Quantitative traits IV

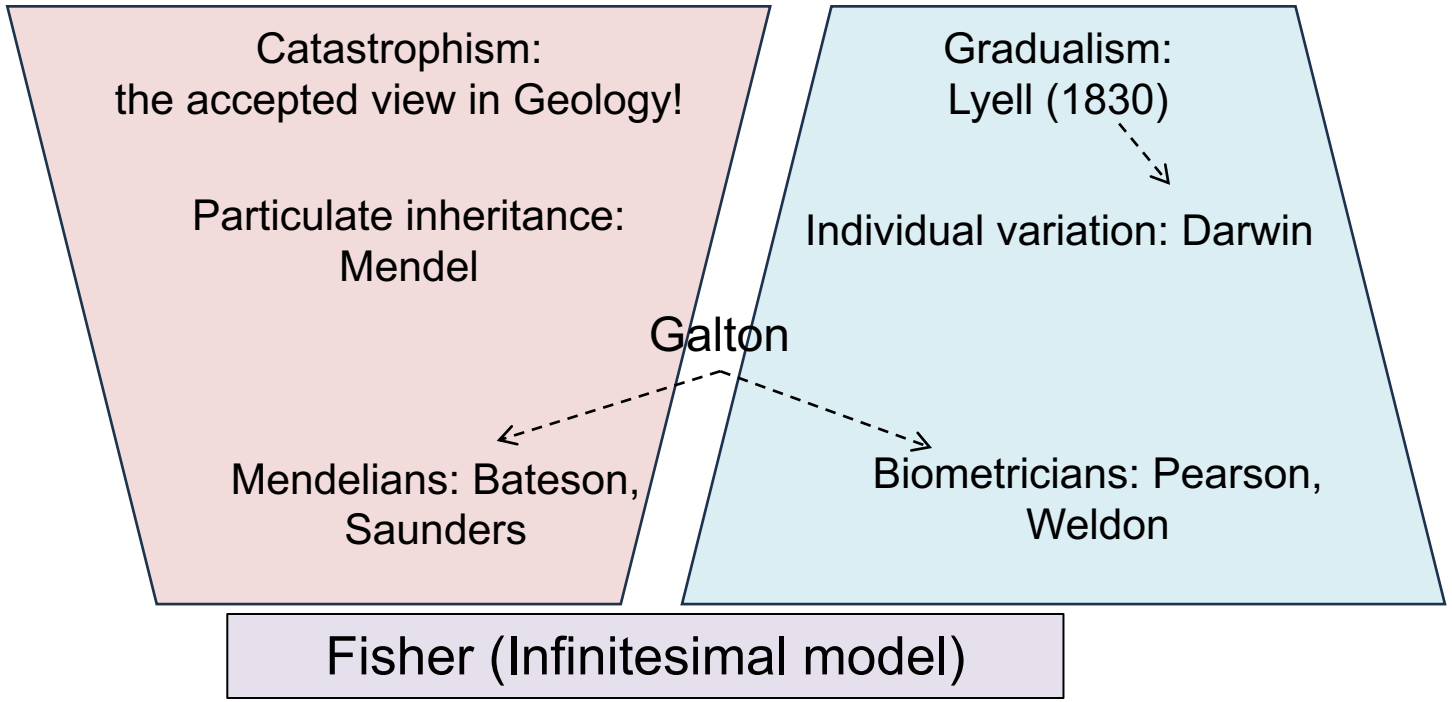
Hancock

16. April 2024

Some background

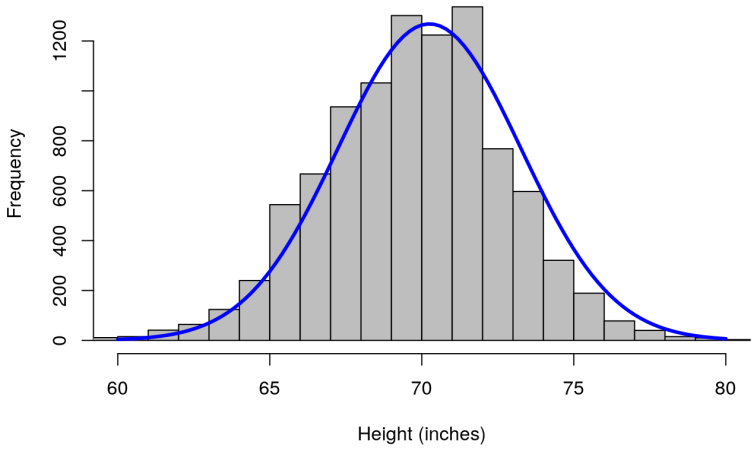
Brief history of thought on evolutionary change

