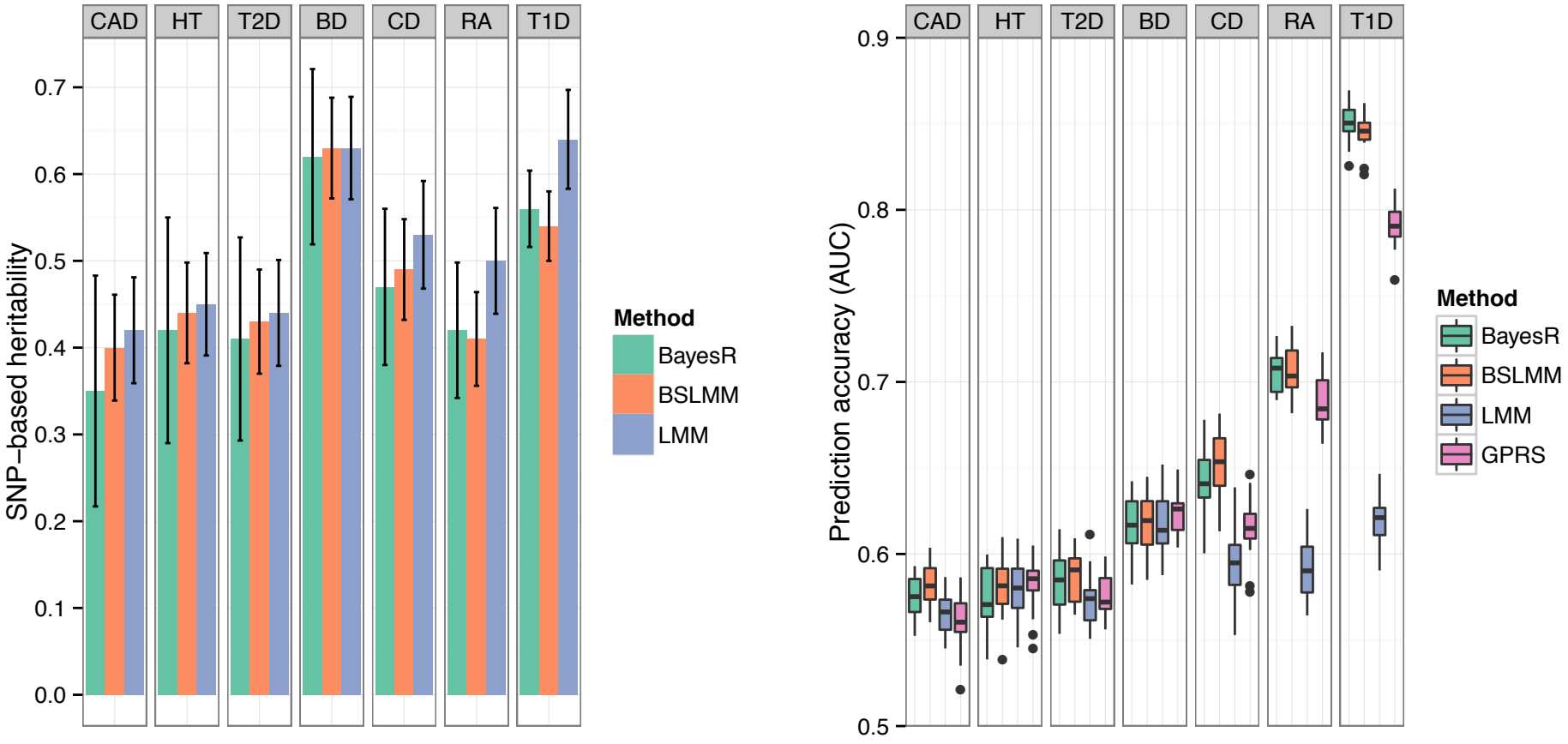
