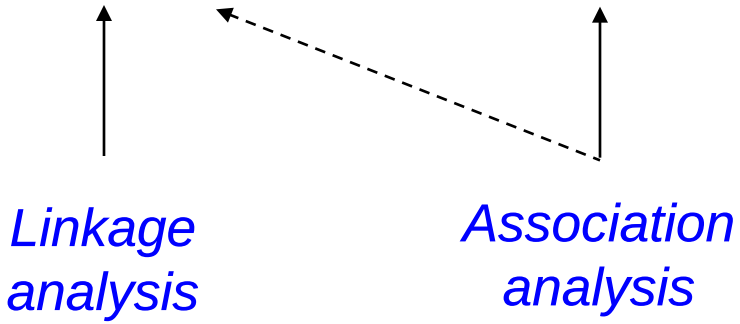
Genetic principles for linkage and association analyses

Manuel Ferreira & Pak Sham

Boulder, 2009

Gene mapping

- LOCALIZE and then IDENTIFY a locus that regulates a trait



Linkage: If a locus regulates a trait, **Trait Variance** and **Covariance between individuals** will be influenced by this locus.

Association: If a locus regulates a trait, **Trait Mean** in the population will also be influenced by this locus.

- Revisit common genetic parameters - such as allele frequencies, genetic effects, dominance, variance components, etc

- Use these parameters to construct a biometric genetic model

↓

Model that expresses the:

(1) Mean

(2) Variance

(3) Covariance between individuals

for a quantitative phenotype as a function of the genetic parameters of a given locus.

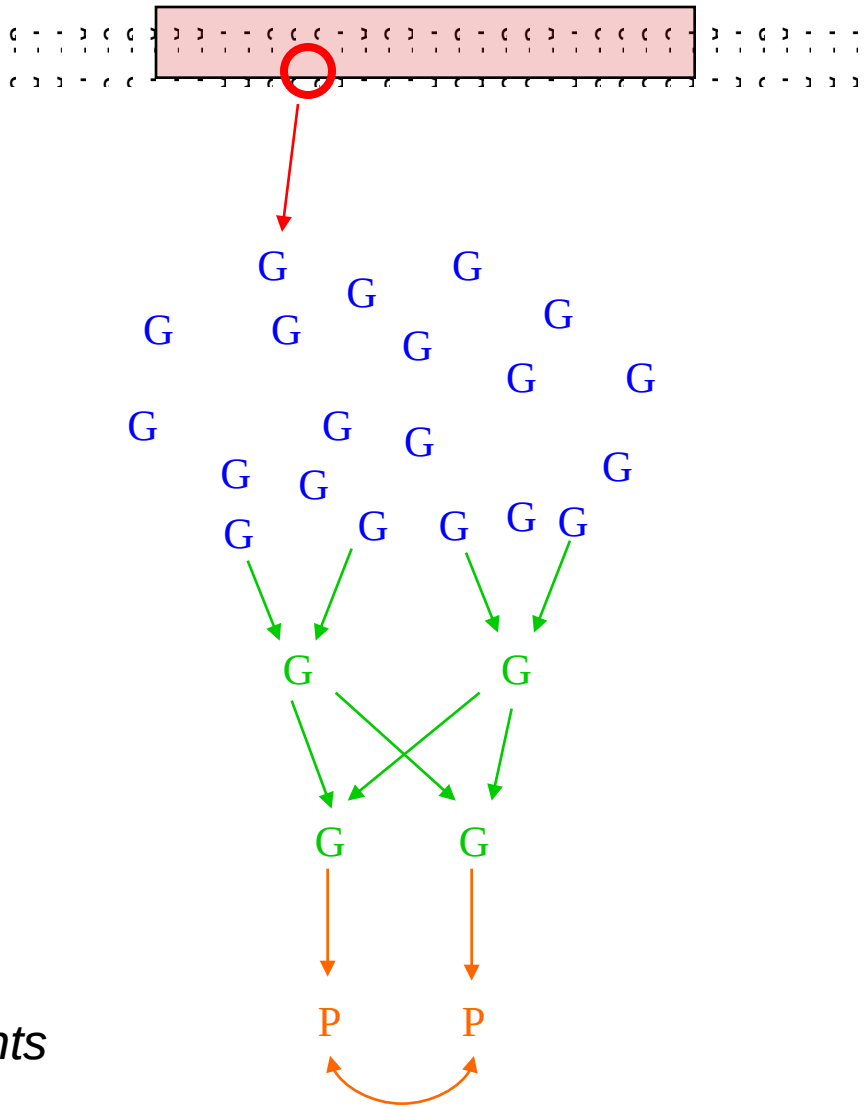
- See how the biometric model provides a useful framework for linkage and association methods.

Outline

1. Genetic concepts
2. Very basic statistical concepts
3. Biometrical model
4. Introduction to linkage analysis

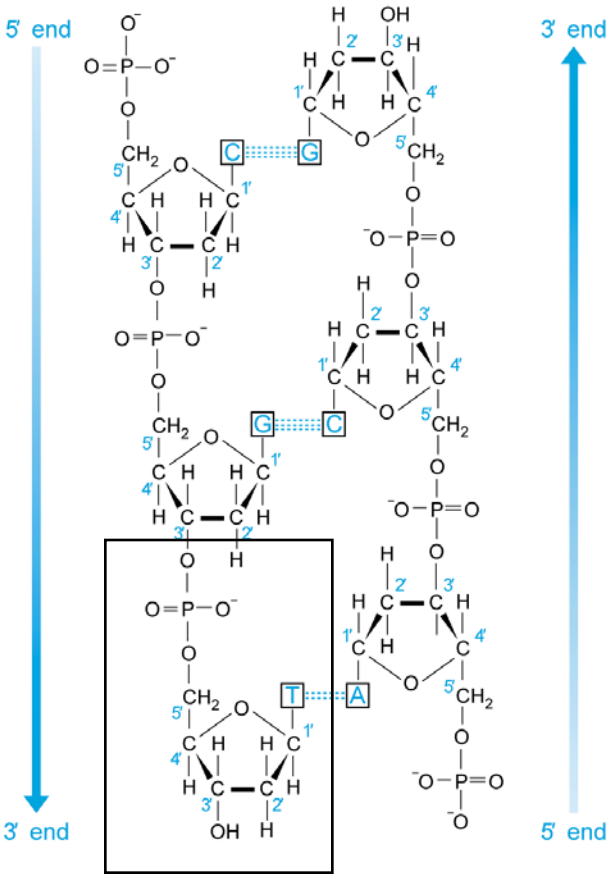
1. Genetic concepts

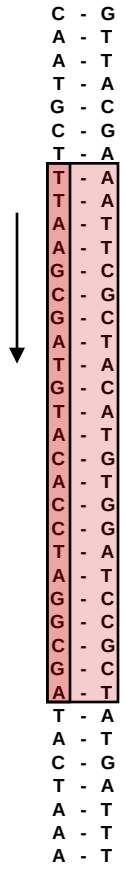
- **A. DNA level**
  - *DNA structure, organization*
  - *recombination*
- **B. Population level**
  - *Allele and genotype frequencies*
- **C. Transmission level**
  - *Mendelian segregation*
  - *Genetic relatedness*
- **D. Phenotype level**
  - *Biometrical model*
  - *Additive and dominance components*



A. DNA level

- 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
- Two DNA molecules are held together in anti-parallel fashion by hydrogen bonds between bases [Watson-Crick rules] *Antiparallel double helix*
- A gene is a segment of DNA which is transcribed to give a protein or RNA product
- Only one strand is read during gene transcription
- Nucleotide: 1 phosphate group + 1 sugar + 1 base



