Biometrical genetics

Pak C Sham

The University of Hong Kong

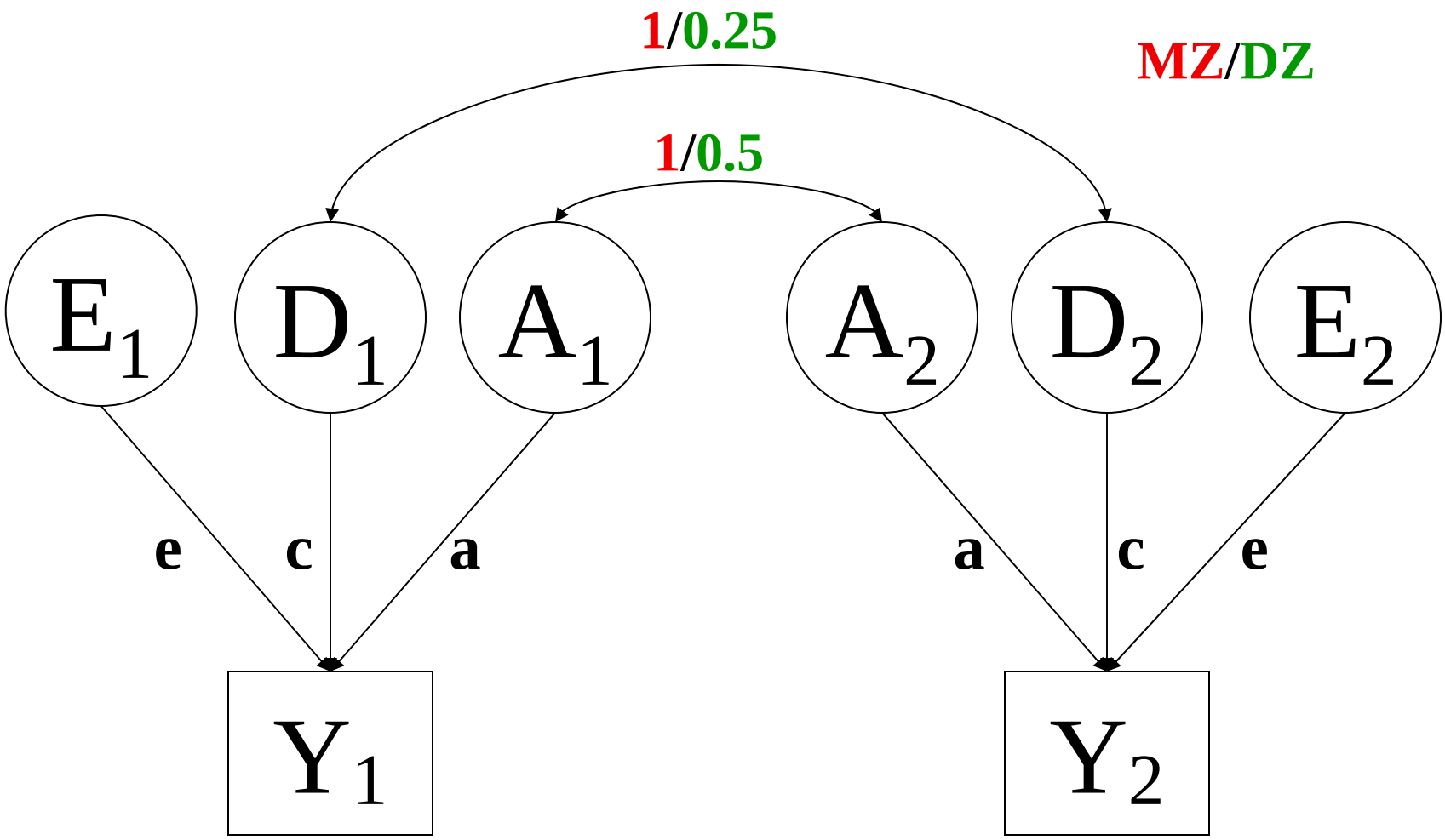
Manuel AR Ferreira

Queensland Institute for Medical Research

23rd International Workshop on Methodology of Twin and Family Studies

2nd March, 2010

ADE Model for twin data



Biometrical Genetics

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.

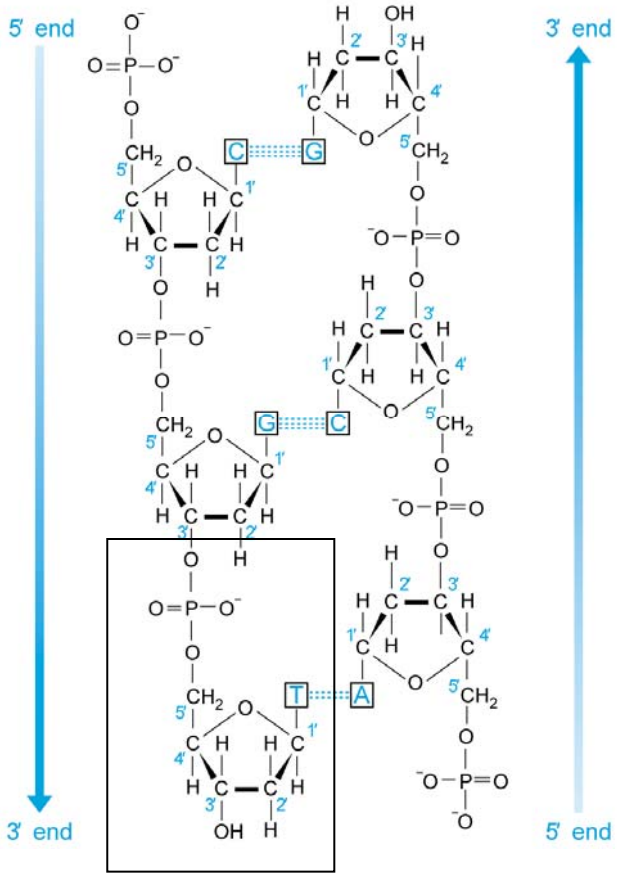
Outline

1. Basic molecular genetics
2. Components of genetic model
3. Biometrical properties for single locus
4. Introduction to linkage analysis

1. Basic molecular genetics

DNA

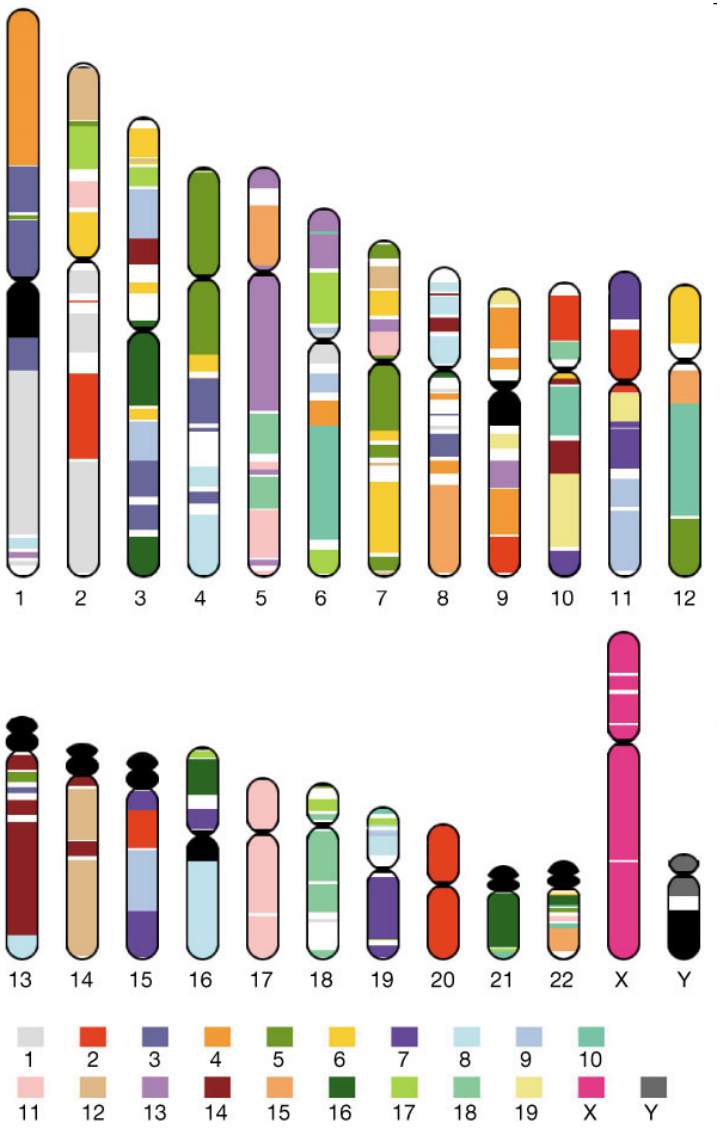
- A DNA molecule is a linear backbone of alternating sugar residues and phosphate groups
- Attached to carbon atom 1’ of each sugar is a nitrogenous base: A, C, G or T
- Complementarity: A always pairs with T, likewise C with G
- A gene is a segment of DNA which is translated to a peptide chain



