Estimating genetic variation within families

Peter M. Visscher

Queensland Institute of Medical Research

Brisbane, Australia

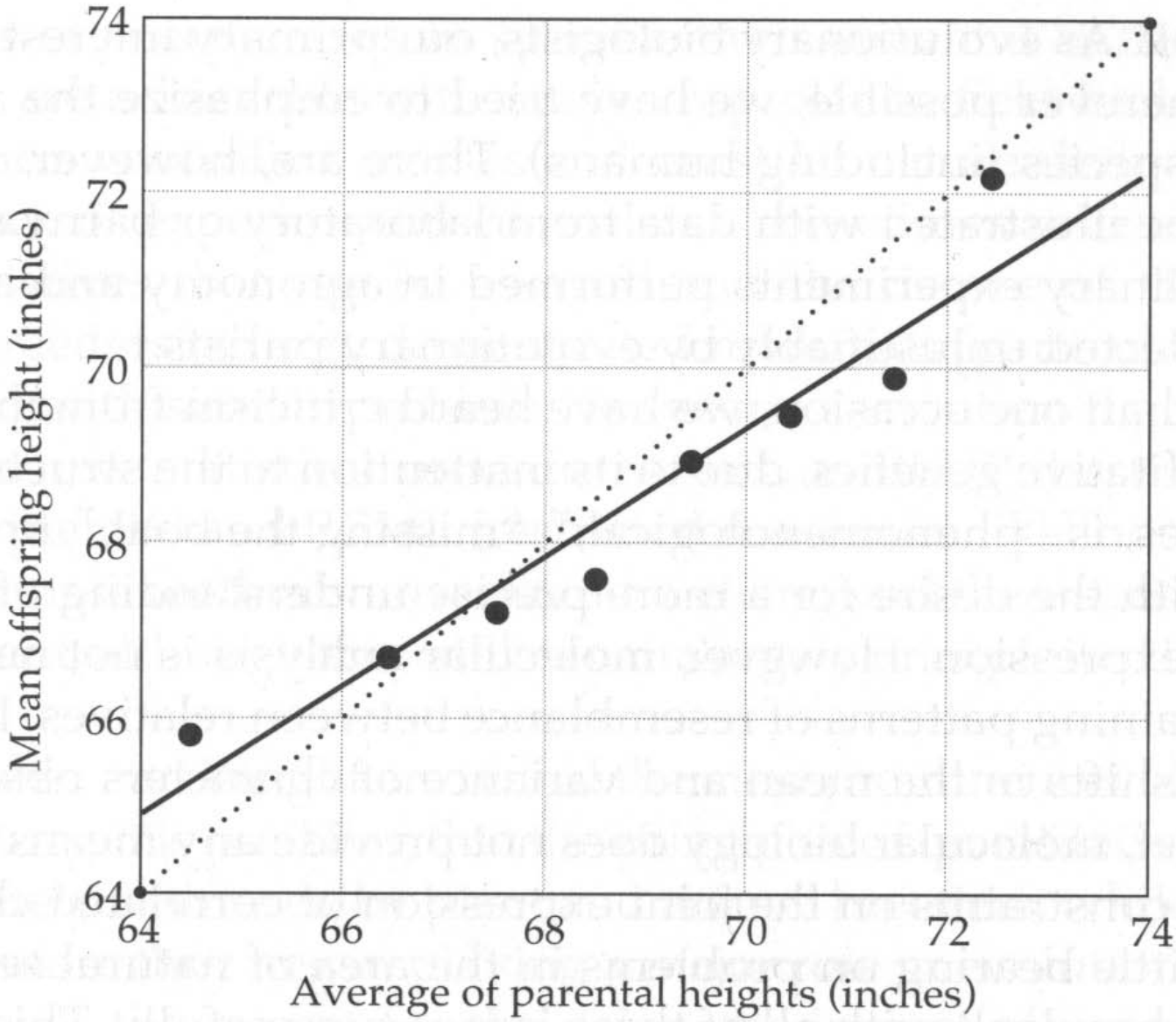
peter.visscher@qimr.edu.au

Overview

- Estimation of genetic parameters
- Variation in identity
- Applications
  - mean and variance of genome-wide IBD sharing for sibpairs
  - estimation of heritability of height
  - genome partitioning of genetic variation

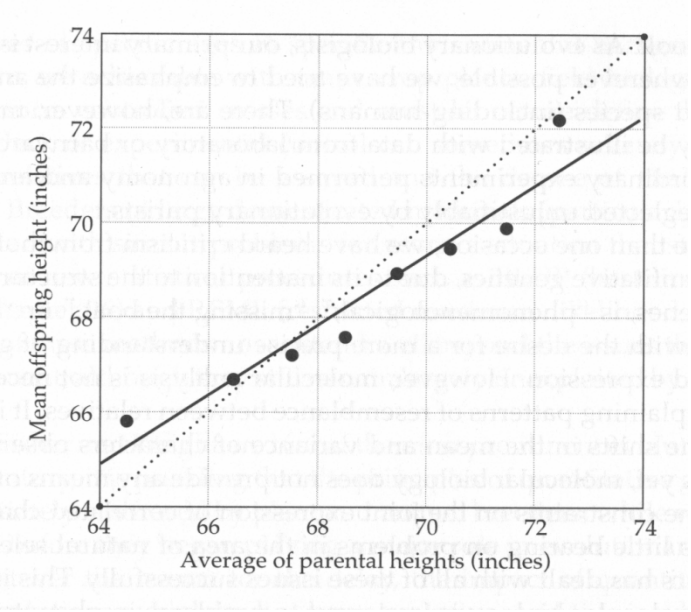
Estimation of genetic parameters

- Model
  - expected covariance between relatives
    - Genetics
    - Environment
- Data
  - correlation/regression of observations between relatives
- Statistical method
  - ANOVA
  - regression
  - maximum likelihood
  - Bayesian analysis


74
72
70
68
66
64
Mean offspring height (inches)
64
66
68
70
72
74
Average of parental heights (inches)

[Galton, 1889]

The height vs. pea debate (early 1900s)



Biometricians

Mendelians

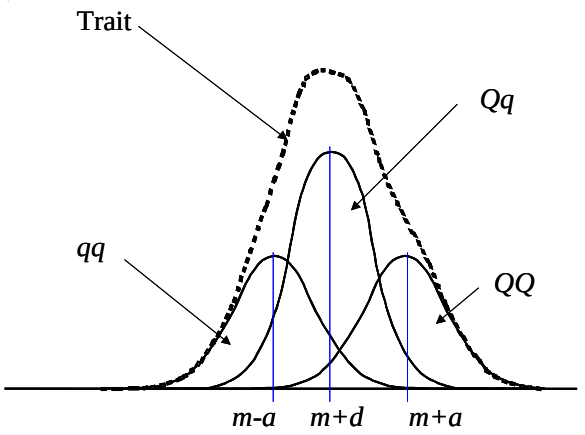
Do quantitative traits have the same hereditary and evolutionary properties as discrete characters?

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

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

CONTENTS.

Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this scheme. It is hoped that in this way it will be possible to make a more exact analysis of the causes of human variability. The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations σ_1 and σ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\sigma_1^2+\sigma_2^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce. It



RA Fisher (1918). *Transactions of the Royal Society of Edinburgh* **52**: 399-433.

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 * coefficient of kinship (= $E(\pi)$)

d = coefficient of fraternity

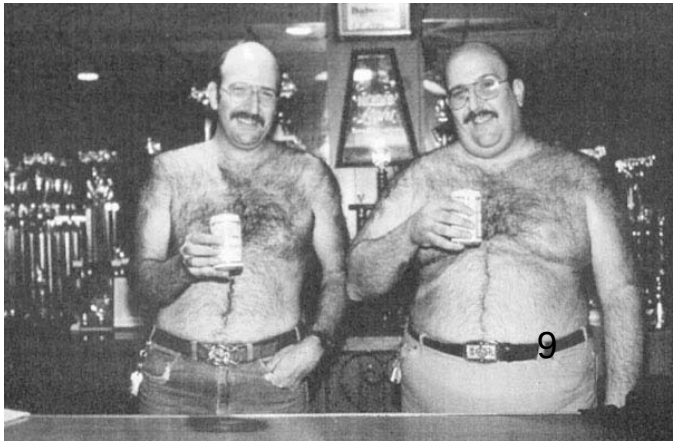
= Prob(2 alleles are IBD)

Examples (no inbreeding)