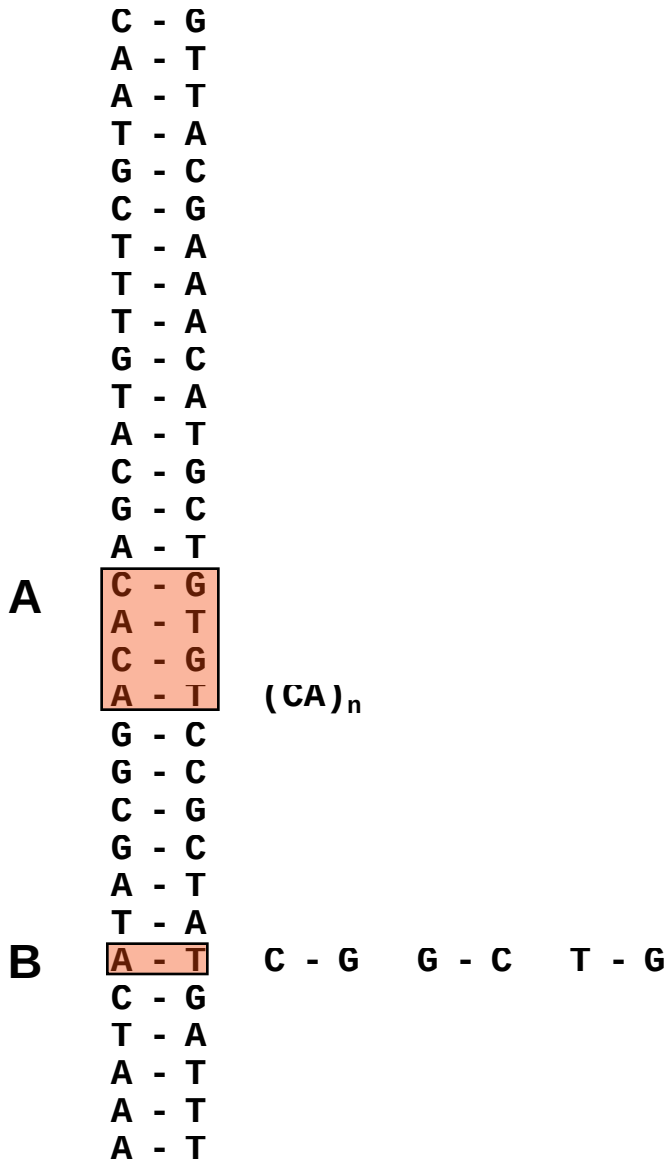


DNA polymorphisms

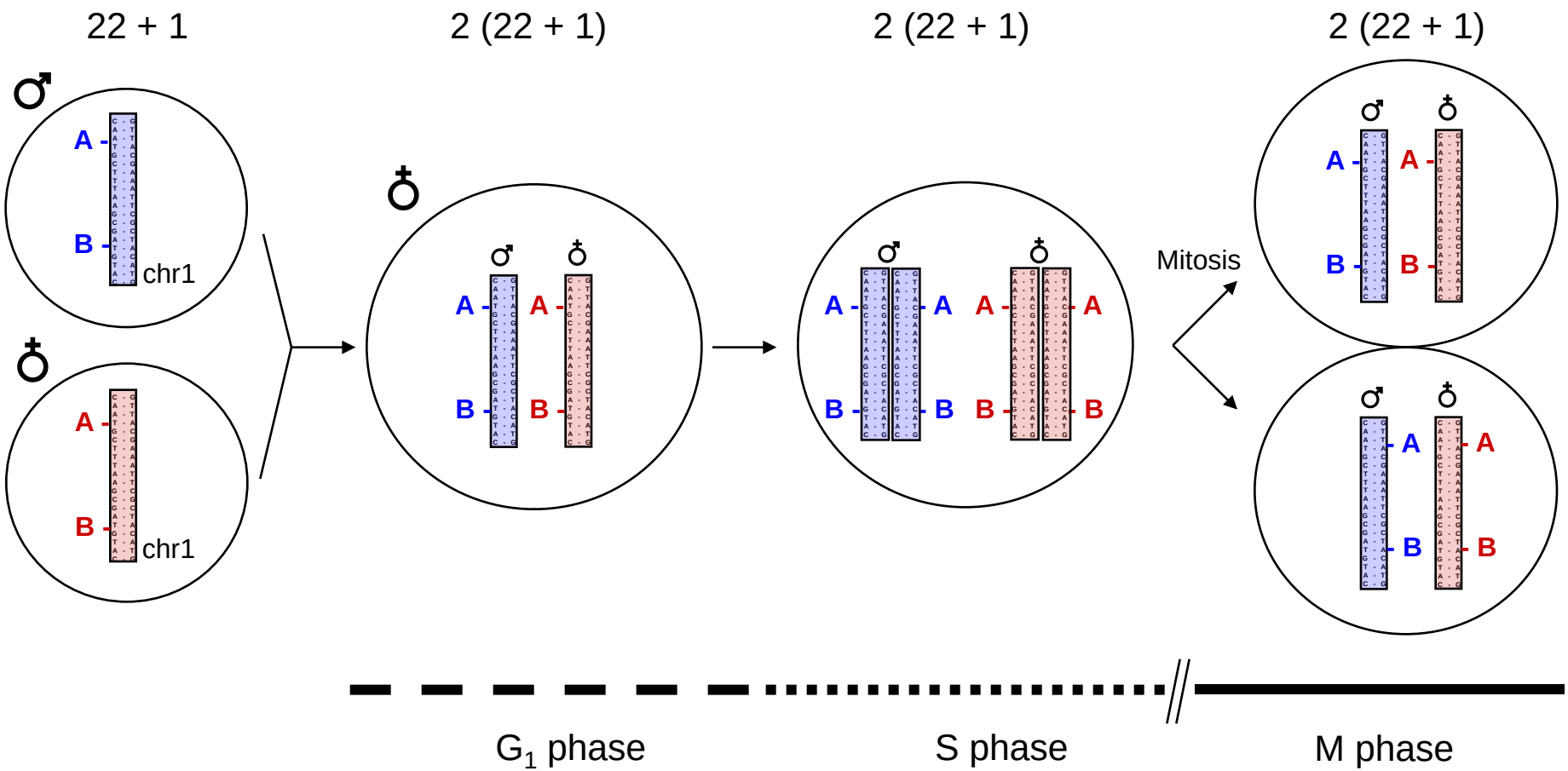
- Microsatellites
  >100,000
  Many alleles, eg. $(CA)_n$ repeats, very informative

- SNPs
  14,708,752 (build 129, 03 Mar ‘09)
  Most with 2 alleles (up to 4), not very informative

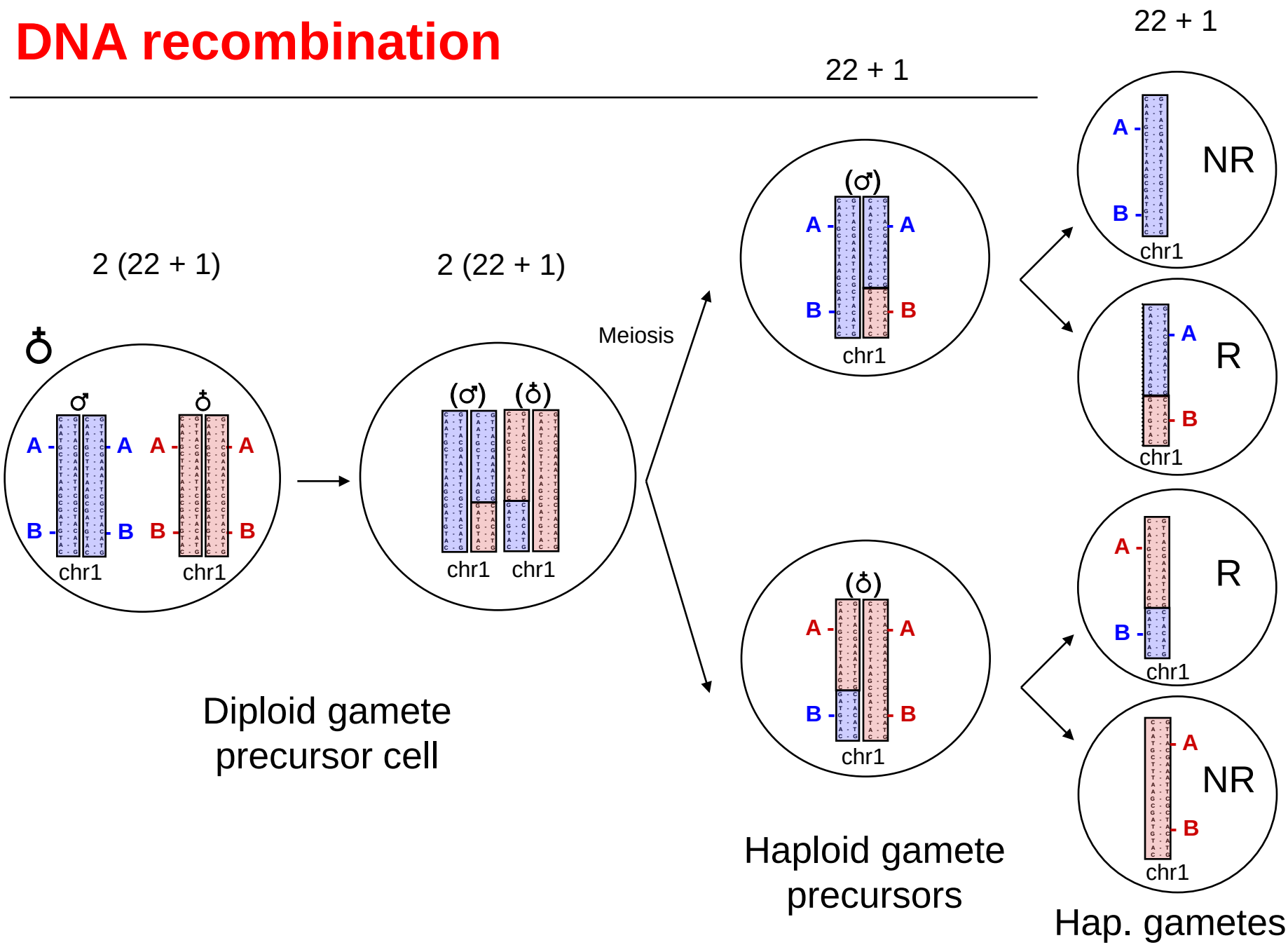
- Copy Number Variants
  >5,000
  Many alleles, just recently automated



