Whole genome approaches to quantitative genetics

Leuven 2008

Overview

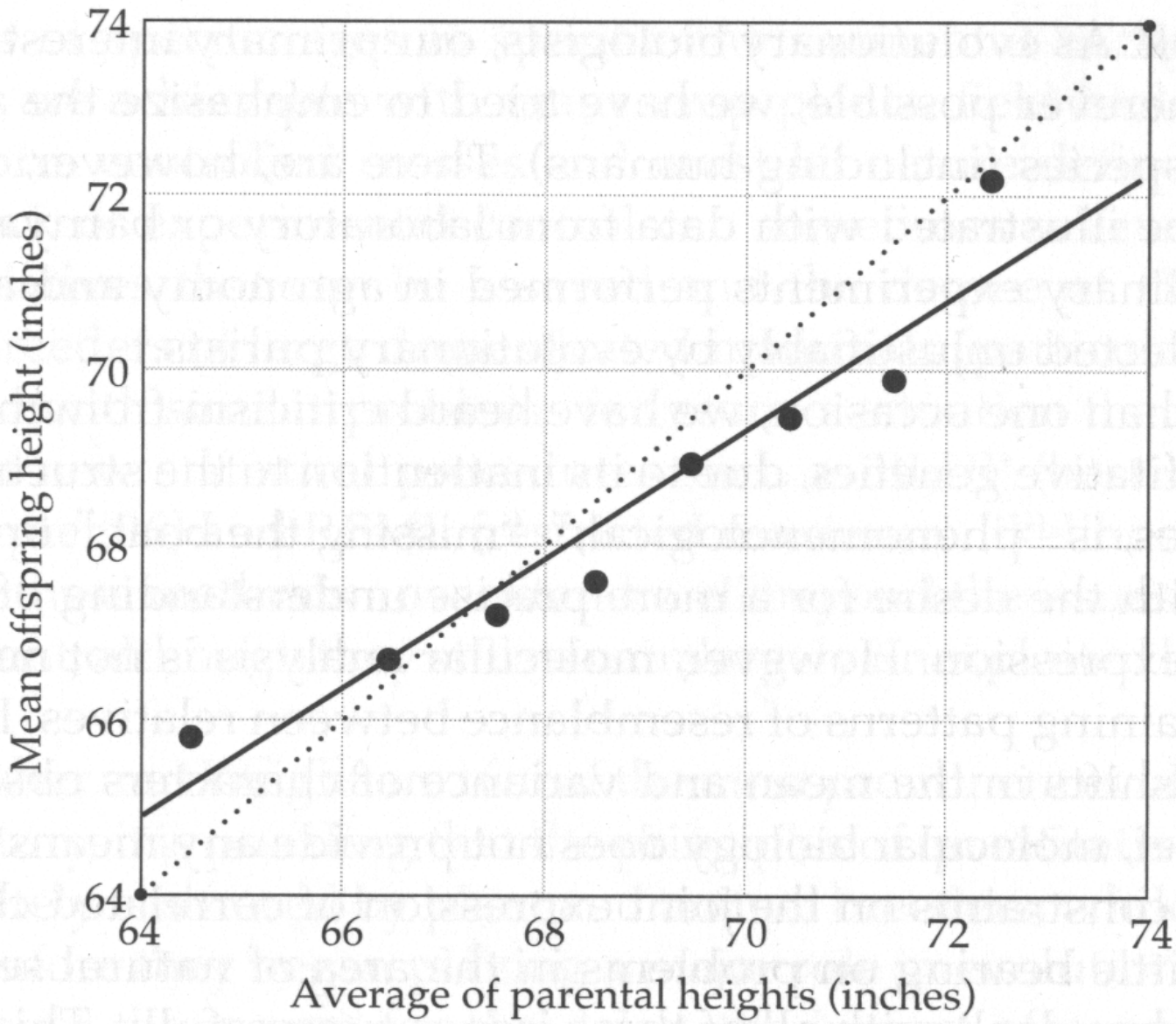
- Rationale/objective of session
- Estimation of genetic parameters
- Variation in identity
- Application/Practical
  - mean and variance of genome-wide IBD sharing for sibpairs
  - estimation of heritability of height
  - genome partitioning of genetic variation