nucleotide

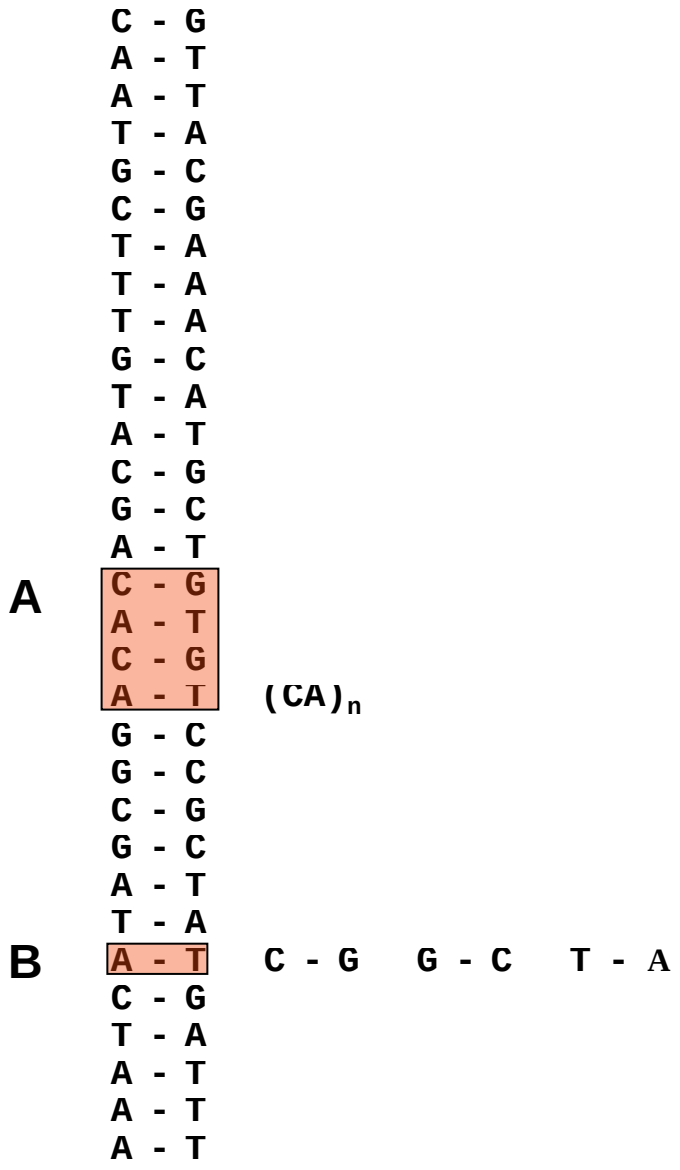
Human genome

- 23 chromosome pairs
  - 22 autosomes, X,Y
- ~ 33,000,000,000 base pairs
- ~ 25,000 translated “genes”
- Other functional sequences
  - non-translated RNA
  - binding sites for regulatory molecules

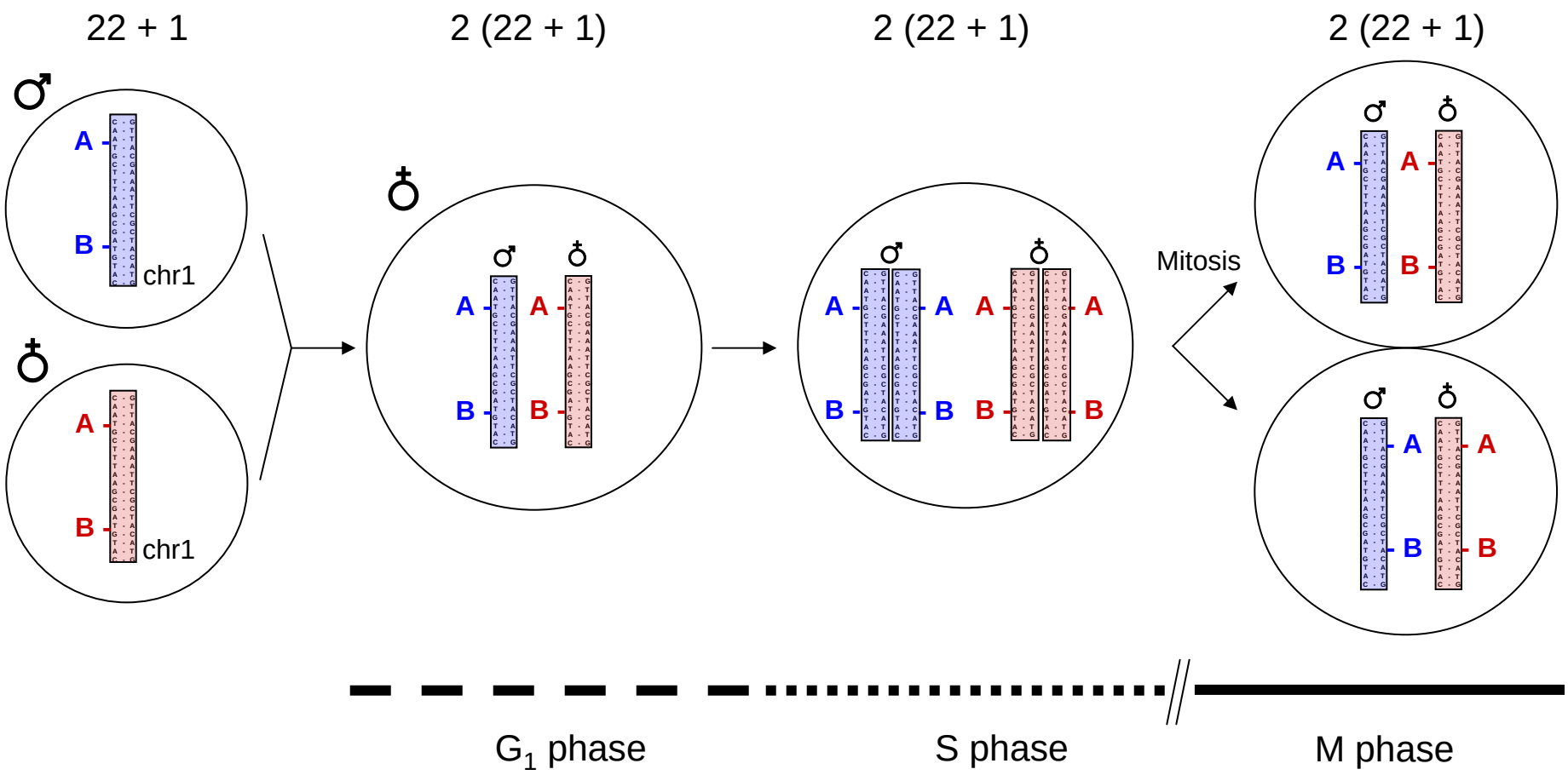


DNA sequence variation (polymorphisms)

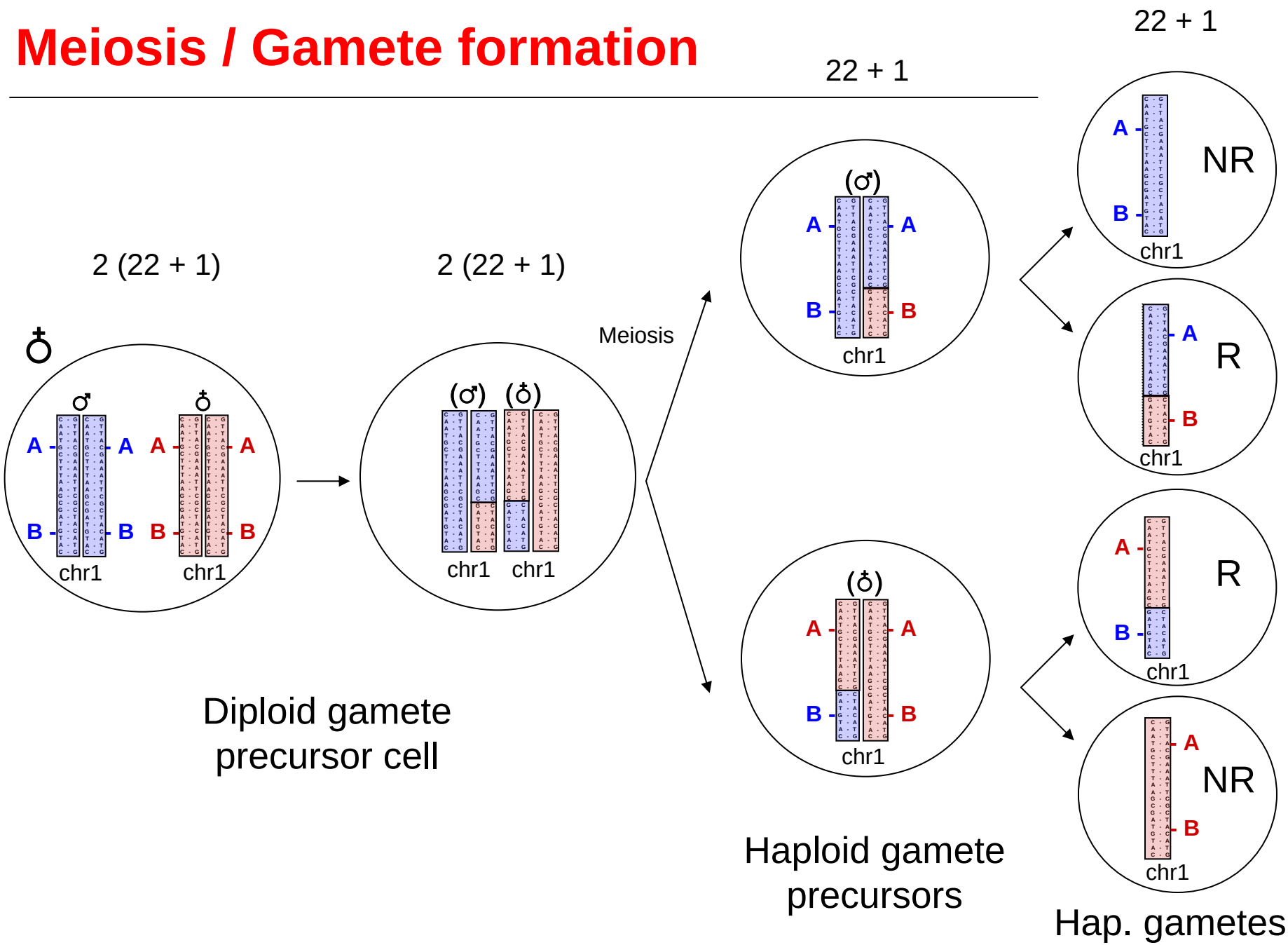
- Microsatellites
  >100,000
  Many alleles, eg. $(CA)_n$ repeats, very informative, easily automated
- Single nucleotide polymorphims (SNPs)
  11,883,685 (build 128, 03 Mar ‘08)
  Most with 2 alleles (up to 4), not very informative, easily automated
- Copy Number polymorphisms
  Large-scale insertions / deletions