Modern synthesis

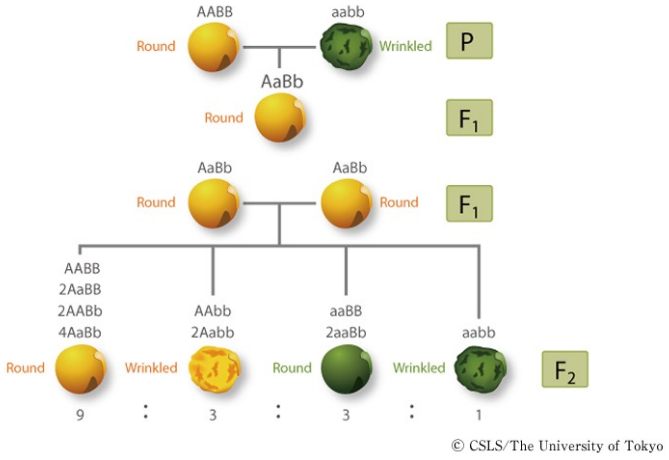


Mendelian versus quantitative traits

Human height



Mendel's peas



Continuous variation

Infinitesimal effects

=

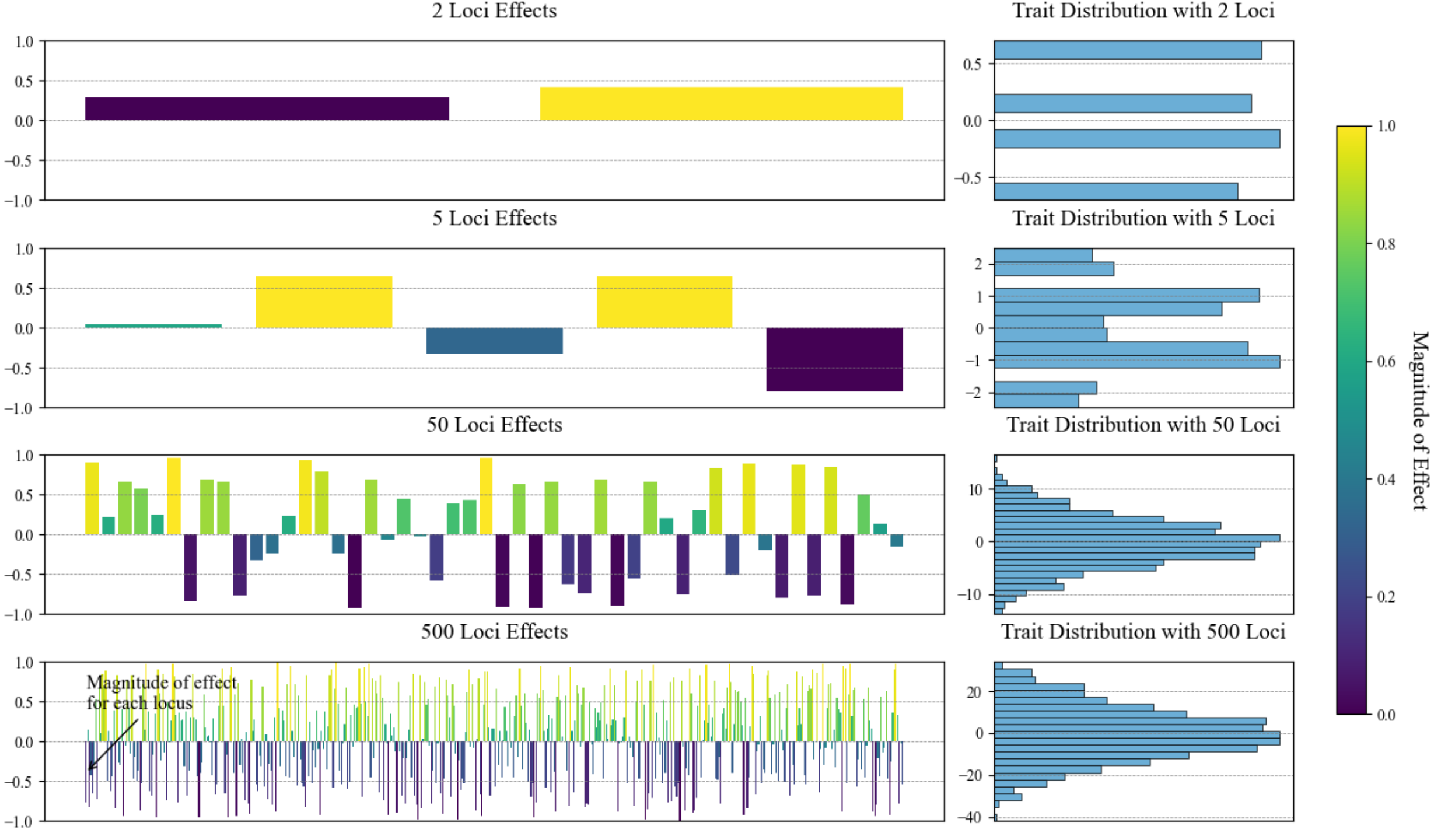
Discrete variation

Mendelian factors

Fisher united the Mendelians and the Biometricians

Fisher's infinitesimal model (1918):
The combined action of many genes, each with an infinitesimally small effect, can provide the basis for continuous variation

Many variants each with a small effect explain the pattern!!!


2 Loci Effects
Trait Distribution with 2 Loci
5 Loci Effects
Trait Distribution with 5 Loci
50 Loci Effects
Trait Distribution with 50 Loci
500 Loci Effects
Trait Distribution with 500 Loci
Magnitude of effect
for each locus
Magnitude of Effect

Basic theory

Trait values and means

$$P = G + E$$

where P is the phenotypic value, G is the genotypic value and E is the environmental deviation

P = G if the environment is held constant (this is not interaction, just the effect of the environment), so

$$G = E(P)$$

Genotypic effects

G can be broken down into additive and dominance effects:

$$G = A + D$$

Dominance effects occur when the effects at a locus are not additive

Model with a single di-allelic genetic locus

Contribution of the variant to the mean phenotypic value (M):

Genotype	**AA**	**Aa**	**aa**
Genotypic Value	a	d	-a
Frequency (in HWE)	p^2	2pq	q^2

Theoretical phenotypic range: (+a, -a)

$$M = a*p^2 + d*2pq - a*q^2$$

$$= \underbrace{a(p-q)}_{A} + \underbrace{2pqd}_{D}$$

The contribution to the phenotype distribution due to dominance depends on frequencies of the alleles, and so it can change over time (e.g., due to selection) in a population

(with multiple loci: $M = \Sigma a(p-q) + 2\Sigma dpq$)

Variance for a model with a single di-allelic locus

Contribution of the allele to the phenotypic variance (V):

Genotype	**AA**	**Aa**	**aa**
Genotypic Value	a	d	-a
Frequency (in HWE)	p^2	2pq	q^2