Objectives

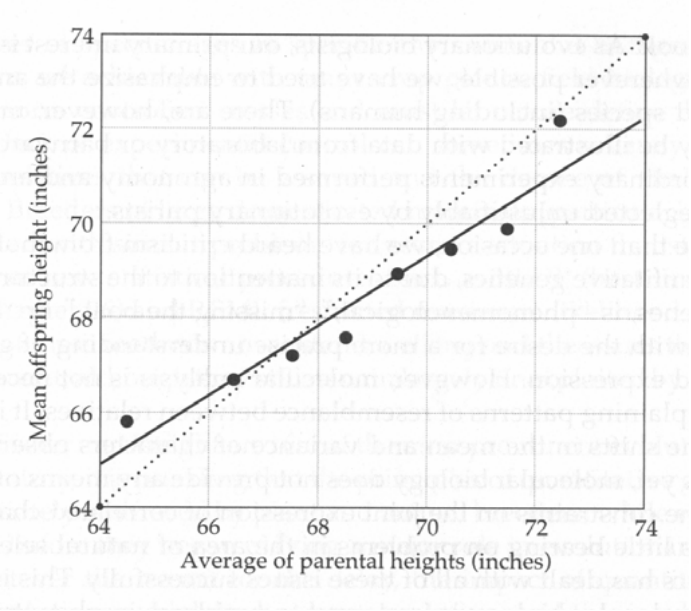
- Understand that there is variation in identity (per locus, chromosome and genome-wide)
- How this can be estimated with genetic markers
- How and why variation in identity changes with the length of the chromosome
- How this can be exploited to estimate genetic variance
- How this relates to linkage analysis

Estimation of genetic parameters

- Model
  - expected covariance between relatives
    - Genetics
    - Environment
- Data
  - correlation/regression of observations between relatives
- Statistical method
  - Least squares (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

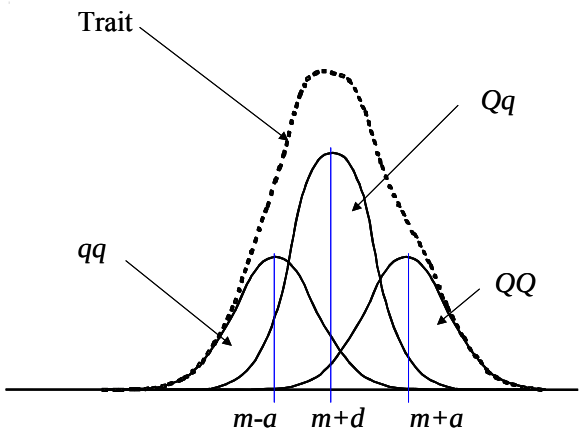
Do quantitative traits have the same hereditary and evolutionary properties as discrete characters?

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

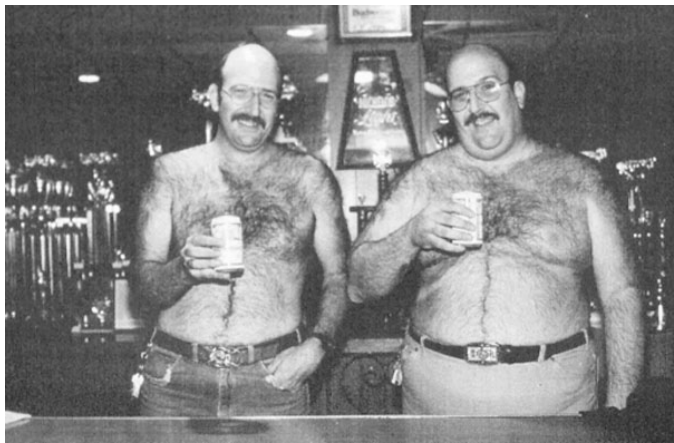
= Prob(2 alleles are IBD)

Examples (no inbreeding)

