Estimation of identity-by-descent relationship from marker genotypes and estimation of within family genetic variation

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

Recap: 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

= 2θ

D = coefficient of fraternity

= Prob(2 alleles are IBD) = Δ

Examples (no inbreeding)

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

Relationships

We use relationship data

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

Relationships

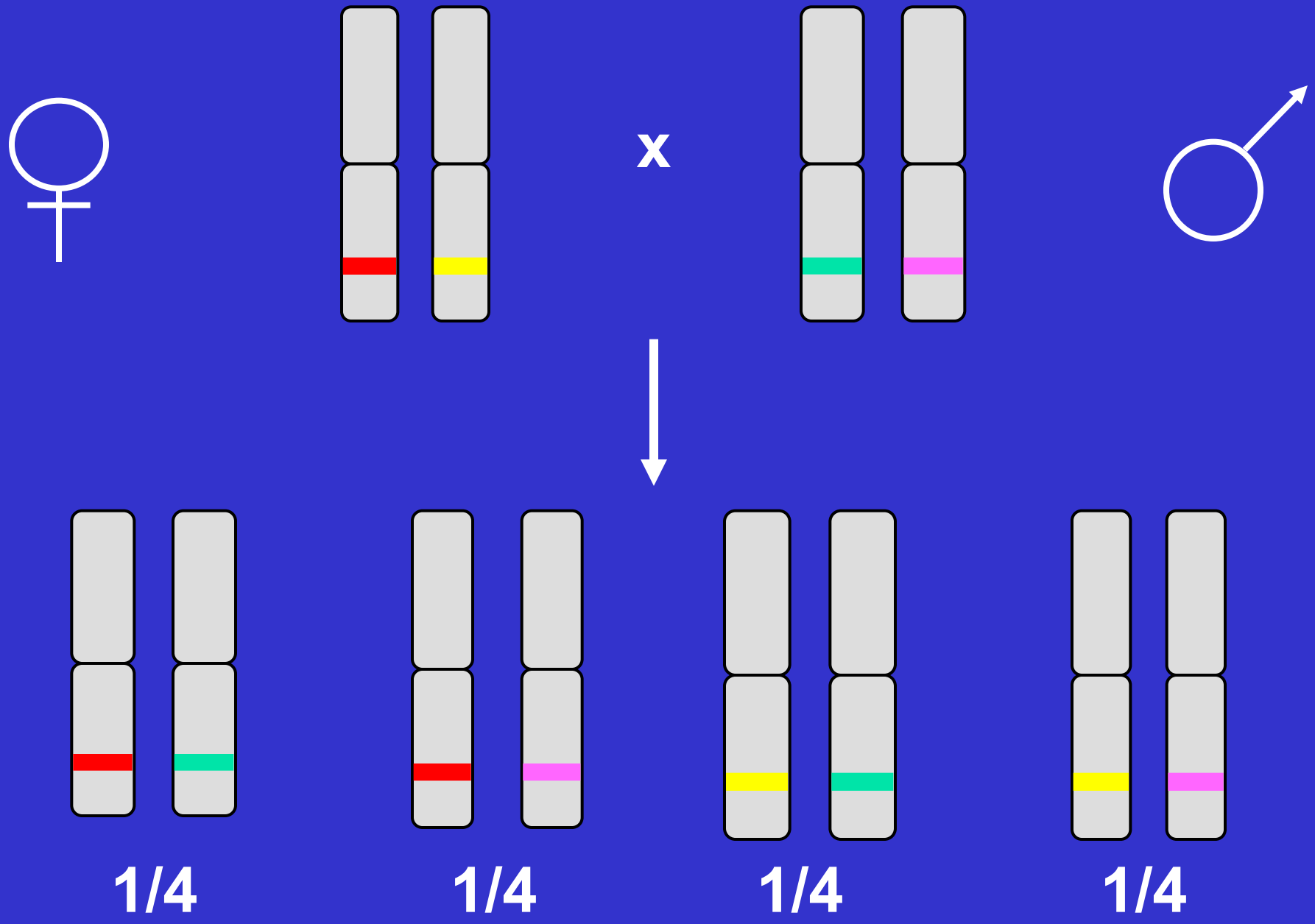
Additive genetic relationship G_{ij}

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

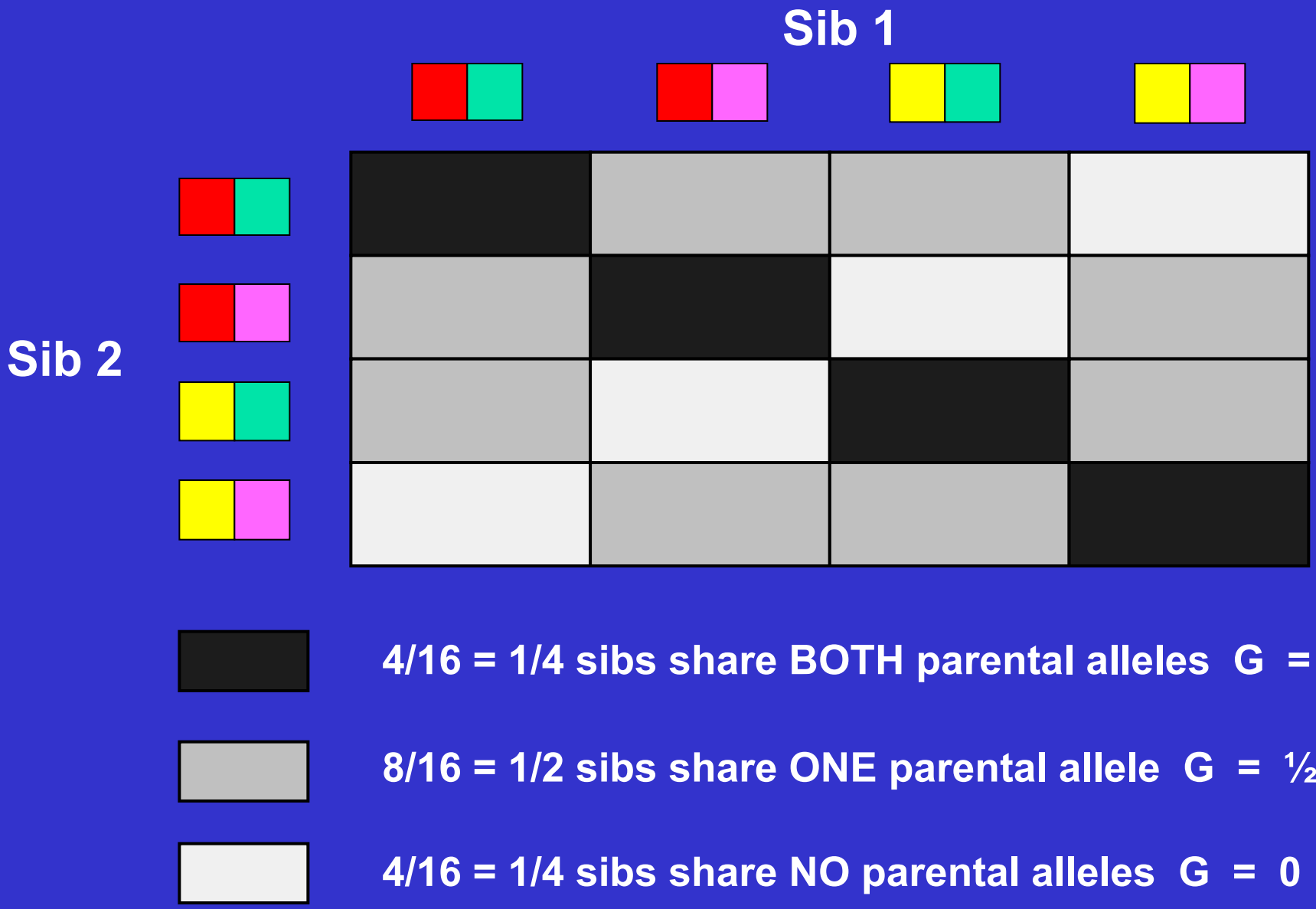
Pedigree relationship A_{ij} = Prob (IBD)