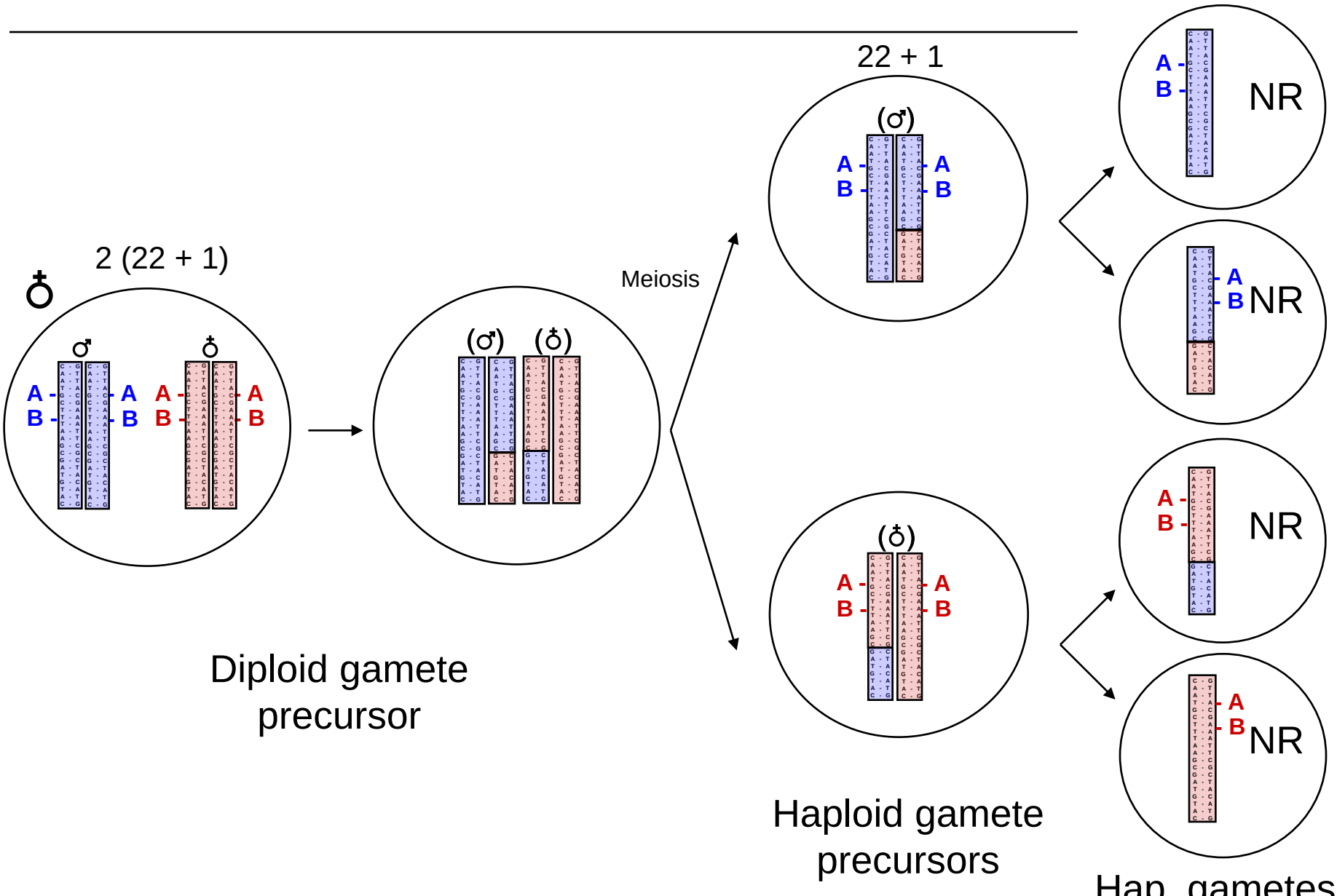
DNA organization



DNA recombination



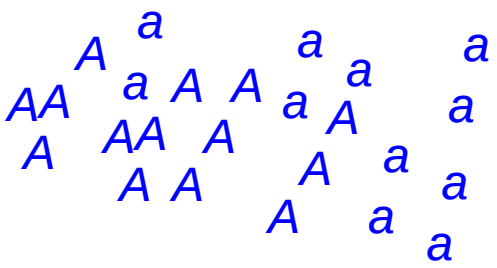
DNA recombination between linked loci



B. Population level

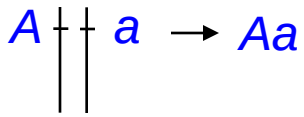
1. Allele frequencies

- A single locus, with two alleles
  - Biallelic
  - Single nucleotide polymorphism, SNP



- Alleles ***A*** and ***a***
  - Frequency of ***A*** is ***p***
  - Frequency of ***a*** is ***q*** = 1 – ***p***

- A genotype is the combination of the two alleles



B. Population level

2. Genotype frequencies (Random mating)

Allele 2 \ Allele 1	***A*** (***p***)	***a*** (***q***)
A (***p***)	***AA*** (***p²***)	***Aa*** (***pq***)
a (***q***)	***aA*** (***qp***)	***aa*** (***q²***)

Hardy-Weinberg Equilibrium frequencies

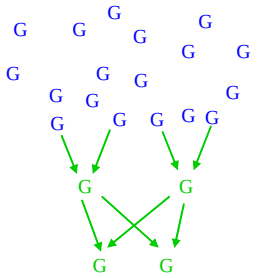
P (***AA***) = p^2

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

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

$p^2 + 2pq + q^2 = 1$

C. Transmission level



Mendel's law of segregation

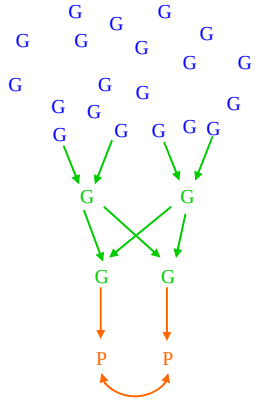
Mother (A_3A_4)

Segregation (Meiosis)

Gametes

Father (A_1A_2)	A_3 (½)	A_4 (½)
A_1 (½)	A_1A_3 (¼)	A_1A_4 (¼)
A_2 (½)	A_2A_3 (¼)	A_2A_4 (¼)

D. Phenotype level



1. Classical Mendelian traits

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

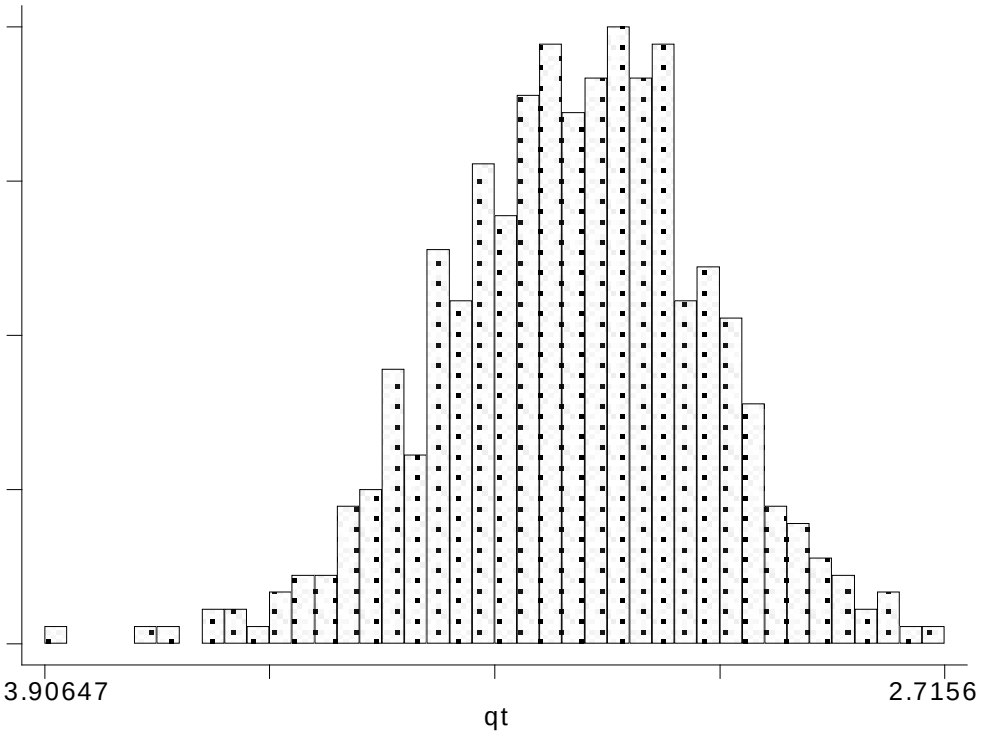
Huntington's disease
(CAG)n repeat, huntingtin gene