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

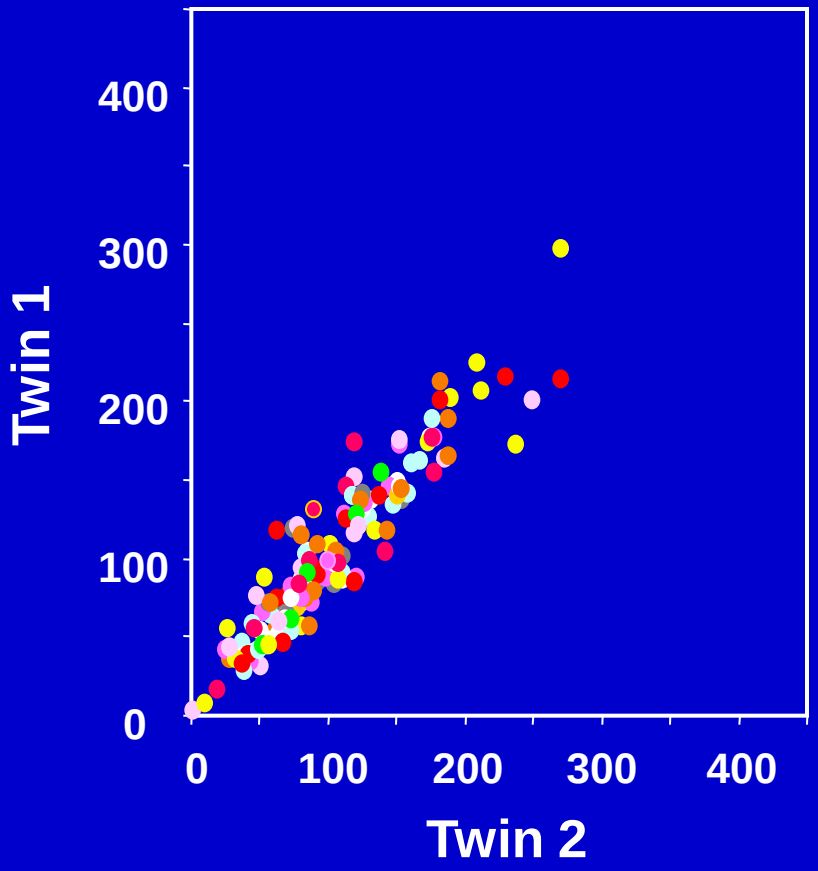
Controversy/confounding: nature vs nurture