= $E(G_{ij})$

Actual relationship deviates randomly from this expectation


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

IDENTITY BY DESCENT



Relationships

Summary of single locus case, full sibs

Pairs of sibs share

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

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

Single locus

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

Several notations in the literature

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

$A = \frac{1}{2}k_1 + k_2 = 2\theta$

$D = k_2 = \Delta_{xy}$

Realisations

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

$G = E(A)$

$\pi_d = E(D)$

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

n multiple unlinked loci

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

But: 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 G

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

[1 M is the genetic distance between loci, the expected number of crossovers during meiosis]

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

Fullsibs: genome-wide (Total length L Morgan)

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

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

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

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

Fullsibs: Correlation additive and dominance relationships

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

Using β(G on π_d) = 1

Difficult but not impossible to disentangle additive and dominance variance

NB Practical

Summary of theory

Additive and dominance (fullsibs)

	SD(G)	SD(π_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 G and π_d)

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

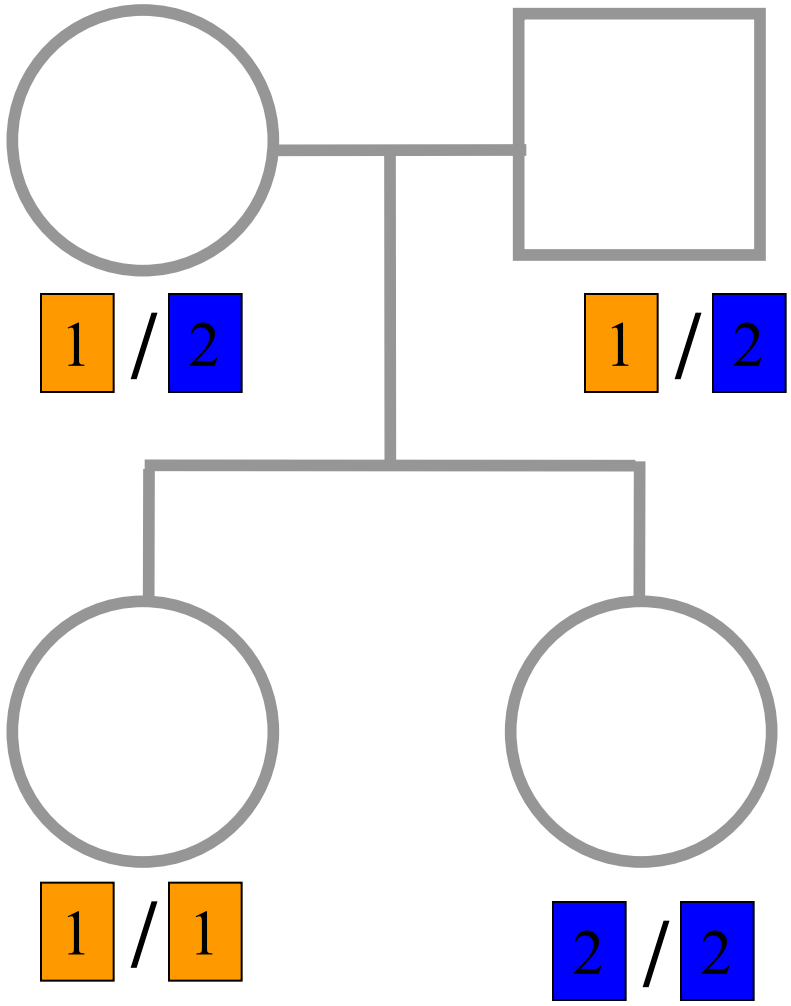
1. Well defined, recent base (reference)

e.g. Data on families of full-sibs and parents of sibs are the base

Estimating relatedness with markers

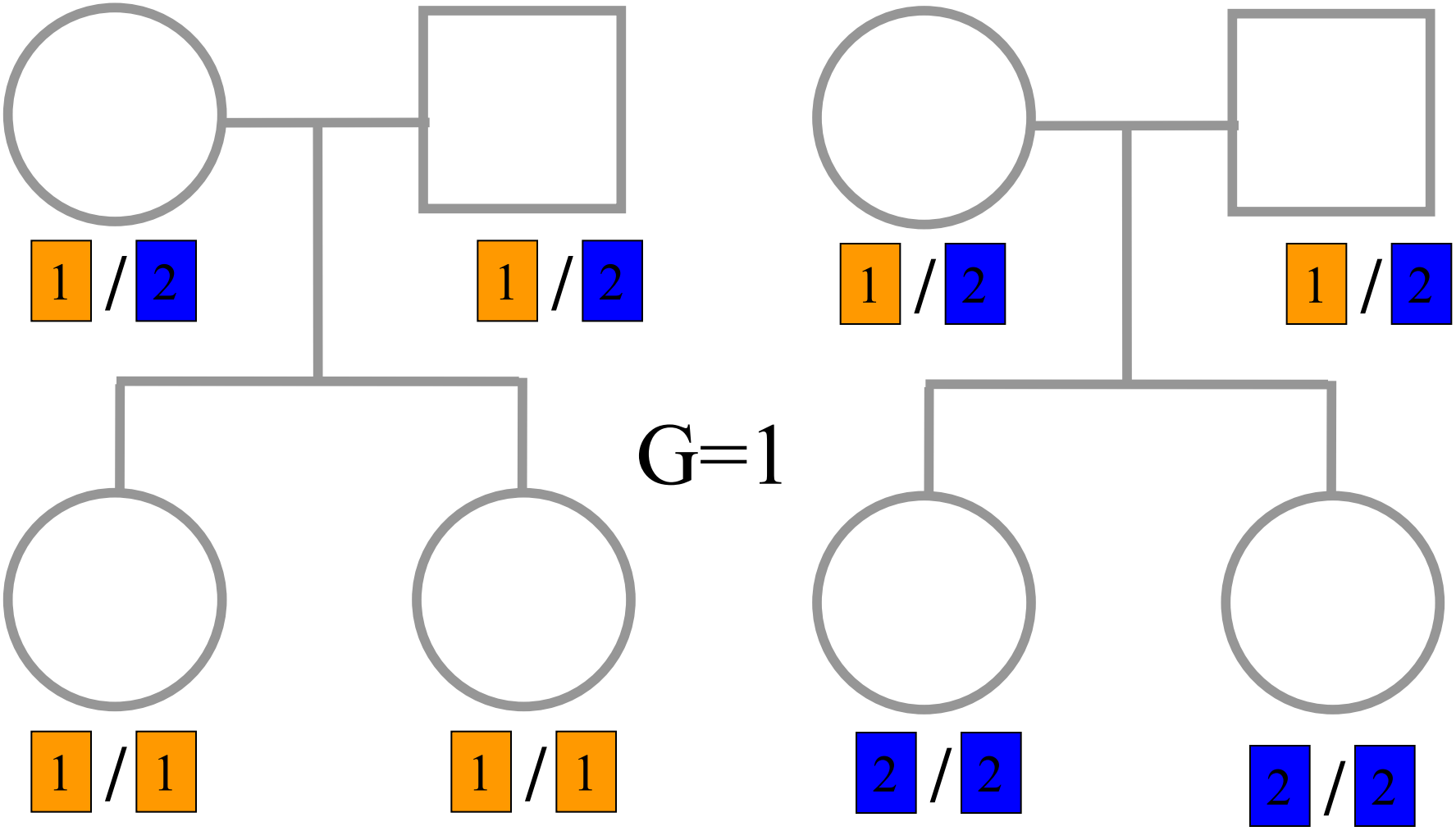
- Using:
  - Observed data (SNP genotypes)
  - Mendelian segregation rules (prior probability of sharing alleles IBD)
  - Marker allele frequencies in the population