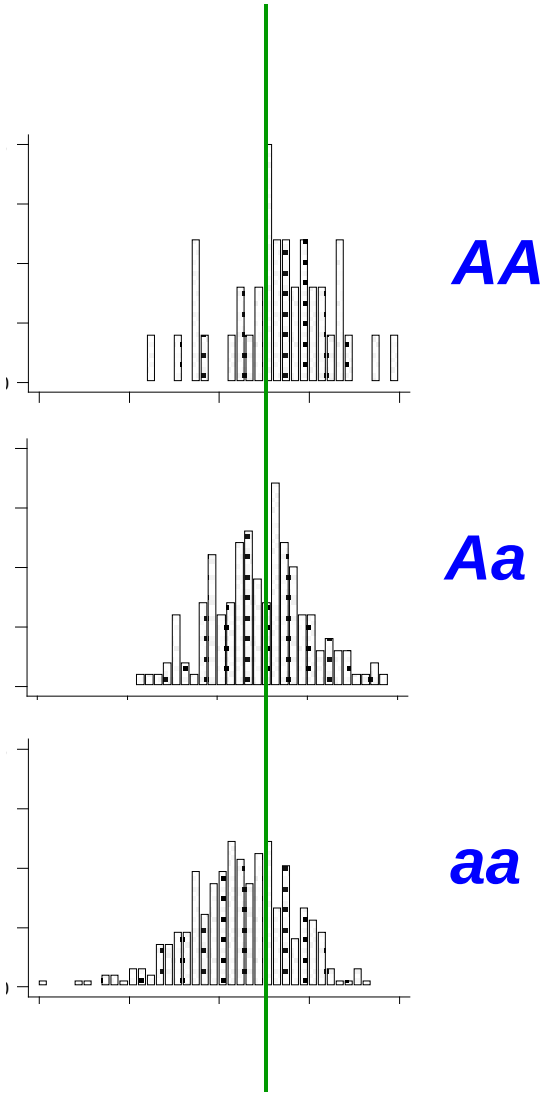
- Recessive trait
  - ***AA*** **1**
  - ***aa, Aa*** **0**

Cystic fibrosis
3 bp deletion exon 10 CFTR gene

D. Phenotype level

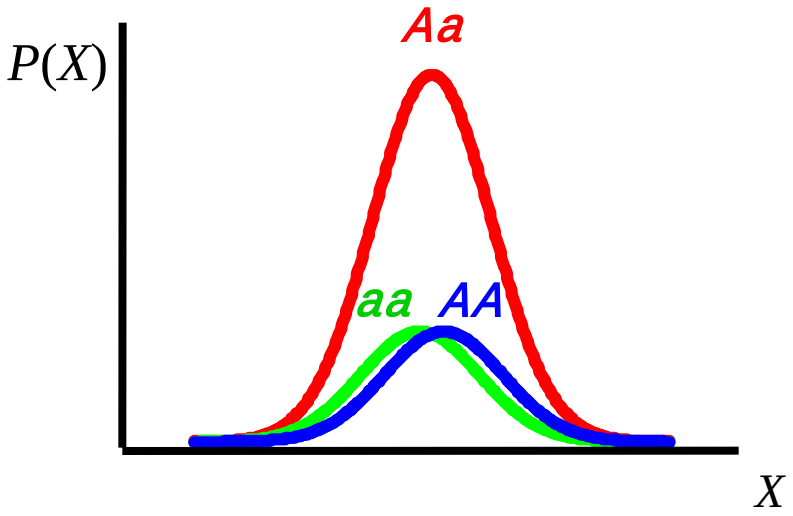
2. Quantitative traits

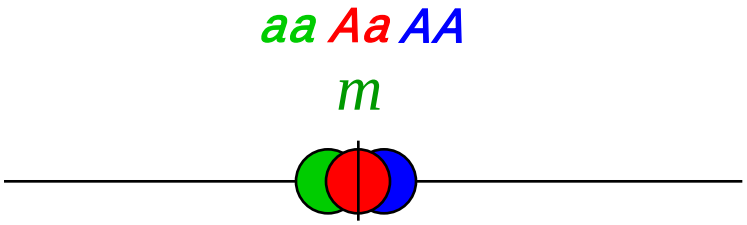




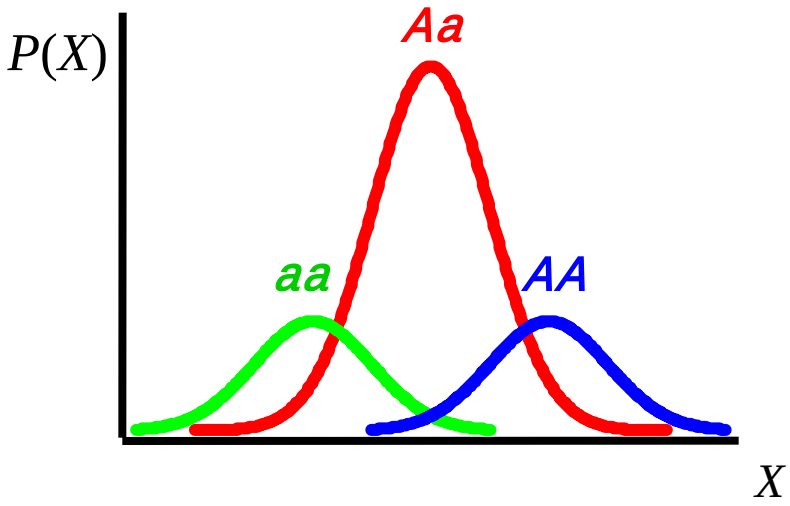
e.g. cholesterol levels

D. Phenotype level

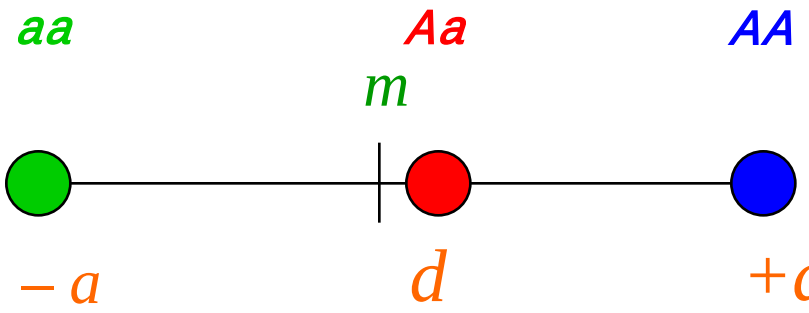




D. Phenotype level



Biometric Model



← Genotypic effect

2. Very basic statistical concepts

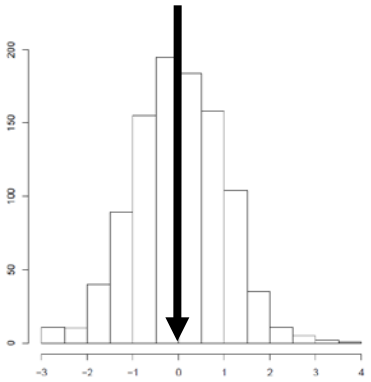
Mean, variance, covariance

1. Mean (X)

$$\mu(X) = \frac{\sum_i x_i}{n}$$

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

X
x_1
x_2
x_3
x_4
…
x_n



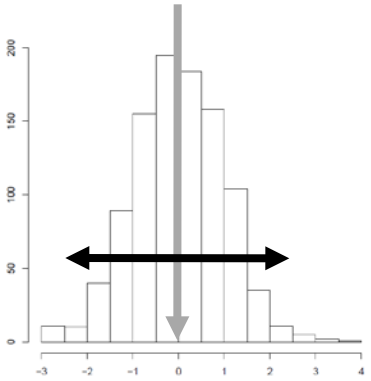
Mean, variance, covariance

2. Variance (X)

$$Var(X) = \frac{\sum_i (x_i - \mu)^2}{n-1}$$

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

X	**X-μ**	**(X-μ)^2**
x_1	x_1-μ	$(x_1$-$\mu)^2$
x_2	x_2-μ	$(x_2$-$\mu)^2$
x_3	x_3-μ	$(x_3$-$\mu)^2$
x_4	x_4-μ	$(x_4$-$\mu)^2$
…	…	…
x_n	x_n-μ	$(x_n$-$\mu)^2$