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

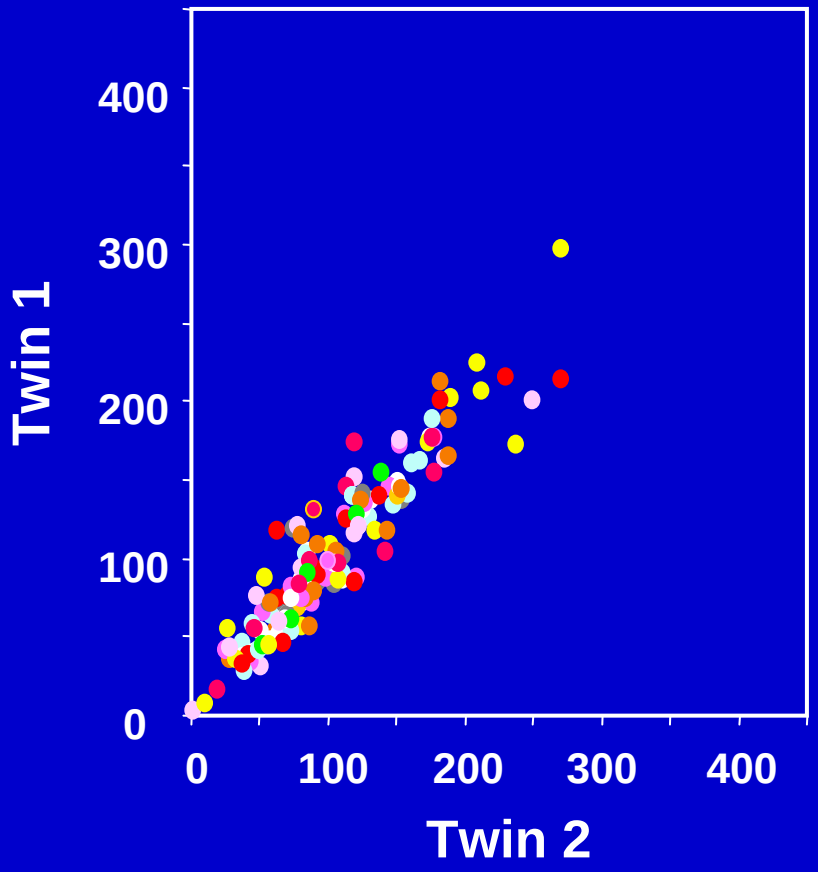
Controversy/confounding: nature vs nurture