IBD can be trivial…

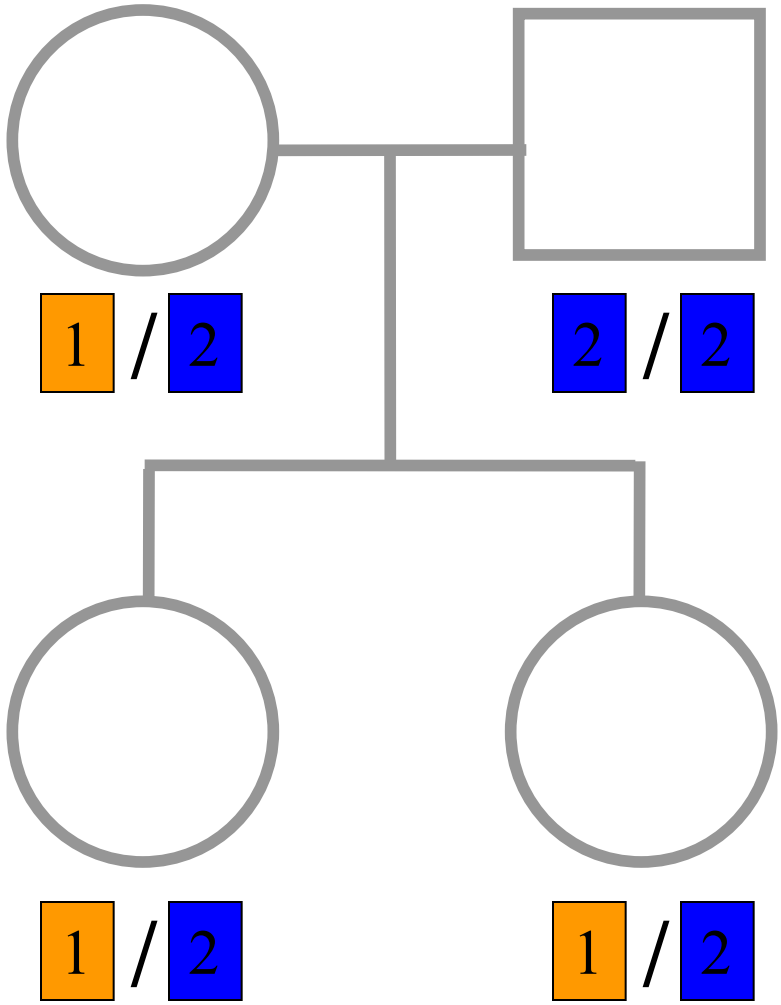


G=0

Two Other Simple Cases…



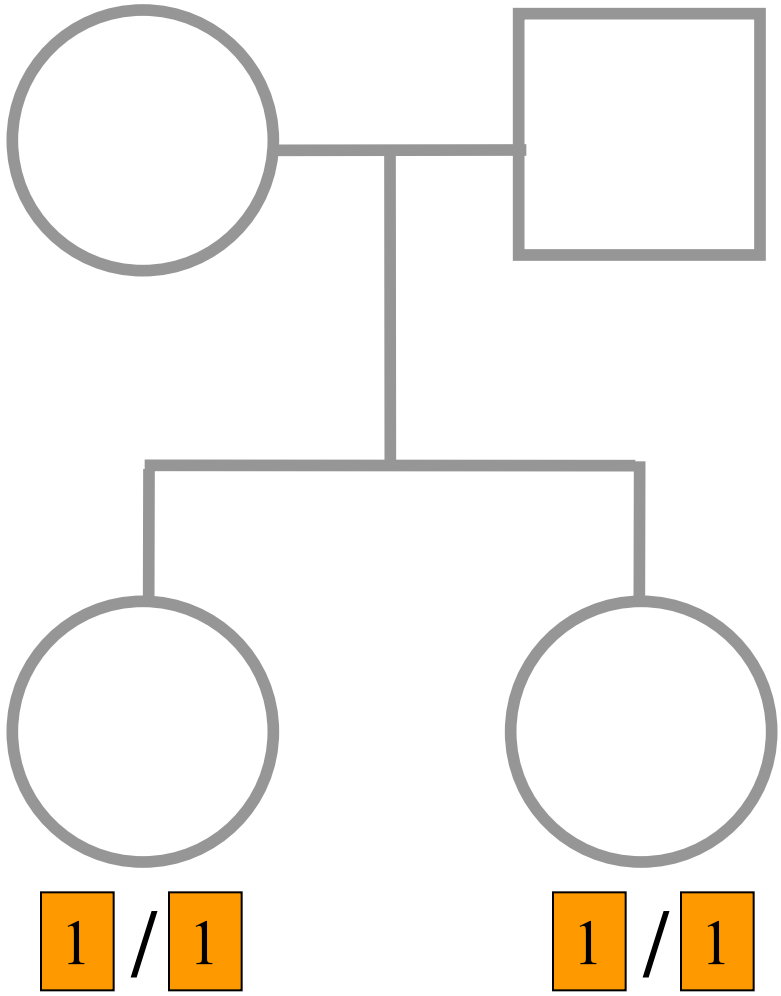
A little more complicated…



G= ½
(50% chance)

G=1
(50% chance)

And even more complicated…



G=?