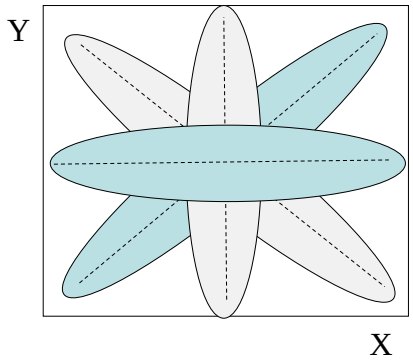
Mean, variance, covariance

3. Covariance (X,Y)

$$Cov(X,Y) = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{n-1}$$

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

X	Y	X-μ_X	Y-μ_Y
x_1	y_1	x_1-μ_X	y_1-μ_Y
x_2	y_2	x_2-μ_X	y_2-μ_Y
x_3	y_3	x_3-μ_X	y_3-μ_Y
x_4	y_4	x_4-μ_X	y_4-μ_Y
…	…	…	…
x_n	y_n	x_n-μ_X	y_n-μ_Y



3. Biometrical model

Biometrical model for single biallelic QTL

- Biallelic locus
  - Genotypes: ***AA*, *Aa*, *aa***
  - Genotype frequencies: p^2, $2pq$, q^2
- Alleles at this locus are transmitted from P-O according to Mendel's law of segregation
- Genotypes for this locus influence the expression of a quantitative trait *X* (i.e. locus *is* a QTL)

Biometrical genetic model that estimates the contribution of this QTL towards the **(1) Mean**, **(2) Variance** and **(3) Covariance between individuals** for this quantitative trait *X*

Biometrical model for single biallelic QTL

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

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

e.g. cholesterol levels in the population

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

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

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

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

$$= V_{QTL}$$

Heritability of *X* at this locus = V_{QTL} / V_{Total}

Biometrical model for single biallelic QTL

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

$m = a(p-q) + 2pqd$

$$= \underline{2pq[a+(q-p)d]^2} + \underline{(2pqd)^2}$$

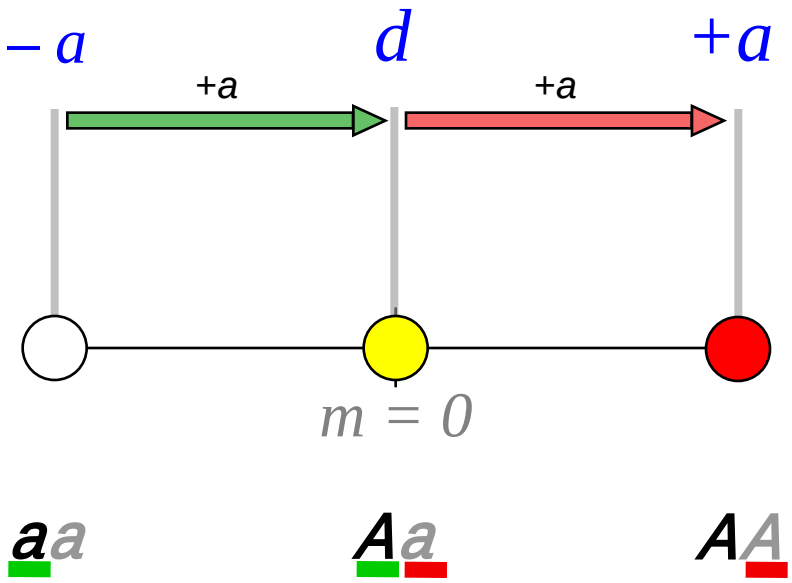
$$= V_{A_{QTL}} + V_{D_{QTL}}$$

Additive effects: the main effects of individual alleles

Dominance effects: represent the interaction between alleles

Biometrical model for single biallelic QTL

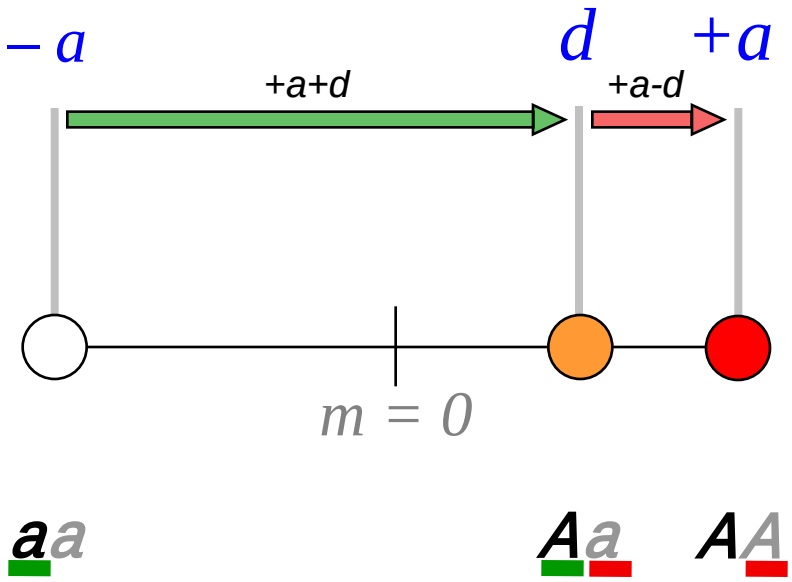
$d = 0$ *(no dominance)*



Additive model

Biometrical model for single biallelic QTL

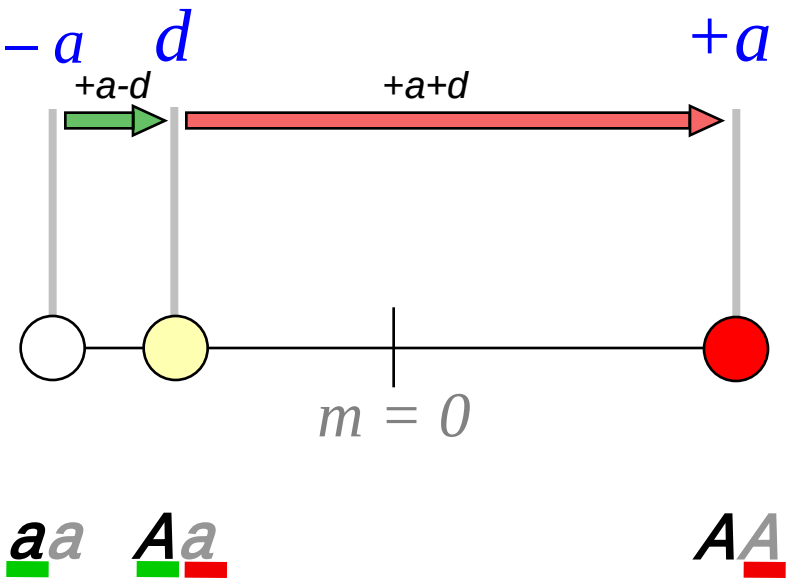
$d > 0$ *(dominance)*



[Dominant model]

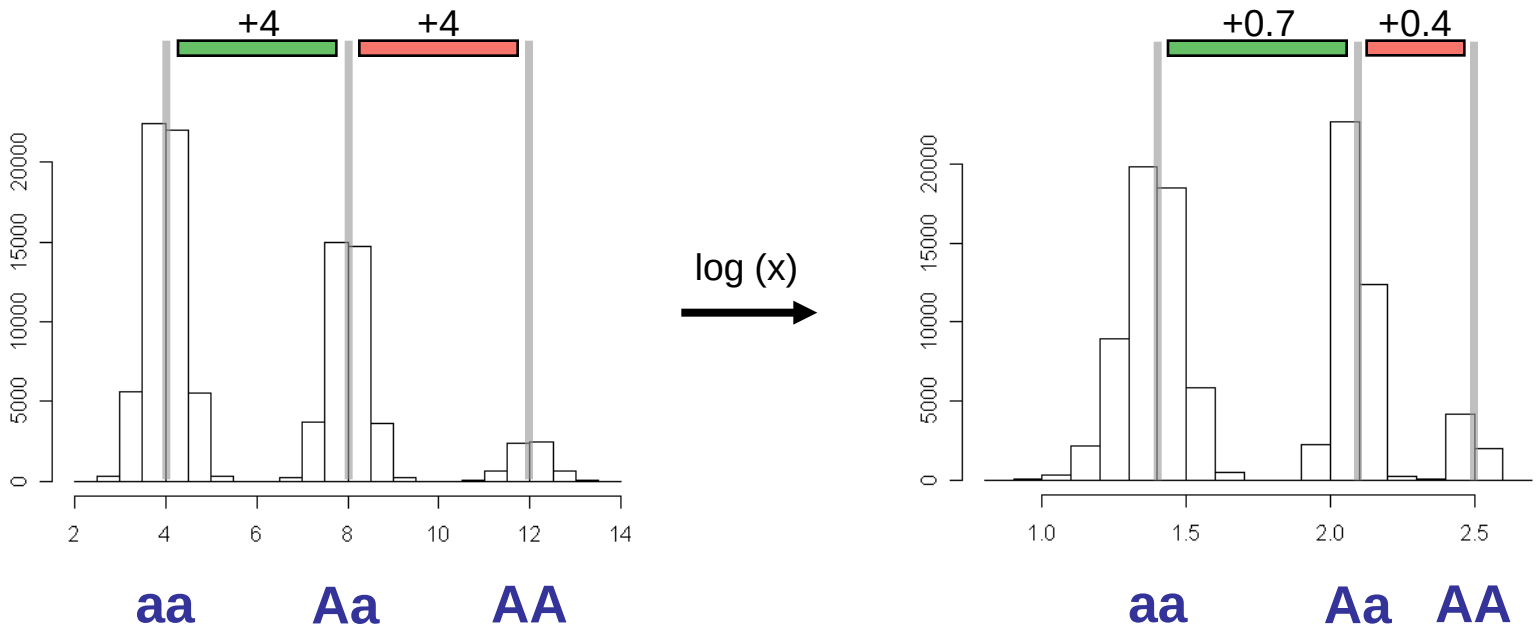
Biometrical model for single biallelic QTL