- 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

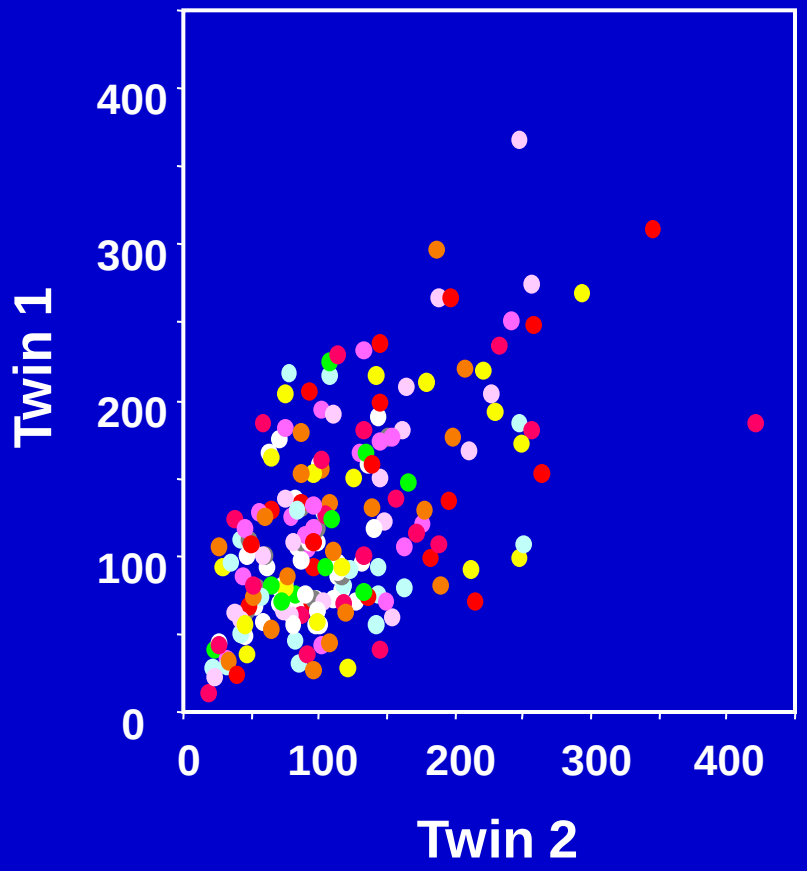


Total mole count for MZ and DZ twins

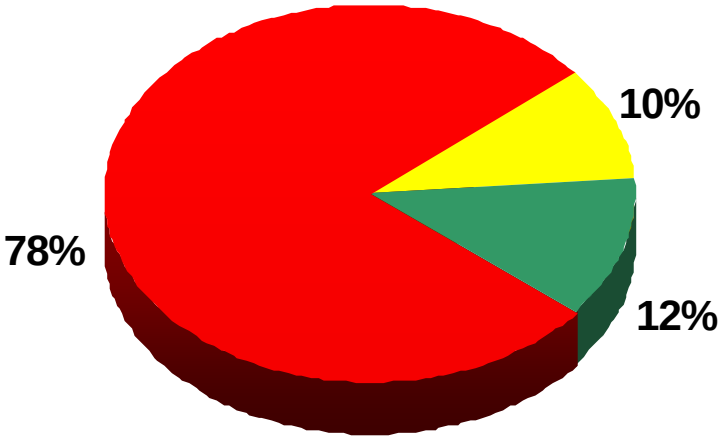
MZ twins - 153 pairs, r = 0.94



DZ twins - 199 pairs, r = 0.60



Sources of variation in Queensland school test results of 16-year olds



Additive genetic

Shared environment

Non-shared environment

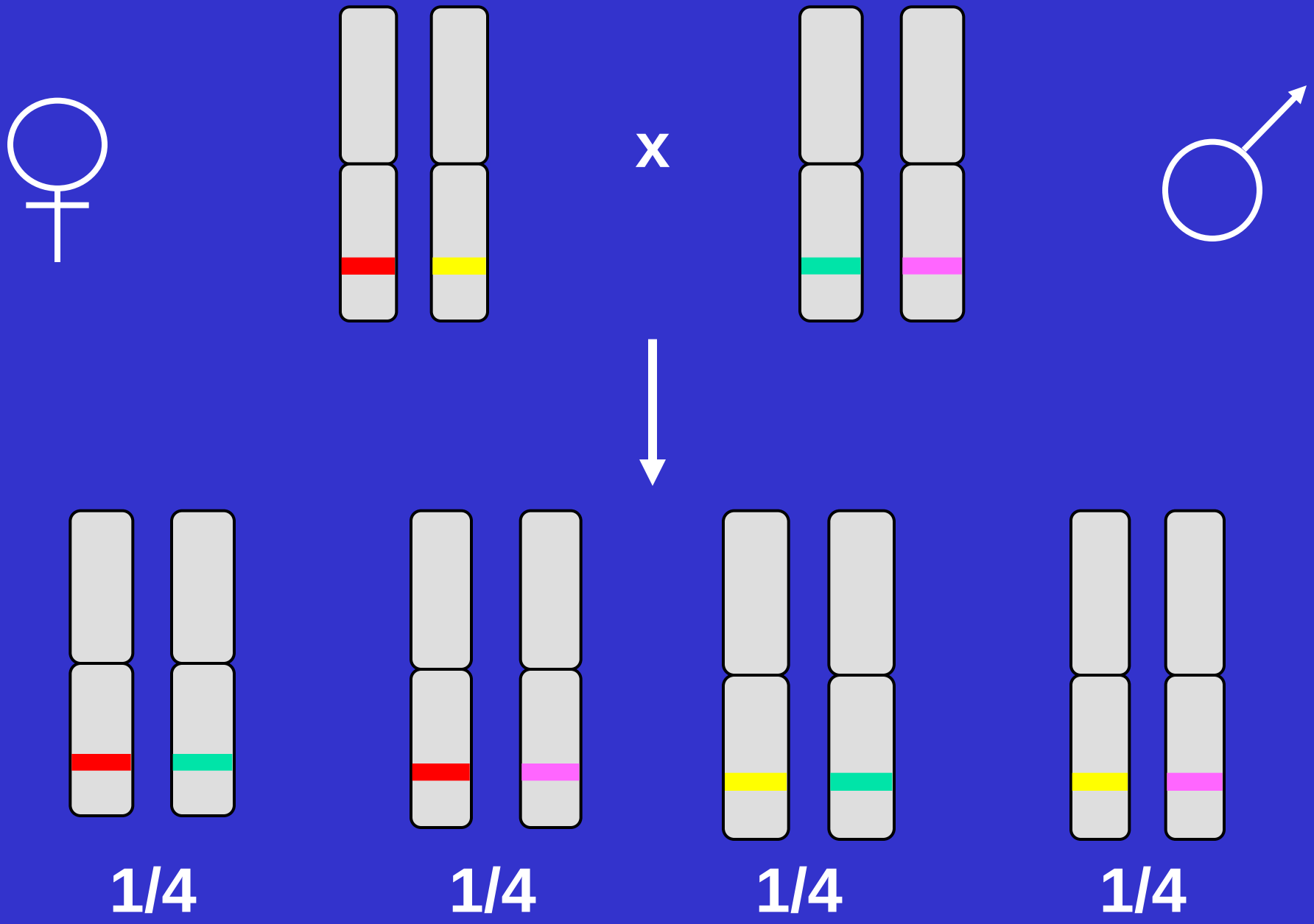
An unbiased approach

Estimate genetic variance within families

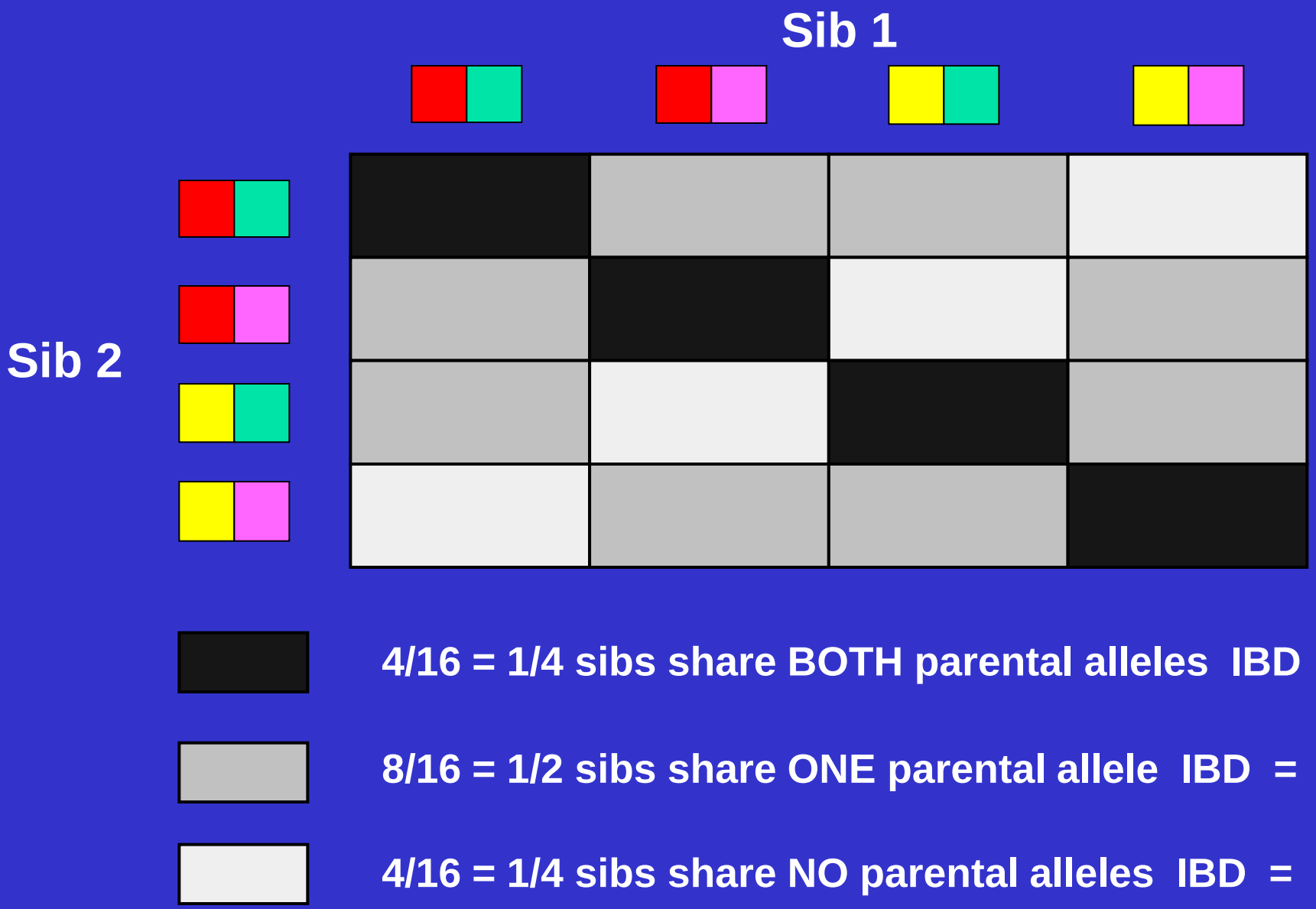
Actual or *realised* genetic relationship

= 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(π_a)	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 π_a and π_d)

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 from Merlin every centimorgan and averaged at all 3491 points