Fertilization and mitotic cell division



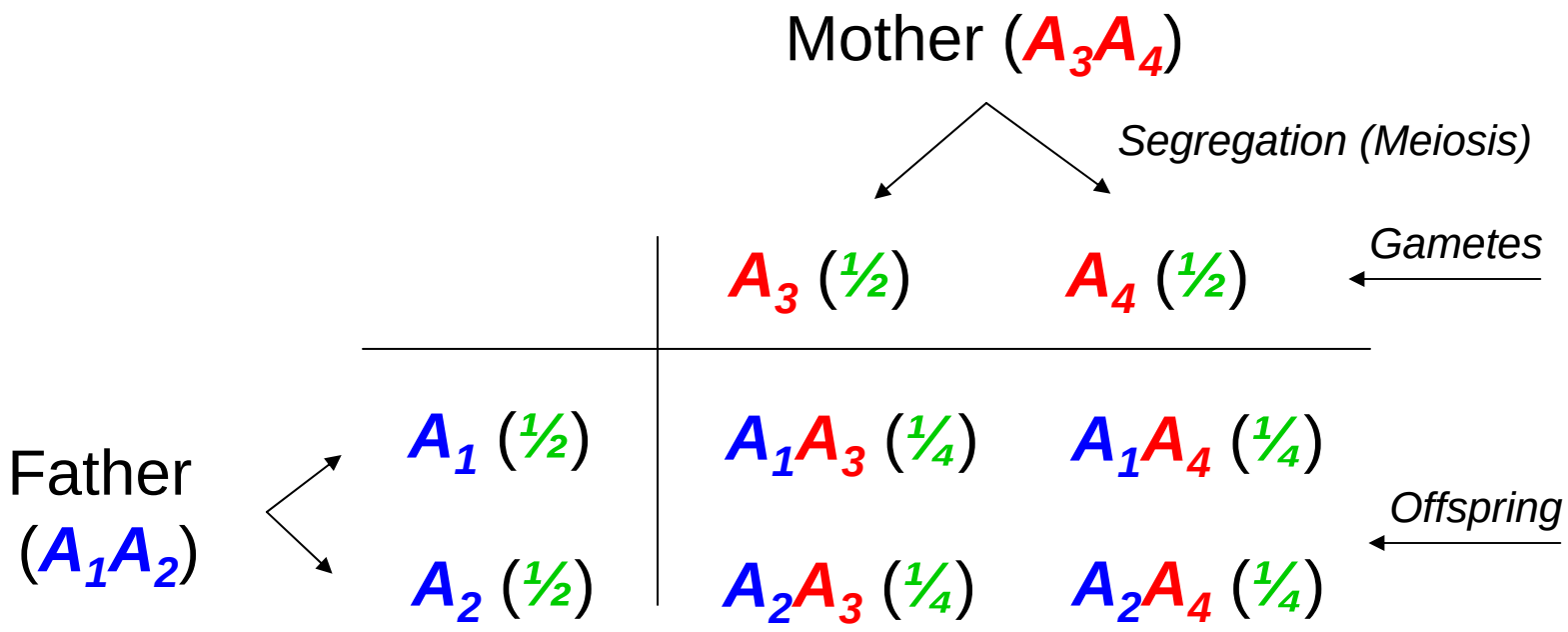
Meiosis / Gamete formation



2. Components of genetic model

A. Transmission model

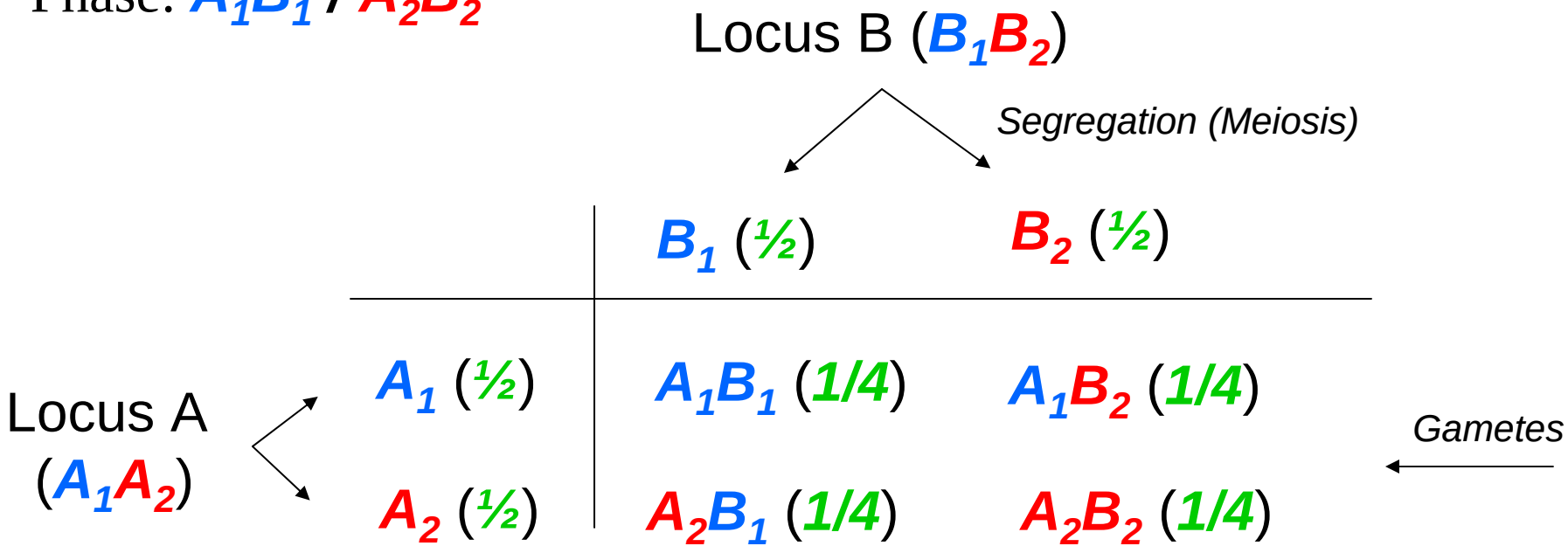
Mendel's law of segregation



Note: 50:50 segregation can be distorted ("meiotic drive")

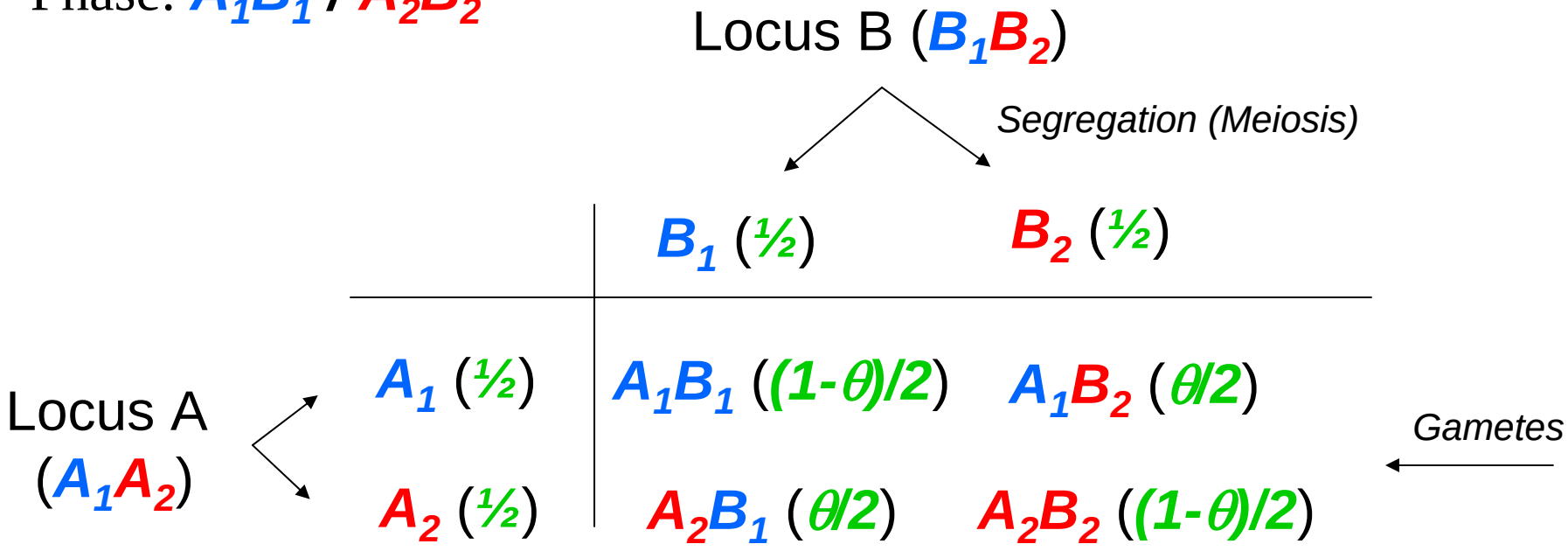
A. Transmission model: two unlinked loci

Phase: A_1B_1 / A_2B_2



A. Transmission model: two linked loci

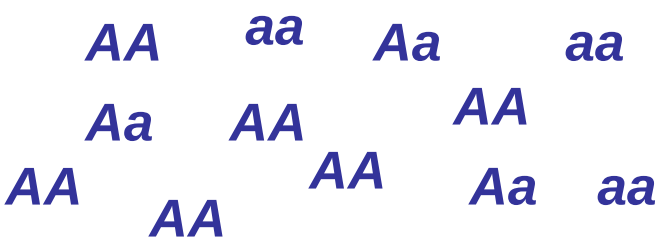
Phase: A_1B_1 / A_2B_2



θ : Recombination fraction, between 0 (complete linkage) and 1/2 (free recombination)