- 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

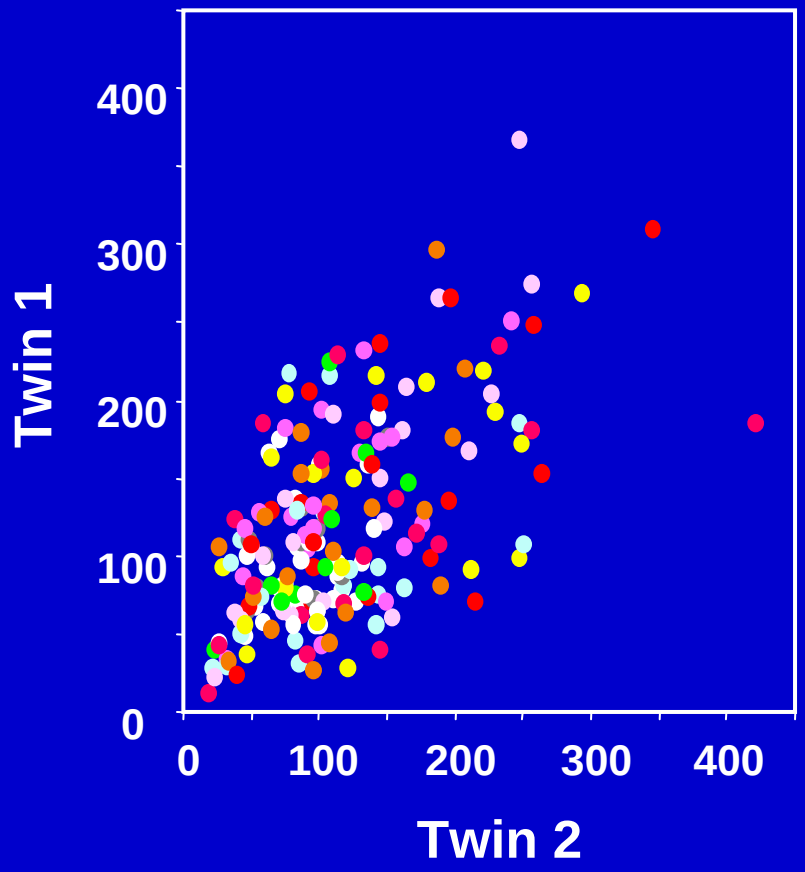


Total mole count for MZ and DZ twins

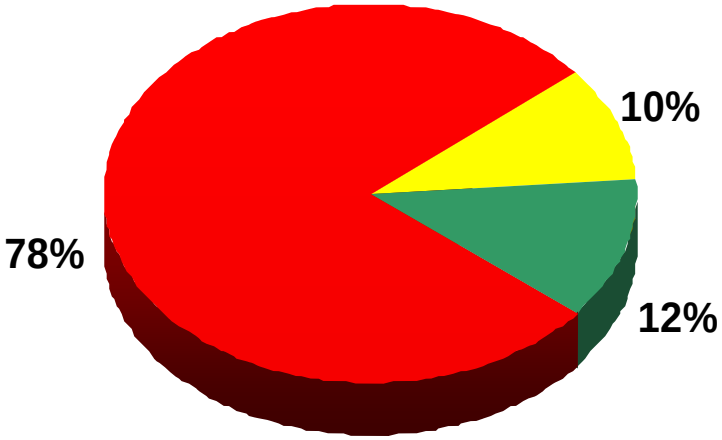
MZ twins - 153 pairs, r = 0.94