additive

$$\overline{\hat{\pi}}_{a(j)} = \sum_{i=1}^{3491} \hat{\pi}_{a(ij)} / 3491$$

dominance

$$\overline{\hat{\pi}}_{d(j)} = \sum_{i=1}^{3491} p_{2(ij)} / 3491$$

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

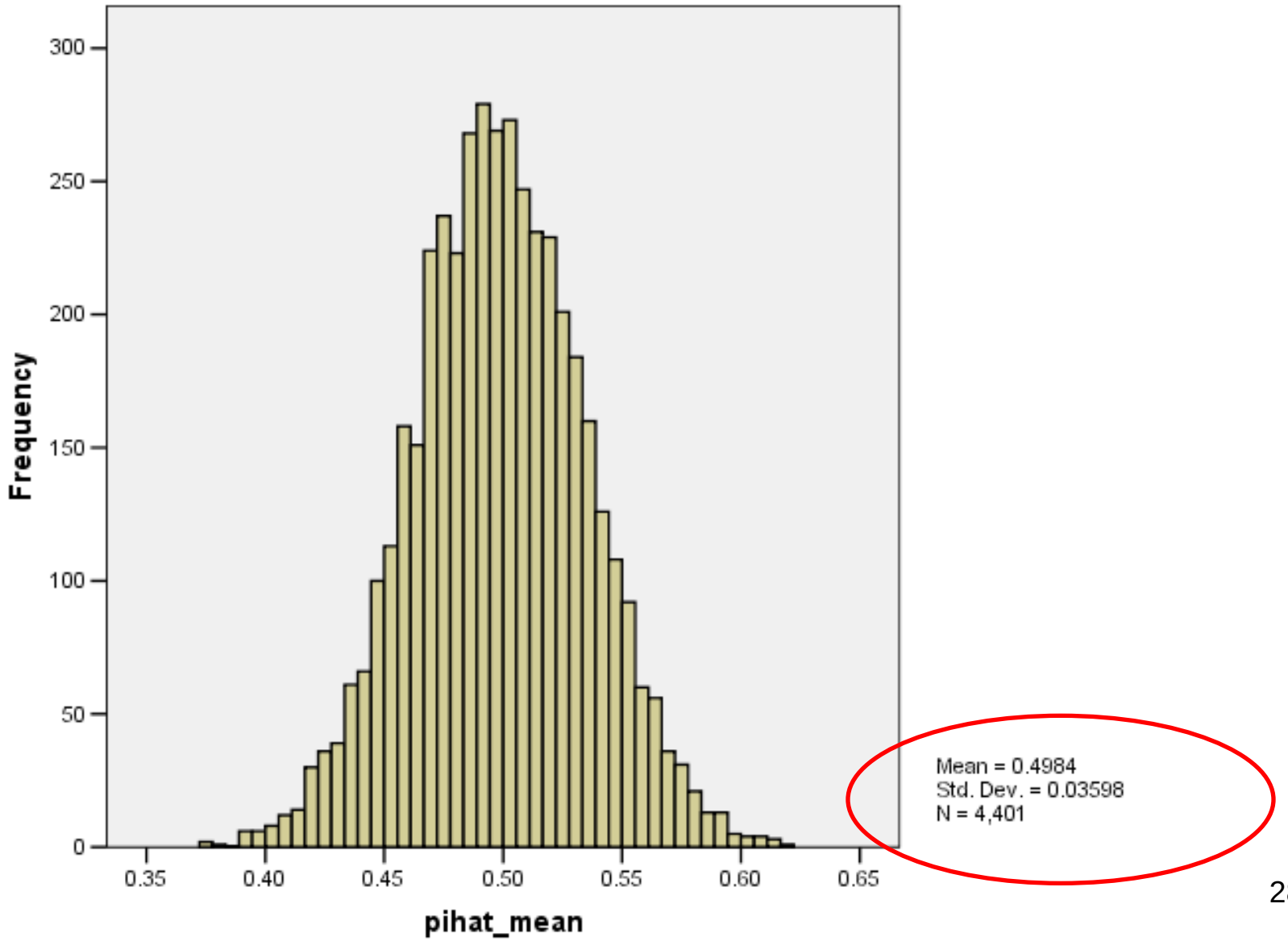
additive

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

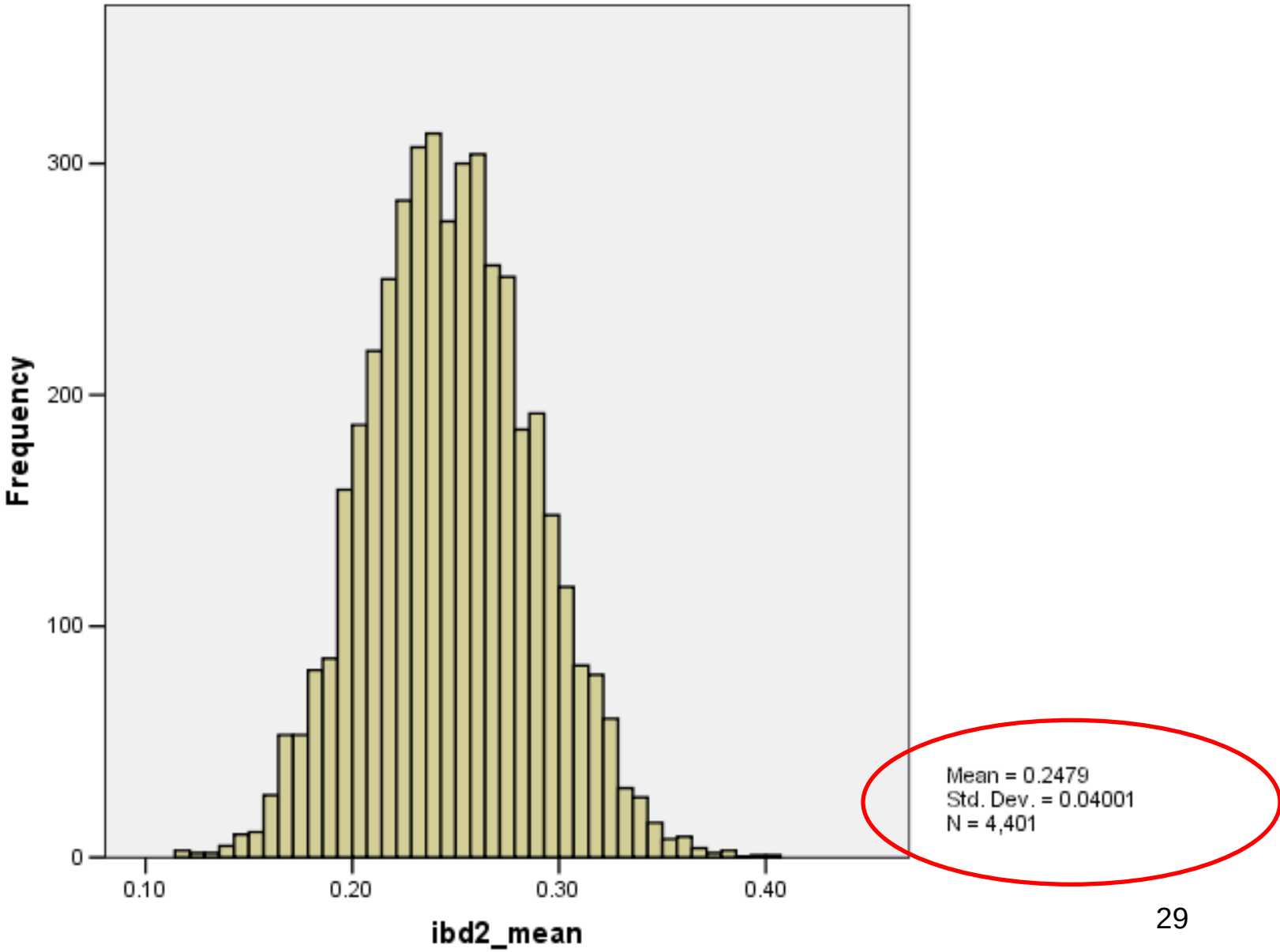
dominance

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

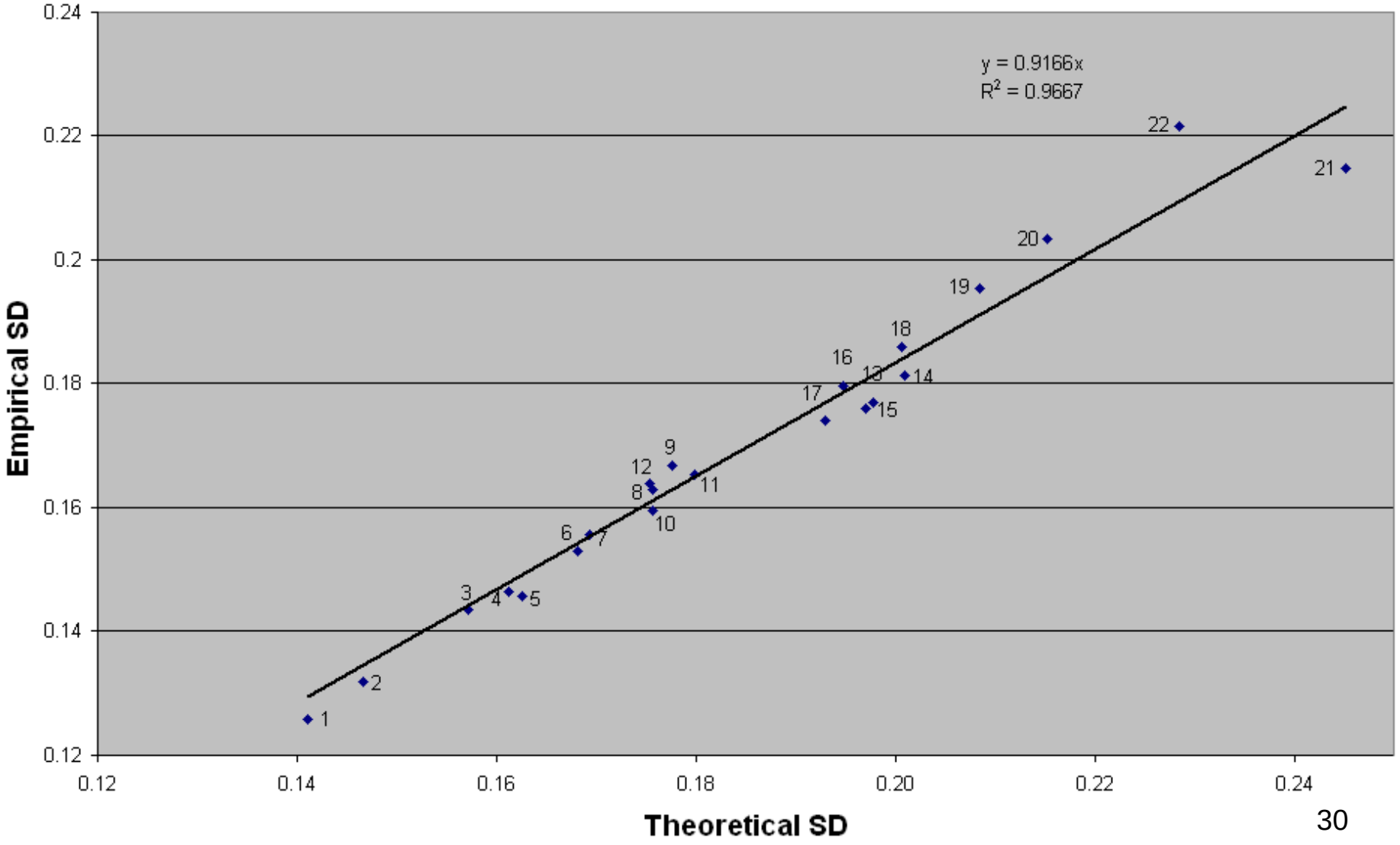
Mean and SD of genome-wide additive relationships