$d < 0$ *(dominance)*



Recessive model

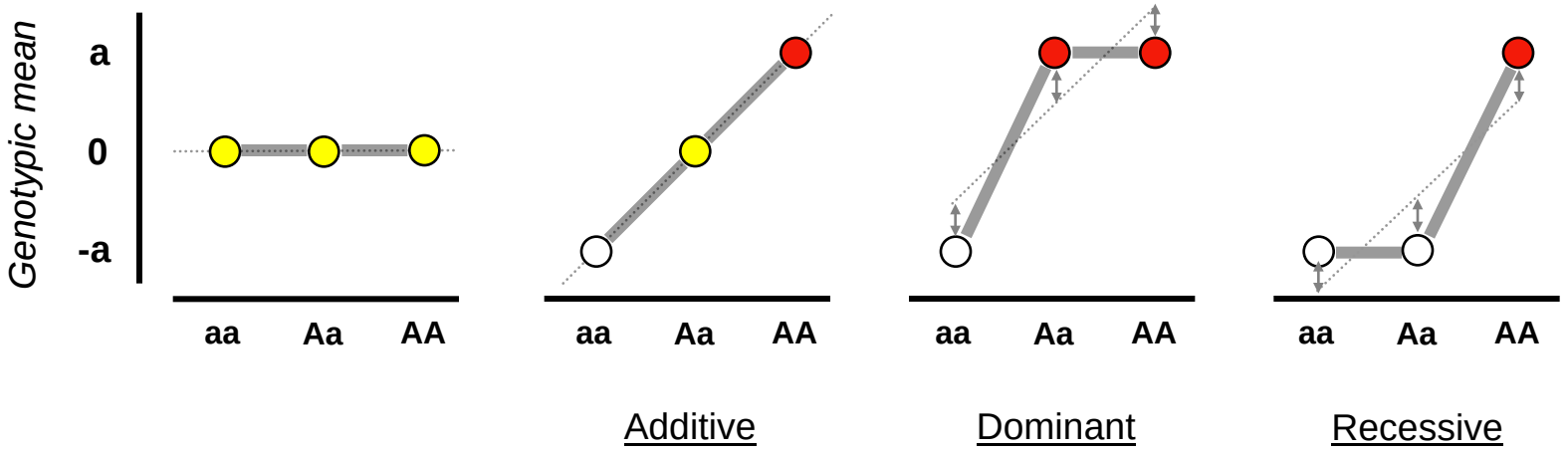
Statistical definition of dominance is scale dependent



No departure from additivity

Significant departure from additivity

Biometrical model for single biallelic QTL



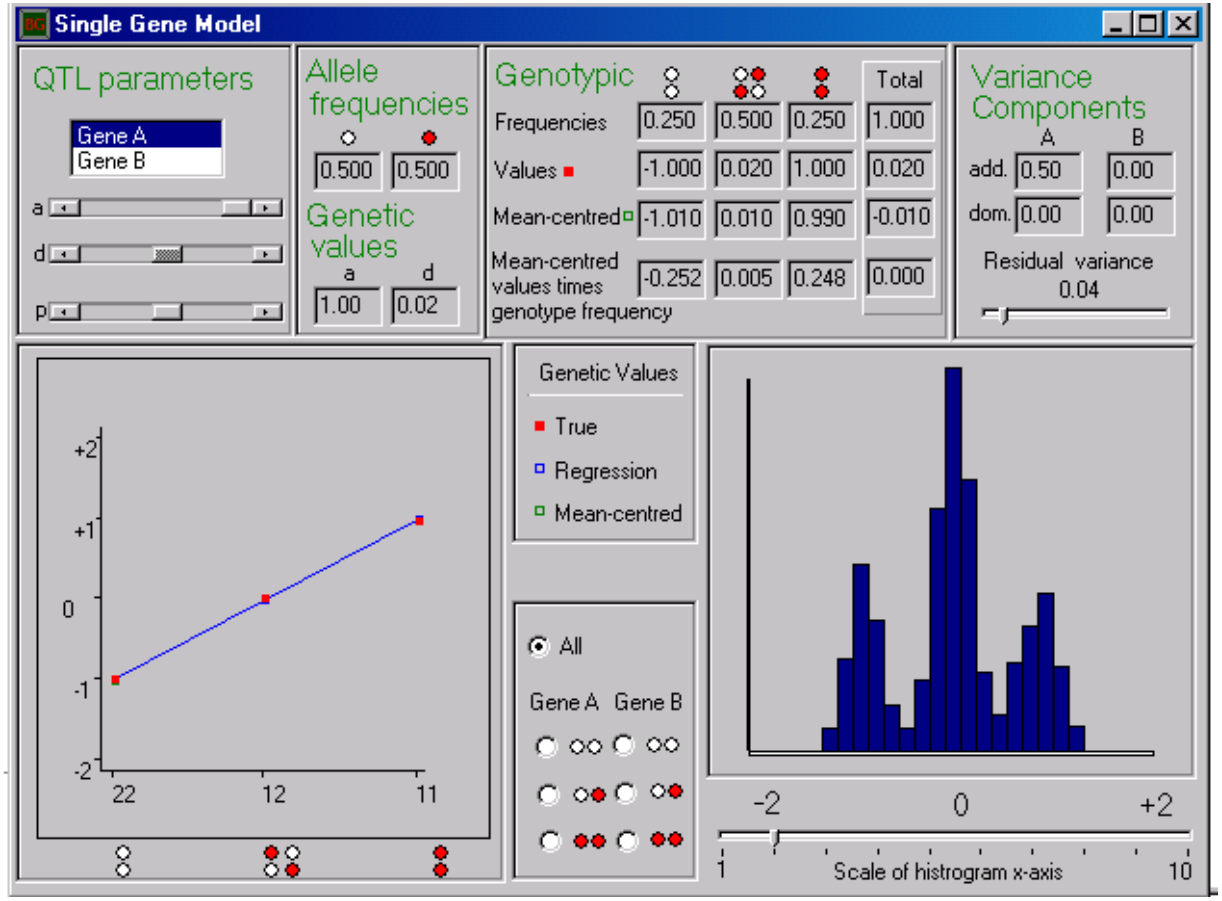
$$\text{Var}(X) = \text{Regression Variance} + \text{Residual Variance}$$