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

$$= \underbrace{2pq[a+(q-p)d]^2}_{V_A} + \underbrace{(2pqd)^2}_{V_D}$$

Additive effects: main effects of individual alleles

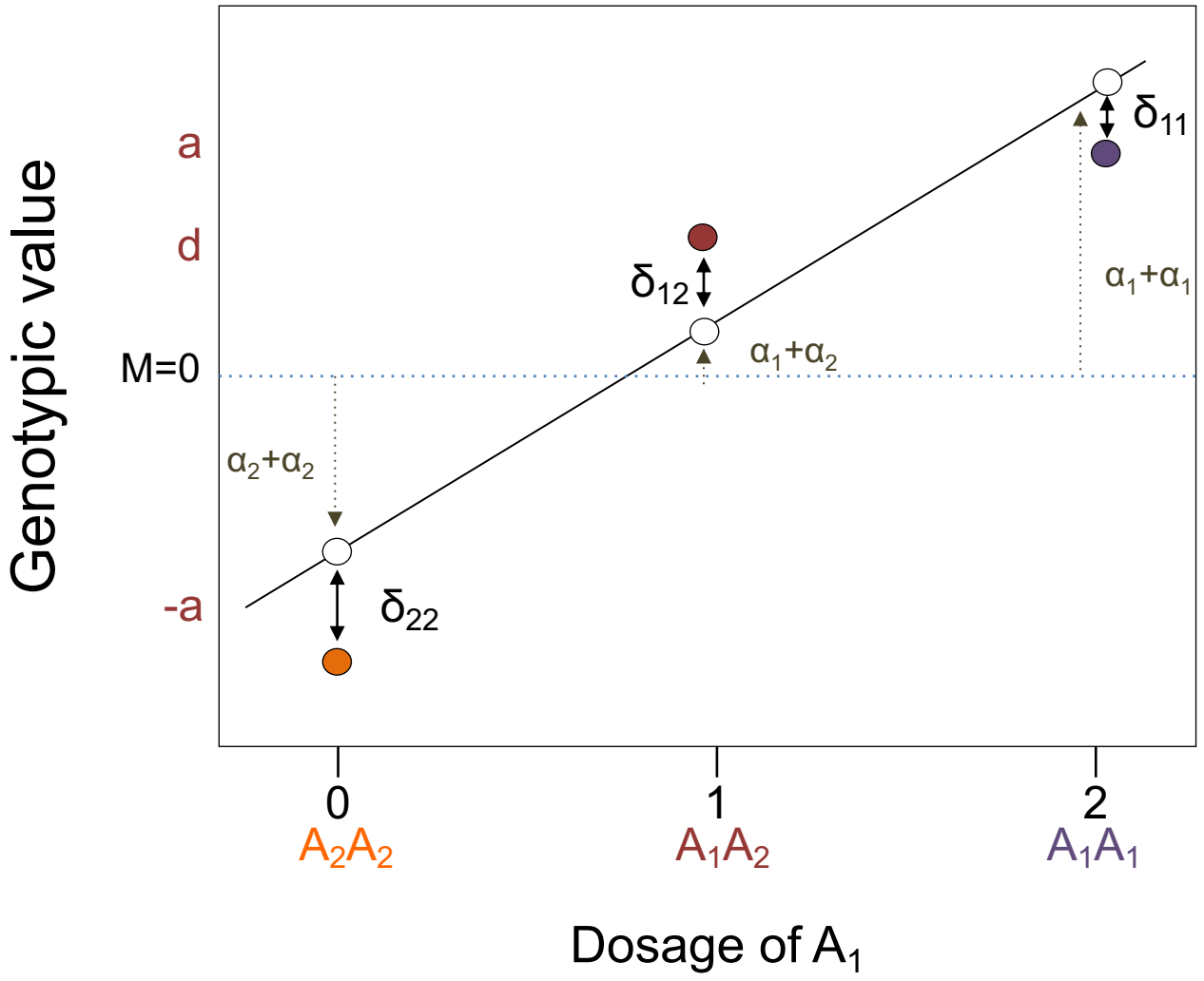
Dominance effects: non-additive effects between alleles

Model with a single di-allelic locus

The average effect for an allele (α) is the difference between the effect of that allele and the population mean, M:

$$\alpha_A = \underbrace{pa}_{AA} + \underbrace{qd}_{Aa} - \underbrace{[a(p-q) + 2dpq]}_{M} = q[a+d(q-p)]$$

Additive effects (α) and dominance deviations (δ)



M: weighted average of effects

α: average effect (relative to M)

δ: dominance deviation

Quantitative traits tend to be shaped by multiple genetic and environmental factors

- Trait variance is shaped by genetic and environmental factors:

$$V_P = V_G + V_E$$

- Trait variance may also be shaped by interactions between genes and environment

$$V_P = V_G + V_E + V_{GxE}$$

- Or by interactions between genetic variants

$$V_P = V_{G1} + V_{G2} + V_{G1}xV_{G2}$$

Mapping quantitative traits

Linkage mapping is family-based mapping

- In sampled families, e.g., large pedigrees or parent-offspring trios
- In constructed ‘families’ or populations (model organisms)

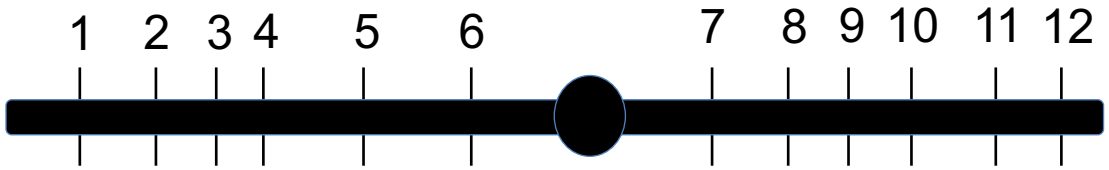
Linkage mapping

Identify the marker that is most tightly linked to the trait of interest

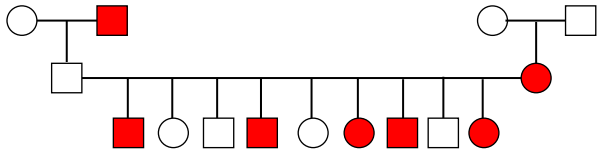
… for each marker, calculate the likelihood of the data under a linkage model relative to a model in which the trait is unlinked from the marker

Linkage Mapping

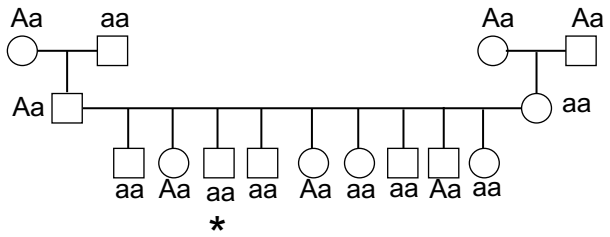
Step 1: Produce a map relating locations of markers across each chromosome