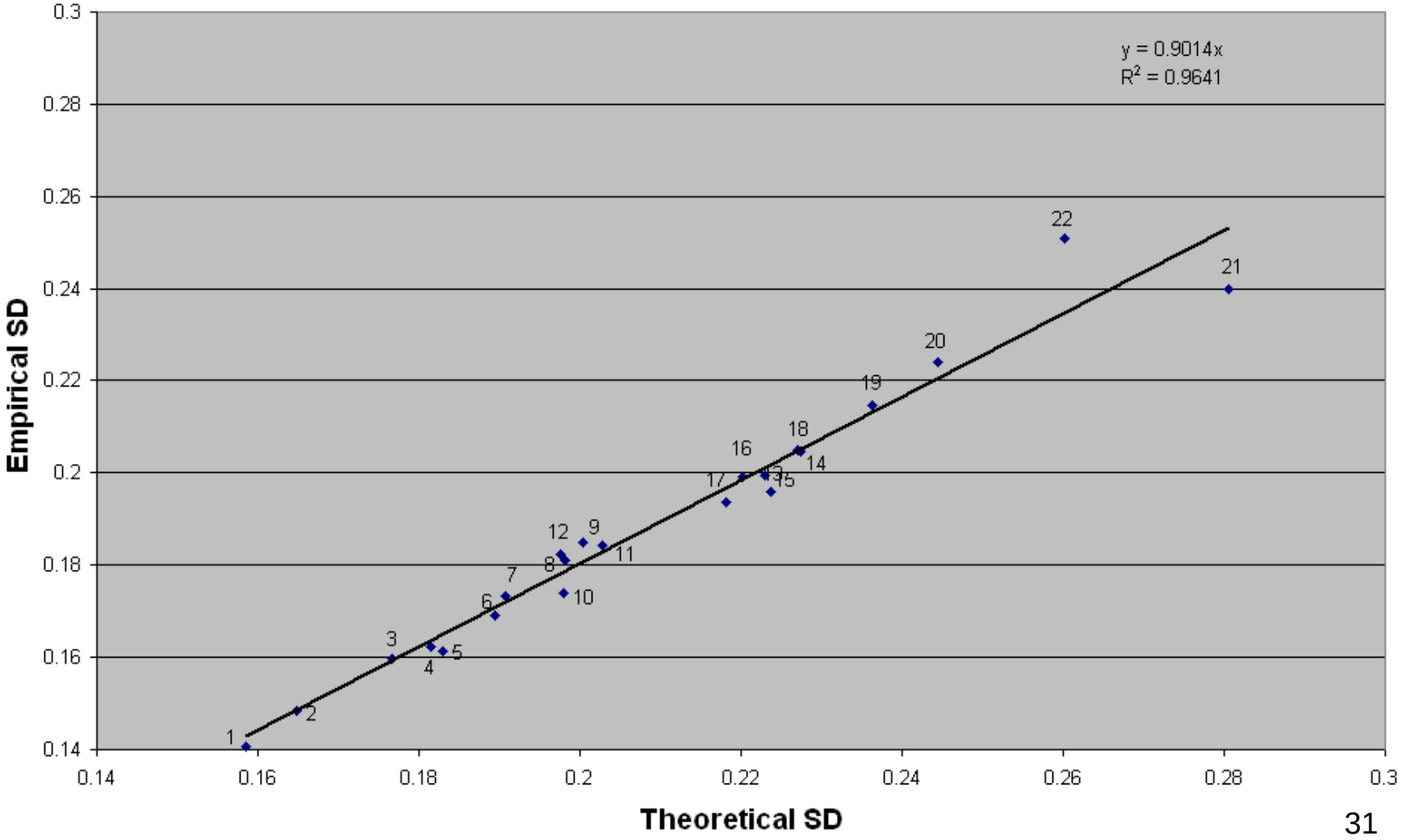
Mean and SD of genome-wide dominance relationships



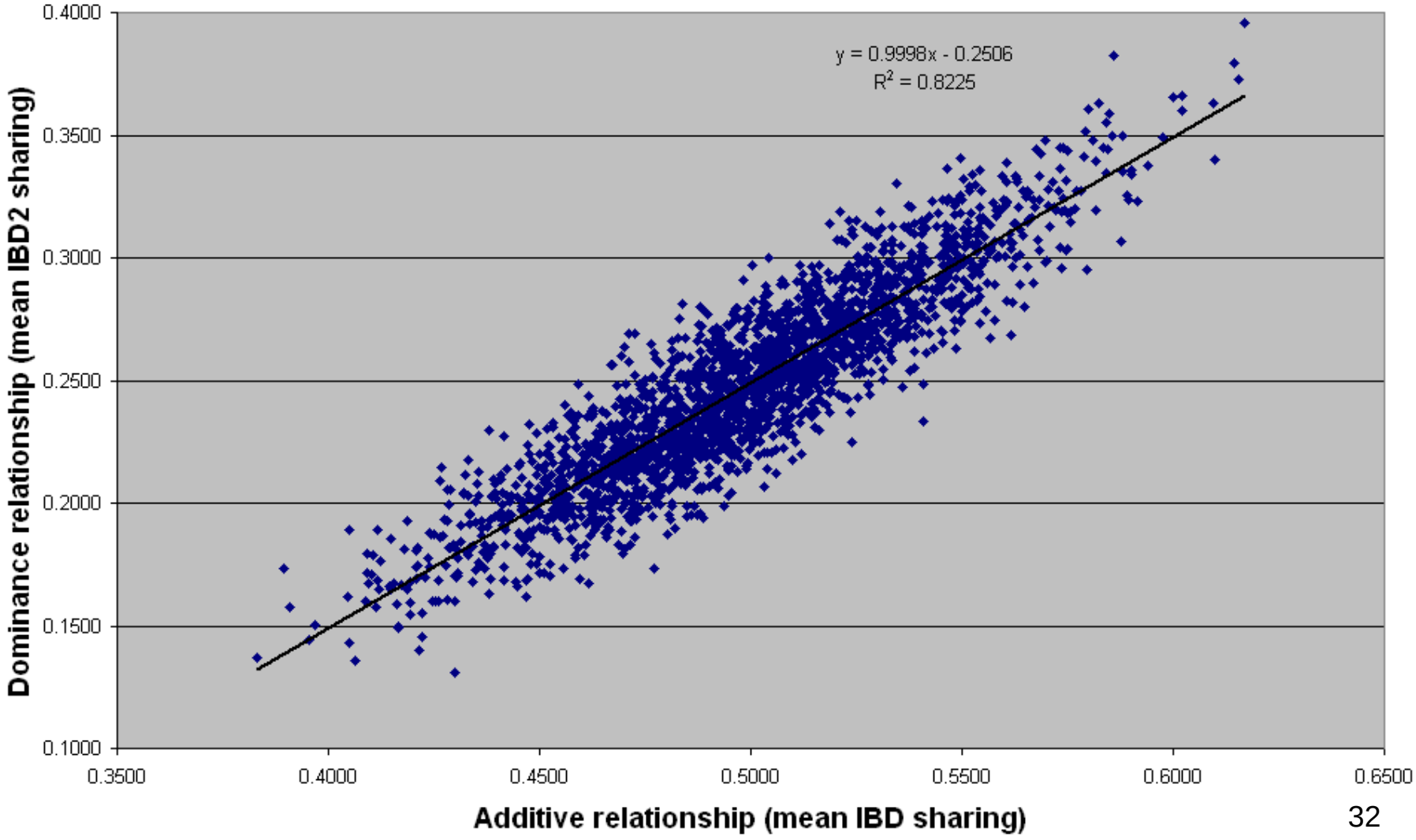
Empirical and theoretical SD of additive relationships correlation = 0.98 (*n* = 4401)



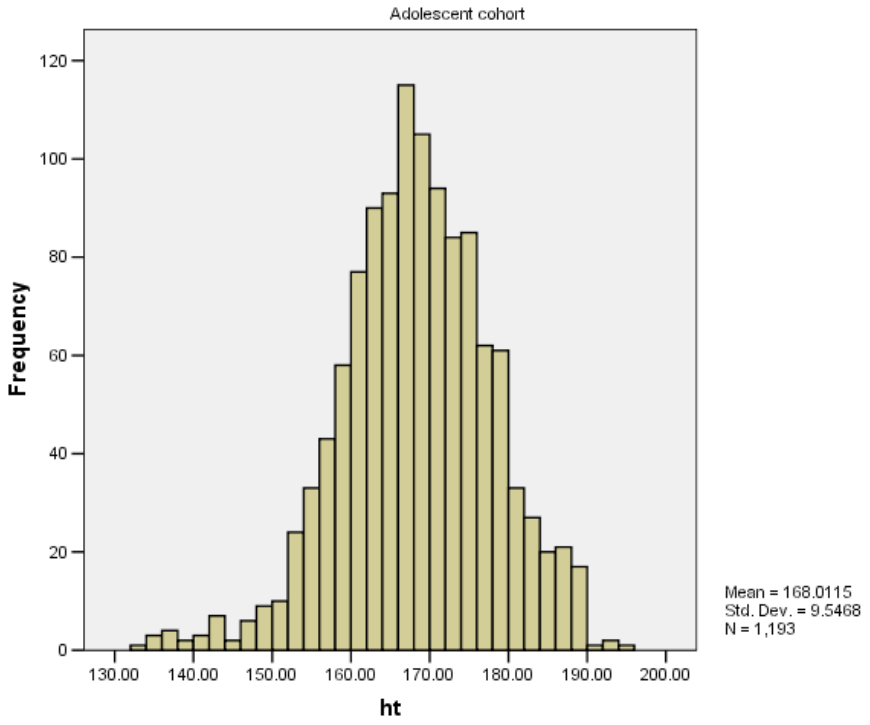
Empirical and theoretical SD of dominance relationships correlation = 0.98 (*n* = 4401)