B: Population model

Frequencies



Genotype frequencies:

AA: P

Aa: Q

aa: R

Allele frequencies:

A: P+Q/2

a: R+Q/2

B: Population model

Hardy-Weinberg Equilibrium (Hardy GH, 1908; Weinberg W, 1908)

	AA	**Aa**	**aa**
AA	P^2	PQ	PR
Aa	PQ	Q^2	QR
aa	PR	QR	R^2

Random mating

	AA	**Aa**	**aa**
AA	AA	AA:Aa 0.5:0.5	Aa
Aa	AA:Aa 0.5:0.5	AA:Aa:aa 0.25:0.5:0.25	Aa:aa 0.5:0.5
aa	Aa	Aa:aa 0.5:0.5	aa

Offspring genotypic distribution

B: Population model

Hardy-Weinberg Equilibrium (Hardy GH, 1908; Weinberg W, 1908)

Offspring genotype frequencies

Genotype	Frequency
AA	$P^2+PQ+Q^2/4 = (P+Q/2)^2$
Aa	$2PR+PQ+QR+Q^2/2 = 2(P+Q/2)(R+Q/2)$
aa	$R2+QR+Q^2/4 = (R+Q/2)^2$

Offspring allele frequencies

Allele	Frequency
A	$(P+Q/2)^2 + (P+Q/2)(R+Q/2) = P+Q/2$
a	$(R+Q/2)^2 + (P+Q/2)(R+Q/2) = R+Q/2$

B. Population model

Panmixia (Random union of gametes)

Paternal allele \ Maternal allele	***A*** (***p***)	***a*** (***q***)
A (***p***)	***AA*** (p^2)	***Aa*** (***pq***)
a (***q***)	***aA*** (***qp***)	***aa*** (q^2)

P (***AA***) = p^2

P (***Aa***) = $2pq$

P (***aa***) = q^2

Deviations from HWE

Assortative mating

Imbreeding

Population stratification

Selection

C. Phenotype model

Classical Mendelian (Single-gene) traits

- Dominant trait
  - ***AA***, ***Aa*** **1**
  - ***aa*** **0**

Huntington's disease
(CAG)n repeat, huntingtin gene

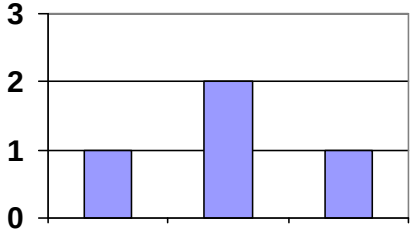
- Recessive trait
  - ***AA*** **1**
  - ***aa, Aa*** **0**

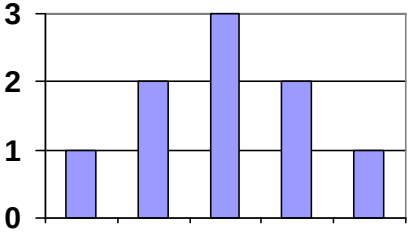
Cystic fibrosis
3 bp deletion exon 10 CFTR gene

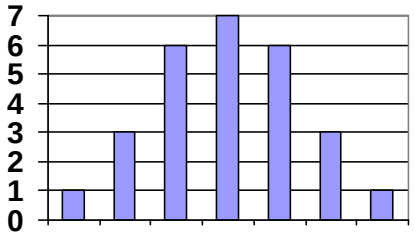
C. Phenotype model

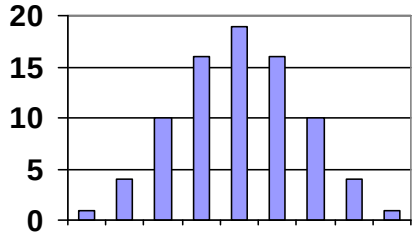
Polygenic model

1 Gene	2 Genes	3 Genes	4 Genes
→ 3 Genotypes	→ 9 Genotypes	→ 27 Genotypes	→ 81 Genotypes
→ 3 Phenotypes	→ 5 Phenotypes	→ 7 Phenotypes	→ 9 Phenotypes