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

A different approach

Estimate genetic variance within families

Actual 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

Notation / concept

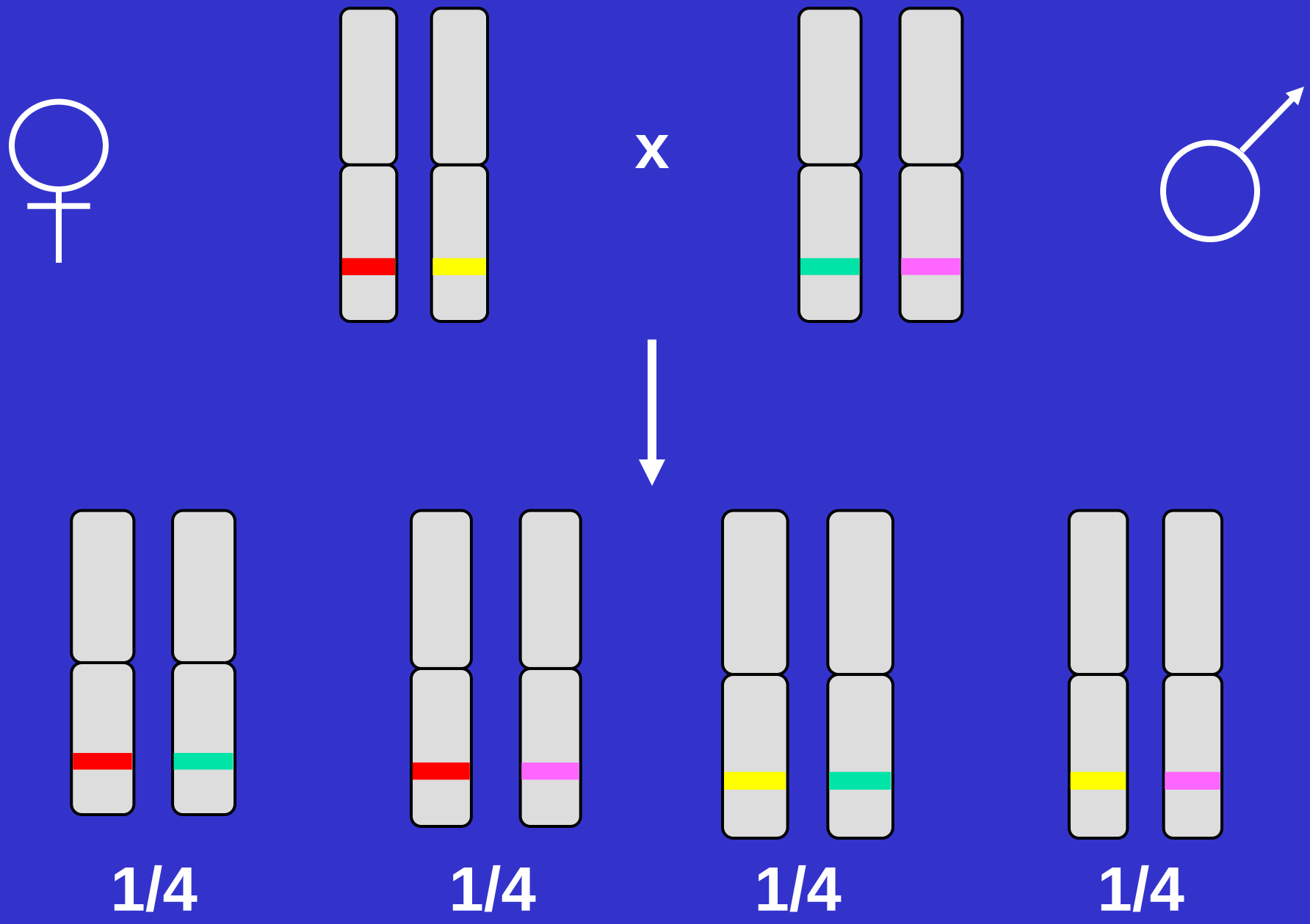
π is a random variable!