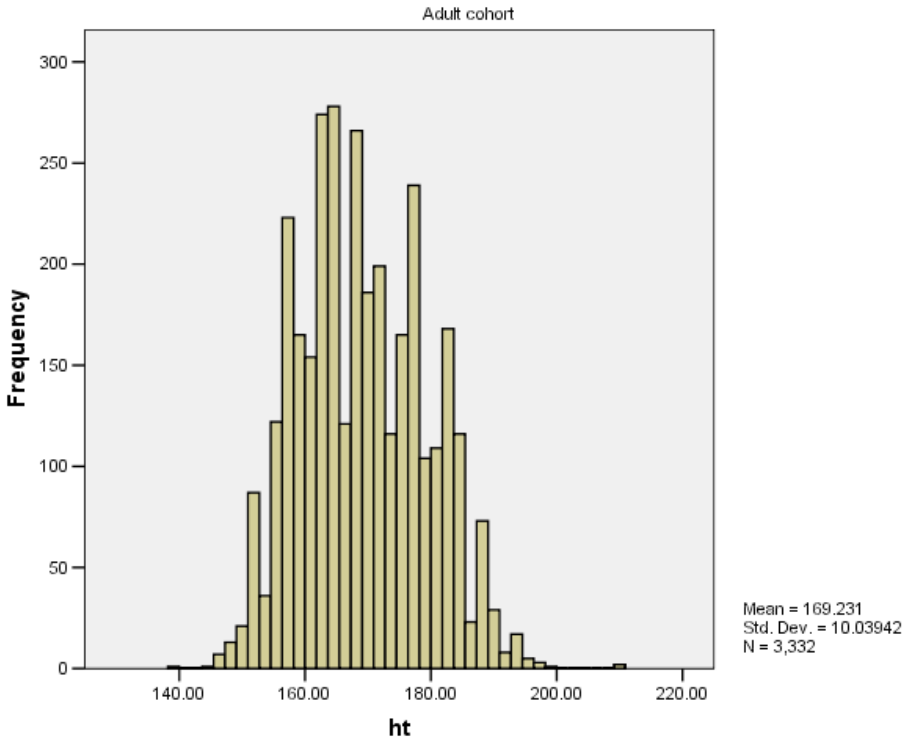


Additive and dominance relationships correlation = 0.91 (*n*= 4401)



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

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)

F+A more accurately estimated

Cohort	C+A (95% CI)
Adolescent	0.80 (0.36 – 0.90)
Adult	0.80 (0.61 – 0.86)
Combined	0.80 (0.62 – 0.85)

► ***Prediction of MZ correlation from fullsibs!***

Power and SE of estimates

- True parameters (*t*)
- Sample size (*n*)
- Variance in genome-wide IBD sharing (var(π))

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

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

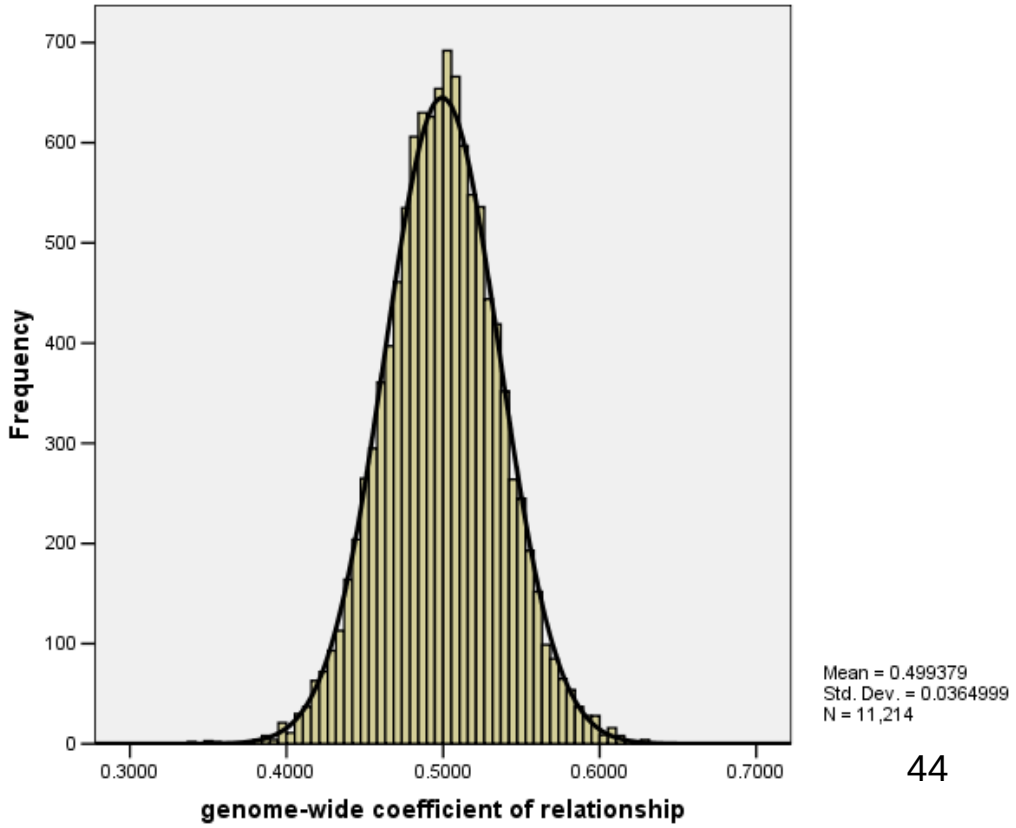
Application (2) Genome partitioning of additive genetic variance for height

- 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



Estimates from genome-wide additive and dominance coefficients

ACE model

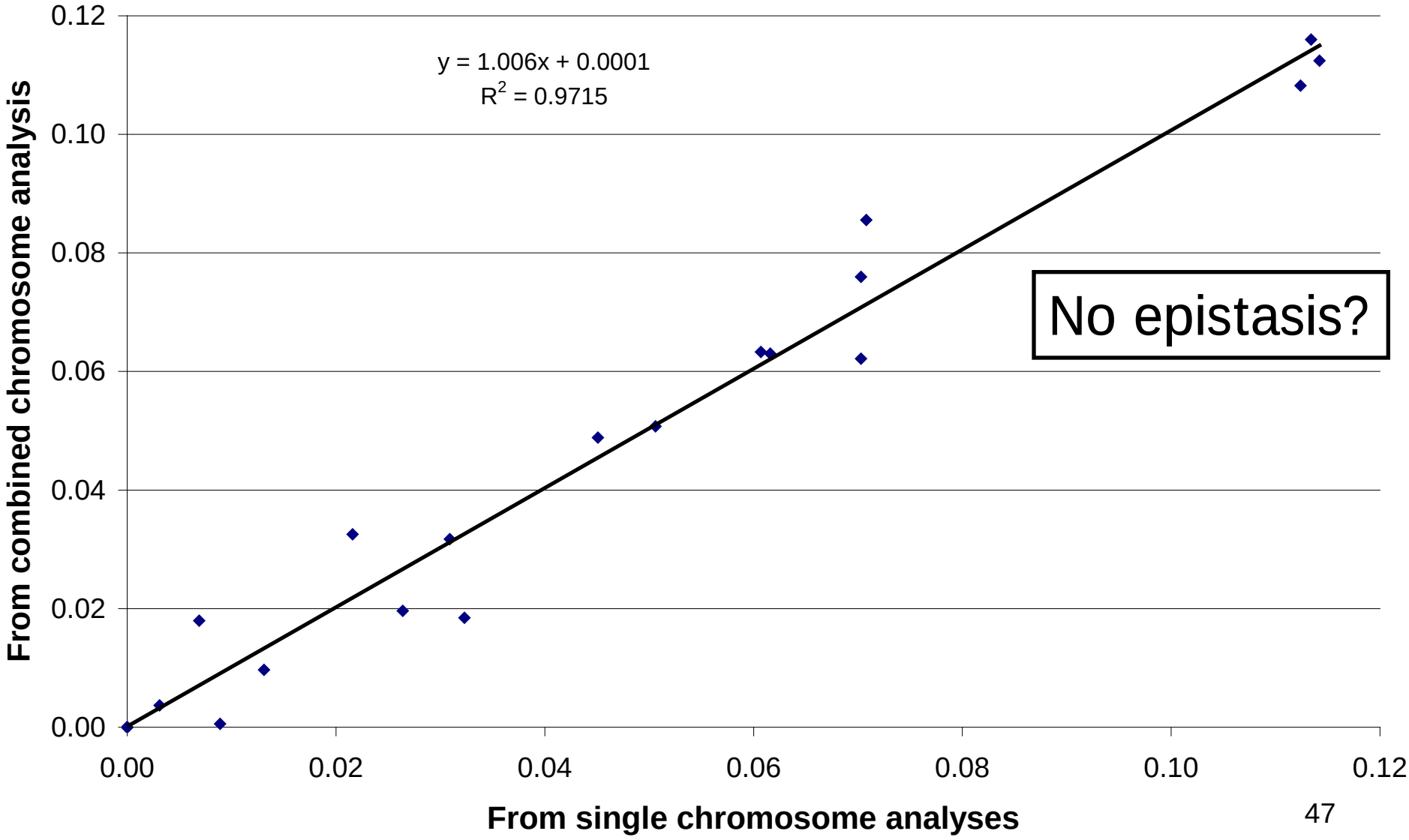
Heritability	0.86 (0.49 – 0.95)	P<0.0001
Family	0.03 (0.00 – 0.03)	P=0.38