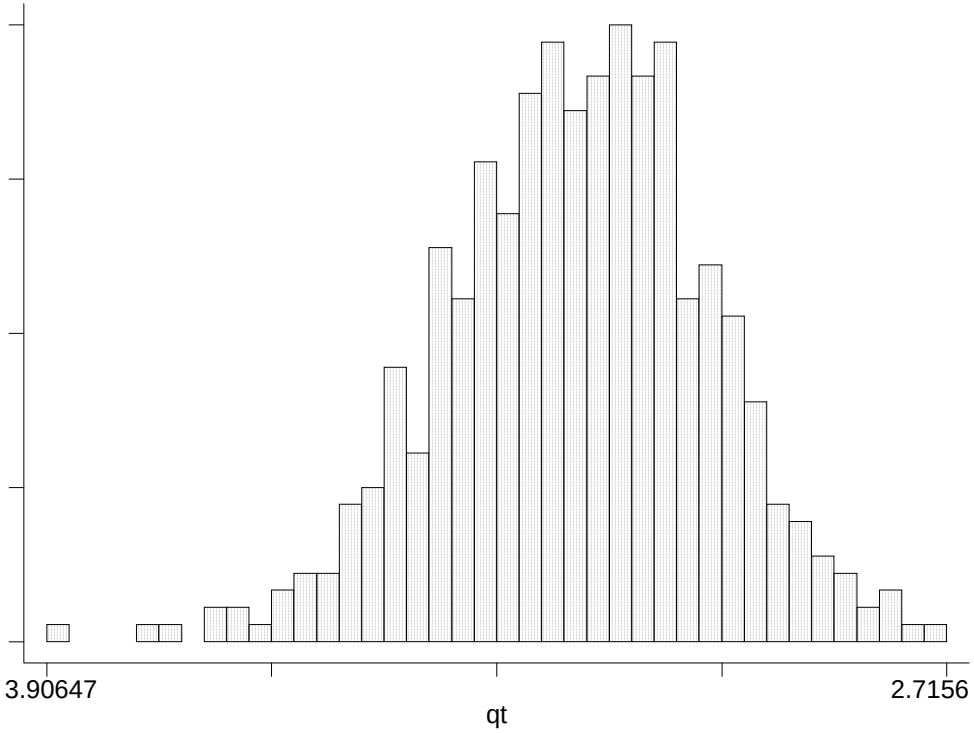


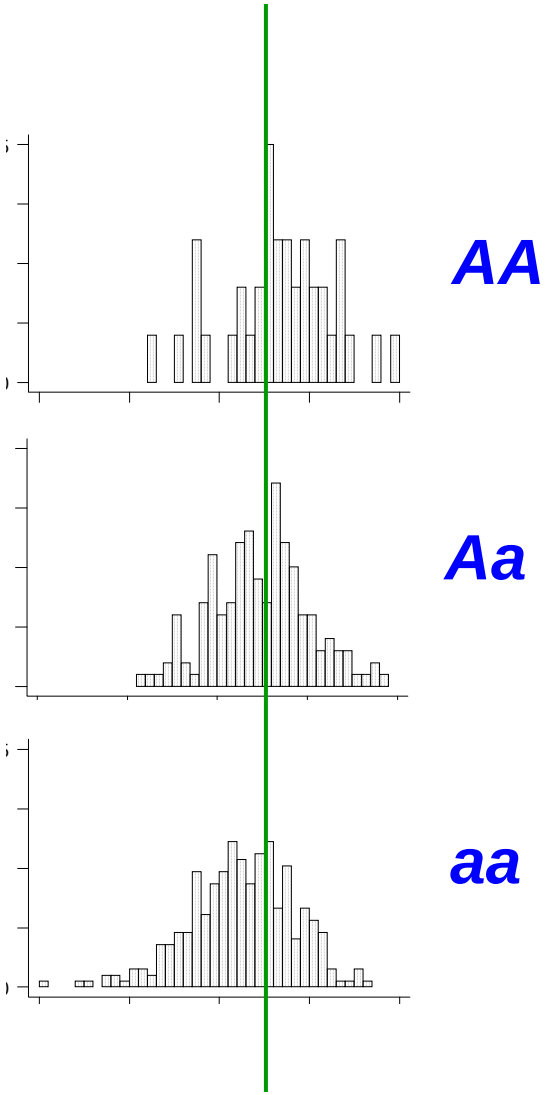


Central Limit Theorem → Normal Distribution

C. Phenotype model

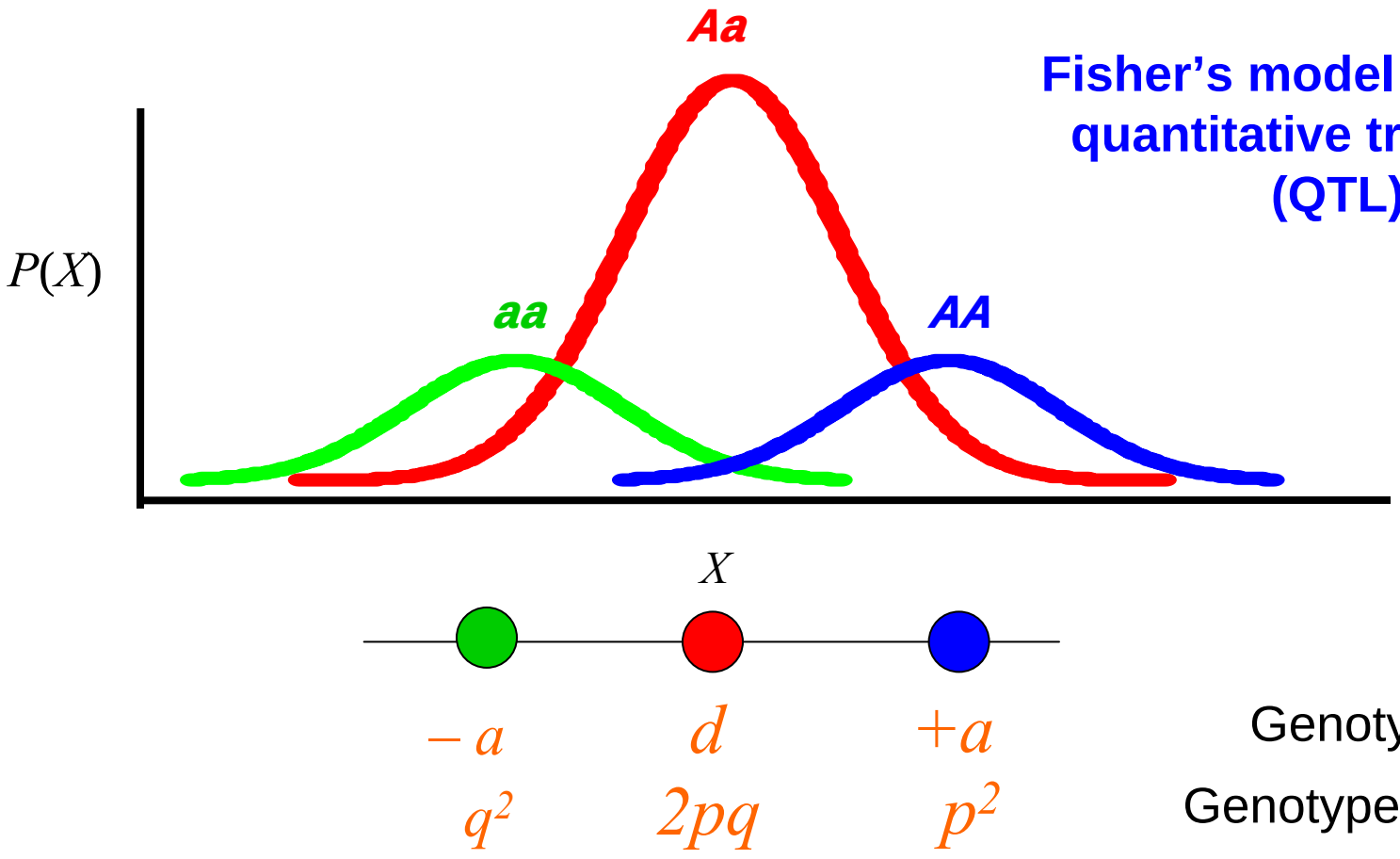
Quantitative traits





e.g. cholesterol levels

D. Phenotype model



Fisher's model for single quantitative trait locus (QTL)

Genotypic effects

Genotype frequencies

Assumption: Effect of allele independent of parental origin (Aa = aA)
Violated in genomic imprinting

3. Biometrical properties of single locus

Biometrical model for single biallelic QTL

1. Contribution of the QTL to the Mean (*X*)

$$\mu = \sum_i x_i f(x_i)$$

Genotypes	***AA***	***Aa***	***aa***
Effect, *x*	***a***	***d***	***-a***
Frequencies, *f*(*x*)	p^2	$2pq$	q^2

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

Note:

If everyone in population has genotype **aa**

then population mean = ***-a***

∴ change in mean due to **A** = $((p-q)a + 2pqd) - (-a) = 2p(a+qd)$

Biometrical model for single biallelic QTL

2. Contribution of the QTL to the Variance (*X*)

$$Var = \sum_i (x_i - \mu)^2 f(x_i)$$

Genotypes	***AA***	***Aa***	***aa***
Effect, *x*	***a***	***d***	***-a***
Frequencies, *f*(*x*)	p^2	$2pq$	q^2

$$Var\,(X) = (a-m)^2p^2 + (d-m)^2 2pq + (-a-m)^2q^2$$

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

$$= V_{QTL}$$

Broad-sense heritability of *X* at this locus = $V_{QTL} / V_{\text{Total}}$

Biometrical model for single biallelic QTL

2. Partitioning of QTL variance: additive component

		Maternal allele: *A* (*p*)	Maternal allele: *a* (*q*)	Average
Paternal allele	*A* (*p*)	*a*	*d*	*pa+qd*
	a (*q*)	*d*	*-a*	*pd-qa*
	Average	*pa+qd*	*pd-qa*	$(p-q)a + 2pqd$

Variance due to a single allele $= p(q(d+a)-2pqd)^2+q(p(d-a)-2pqd)^2$

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

For both alleles, additive variance $= 2pq(a+(q-p)d)^2$

Biometrical model for single biallelic QTL

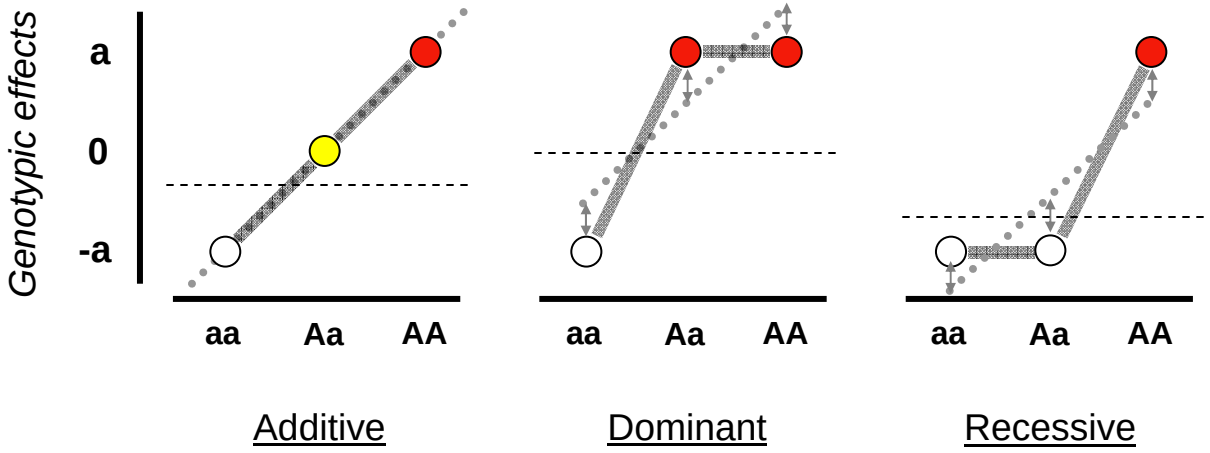
2. Partitioning of QTL variance: dominance component

Genotype	*Effect*	*Additive effect*
AA (p^2)	a	$2(pa+qd)$
Aa ($2pq$)	d	$(pa+qd)+(pd-qa)$
aa (q^2)	$-a$	$2(pd-qa)$

Dominance variance due to QTL = $p^2(a-2(pa+qd))^2$
$+2pq(d-(pa+qd+pd-qa))^2$
$+q^2(-a-2(pd-qa)$

$= (2pqd)^2$

Biometrical model for single biallelic QTL

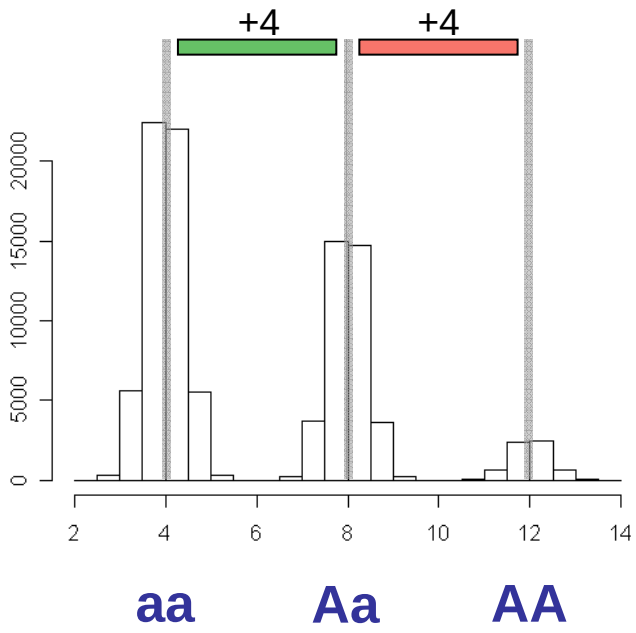


$$\text{Var}(X) = \text{Regression Variance} + \text{Residual Variance}$$

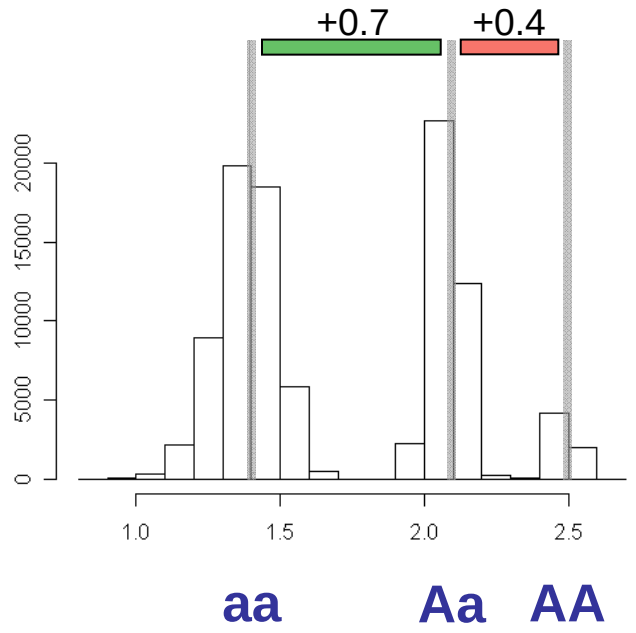
$$= \text{Additive Variance} + \text{Dominance Variance}$$

$$= V_{A_{QTL}} + V_{D_{QTL}}$$

Statistical definition of dominance is scale dependent



log (x)