Bayes Theorem for IBD Probabilities

posterior

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

prior

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

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

Prob(data)

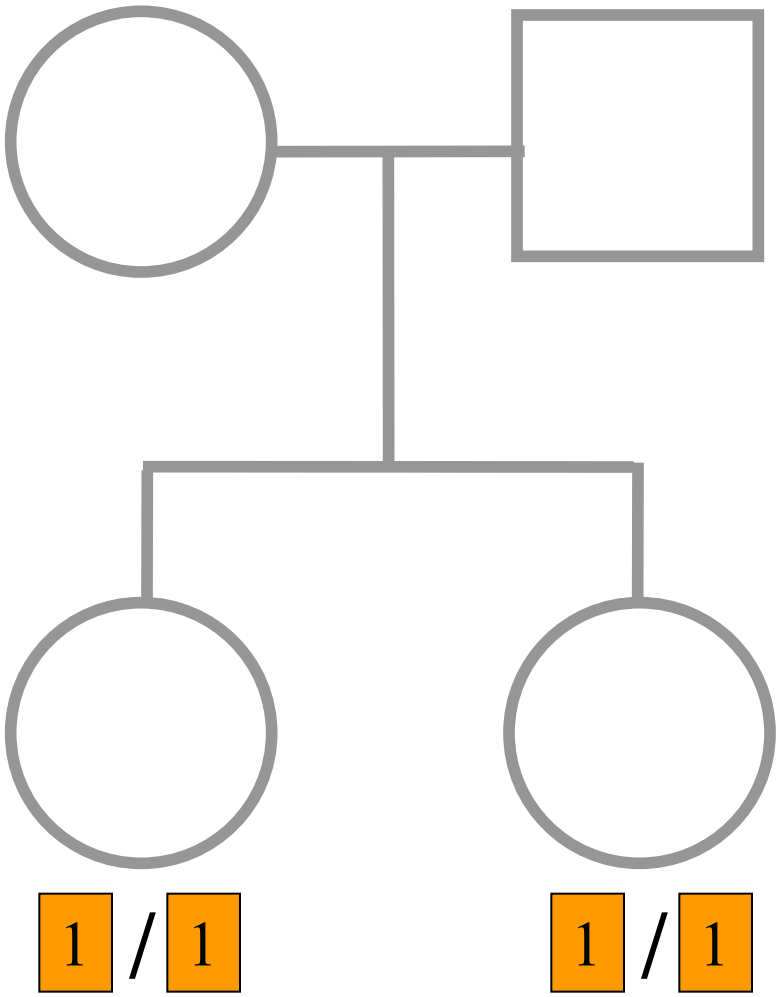
E(G) = ½P(IBD=1|Genotypes) + P(IBD=2|Genotypes)

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



$$p_1 = 0.5$$

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

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

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

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

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

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

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

E(G) = 2/3

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

Estimate relationship from markers

1. **Well defined, recent base**

e.g. 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

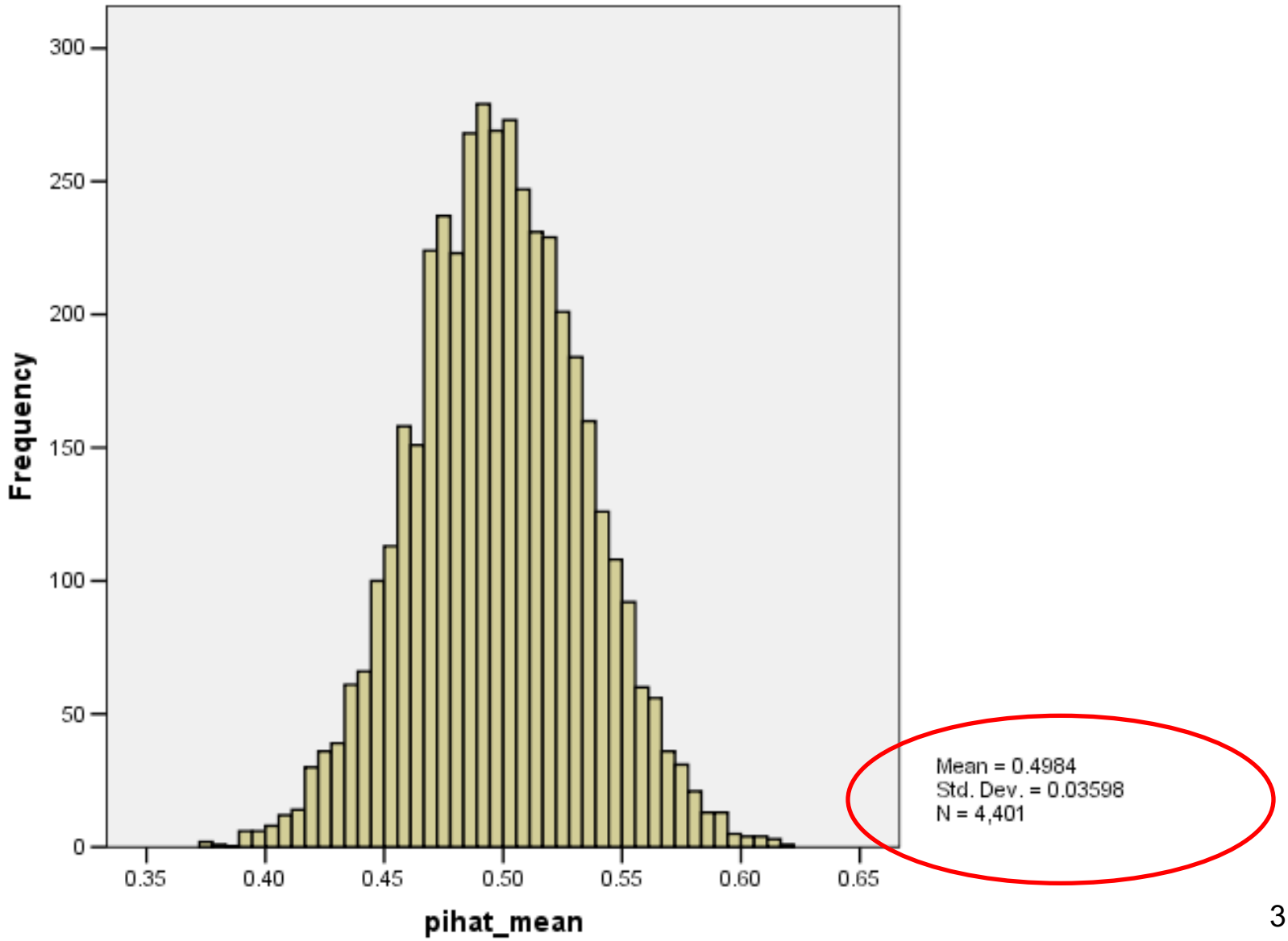
average over SNPs

b) Use haplotypes ?