$\hat{\pi}$ (pihat) is an estimate of π

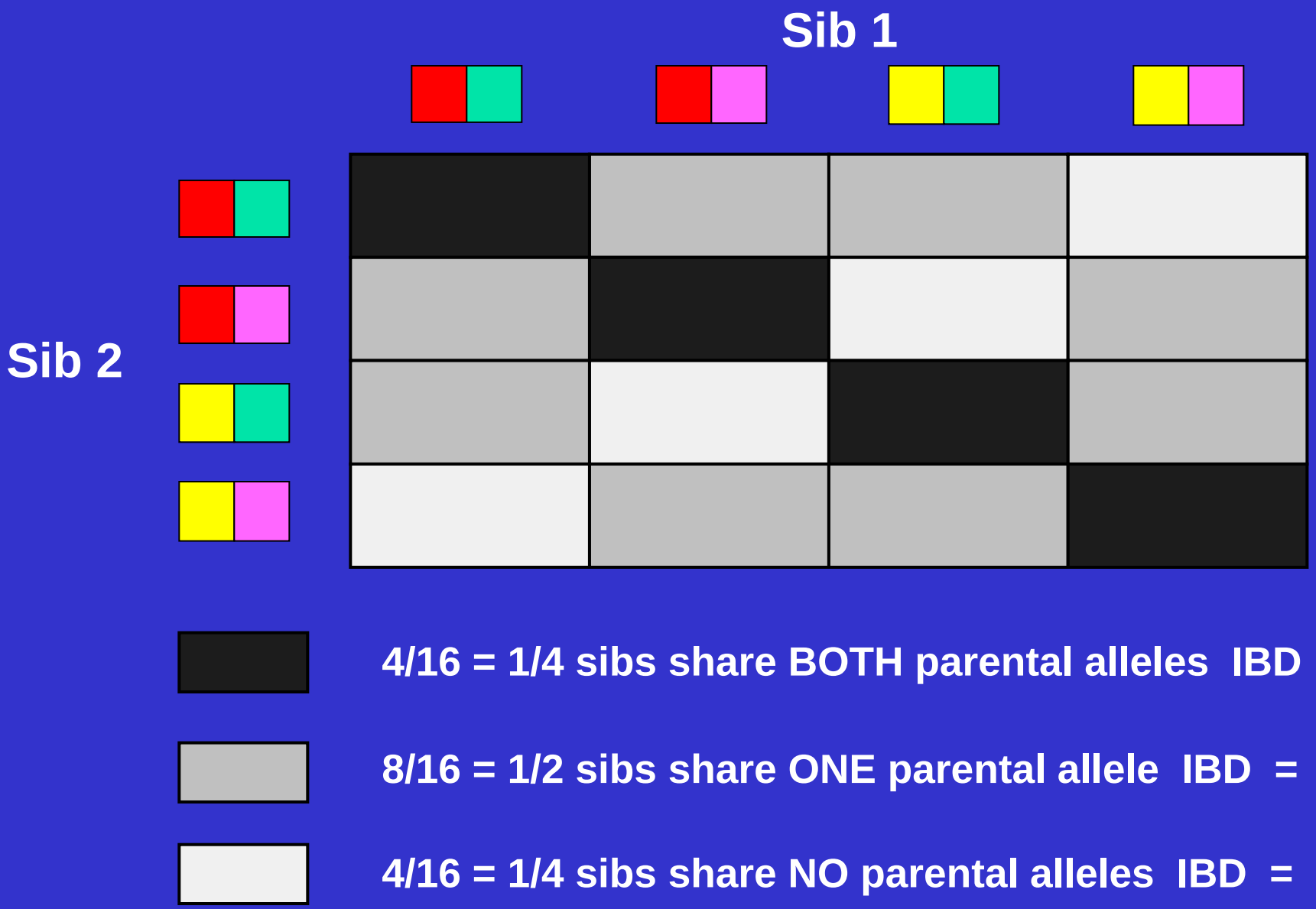
If the estimate is unbiased then

E(π|pihat) = pihat: the regression of true on estimated values is 1.0

E(pihat) ≠ π


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	$1/_{16}$

n unlinked loci

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

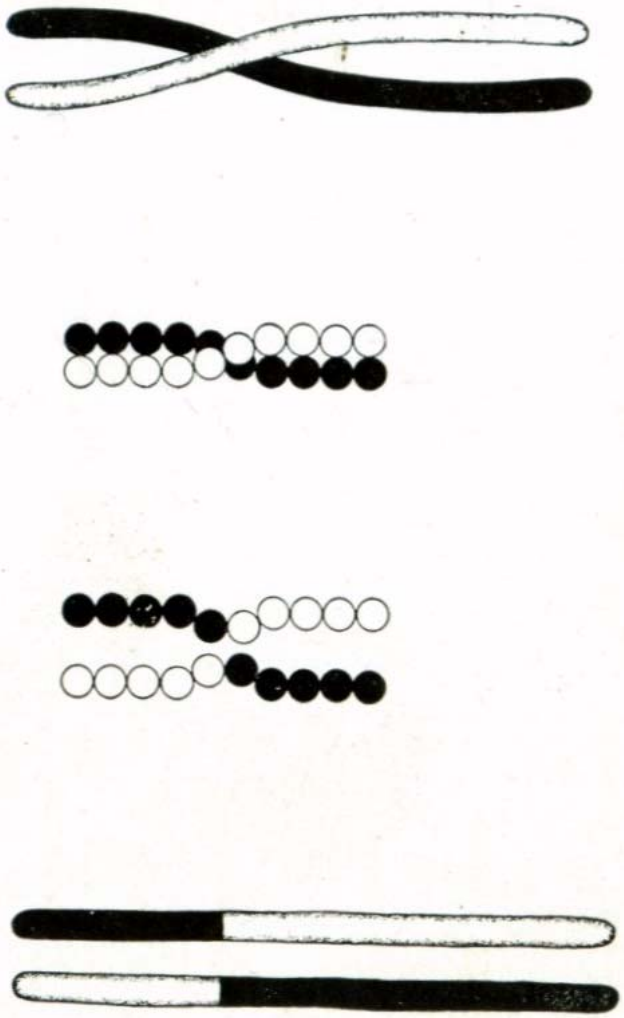


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

[Thomas Hunt Morgan, 1916]

Loci are on chromosomes

- The cross-over rate per meiosis is ~low: segregation of large chromosome segments within families
  - increases variance of IBD sharing
- Independent segregation of chromosomes
  - decreases variance of IBD sharing

Chromosome length

- Longer chromosomes have more recombination
  - more ‘independent’ segments
  - smaller variance in mean IBD sharing
- Smaller chromosomes have less recombination
  - more like single loci
  - larger variance in mean IBD sharing