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

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

Practical

`H:\manuel\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*

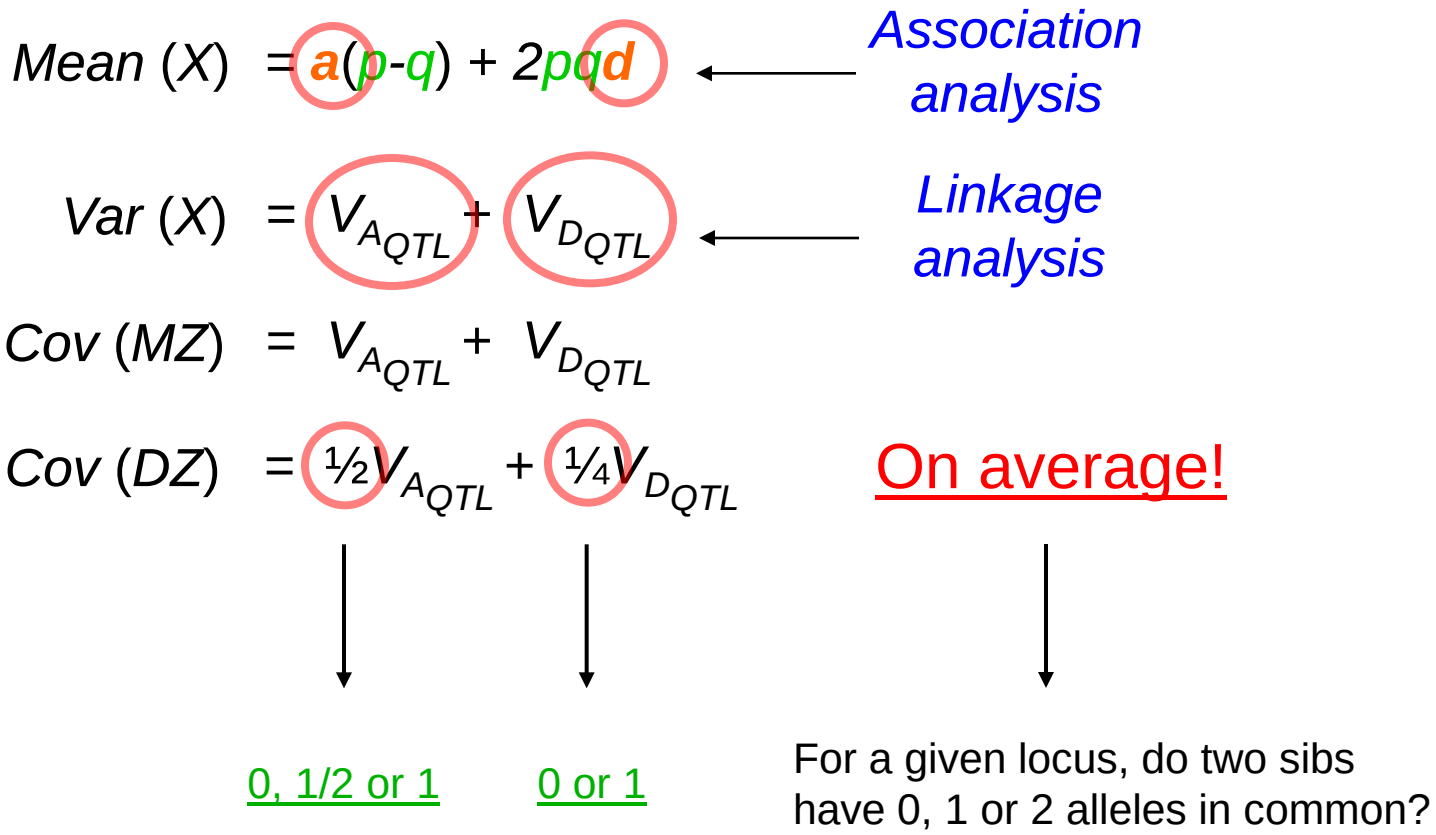
$$Cov(X,Y) = \tfrac{1}{4} Cov(MZ) + \tfrac{1}{2} Cov(P\text{-}O) + \tfrac{1}{4} Cov(Unrel)$$

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

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

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

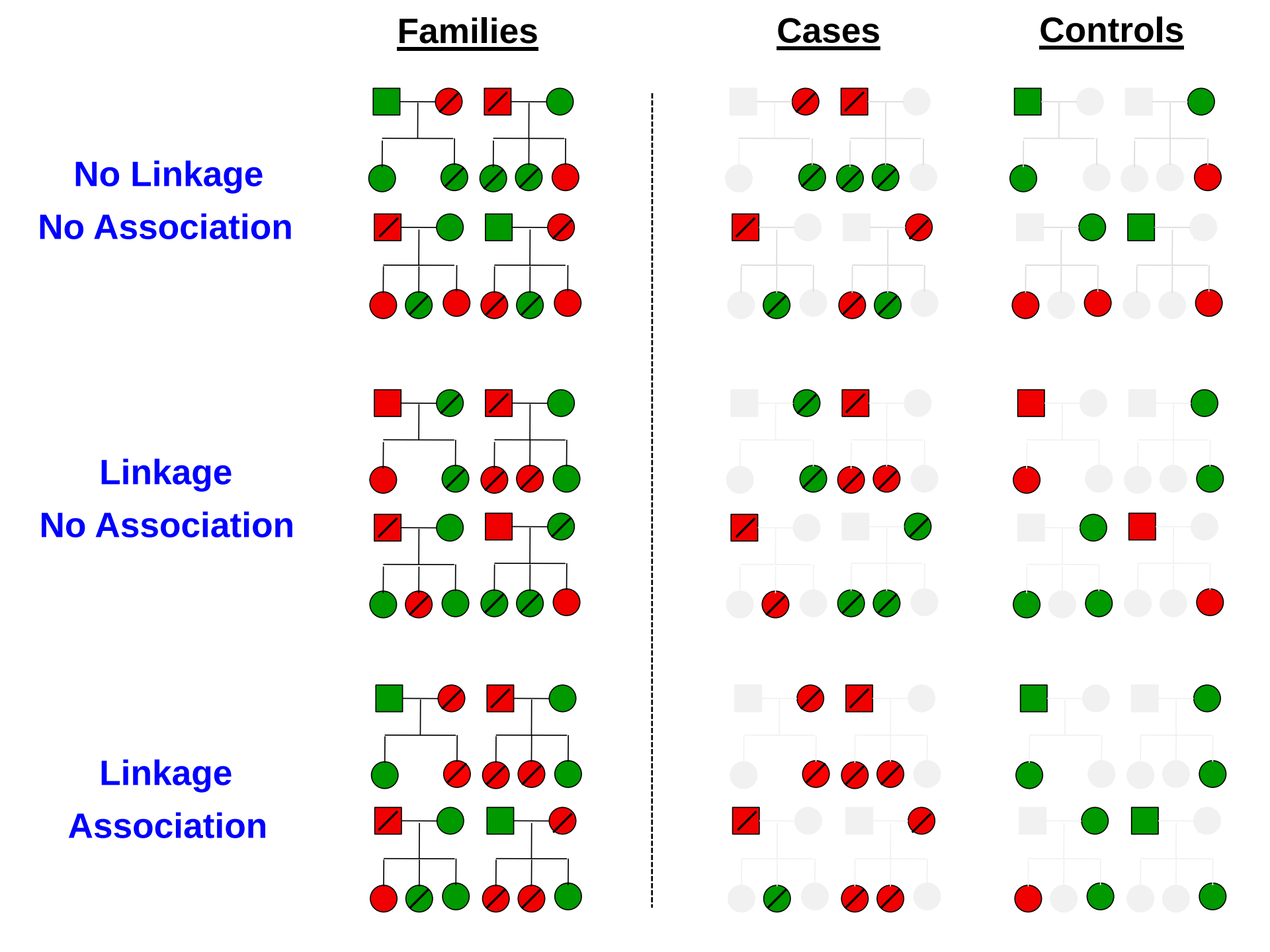
Family-specific phenomenon:
Affected individuals in a family share the same ancestral predisposing DNA segment at a given QTL

Can only detect very large effects

Association: identify a QTL that regulates the trait

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

Can detect weaker effects


Families
Cases
Controls
No Linkage
No Association
Linkage
No Association
Linkage
Association

Non-parametric linkage approach

Linkage tests co-segregation between a marker and a trait

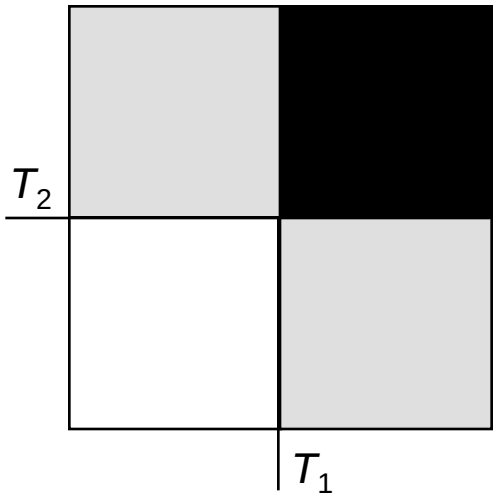
If a trait locus truly regulates the expression of a phenotype, then two relatives with similar phenotypes should have inherited *from a common ancestor* the same predisposing allele at a marker near the trait locus, and vice-versa.