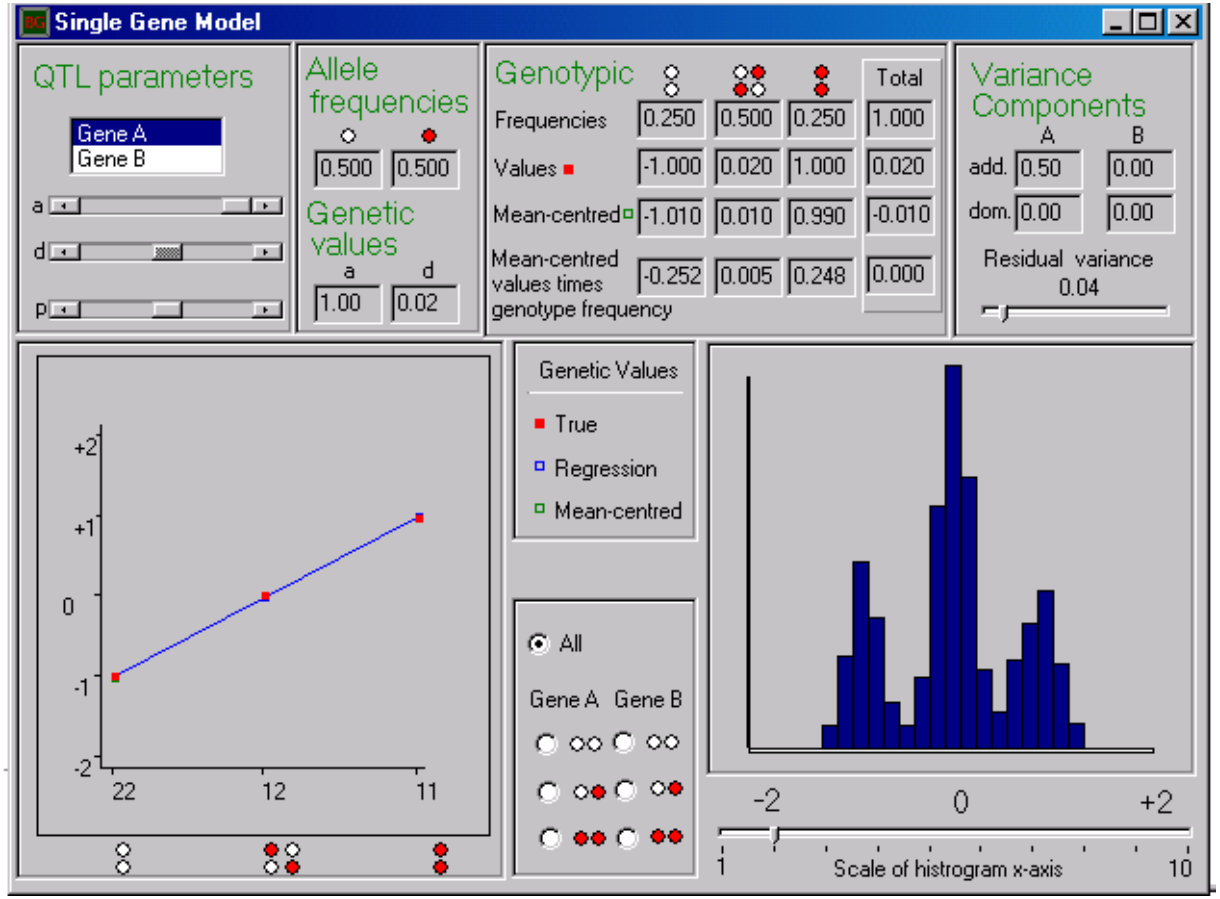


No departure from additivity

Significant departure from additivity

Practical

`H:\ferreira\biometric\sgene.exe`



Practical

- **Aim** Visualize graphically how allele frequencies, genetic effects, dominance, etc, influence trait mean and variance

Ex1

a=0, d=0, p=0.4, Residual Variance = 0.04, Scale = 2.
Vary a from 0 to 1.

Ex2

a=1, d=0, p=0.4, Residual Variance = 0.04, Scale = 2.
Vary d from -1 to 1.

Ex3

a=1, d=0, p=0.4, Residual Variance = 0.04, Scale = 2.
Vary p from 0 to 1.

Look at scatter-plot, histogram and variance components.

Some conclusions

1. Additive genetic variance depends on

   allele frequency p

   & additive genetic value a

   as well as

   dominance deviation d

2. Additive genetic variance typically greater than dominance variance

Biometrical model for single biallelic QTL

1. Contribution of the QTL to the Mean (X)

2. Contribution of the QTL to the Variance (X)

3. Contribution of the QTL to the Covariance (X, Y)

Biometrical model for single biallelic QTL

3. Contribution of the QTL to the Cov (*X*,*Y*)

$$Cov(X,Y) = \sum_i (x_i - \mu_X)(y_i - \mu_Y) f(x_i, y_i)$$

	***AA* (*a*-*m*)**	***Aa* (*d*-*m*)**	***aa* (*-a*-*m*)**
***AA* (*a*-*m*)**	$(a-m)^2$		
***Aa* (*d*-*m*)**	$(a-m)(d-m)$	$(d-m)^2$	
***aa* (*-a*-*m*)**	$(a-m)(-a-m)$	$(d-m)(-a-m)$	$(-a-m)^2$

Biometrical model for single biallelic QTL

3A. Contribution of the QTL to the Cov *(X,Y)* – MZ twins

$$Cov(X,Y) = \sum_i (x_i - \mu_X)(y_i - \mu_Y) f(x_i, y_i)$$

	AA *(a-m)*	***Aa*** *(d-m)*	***aa*** *(-a-m)*
AA *(a-m)*	$p^2(a-m)^2$		
Aa *(d-m)*	$0\ (a-m)\ (d-m)$	$2pq(d-m)^2$	
aa *(-a-m)*	$0\ (a-m)(-a-m)$	$0\ (d-m)(-a-m)$	$q^2(-a-m)^2$

$$Cov(X,Y) = (a-m)^2p^2 + (d-m)^2 2pq + (-a-m)^2q^2$$

$$= 2pq[a+(q-p)d]^2 + (2pqd)^2 = V_{A_{QTL}} + V_{D_{QTL}}$$

Biometrical model for single biallelic QTL

3B. Contribution of the QTL to the Cov (X,Y) – Parent-Offspring

	AA $(a-m)$	***Aa*** $(d-m)$	***aa*** $(-a-m)$
AA $(a-m)$	$p^3(a-m)^2$		
Aa $(d-m)$	$p^2q\,(a-m)\,(d-m)$	$pq\,(d-m)^2$	
aa $(-a-m)$	$0\,(a-m)\,(-a-m)$	$pq^2\,(d-m)\,(-a-m)$	$q^3\,(-a-m)^2$

- *e.g.* given an *AA* father, an *AA* offspring can come from either *AA* x *AA* or *AA* x *Aa* parental mating types

AA x *AA*	will occur $p^2 \times p^2 = p^4$
	and have *AA* offspring Prob()=1
AA x *Aa*	will occur $p^2 \times 2pq = 2p^3q$
	and have *AA* offspring Prob()=0.5
	and have *Aa* offspring Prob()=0.5

Therefore, P(*AA* father & *AA* offspring) $= p^4 + p^3q$

$= p^3(p+q)$

$= p^3$

Biometrical model for single biallelic QTL

3B. Contribution of the QTL to the Cov (X,Y) – Parent-Offspring

	AA $(a-m)$	***Aa*** $(d-m)$	***aa*** $(-a-m)$
AA $(a-m)$	$p^3(a-m)^2$		
Aa $(d-m)$	$p^2q\,(a-m)\,(d-m)$	$pq\,(d-m)^2$	
aa $(-a-m)$	$0\,(a-m)\,(-a-m)$	$pq^2\,(d-m)\,(-a-m)$	$q^3\,(-a-m)^2$

$$Cov\,(X,Y) = (a-m)^2p^3 + \ldots + (-a-m)^2q^3$$

$$= pq[a+(q-p)d]^2 = \tfrac{1}{2}V_{A_{QTL}}$$

Biometrical model for single biallelic QTL

3C. Contribution of the QTL to the Cov (X,Y) – Unrelated individuals

	AA $(a-m)$	***Aa*** $(d-m)$	***aa*** $(-a-m)$
AA $(a-m)$	$p^4(a-m)^2$		
Aa $(d-m)$	$2p^3q\,(a-m)\,(d-m)$	$4p^2q^2\,(d-m)^2$	
aa $(-a-m)$	$p^2q^2(a-m)(-a-m)$	$2pq^3\,(d-m)(-a-m)$	$q^4\,(-a-m)^2$

$$Cov\,(X,Y) = (a-m)^2p^4 + \ldots + (-a-m)^2q^4$$

$$= 0$$

Biometrical model for single biallelic QTL

3D. Contribution of the QTL to the Cov (*X,Y*) – DZ twins and full sibs

	¼ genome	¼ genome	¼ genome	¼ genome
# identical alleles inherited from parents	**2**	**1** (father)	**1** (mother)	**0**

¼ (2 alleles) + ½ (1 allele) + ¼ (0 alleles)