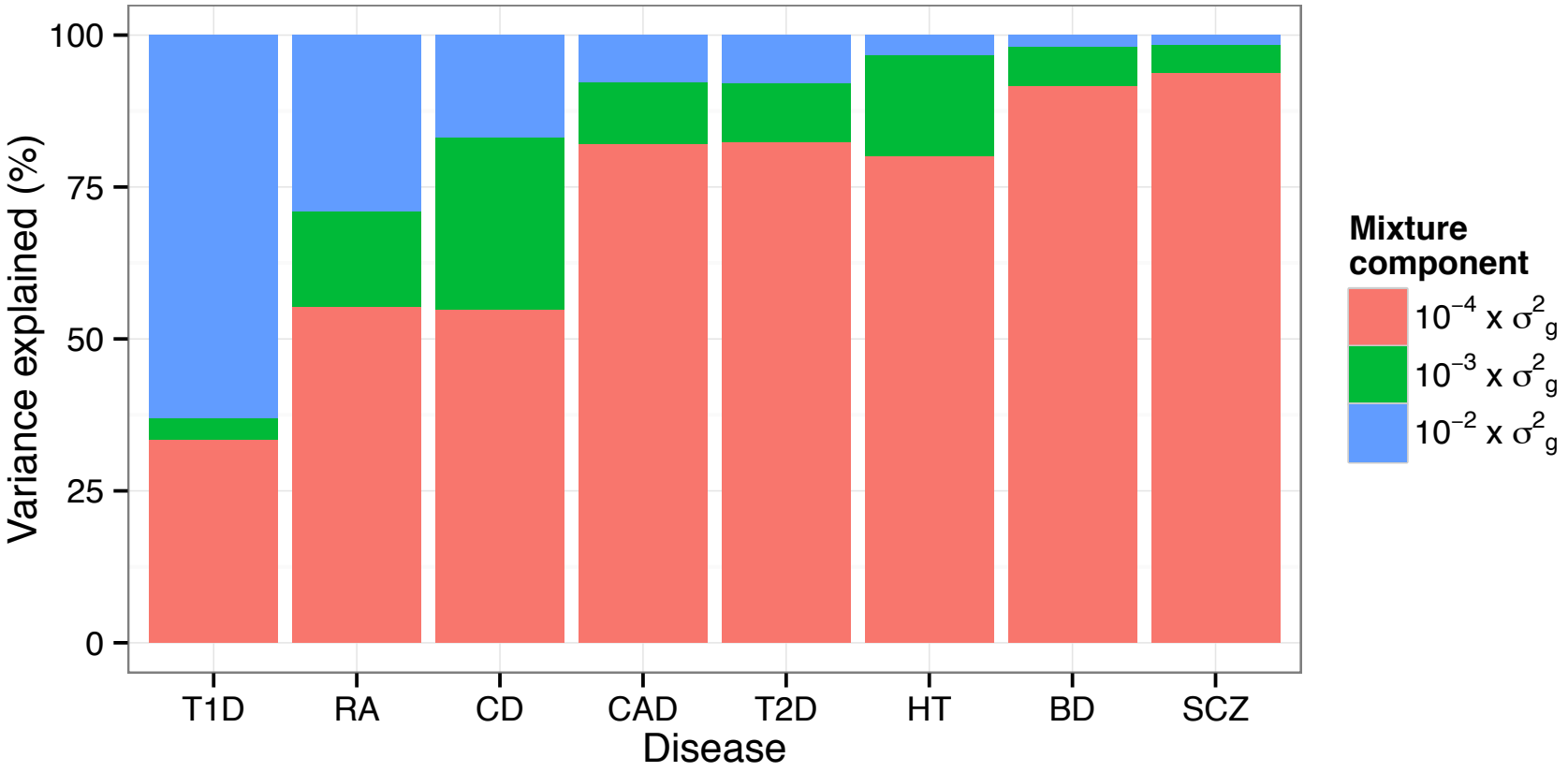