ADCE model

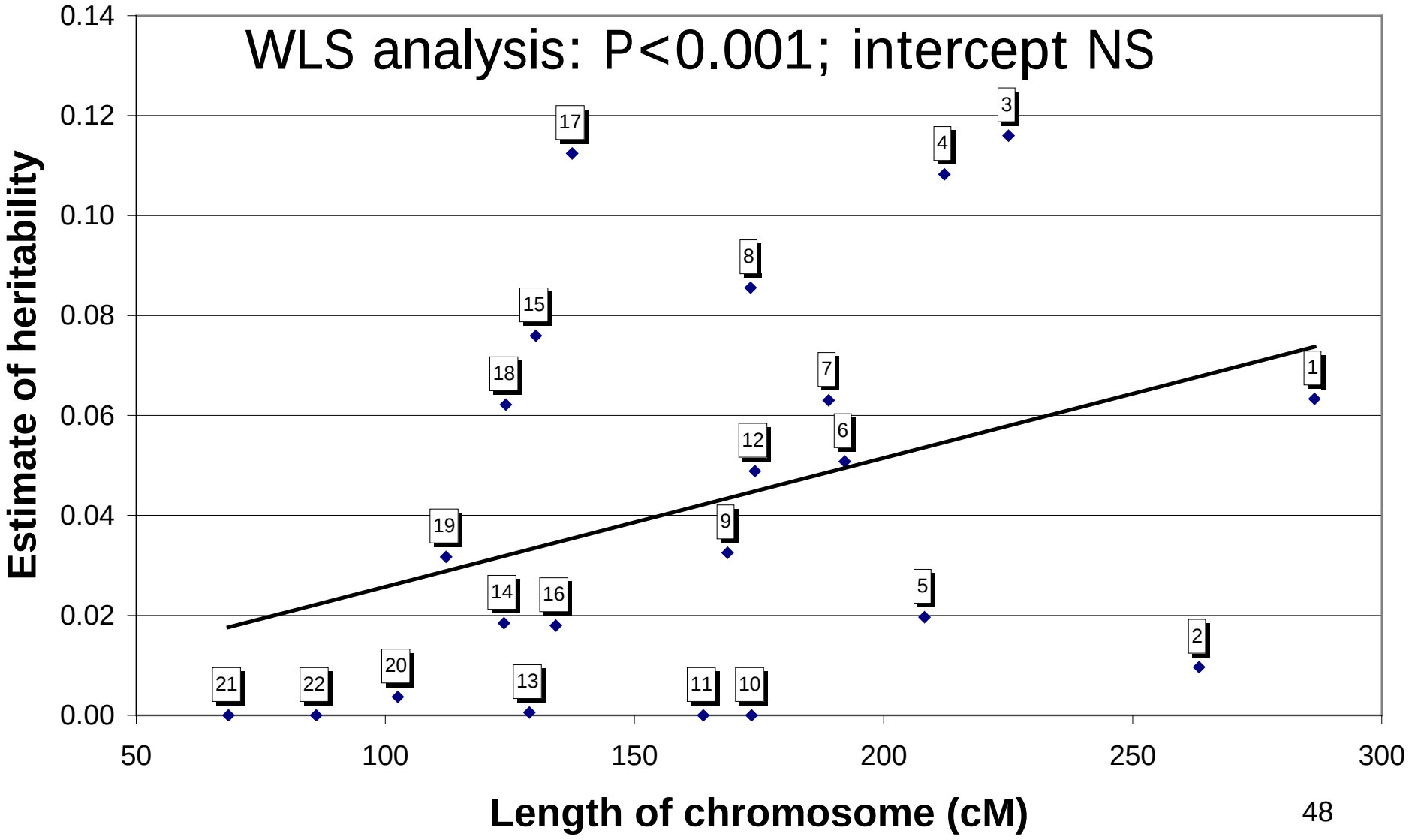
Additive component	0.70
Dominance component	0.16 (P=0.35)