Step 2: Determine phenotypes across one or more pedigrees



Step 3: Determine marker status across the same pedigree(s)



Step 4: Estimate recombination frequencies between the trait and each chromosomal locus to identify ‘linked’ loci

Estimate recombination frequencies between marker and trait to identify linked loci

In practice: Calculate the LOD score (log of odds ratio) at each marker for a variety of recombination fractions

$$\text{LOD} = \log_{10} \frac{likelihood\ of\ observed\ data\ with\ r = r'}{likelihood\ of\ observed\ data\ with\ r = 0.5}$$

$$= \log_{10} \frac{(1-\theta)^{NR} * \theta^{R}}{0.5^{(NR+R)}}$$

Iteratively calculate for different genetic distances to find the best fit

where NR denotes number of non-recombinant offspring and R denotes number of recombinants, and θ is the recombination fraction (R/(NR+R))

*Under a model of no linkage θ=0.5

LOD of 3 (1000:1 odds of linkage) often used as a cutoff for evidence of linkage

Building a physical map

Identify landmarks and determine their locations in the genome

Low Resolution	Cytogenetic Map *Based on banding patterns on stained chromosomes*
	Radiation Hybrid map *Markers based on patterns of X-ray breakage along the chromosome*
	Contig Map *Order of DNA fragments across the genome*
	Restriction Fragment Map *Landmarks based on restriction sites across the genome*
	STS Map *Based on 200-500 bp PCR amplified sequences across the genome*
High Resolution	Sequence-based map *Uses complete sequence data – now possible for diverse species due to 'NGS' methods*

Proof of principle: RFLPs open the path for mapping ‘densely’ within chromosomes!

Am J Hum Genet 32:314–331, 1980

Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms

DAVID BOTSTEIN,[1] RAYMOND L. WHITE,[2] MARK SKOLNICK,[3] AND RONALD W. DAVIS[4]

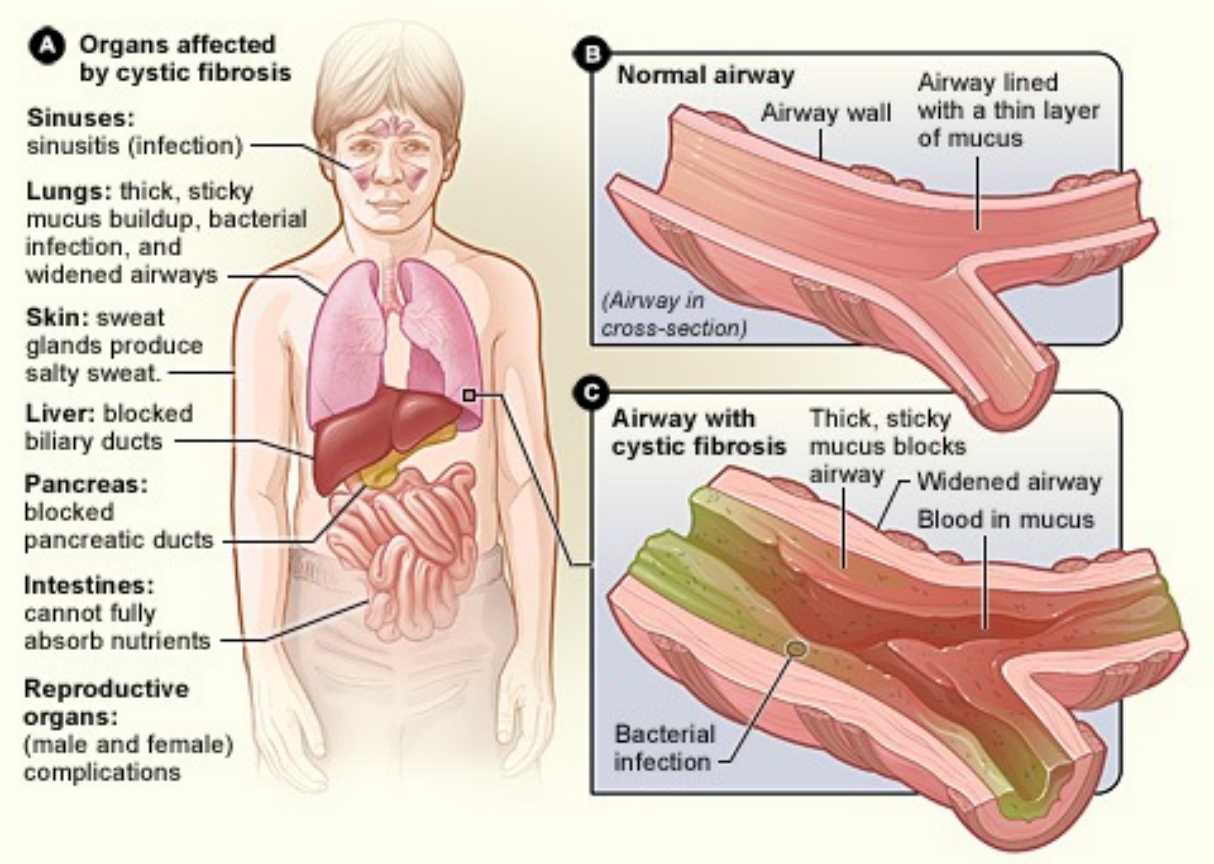
SUMMARY

We describe a new basis for the construction of a genetic linkage map of the human genome. The basic principle of the mapping scheme is to develop, by recombinant DNA techniques, random single-copy DNA probes capable of detecting DNA sequence polymorphisms, when hybridized to restriction digests of an individual's DNA. Each of these probes will define a locus. Loci can be expanded or contracted to include more or less polymorphism by further application of recombinant DNA technology. Suitably polymorphic loci can be tested for linkage relationships in human pedigrees by established methods; and loci can be arranged into linkage groups to form a true genetic map of ''DNA marker loci.'' Pedigrees in which inherited traits are known to be segregating can then be analyzed, making possible the mapping of the gene(s) responsible for the trait with respect to the DNA marker loci, without requiring direct access to a specified gene's DNA. For inherited diseases mapped in this way, linked DNA marker loci can be used predictively for genetic counseling.