Practical: test empirically

Dominance (fullsibs): π_d

Prob(2 alleles IBD) = ¼

Prob(2 alleles non-IBD)= ¾

Mean(IBD2)= ¼

Variance(IBD2) = ¼ - ¼^2 = $^3/_{16}$

→ Variation in (mean) π_d is larger than variation in (mean) π_a

Practical: test empirically

Theoretical SD of π_a

Relatives	1 chrom (1 M)	genome (35 M)
Fullsibs	0.217	0.038
Halfsibs	0.154	0.027

[Stam 1980; Hill 1993; Guo 1996]

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

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

Difficult but not impossible to disentangle additive and dominance variance

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 (genome-wide π_a and π_d) 0.89

Practical: test empirically

Analysis (fullsibs)

$Y = \mu + A + C + E$

$var(Y) = \sigma^2(A) + \sigma^2(C) + \sigma^2(E)$

$cov(Y_1,Y_2) = \pi_a\sigma^2(A) + \sigma^2(C)$

Full model: ACE

Reduced model: CE

- Need software that can handle VC and 'user-defined' covariance structure
  - e.g. Mx, QTDT, ASREML

Idea not new

Ritland, K (1996). A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* 50: 1062-1073.

Thomas SC, Pemberton JM, Hill WG (2000). Estimating variance components in natural populations using inferred relationships. *Heredity* 84:427-36.

Practical

Data from:

REPORT

Genome Partitioning of Genetic Variation for Height from 11,214 Sibling Pairs

Peter M. Visscher, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G. Hill, Jouke-Jan Hottenga, Gonneke Willemsen, Dorret I. Boomsma, Yao-Zhong Liu, Hong-Wen Deng, Grant W. Montgomery, and Nicholas G. Martin

Marker data summarised into average 'pihats' and IBD2 coefficients per chromosome and genome wide, per sibling pair

Files

```
data.txt
data.xls
a_genome.mx

qtdt.ped
qtdt.dat
qtdt.ibd
```

Data set (`data.txt`, `data.xls`)

Column	What
1	Pair ID
2-24	Chromosomal mean pihats
25-47	Chromosomal mean IBD2
48	Genome-wide mean pihat
49	Genome-wide mean IBD2

Data set

Column	What
50	sex sib1 (1=male)
51	age sib1
52	raw height sib1
53	Z-score sib1
54-57	and for sib2
58	code for sex of sibling pair
59	country code (1+2=OZ, 3=US, 4=NL)

Part of a_genomewide.mx

```
Rectangular File=data.txt
Labels
 famid
 a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17 a18 a19
 a20 a21 a22 a23
 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17 d18 d19
 d20 d21 d22 d23
 meana meand
 sex1 age1 ht1 y1 sex2 age2 ht2 y2 sexboth code

SElect y1 y2 meana age1 sex1 age2 sex2;
Definition_variables  meana age1 sex1 age2 sex2;
```

Output Mx

```
MATRIX I
 This is a computed FULL matrix of order
  1 by    4
  [=F%T|K%T|(F+K)%T|E%T]
              1           2           3           4
 1     0.0292    0.8606    0.8897    0.1103
```

C A C+A E

With C+A+E = 1

qtdt.ped

- Pedigree + phenotypes + covariates + markers
- Dummy markers used: ignore!

```
1 3 4 C1  0.4096 0.5758 0.0146
1 3 4 C2  0.2750 0.5674 0.1576
1 3 4 C3  0.2222 0.6350 0.1428
1 3 4 C4  0.3557 0.6242 0.0201
1 3 4 C5  0.2445 0.5508 0.2047
1 3 4 C6  0.6012 0.3886 0.0102
1 3 4 C7  0.1051 0.4940 0.4009
1 3 4 C8  0.6970 0.2712 0.0318
1 3 4 C9  0.3490 0.6052 0.0458
1 3 4 C10 0.3468 0.4616 0.1916
1 3 4 C11 0.2224 0.6452 0.1324
1 3 4 C12 0.5152 0.3758 0.1090
1 3 4 C13 0.3540 0.1952 0.4508
1 3 4 C14 0.4815 0.5078 0.0107
1 3 4 C15 0.0786 0.5460 0.3754
1 3 4 C16 0.0097 0.4656 0.5247
1 3 4 C17 0.1878 0.5656 0.2466
1 3 4 C18 0.0070 0.4370 0.5560
1 3 4 C19 0.0168 0.6804 0.3028
1 3 4 C20 0.0099 0.6166 0.3735
1 3 4 C21 0.0069 0.2242 0.7689
1 3 4 C22 0.0486 0.8276 0.1238
1 3 4 G   0.2520 0.5119 0.2361
```