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	

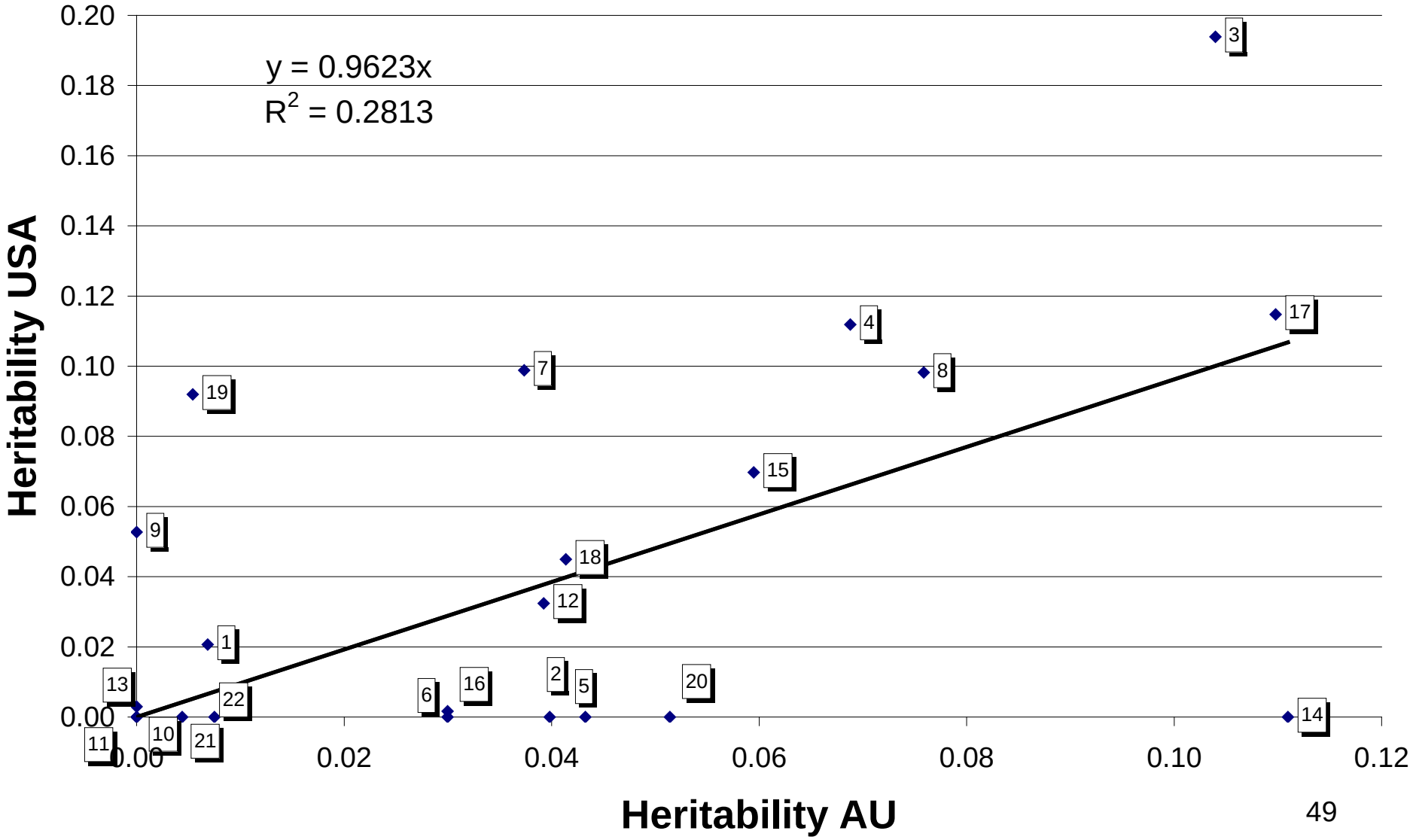
Estimates of chromosomal heritabilities



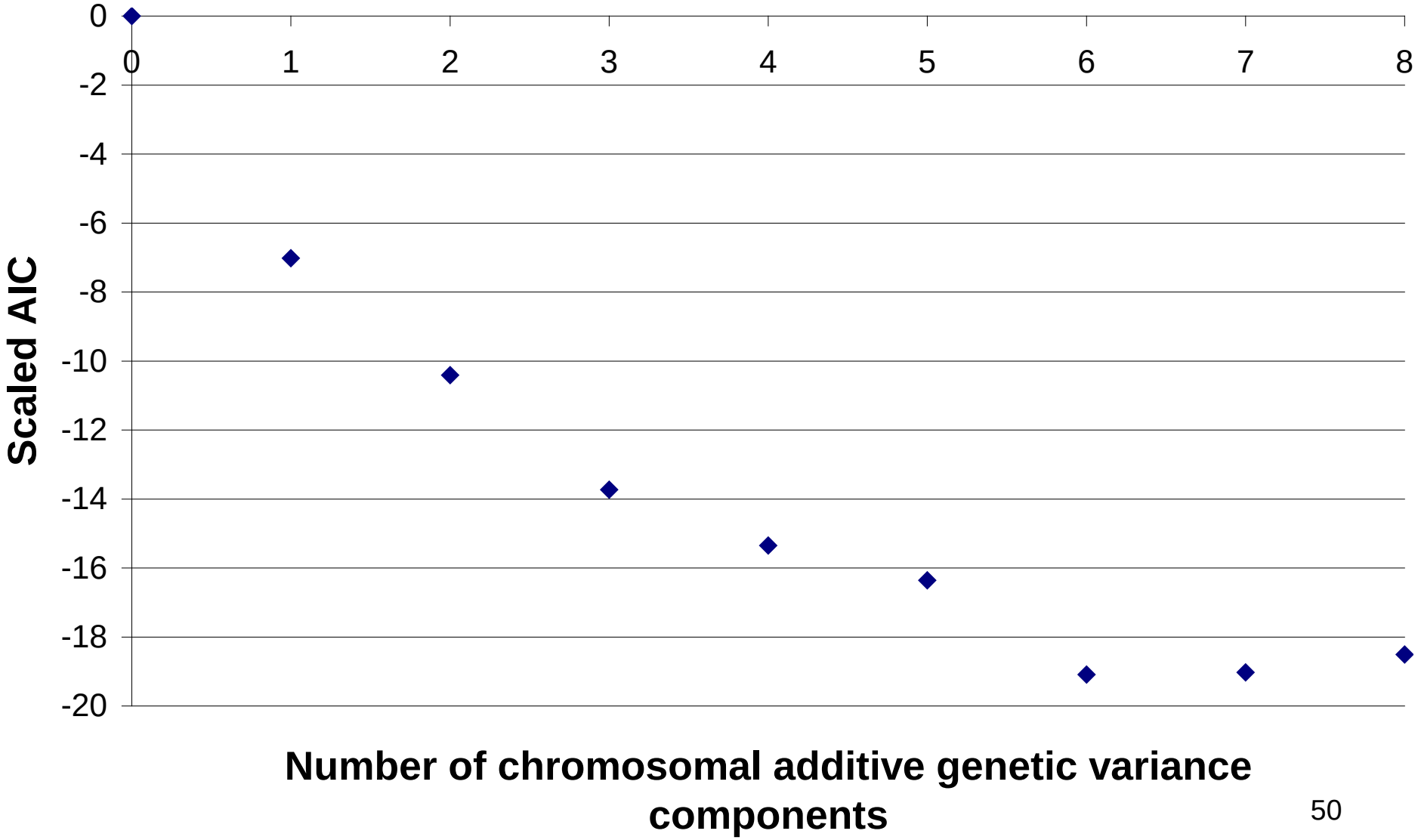
Longer chromosomes explain more additive genetic variance: ~0.03 per 100 cM



Estimates are consistent across countries



Stepwise analyses: at least 6 chromosomes are needed to explain the additive genetic variance

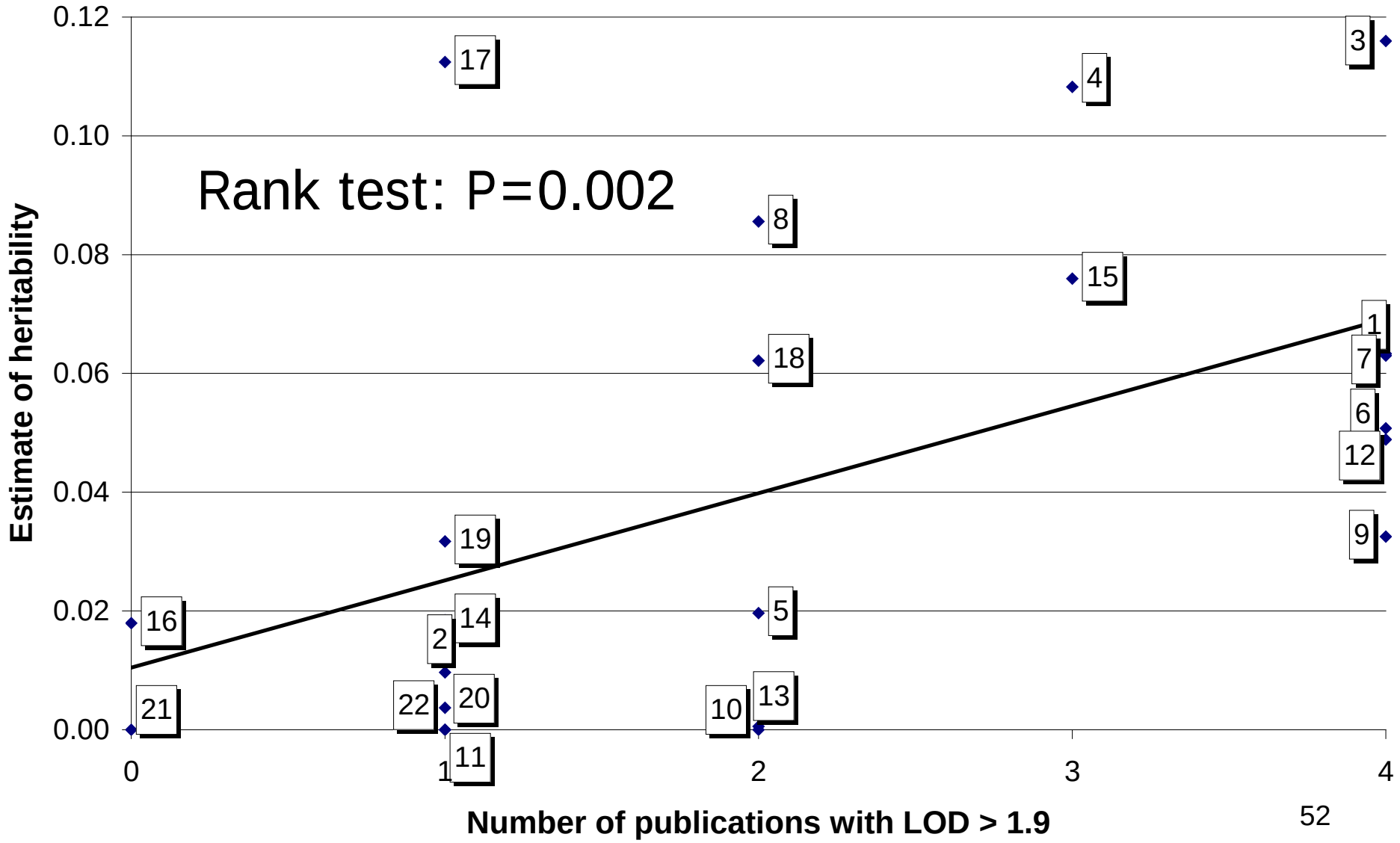


Hypothesis test

Model	h^2	c^2	df	LRT
Full (22 chrom.)	0.92	0.00	22	
Genome-wide	0.86	0.03	1	19.2

Additive genetic variance in proportion to lenght not rejected

Data consistent with published QTL results



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

Conclusions

- With large sample sizes it will become possible to estimate
  - dominance variance
  - epistatic variance
  - genome-wide parent-of-origin variance

  - genetic relative risk to disease

Genetic architecture for height

- Additive genetic variance
- No QTL of large effects
- Chromosomes explain ~10% of genetic variance
- Consequences for genome-wide association

Other applications: breeding programmes

- Exploit variance in genome-wide IBD by using the realised A-matrix
  - large increase in accuracy of selection if
    - variance in identity is large
    - family size is large
- “Genomic Selection”

Using the realised A-matrix: Reliability of EBV for an unphenotyped individual from *n*-1 phenotyped relatives (a simulation study)

Table 1. *Comparison of analytical (in bold) and simulated results for reliability of GEBVs calculated using either predicted or realized relationship matrices*

Design	h^2	N progeny	Reliability A-matrix	Reliability G-matrix	Ratio (SE)
Full sibs	1·0	10	0·450	0·480	1·067 (0·006)
			0·454	**0·466**	
		100	0·495	0·741	1·496 (0·013)
			0·495	**0·625**	
		1000	0·499	0·968	1·939 (0·002)
			0·499	**0·890**	
	0·50	10	0·374	0·384	1·025 (0·005)
			0·385	**0·383**	
		100	0·486	0·582	1·201 (0·009)
			0·485	**0·560**	
		1000	0·498	0·816	1·535 (0·009)
			0·499	**0·819**	
	0·10	10	0·161	0·163	1·011 (0·006)
			0·161	**0·162**	
		100	0·419	0·438	1·042 (0·002)
			0·419	**0·437**	
		1000	0·491	0·619	1·261 (0·008)
			0·491	**0·622**	
Half sibs	1·0	10	0·188	0·194	1·036 (0·008)
			0·189	**0·195**	
		100	0·243	0·320	1·320 (0·004)
			0·243	**0·308**	
		1000	0·249	0·433	1·736 (0·014)
			0·249	**0·444**	
	0·50	10	0·140	0·146	1·039 (0·008)
			0·147	**0·145**	
		100	0·234	0·272	1·168 (0·003)
			0·234	**0·271**	
		1000	0·248	0·392	1·577 (0·010)
			0·248	**0·408**	
	0·10	10	0·047	0·048	1·021 (0·008)
			0·047	**0·048**	
		100	0·179	0·188	1·046 (0·005)
			0·179	**0·188**	
		1000	0·241	0·300	1·249 (0·007)
			0·241	**0·306**	

Genet. Res., Camb. (2009), **91**, *pp.* 47–60. © 2009 Cambridge University Press
doi:10.1017/S0016672308009981 Printed in the United Kingdom

Increased accuracy of artificial selection by using the realized relationship matrix

B. J. HAYES[1]*, P. M. VISSCHER[2] AND M. E. GODDARD[1,3]