Interest: correlation between phenotypic similarity and genetic similarity at a locus

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 $\begin{Bmatrix} Var(X_1) & Cov(X_1X_2) \\ Cov(X_1X_2) & Var(X_2) \end{Bmatrix}$

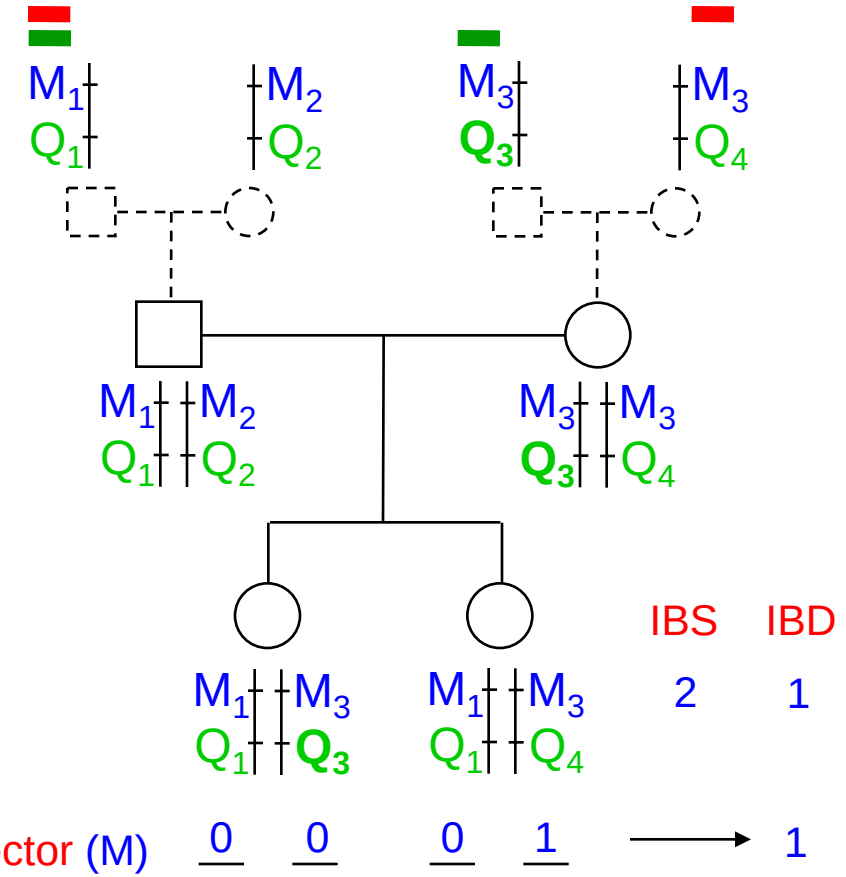
- Affection concordance



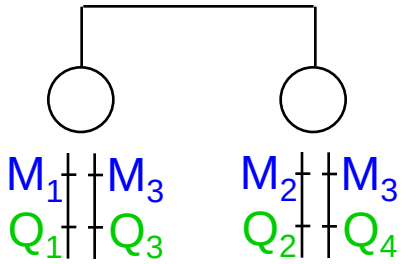
Genotypic similarity between relatives

- 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

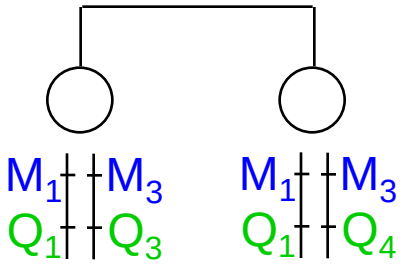
- IBD — Alleles shared Identical By Descent are a copy of the same ancestor allele

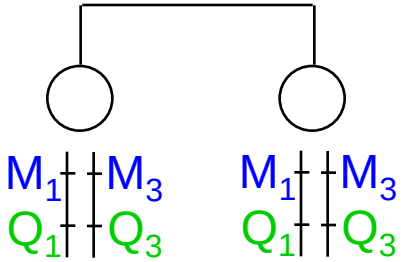


Genotypic similarity between relatives - π



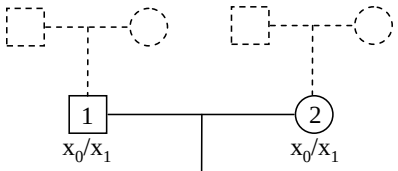
Inheritance vector (M)	Number of alleles shared IBD	Proportion of alleles shared IBD - π
0 0 1 1	0	0
0 0 0 1	1	0.5
0 0 0 0	2	1





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)

$$\hat{\pi} =$$

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

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

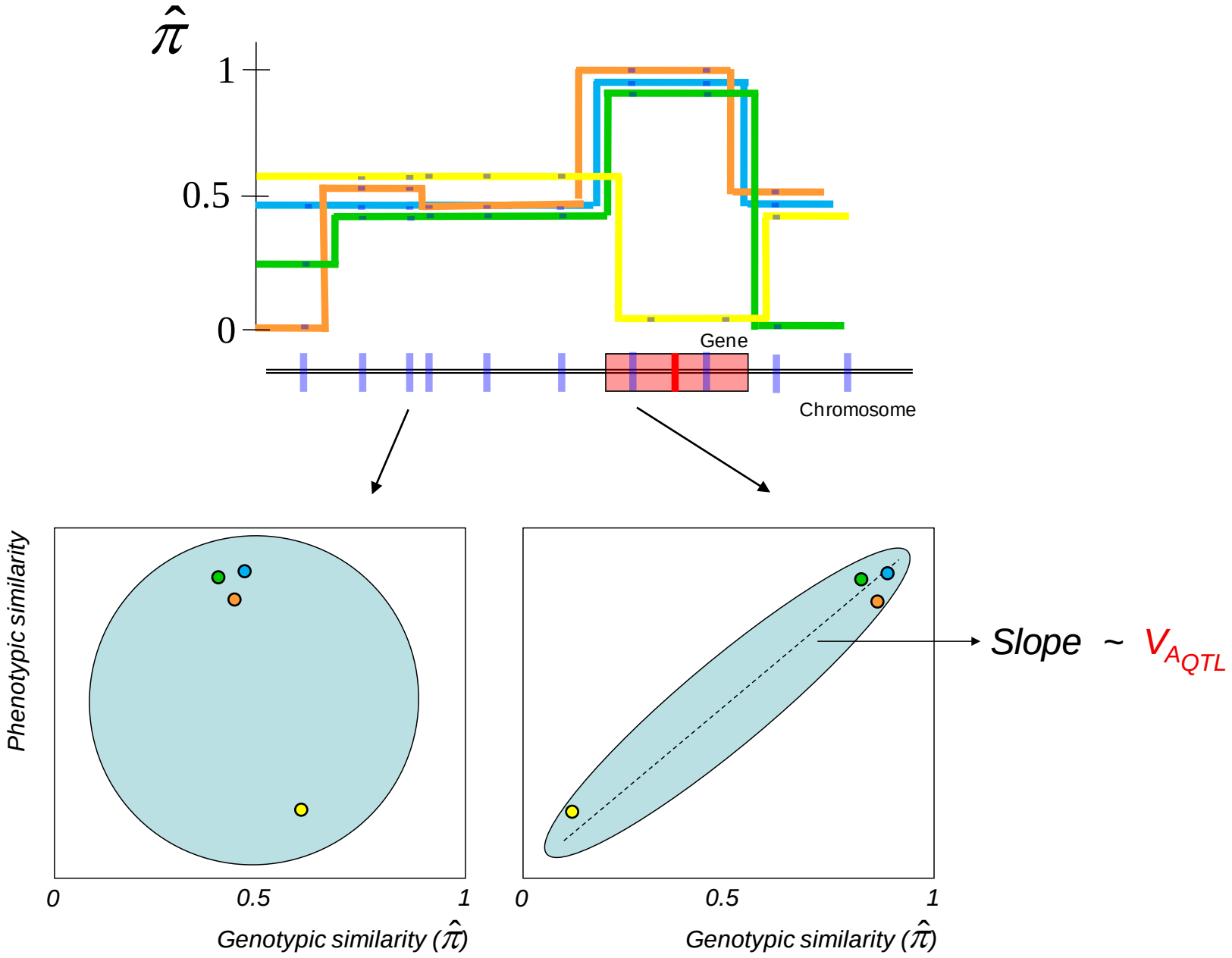
$$Cov\,(DZ) = \tfrac{1}{2}V_{A_{QTL}} + \tfrac{1}{4}V_{D_{QTL}}$$

On average!

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

For a given locus

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


$\hat{\pi}$
1
0.5
0
Gene
Chromosome
Phenotypic similarity
0
0.5
1
Genotypic similarity ($\hat{\pi}$)
0
0.5
1
Genotypic similarity ($\hat{\pi}$)
Slope ~ $V_{A_{QTL}}$

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

- Regression-based methods

  Haseman-Elston, MERLIN-regress

$$(X_1 - X_2)^2 = -2 * V_{A_{QTL}} \cdot \hat{\pi}$$

- Variance components methods

  Mx, MERLIN, SOLAR, GENEHUNTER

$$\Sigma_{jk} \begin{cases} V_{A_{QTL}} + \frac{1}{2} \cdot V_A + V_E \text{ , for j = k} \\ \hat{\pi} \cdot V_{A_{QTL}} + \frac{1}{2} \cdot V_A + V_E \text{ , for j} \neq \text{k} \end{cases}$$

Should we still use linkage analysis?

Given dense SNP data

- Rare genetic variant (not covered by the genotyping platform)
- ... or allelic heterogeneity (multiple disease variants in the same gene)
- *AND* strong effect on phenotype...

Linkage analysis can complement association and provide an additional approach to localise a disease locus (with no loss).