Top of `qtdt.ped`

$P_0 + P_1 + P_2 = 1$

pihat = $\frac{1}{2}P_1 + P_2$

IBD2 = P_2

$\rightarrow P_1 = 2(\text{pihat} - \text{IBD2})$

$\rightarrow P_2 = \text{IBD2}$

$\rightarrow P_0 = 1 - P_1 - P_2$

qtdt.dat

```
T Y
C SEX
C AGE
S2 C1
S2 C2
S2 C3
S2 C4
S2 C5
S2 C6
S2 C7
S2 C8
S2 C9
S2 C10
S2 C11
S2 C12
S2 C13
S2 C14
S2 C15
S2 C16
S2 C17
S2 C18
S2 C19
S2 C20
S2 C21
S2 C22
M  G
```

T = Trait

C = covariate

S2 = skip ‘marker’

M = marker

```
NULL HYPOTHESIS
---------------

Family #1 var-covar matrix terms [2]...[[Ve]][[Vg]]
Family #1 regression matrix...
       [linear] =
        [2 x 3]      Mu     SEX     AGE
            1.3   1.000   1.000  16.000
            1.4   1.000   1.000  16.000

Some useful information...
               df : 22423
  log(likelihood) : 30196.57
        variances :   0.080   0.894
            means :   0.079   0.019  -0.002

FULL HYPOTHESIS
---------------

Family #1 var-covar matrix terms
[3]...[[Ve]][[Vg]][[Va]]
Family #1 regression matrix...
       [linear] =
        [2 x 3]      Mu     SEX     AGE
            1.3   1.000   1.000  16.000
            1.4   1.000   1.000  16.000

Some useful information...
               df : 22422
  log(likelihood) : 30186.27
        variances :   0.079   0.056   0.839
            means :   0.079   0.019  -0.002
```

Output QTDT in regress.tbl

Test statistic
= 2(-30186.27- -30196.57)
= 20.6

E C A

What to do (1)

- Marker data only:
  - Calculate mean and SD of chromosomal pihats and IBD2 (use `Excel`, `R`, or whatever)
  - Calculate mean and SD of genome-wide pihat and IBD2
  - Plot mean genome-wide pihat against mean genome-wide IBD2 for each sibling pair
  - Use autosomes only (1-22)

What to do (2)

- Phenotype data only:
  - What is the sib correlation for the standardised Z-scores?

Use Z-scores because the unit of measurement for Height varies between cohorts!

What to do (3)

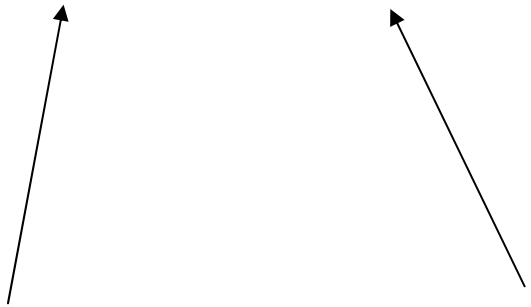
- Marker data plus phenotypes
  - Estimate additive variance from genome-wide pihat using `Mx` or `QTDT`
  - Estimate additive variances for each chromosome
  - Note the test statistic for A

- You need to edit `a_genome.mx` and `qtdt.dat` to run different chromosomes
  - Use e.g. `Notepad`, `Wordpad`, `Word`, `vi`, `emacs`, ...

Analysis examples

Run `a_genome.mx` using `Mx`

```
QTDT –weg –vega –a-
```



Reduced model
e = error
g = polygenic

Full model
e = error
g = ‘polygenic’ (here C!)
a = ‘marker’