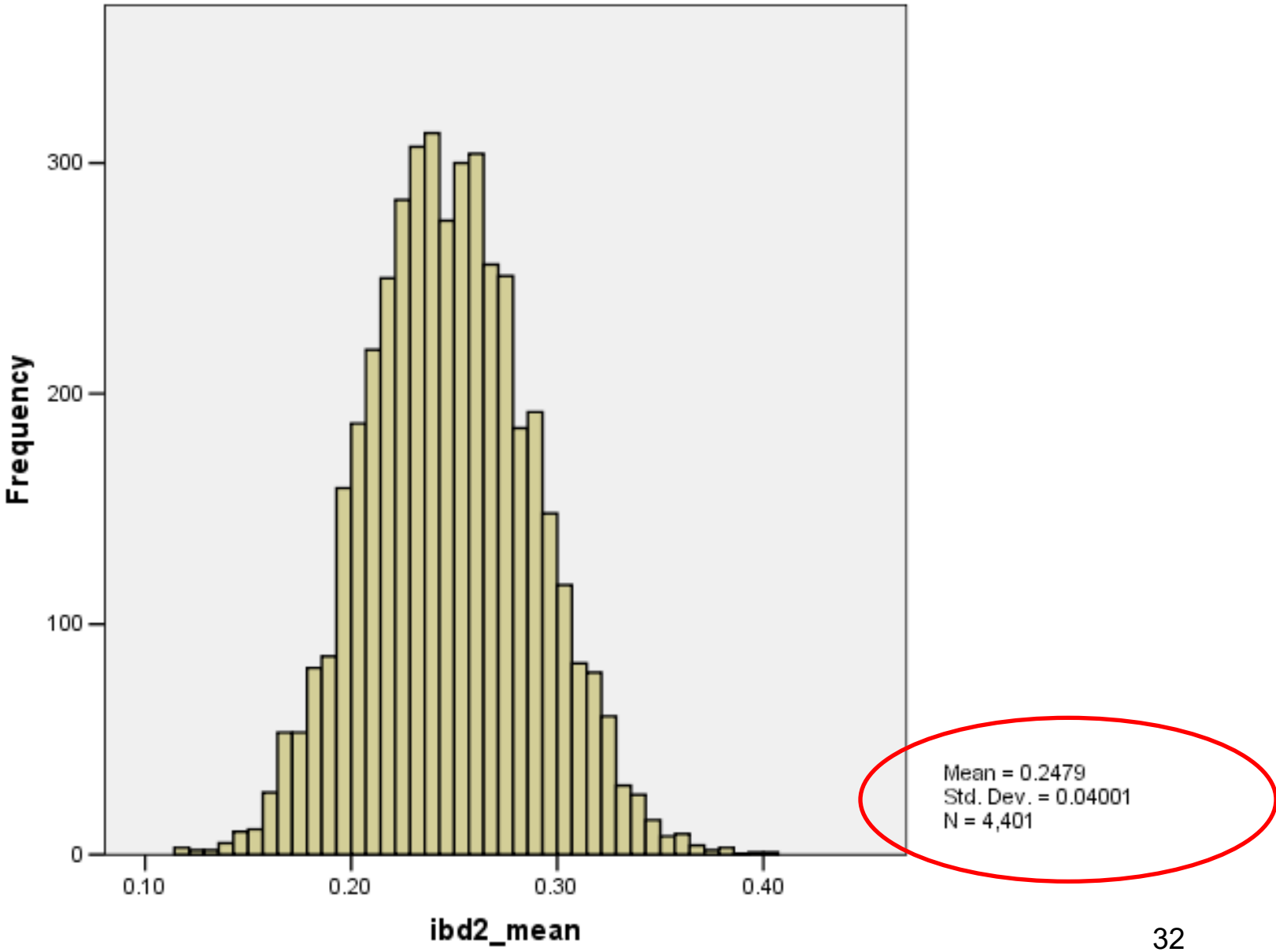
Applications (3x)

- Estimation and partitioning of additive genetic variation within families
  - using realised relationship between fullsibs and their phenotypic covariance
  - phenotype = height

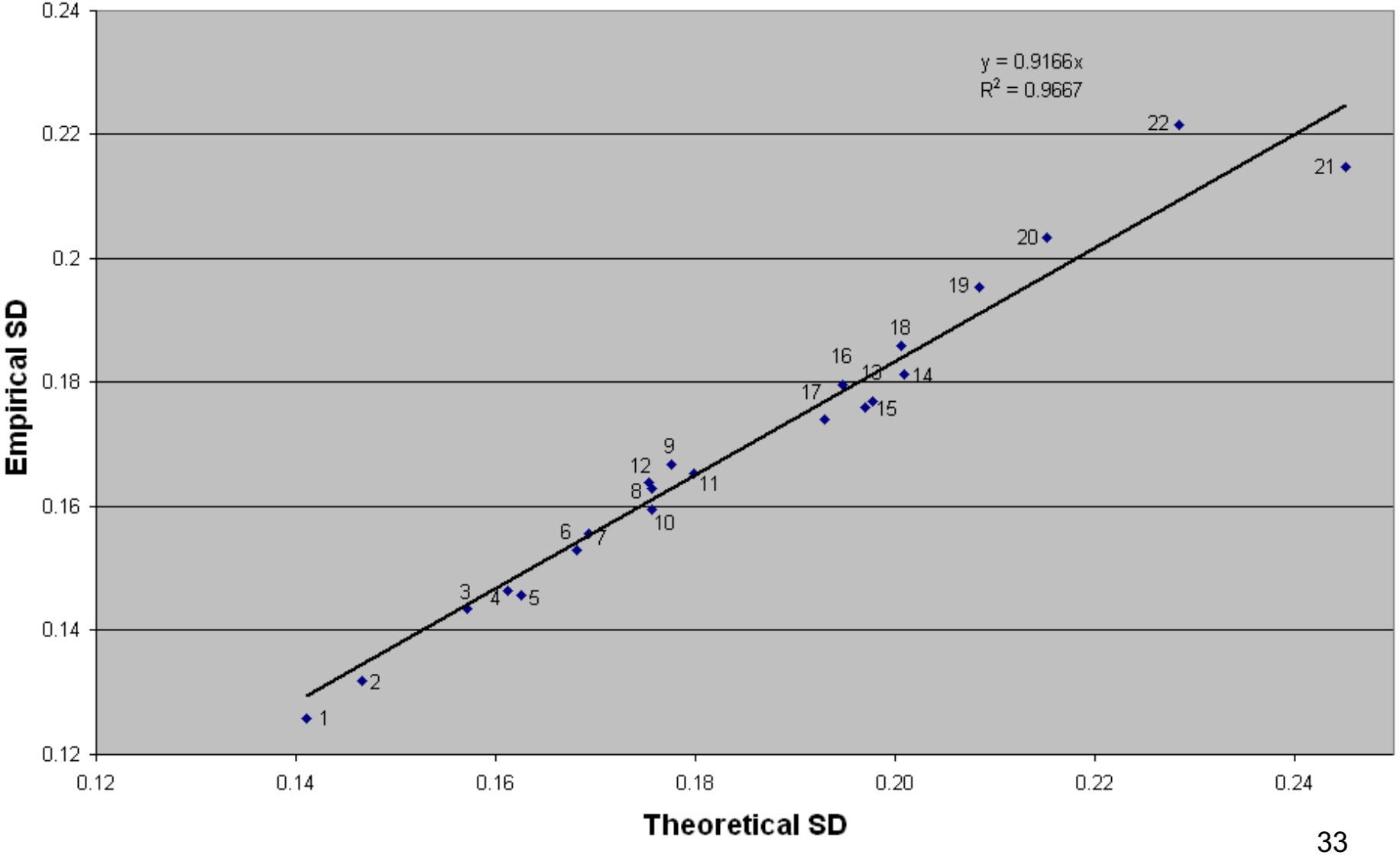
Mean and SD of genome-wide additive relationships