A mapping example:

Cystic Fibrosis

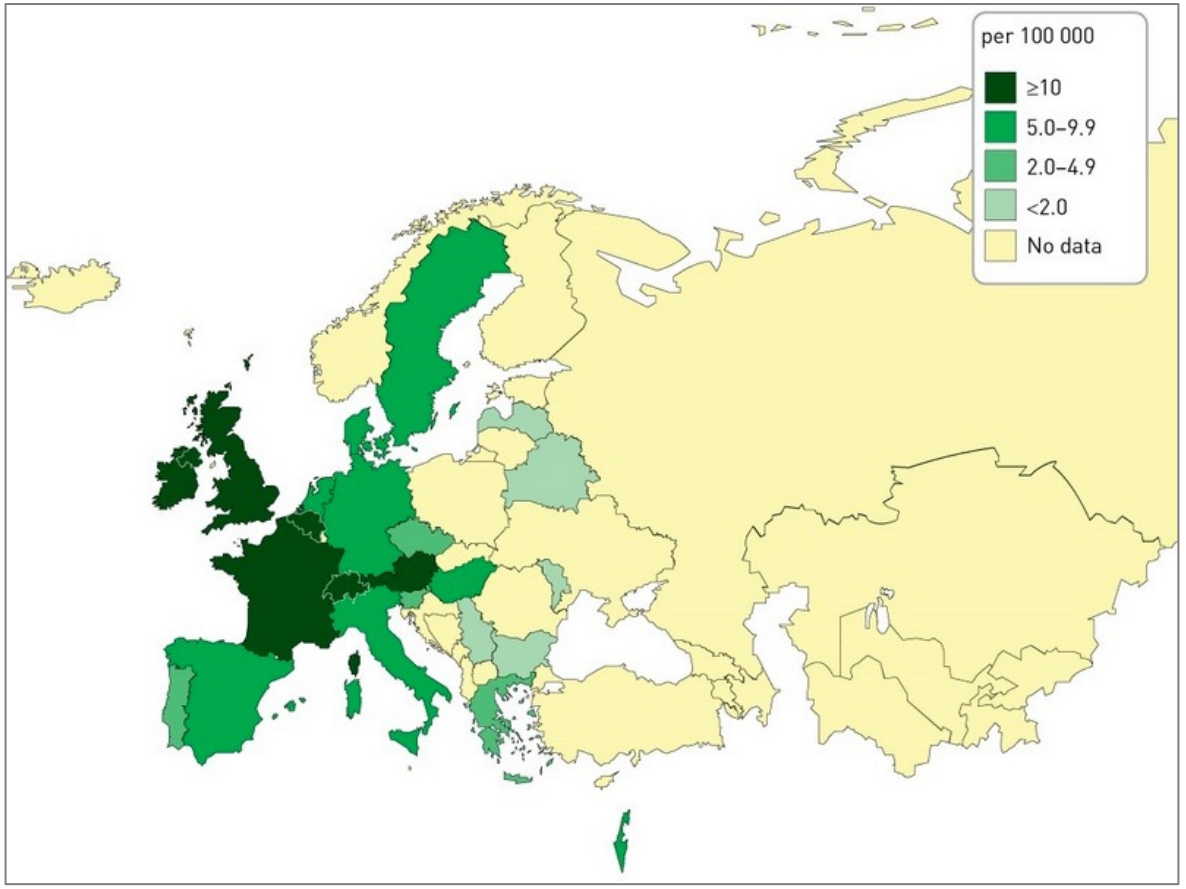
- Causes increased susceptibility to lung infections, high salt excretion in sweat ducts
- Frequency of around 1/2000 births in Europeans and carrier frequency around 5%
- Autosomal recessive inheritance pattern

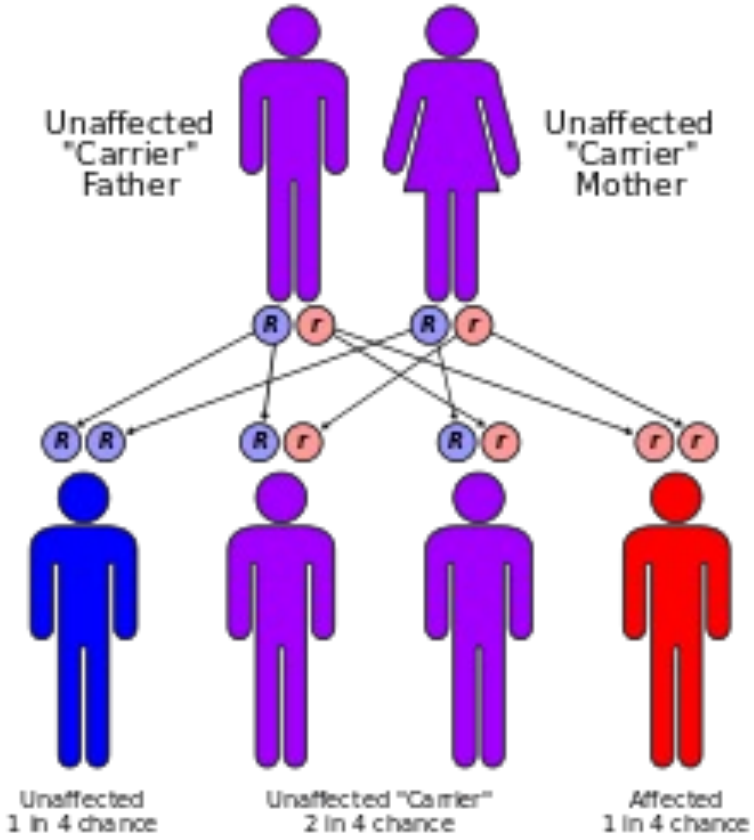


A mapping example:

Cystic Fibrosis

Geographic distribution of the cystic fibrosis disease





A mapping example:
Cystic Fibrosis

A closely linked genetic marker for cystic fibrosis

Ray White*, Scott Woodward*, Mark Leppert*, Peter O'Connell*, Mark Hoff*, John Herbst†, Jean-Marc Lalouel*, Michael Dean‡ & George Vande Woude‡