Figure 4. Comparison of performance of BayesR, BSLMM, LMM and GPRS in WTCCC data. (A) Estimates of SNP-based heritability on the observed scale. Antennas are standard deviations of posterior samples for BayesR and BSLMM or standard errors for LMM. GPRS does not provide estimates of heritability. (B) Distribution of the area under the curve (AUC). The single boxplots display the variation in estimates among 20 replicates. In each replicate, the data set was randomly split into a training sample containing 80% of individuals and a validation sample containing the remaining 20%.

Expected proportion of total SNP variance explained by each mixture

(Number of SNPs in class × variance assigned to SNP) / sum of marker variance



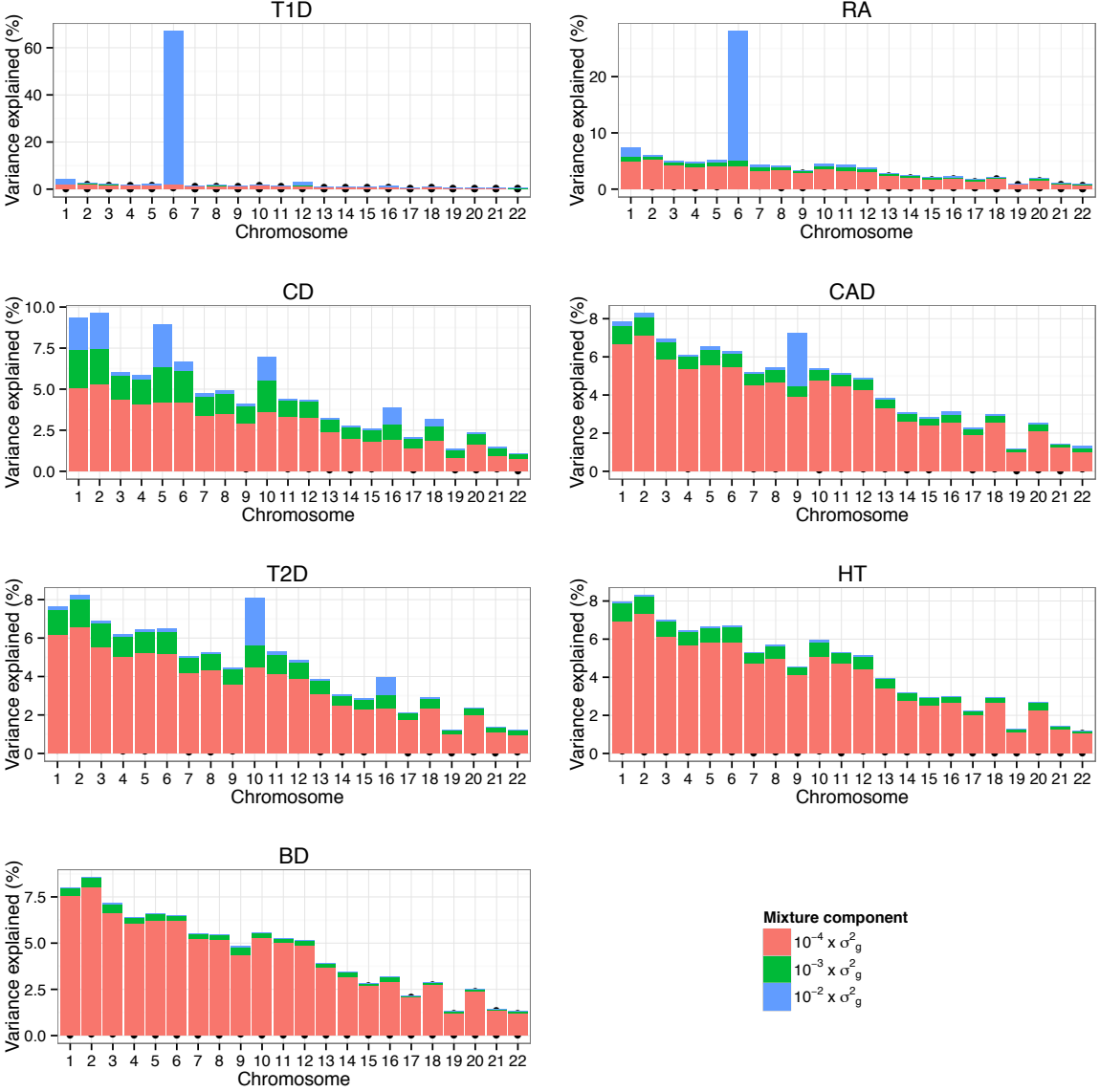
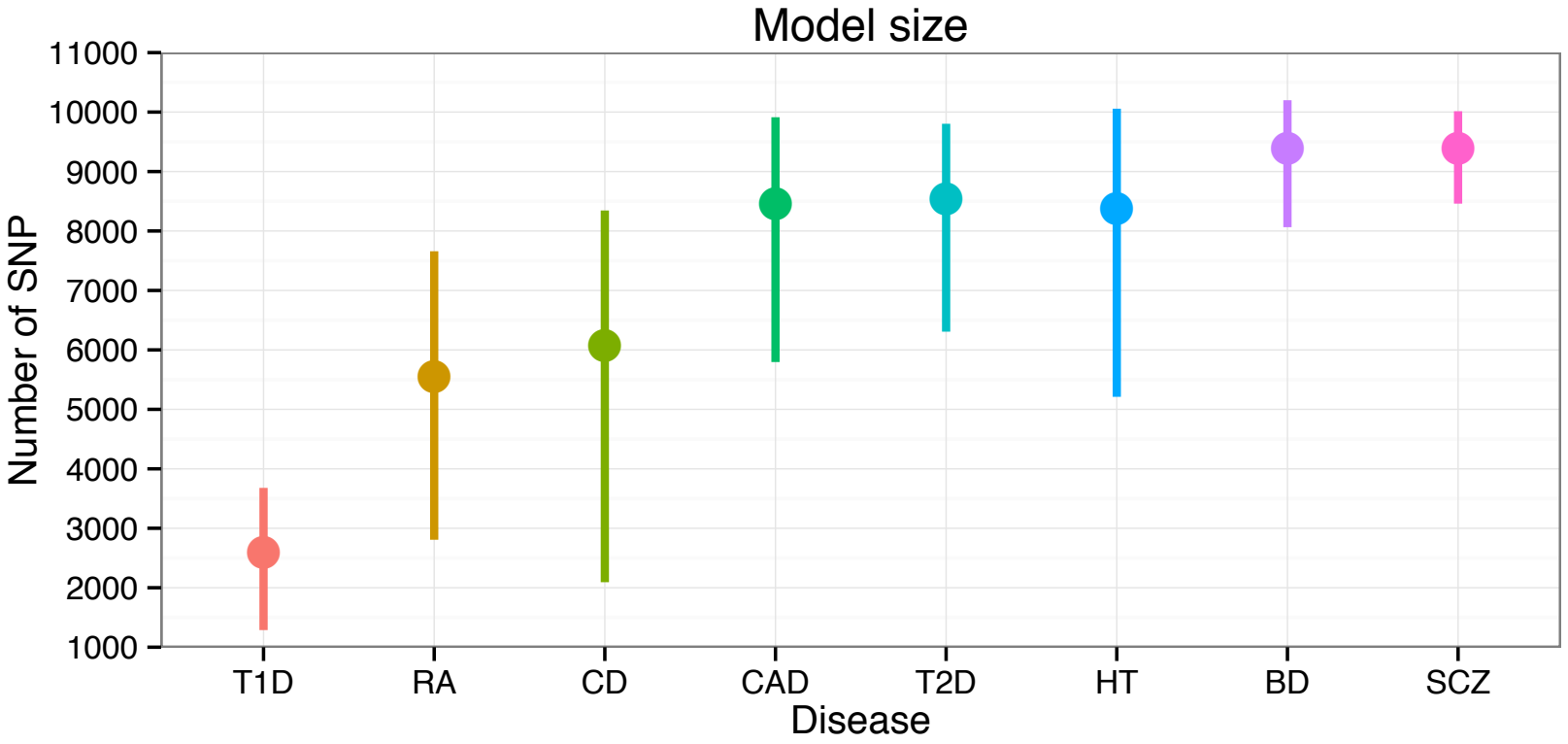