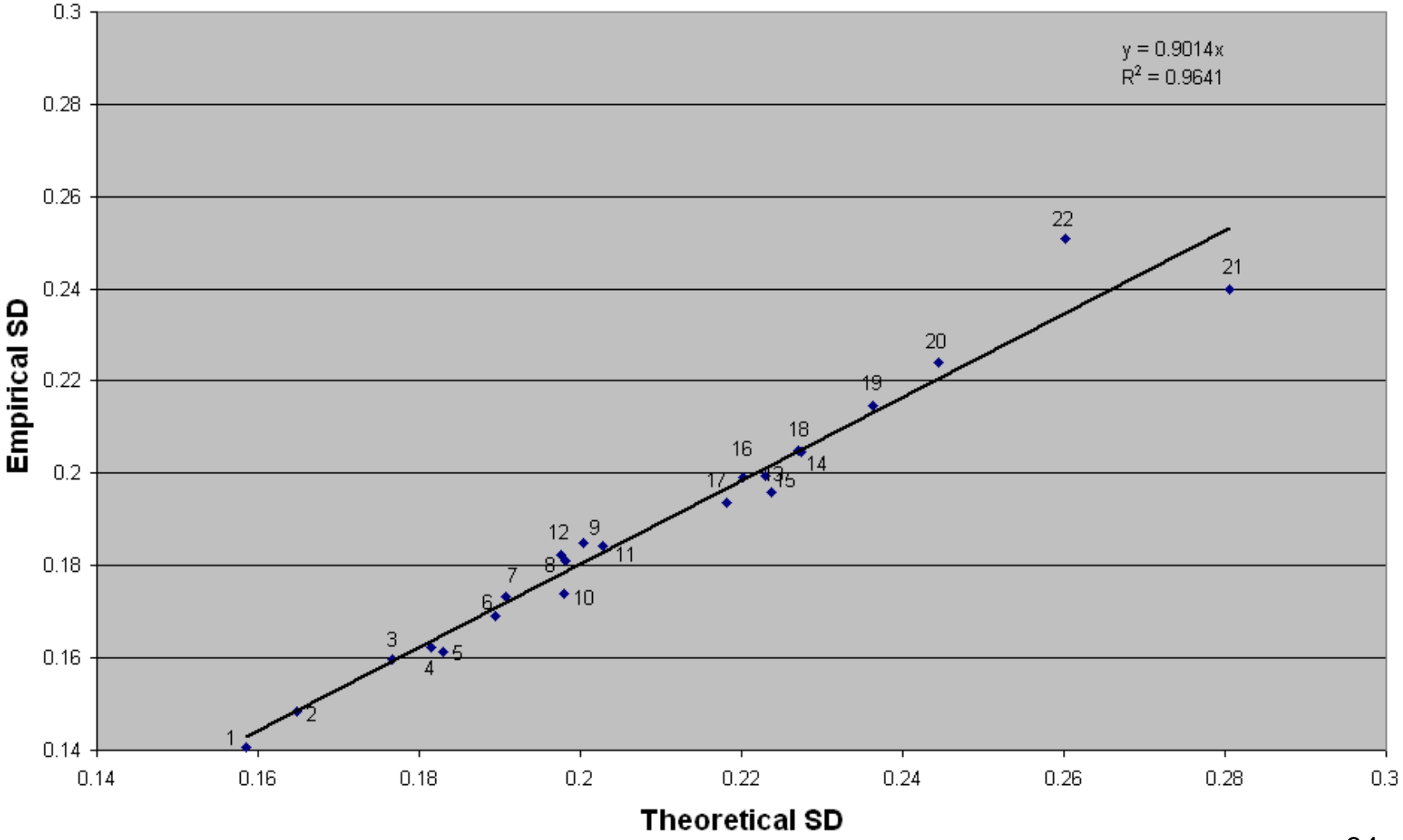
Mean and SD of genome-wide dominance relationships



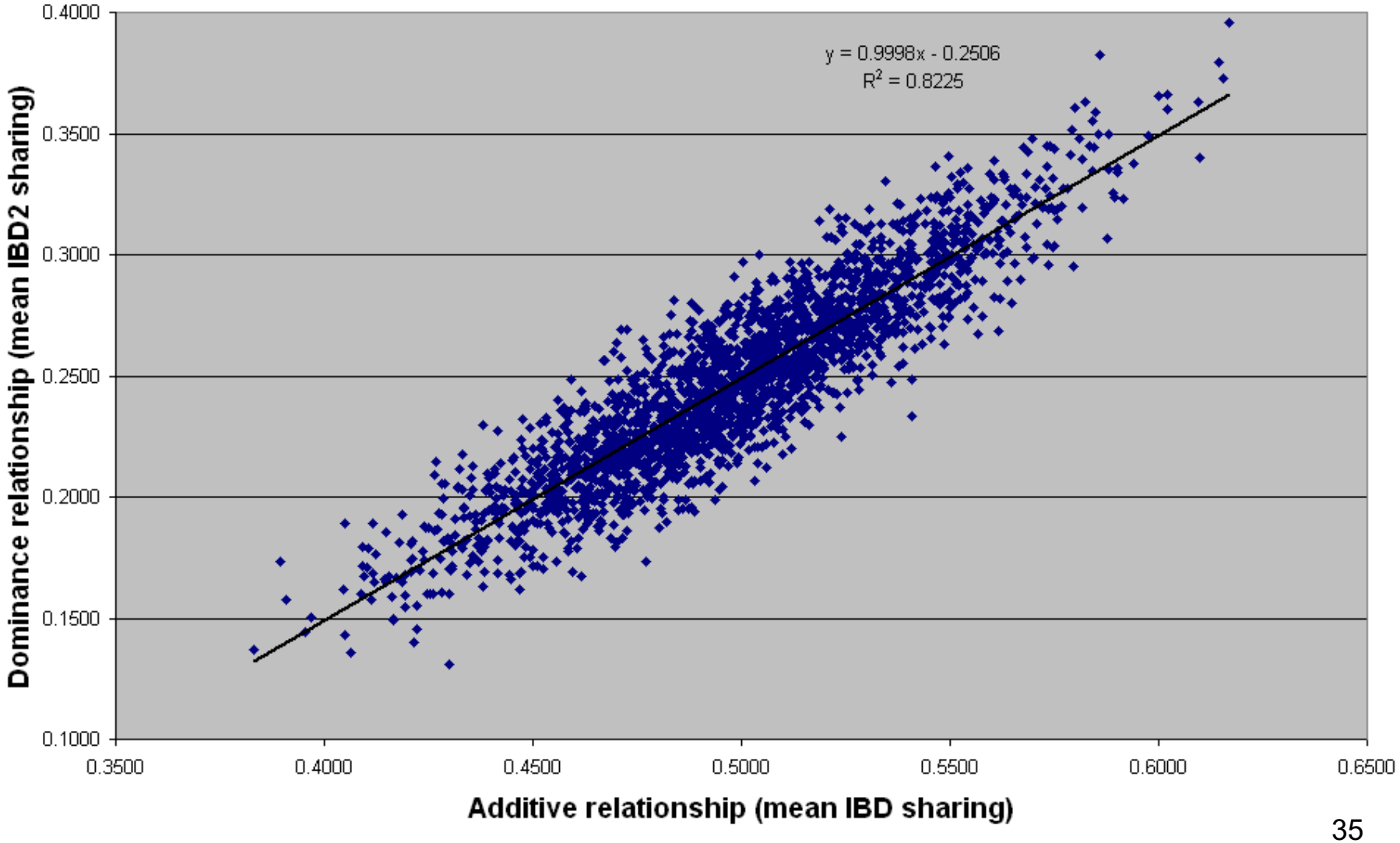
# Empirical and theoretical SD of additive relationships
correlation = 0.98 (*n* = 4401)