* Howard Hughes Medical Institute and the Cystic Fibrosis Research Center, Department of Human Genetics and † Department of Pediatrics, University of Utah College of Medicine, Salt Lake City, Utah 84132, USA
‡ Litton Bionetics, Inc., Basic Research Program, NCI-Frederick Cancer Research Facility, Frederick, Maryland 21701, USA

Localization of cystic fibrosis locus to human chromosome 7cen–q22

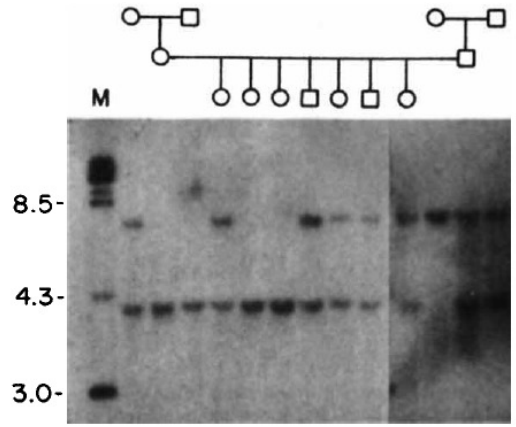
Brandon J. Wainwright*, Peter J. Scambler*, Jorg Schmidtke†, Eila A. Watson*, Hai-Yang Law*, Martin Farrall*, Howard J. Cooke‡, Hans Eiberg§ & Robert Williamson*

* Department of Biochemistry, St Mary's Hospital Medical School, University of London, London W2 1PG, UK
† Institut fur Humangenetik der Universitat, Gosslerstrasse 12d, D-3400 Gottingen, FRG
‡ MRC Mammalian Genome Unit, West Mains Road, Edinburgh EH9 3JJ, UK
§ Institute of Medical Genetics, Panum Institute, Copenhagen, Denmark

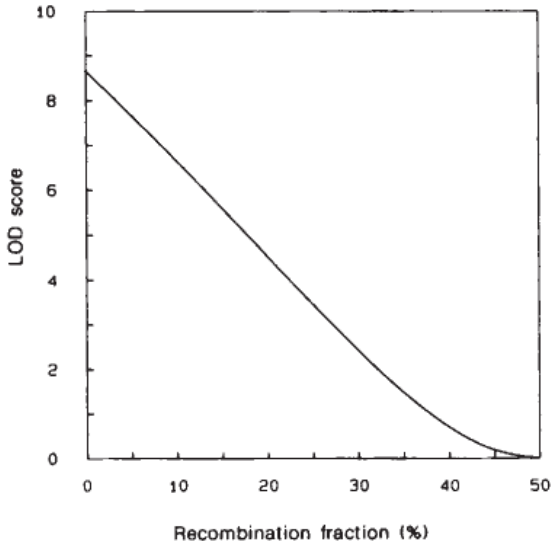
A mapping example:

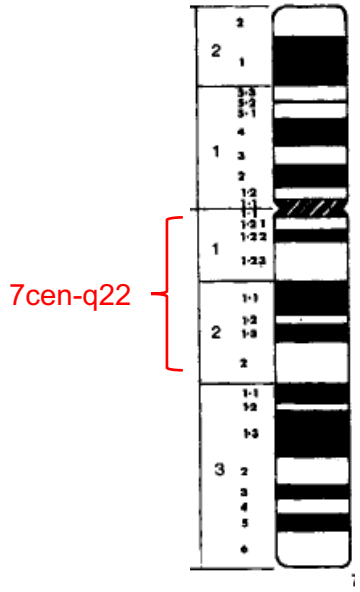
Cystic Fibrosis

Use familial segregation patterns of the trait and markers



…to scan the genome to identify markers that are linked to the trait

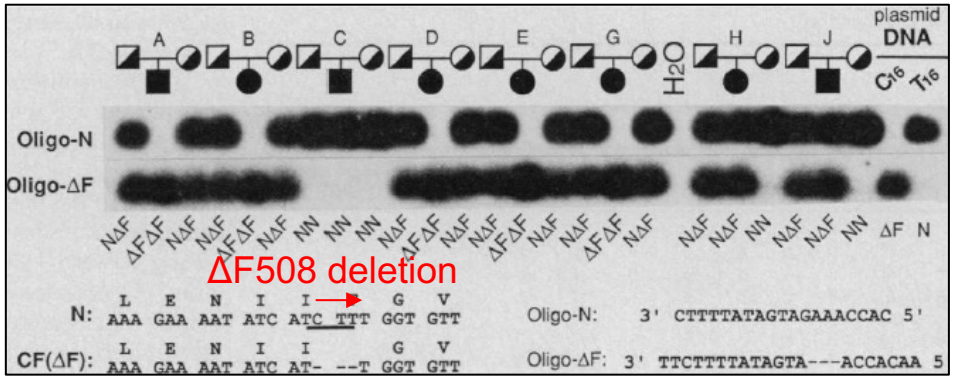




A mapping example:
Cystic Fibrosis

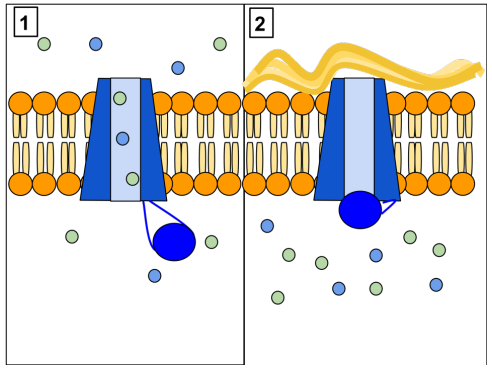
Fine-mapping of the region pinpointed a 3 bp deletion (ΔF508) in a transmembrane conductance regulator (CFTR)