MZ twins *P-O* *Unrelateds*

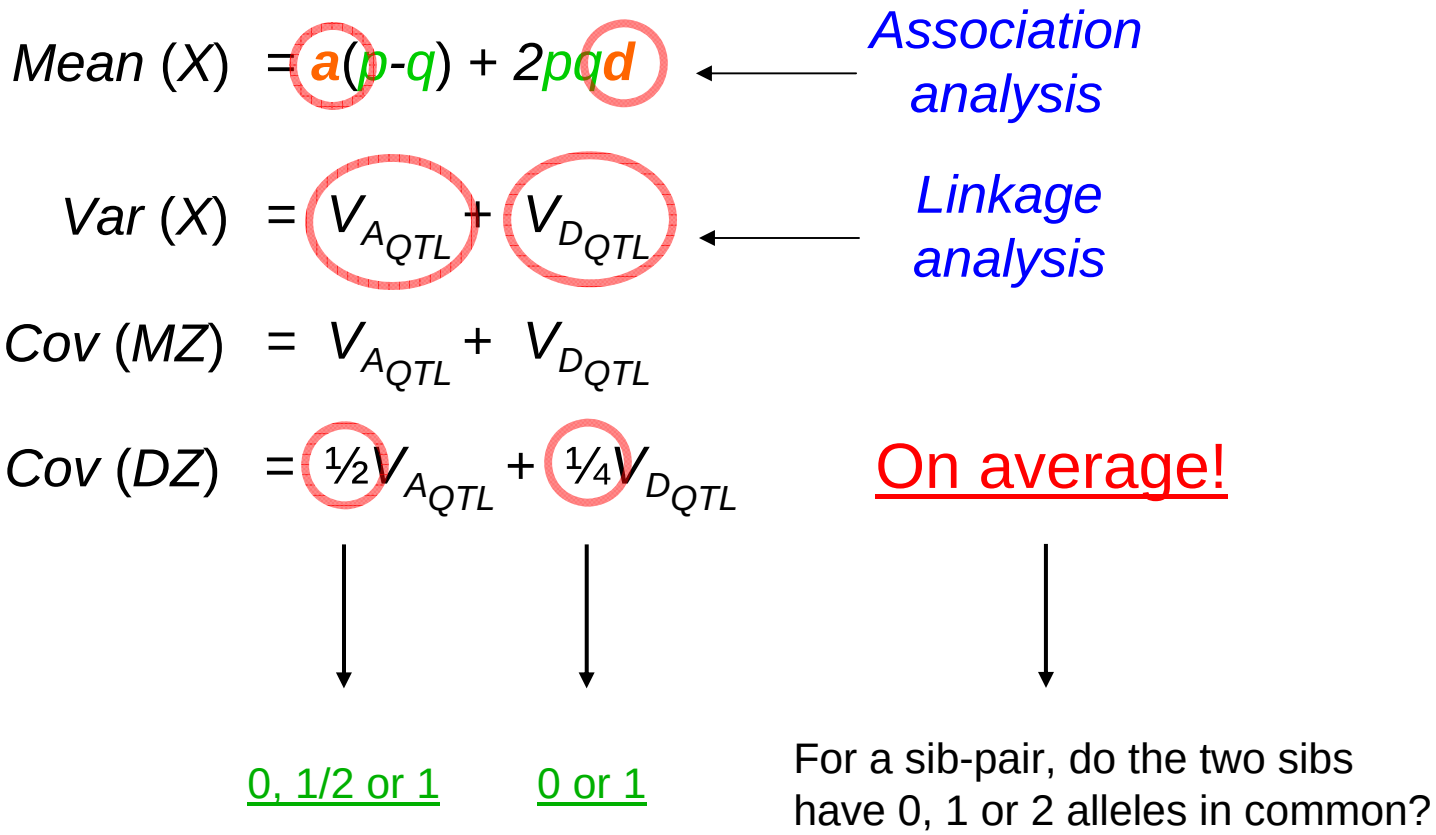
$$Cov(X,Y) = \tfrac{1}{4} Cov(MZ) + \tfrac{1}{2} Cov(P\text{-}O) + \tfrac{1}{4} Cov(Unrel)$$

$$= \tfrac{1}{4}(V_{A_{QTL}} + V_{D_{QTL}}) + \tfrac{1}{2}(\tfrac{1}{2} V_{A_{QTL}}) + \tfrac{1}{4}(0)$$

$$= \tfrac{1}{2} V_{A_{QTL}} + \tfrac{1}{4} V_{D_{QTL}}$$

Summary so far…

- Biometrical model predicts contribution of a QTL to the mean, variance and covariances of a trait



IBD estimation / Linkage

4. Introduction to Linkage Analysis

For a heritable trait...

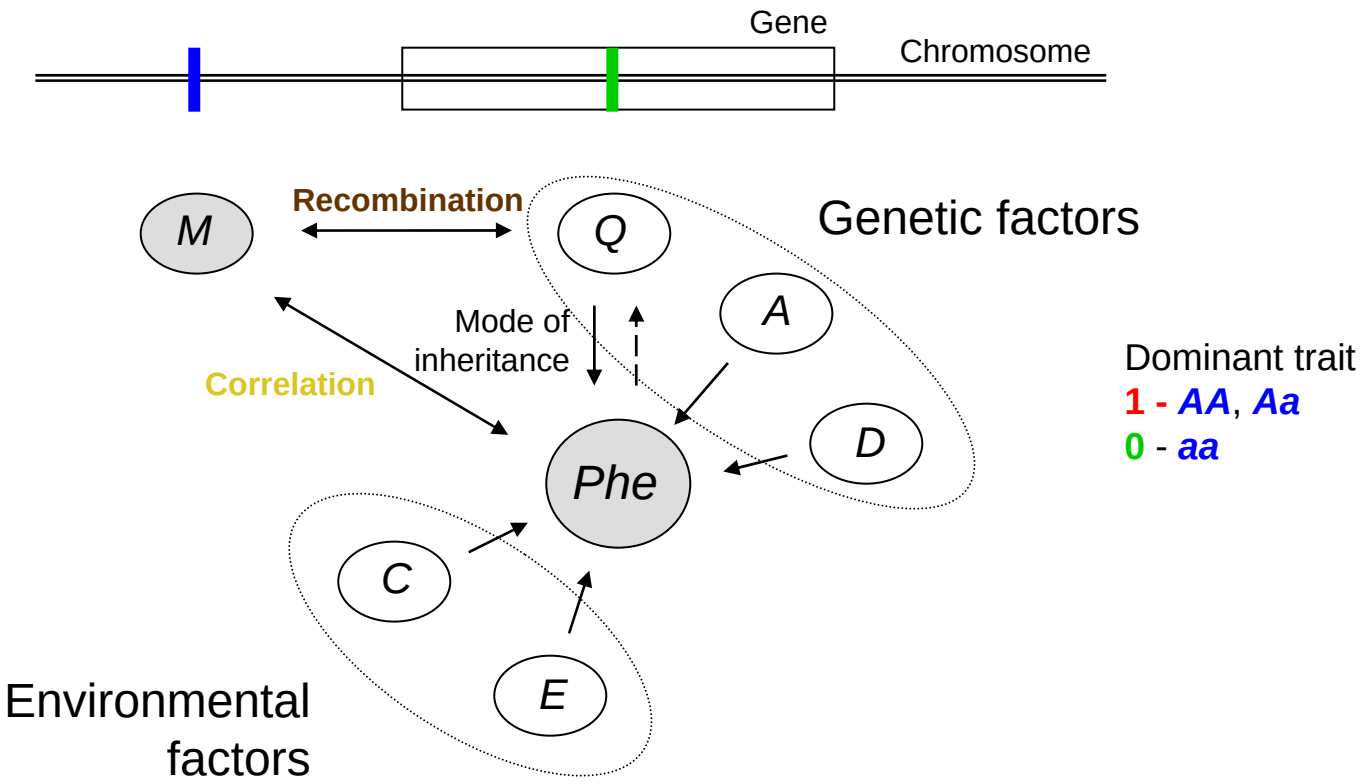
Linkage: localize region of the genome where a QTL that regulates the trait is likely to be harboured

Family-specific phenomenon:
Affected individuals in a family share the same ancestral predisposing DNA segment at a given QTL

Association: identify a QTL that regulates the trait

Population-specific phenomenon:
Affected individuals in a population share the same ancestral predisposing DNA segment at a given QTL

Linkage Analysis: Parametric vs. Nonparametric



Adapted from Weiss & Terwilliger 2000

Approach

- Parametric: genotypes marker locus & genotypes trait locus

  (latter inferred from phenotype according to a specific disease model)

  Parameter of interest: θ between marker and trait loci

- Nonparametric: genotypes marker locus & phenotype

  If a trait locus truly regulates the expression of a phenotype, then two relatives with similar phenotypes should have similar genotypes at a marker in the vicinity of the trait locus, and vice-versa.

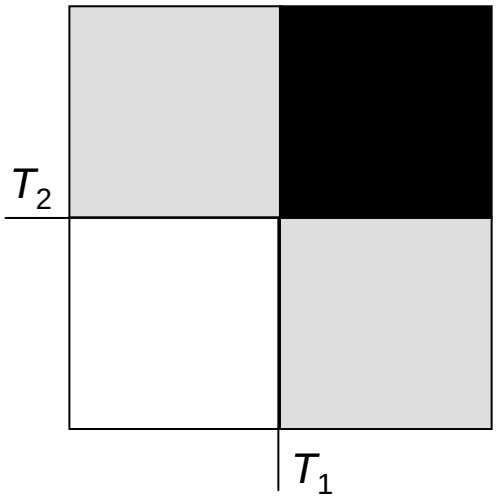
  Interest: correlation between phenotypic similarity and marker genotypic similarity

  No need to specify mode of inheritance, allele frequencies, etc...