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



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



Phenotypic correlation between siblings

	Raw	After age & sex
Adolescents	0.33	0.40
Adults	0.24	0.39

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

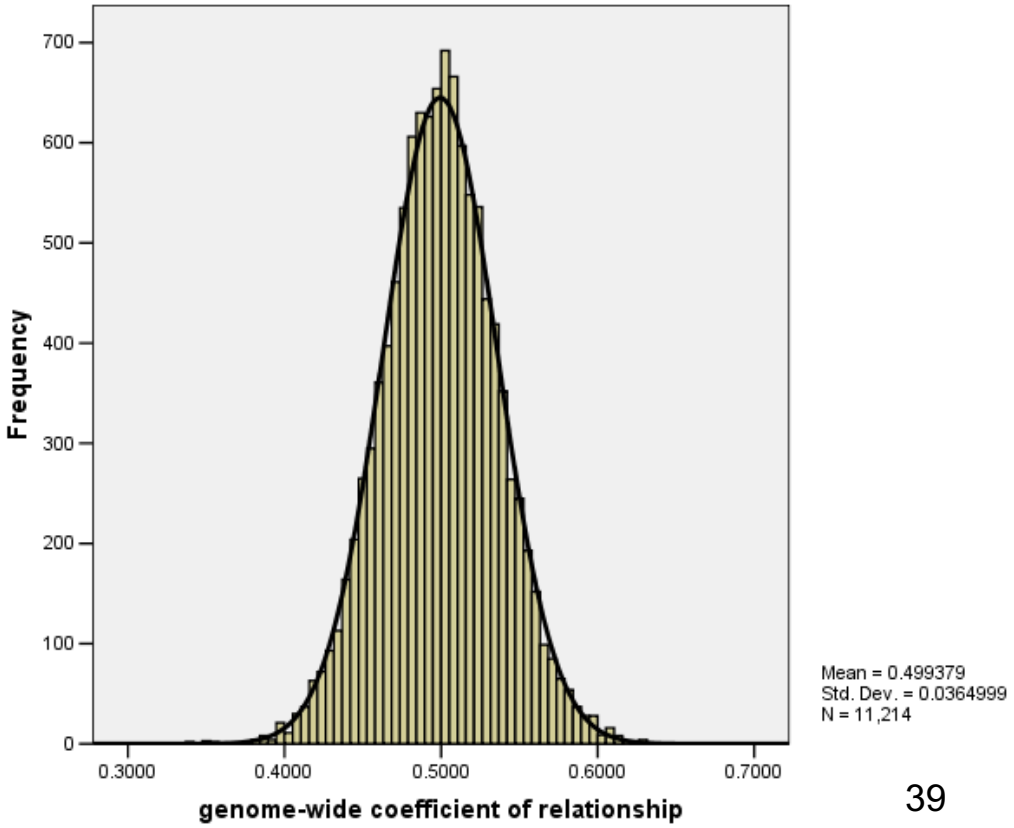
- ***All family resemblance due to additive genetic variation***

Sampling variances are large

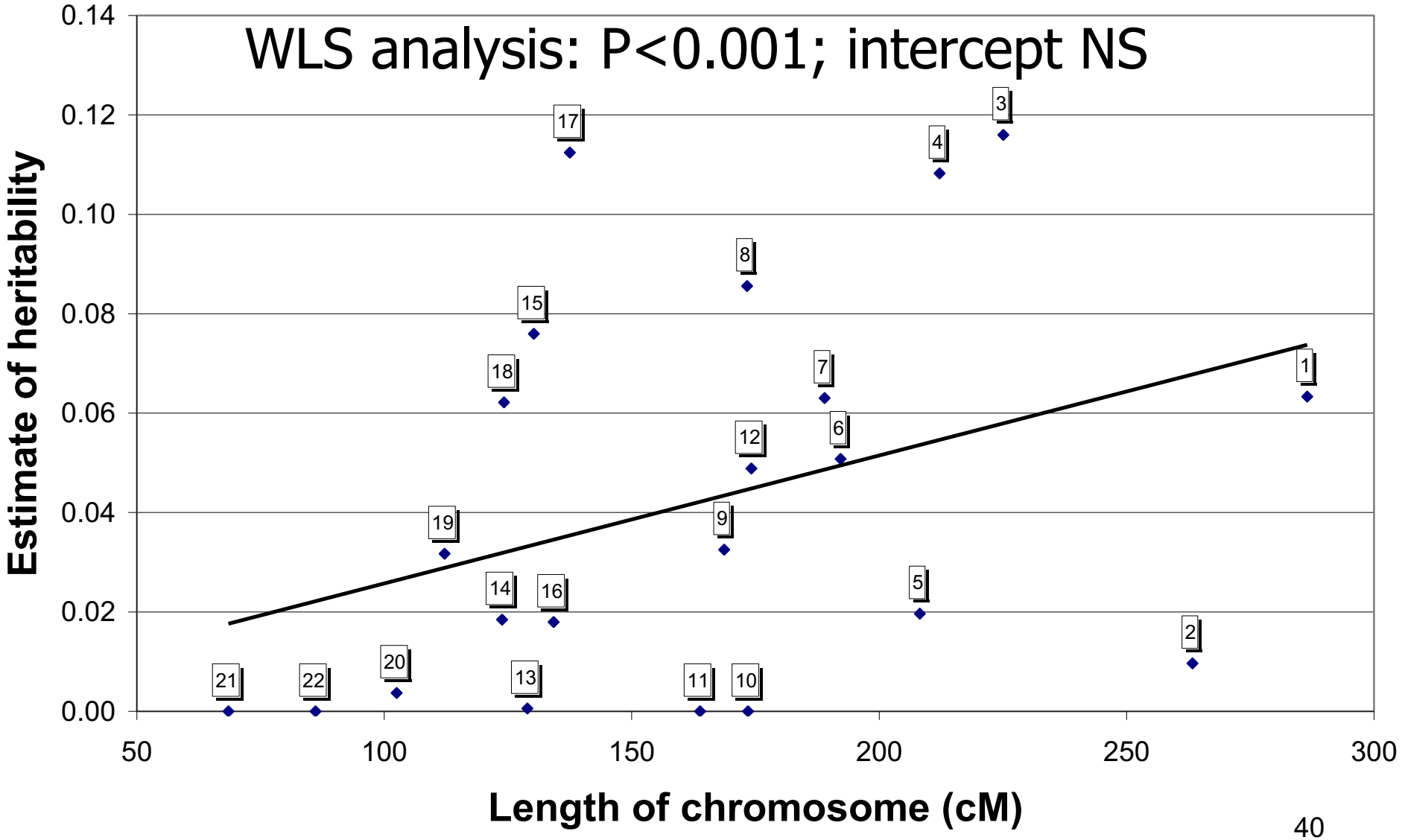
Cohort	A (95% CI)
Adolescent	0.80 (0.00 – 0.90)
Adult	0.80 (0.43 – 0.86)
Combined	0.80 (0.46 – 0.85)