ΔF508 is segregating with the disease in around 70% of CF families



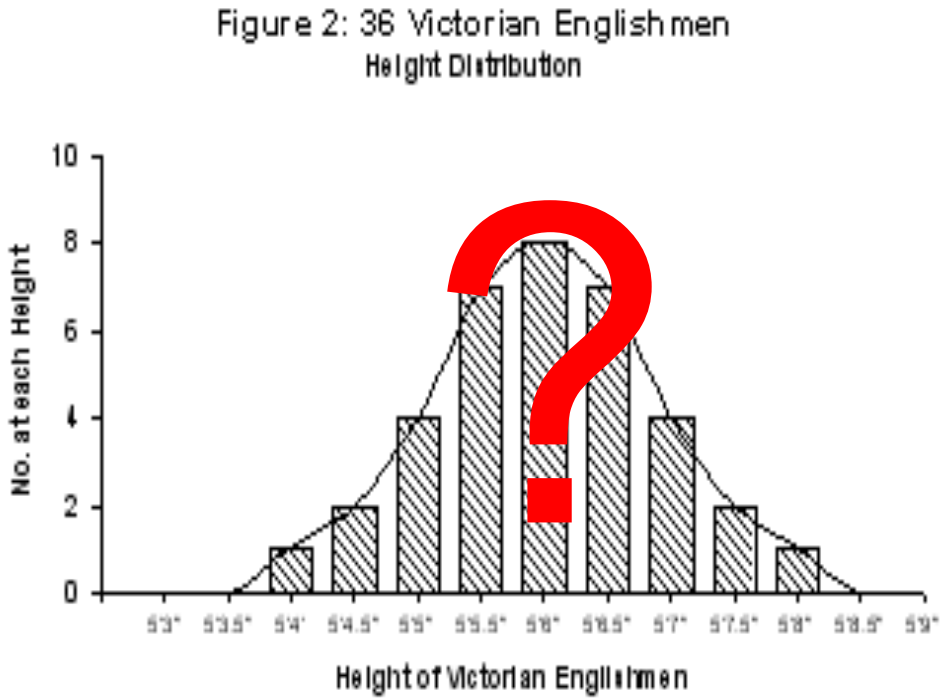
Karem et al., Science 1989

The CFTR protein is a channel protein



When CFTR does not function properly, Cl^- ions build up in cells

What about more complex, continuous traits?



An example: identifying the ‘Mendelian factors’ that underlie quantitative traits in tomato

Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms

Andrew H. Paterson*, Eric S. Lander†‡, John D. Hewitt§, Susan Peterson*, Stephen E. Lincoln† & Steven D. Tanksley*

* Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853, USA
† Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA
‡ Harvard University, Cambridge, Massachusetts 02138, USA
§ Department of Vegetable Crops, University of California, Davis, California 95616, USA

The conflict between the Mendelian theory of particulate inheritance[1] and the observation of continuous variation for most traits in nature was resolved in the early 1900s by the concept that quantitative traits can result from segregation of multiple genes, modified by environmental effects[2–5]. Although pioneering experiments[6–9] showed that linkage could occasionally be detected to such quantitative trait loci (QTLs), accurate and systematic mapping of QTLs has not been possible because the inheritance of an entire genome could not be studied with genetic markers[7]. The use of restriction fragment length polymorphisms[10] (RFLPs) has made such investigations possible, at least in principle. Here, we report the first use of a complete RFLP linkage map to resolve quantitative traits into discrete Mendelian factors, in an interspecific back-cross of tomato. Applying new analytical methods, we mapped at least six QTLs controlling fruit mass, four QTLs for the concentration of soluble solids and five QTLs for fruit *p*H. This approach is broadly applicable to the genetic dissection of quantitative inheritance of physiological, morphological and behavioural traits in any higher plant or animal.

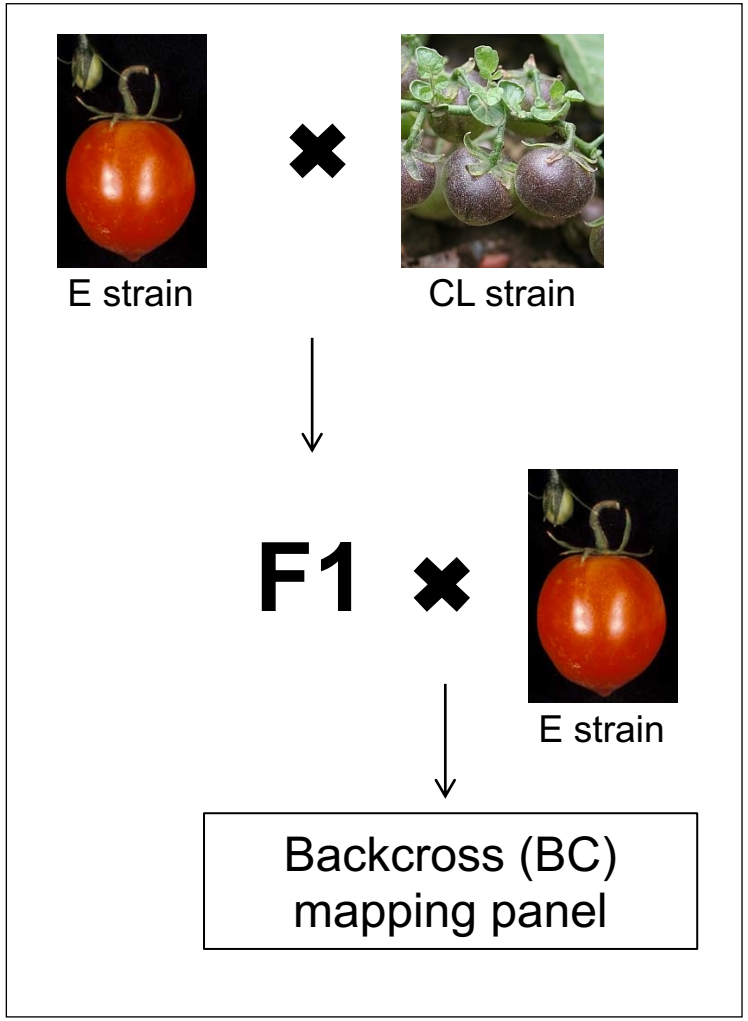
importance, because they jointly determine the yield of tomato paste. In addition, the strains are known to be polymorphic for genes affecting fruit *p*H, which is important for the optimal preservation of tomato products[13]; the difference in *p*H between the parental strains is, however, small.