Phenotypic similarity between relatives

- Squared trait differences $(X_1 - X_2)^2$
- Squared trait sums $(X_1 + X_2)^2$
- Trait cross-product $[(X_1 - \mu)\cdot(X_2 - \mu)]$
- Trait variance-covariance matrix $\begin{Bmatrix} Var(X_1) & Cov(X_1X_2) \\ Cov(X_1X_2) & Var(X_2) \end{Bmatrix}$

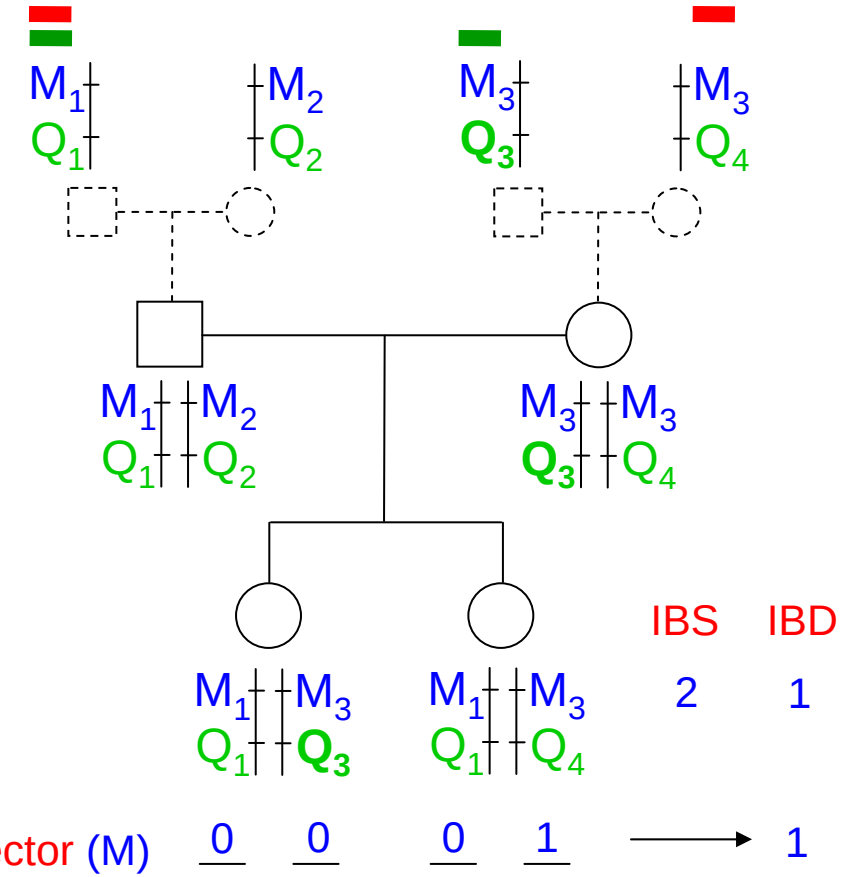
- Affection concordance



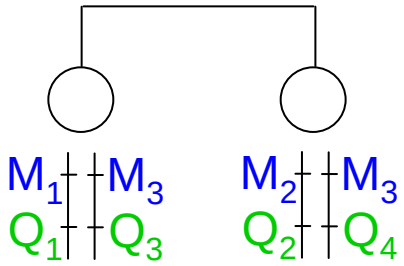
Genotypic similarity between relatives

- IBS — Alleles shared Identical By State "look the same", may have the same DNA sequence but they are not necessarily derived from a known common ancestor

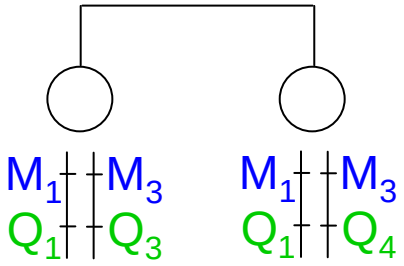
- IBD — Alleles shared Identical By Descent are a copy of the same ancestor allele



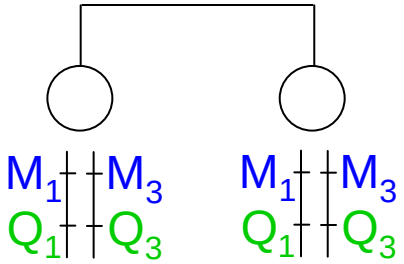
Genotypic similarity between relatives - π



Inheritance vector (M)	Number of alleles shared IBD	Proportion of alleles shared IBD - π
0 0 1 1	0	0



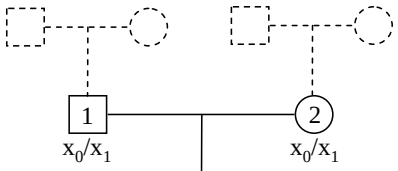
Inheritance vector (M)	Number of alleles shared IBD	Proportion of alleles shared IBD - π
0 0 0 1	1	0.5



Inheritance vector (M)	Number of alleles shared IBD	Proportion of alleles shared IBD - π
0 0 0 0	2	1

Genotypic similarity between relatives - $\hat{\pi}$

A B C D



2^{2n}

3	4	Inheritance vector	IBD
x_0/x_0	x_0/x_0	0000	2
x_0/x_0	x_0/x_1	0001	1
x_0/x_0	x_1/x_0	0010	1
x_0/x_0	x_1/x_1	0011	0
x_0/x_1	x_0/x_0	0100	1
x_0/x_1	x_0/x_1	0101	2
x_0/x_1	x_1/x_0	0110	0
x_0/x_1	x_1/x_1	0111	1
x_1/x_0	x_0/x_0	1000	1
x_1/x_0	x_0/x_1	1001	0
x_1/x_0	x_1/x_0	1010	2
x_1/x_0	x_1/x_1	1011	1
x_1/x_1	x_0/x_0	1100	0
x_1/x_1	x_0/x_1	1101	1
x_1/x_1	x_1/x_0	1110	1
x_1/x_1	x_1/x_1	1111	2

P (IBD=0)
P (IBD=1)
P (IBD=2)

$$\hat{\pi} =$$

$$Var\,(X) = V_{A_{QTL}} + V_{D_{QTL}}$$

$$Cov\,(MZ) = V_{A_{QTL}} + V_{D_{QTL}}$$

$$Cov\,(DZ) = \tfrac{1}{2}V_{A_{QTL}} + \tfrac{1}{4}V_{D_{QTL}}$$

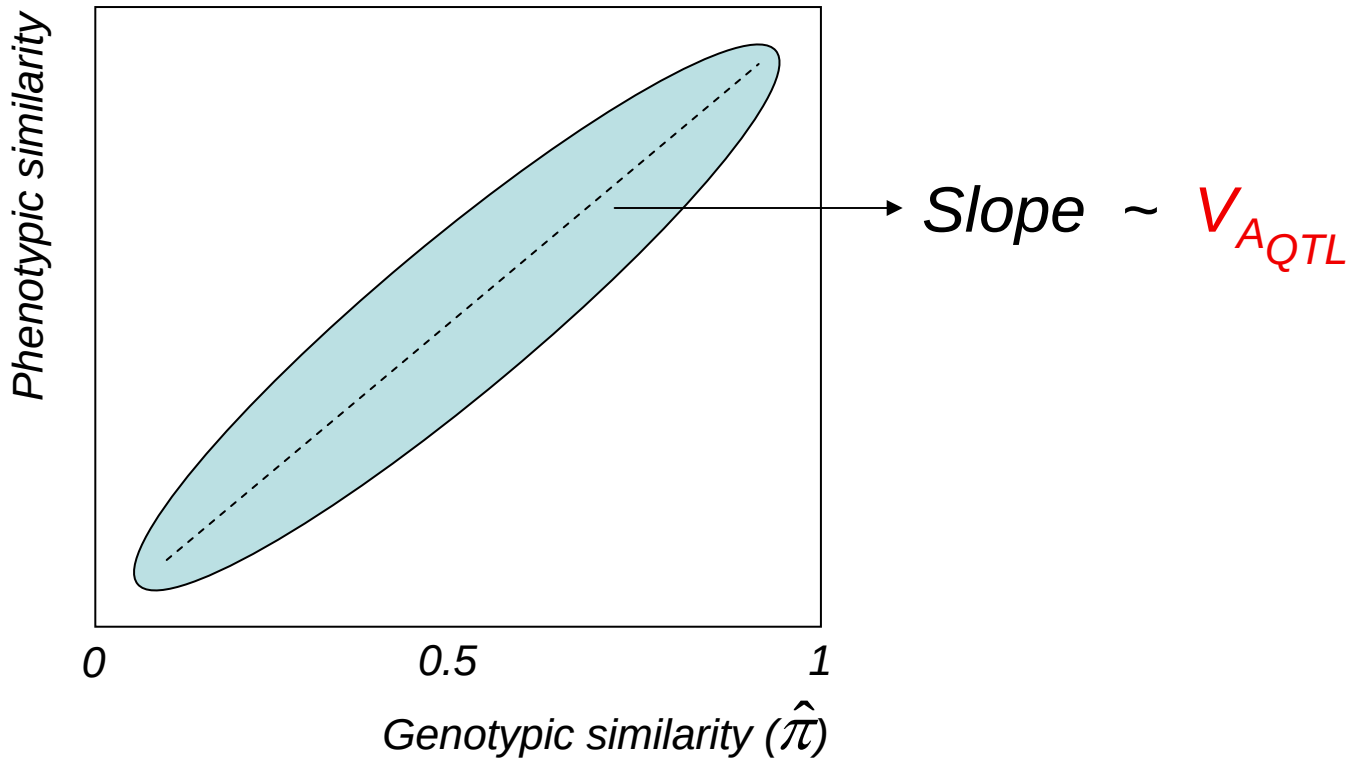
On average!

$$Cov\,(DZ) = \hat{\pi} \cdot V_{A_{QTL}} + \pi_2 \cdot V_{D_{QTL}}$$

For a given twin pair

$$Cov\,(DZ) = \hat{\pi} \cdot V_{A_{QTL}}$$

$$Cov\,(DZ) = V_{A_{QTL}} \cdot \hat{\pi}$$



Statistics that incorporate both phenotypic and genotypic similarities to test V_{QTL}

- Regression-based methods

  Haseman-Elston, MERLIN-regress

- Variance components methods

  Mx, MERLIN, SOLAR, GENEHUNTER