Application (2)

Mean	0.499
Range	0.31 – 0.64
SD	0.036



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



Application (3)

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,*]

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

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]**

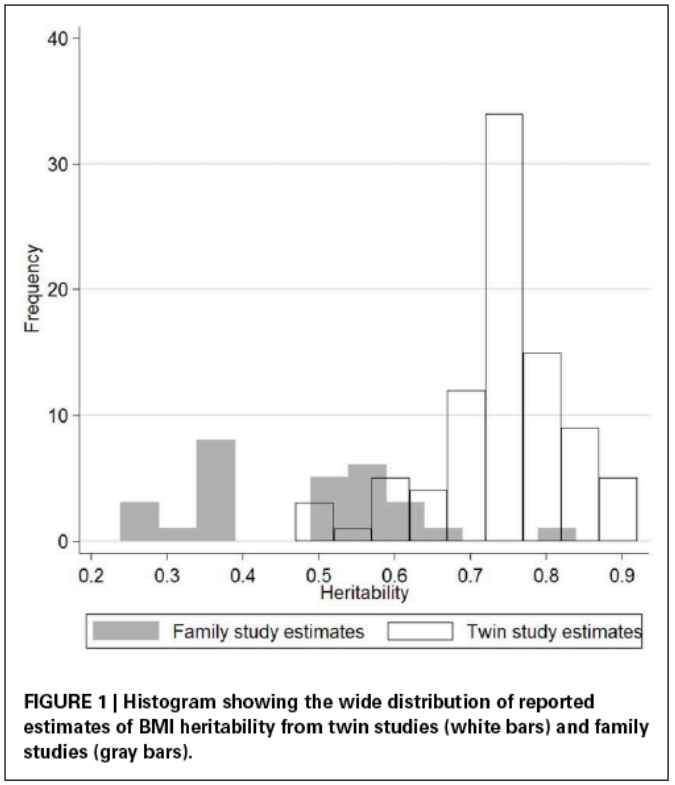
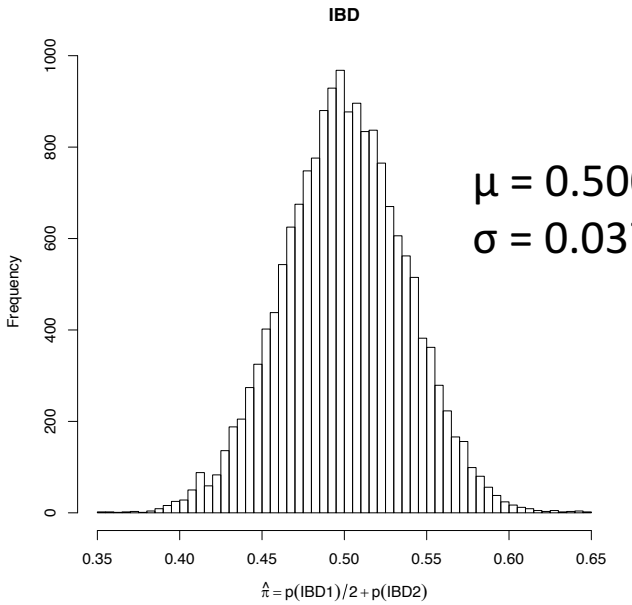


FIGURE 1 | Histogram showing the wide distribution of reported estimates of BMI heritability from twin studies (white bars) and family studies (gray bars).

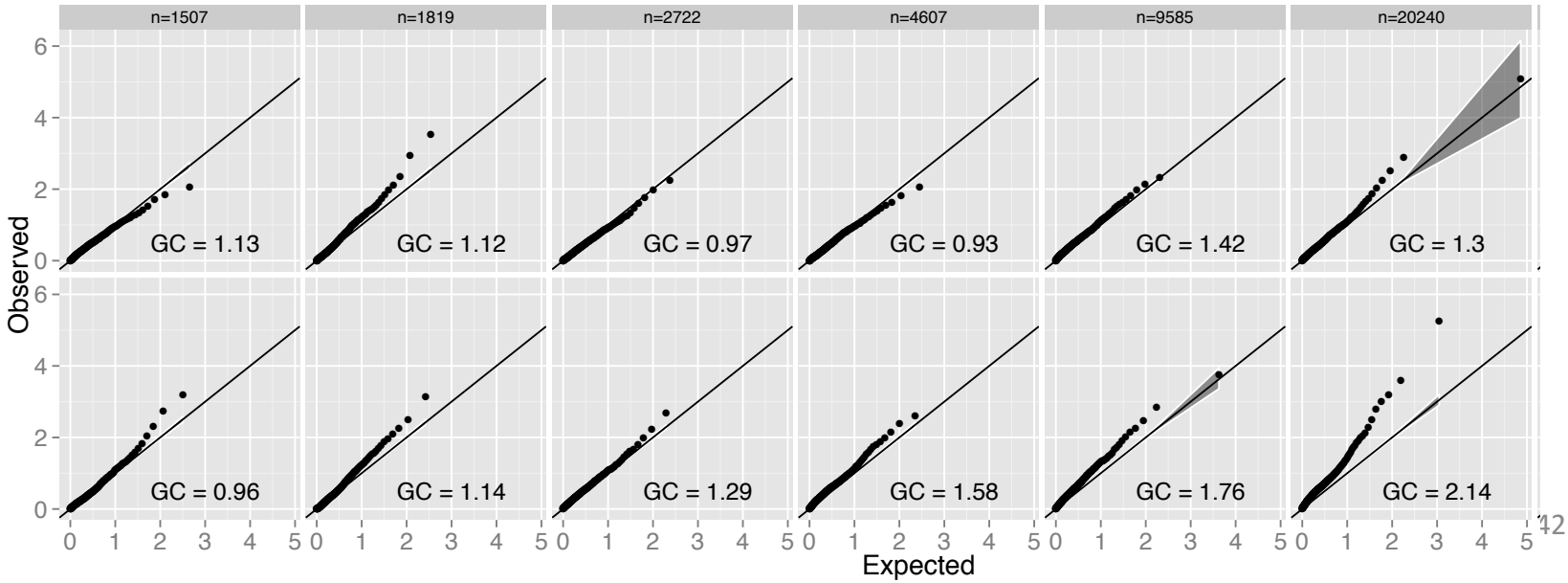
Genetic variation within families using SNP data



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

Heritability estimates from ~20,000 fullsib pairs:

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



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