Figure 6. Proportion of genetic variance on each chromosome explained by SNPs with different effect sizes underlying seven traits in WTCCC. Proportion of additive genetic variation contributed by individual chromosomes and the proportion of variance on each chromosome explained by SNPs with different effect sizes. For each chromosome we calculated the proportion of variance in each mixture component as the sum of the square of the sampled effect sizes of the SNPs allocated to each component divided by the sum of the total variance explained by SNPs. The colored bars partition the genetic variance in contributions from each mixture class.

Posterior mean of number of SNPs estimated by BayesR

- Posterior mean and 95% posterior credible interval
- WTCC1+SCZ swedish



Prediction of genetic value Summary

Best prediction is g-hat = E(g | y)

Genetic values treated as random effects

Eg $g \sim N(0, G\sigma_g^2)$

Equivalent model to predict SNP effects u

E(u | y) depends on prior distribution of u

→ Bayesian models

g-hat = Z u-hat gives higher accuracy than assuming

$g \sim N(0, G\sigma_g^2)$

Bayesian models integrate prediction and mapping of causal variants

Key concepts

- Prediction of phenotypic values is limited by heritability
- Accuracy of prediction depends on
  - how well marker effects are estimated (sample size)
  - how well marker effects are correlated with causal variants (LD)
- Estimation of marker effects and prediction in the same data leads to (severe) bias
  - winner's curse; over-fitting
- Variance explained by a SNP-based predictor is not the same as the variance explained by those SNPs
- Marker data captures both between and within family genetic variation
- Best prediction methods take genetic values as random effects

Supplementary derivations

Theory (additive model)
m unlinked causal variants

$$y_i = \sum_{j=1}^{m} x_{ij} b_j + e_i = a_i + e_i$$

$$\mathrm{var}(y) = \sum_{j=1}^{m} \mathrm{var}(x_j) b_j^2 + \mathrm{var}(e) = \mathrm{var}(a) + \mathrm{var}(e)$$

$$\mathrm{cov}(y_i, y_k) = \sum_{j=1}^{m} \mathrm{cov}(x_{ij}, x_{kj}) b_j^2 + \mathrm{cov}(e_i, e_k)$$

$$= \mathrm{cov}(a_i, a_k) + \mathrm{cov}(e_i, e_k)$$

$$= \mathrm{cov}(a_i, a_k) \text{ if } \mathrm{cov}(e_i, e_k) = 0$$

Prediction

$$\hat{y}_i = \sum_{j=1}^{m} x_{ij}\hat{b}_j = \hat{a}_i$$

$$\mathrm{var}(\hat{y}) = \sum_{j=1}^{m} \mathrm{var}(x_j)\hat{b}_j^2 = \mathrm{var}(\hat{a})$$

$$\mathrm{cov}(\hat{y}_i, \hat{y}_k) = \sum_{j=1}^{m} \mathrm{cov}(x_{ij}, x_{kj})\hat{b}_j^2 = \mathrm{cov}(\hat{a}_i, \hat{a}_k)$$

- theory -

$$\operatorname{cov}(\hat{y}_i, y_i) = \operatorname{cov}\{\sum_{j=1}^{m}(x_{ij}\hat{b}_j), \sum_{j=1}^{m} x_{ij} b_j + e_i\}$$

$$= \sum_{j=1}^{m} \operatorname{var}(x_{ij})\hat{b}_j b_j + \sum_{j=1}^{m} x_{ij} \operatorname{cov}(\hat{b}_j, e_i)$$

If *b* estimated from the same data in which prediction is made, then the second term is non-zero

Effect of errors in estimating SNP effects (least squares; single SNP)

$$y_i = x_i b + e_i$$

$$\hat{b} = b + \varepsilon$$

$$E(\hat{b}) = b$$

$$\mathrm{var}(\hat{b}) = \mathrm{var}(\varepsilon) = \sigma_e^2 / \Sigma x^2 \approx \mathrm{var}(y) / \{N \,\mathrm{var}(x)\}$$

$$\mathrm{var}(x) = 2p(1-p) \text{ under HWE}$$

$$\text{Define } \mathrm{R}^2_{SNP} = \mathrm{var}(x) b^2 / \mathrm{var}(y)$$

= contribution of single SNP to heritability

- effects of errors -

$$\hat{R}^2_{y,\hat{y}} = \mathrm{cov}(y,\hat{y})^2 / \{\mathrm{var}(y)\,\mathrm{var}(\hat{y})\}$$

$$E[\mathrm{cov}(y,\hat{y})] = E[\mathrm{cov}(xb, x\hat{b})] = \mathrm{var}(x_i)E(\hat{b})b$$

$$= \mathrm{var}(x)b^2$$

$$E[\mathrm{var}(\hat{y})] = E[\mathrm{var}(x\hat{b})] = \mathrm{var}(x)E[\hat{b}^2]$$

$$= \mathrm{var}(x)[b^2 + \mathrm{var}(\hat{b})] \approx \mathrm{var}(x)b^2 + \mathrm{var}(x)\,\mathrm{var}(y)/[N\,\mathrm{var}(x)]$$

$$= \mathrm{var}(x)b^2 + \mathrm{var}(y)/N$$

$$E(\hat{R}^2_{y,\hat{y}}) \approx R^2_{SNP} / [1 + 1/\{NR^2_{SNP}\}]$$