A total of 237 back-cross plants, with E as the recurrent parent, were grown in the field at Davis, California. Between five and 20 fruit from each plant were assayed[13] for fruit mass, soluble-solids concentration (°Brix; see Fig. 1 legend for definition) and *p*H, each of which showed continuous variation (Fig. 1). Soluble-solids concentration correlated negatively with fruit mass ($r = -0.42$) and positively with *p*H ($r = +0.33$).

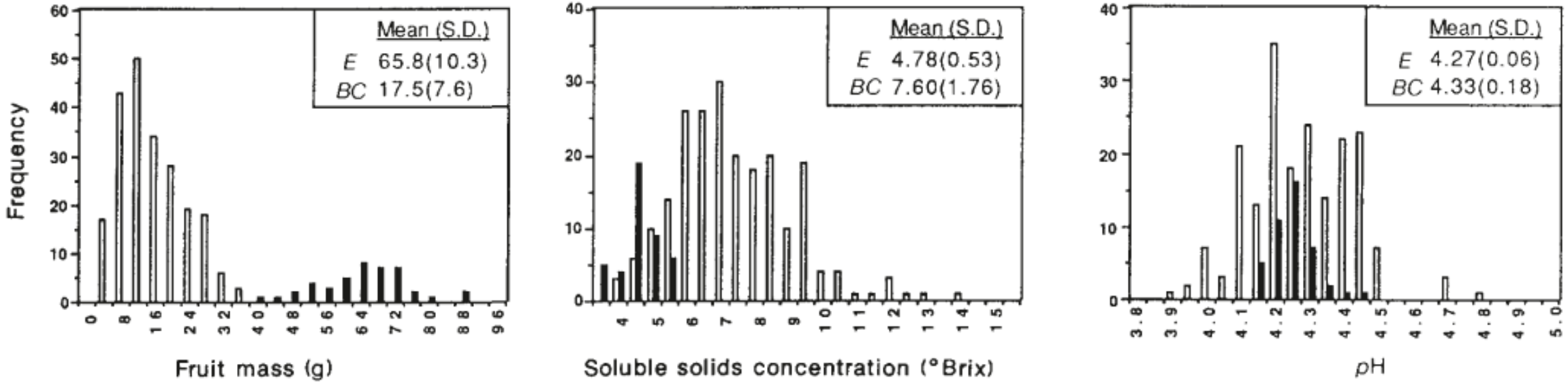
We had previously constructed a genetic linkage map of tomato[14] with over 300 RFLPs and 20 isozyme markers, by analysing 46 F_2 individuals derived from *L. esculentum* cv. VF36 × *L. pennellii* accession LA716 (E × P). The map is essentially complete: it has linkage groups covering all 12 tomato chromosomes with an average spacing of 5 cM between markers (1 cM is the distance along the chromosome which gives a recombination frequency of one per cent). For QTL mapping, we selected a subset of markers spaced at approximately 20 cM intervals and displaying polymorphism between the E and CL strains. These included 63 RFLPs and five isozyme markers. In addition, the E and CL strains differ in two easily-scored, simply-inherited morphological traits: determinacy (described below) and uniform ripening, controlled by the *sp* and *u* genes, respectively. Although a few distal regions did not contain appropriate markers, we estimate that about 95% of the tomato genome was detectably linked to the markers used.

These 70 genetic markers were scored for each of the 237 E × CL back-cross progeny (as described in ref. 13), and a linkage map was constructed *de novo* using MAPMAKER[15]. The map covers all 12 chromosomes with an average spacing of 14.3 cM. Although the linear order of markers inferred from the E × CL cross essentially agreed with that inferred from the E × P cross (but see Fig. 3 legend), genetic distances differed markedly in certain intervals (for example, 51 cM in E × P and 11 cM in E × CL, for the distance between the 45S ribosomal

Resolution of tomato quantitative traits into ‘Mendelian factors’

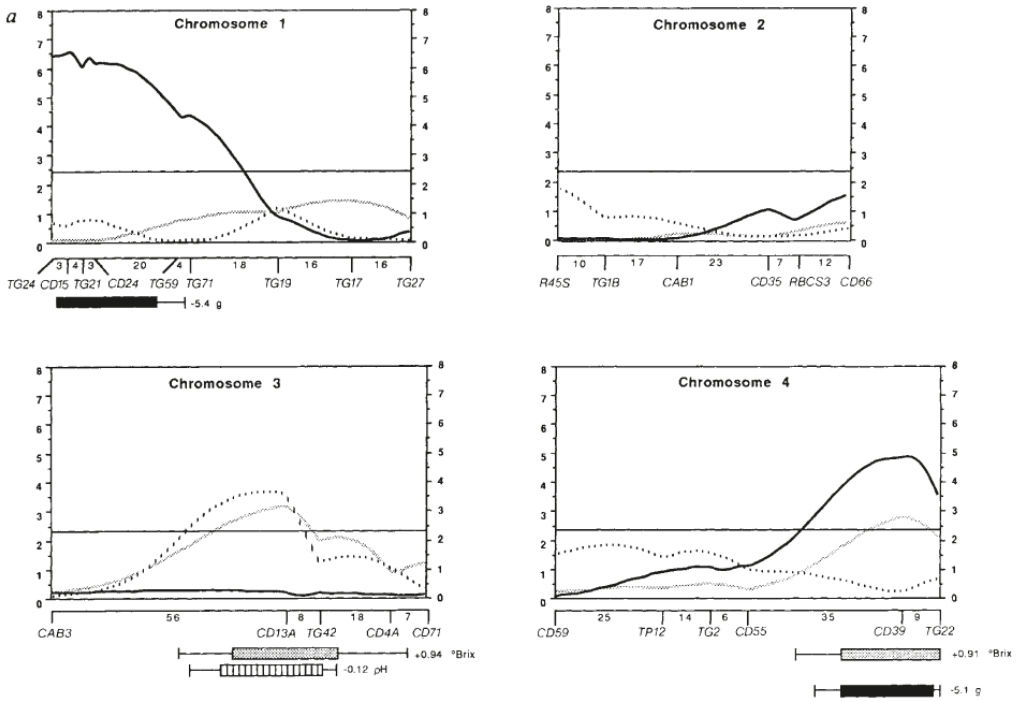


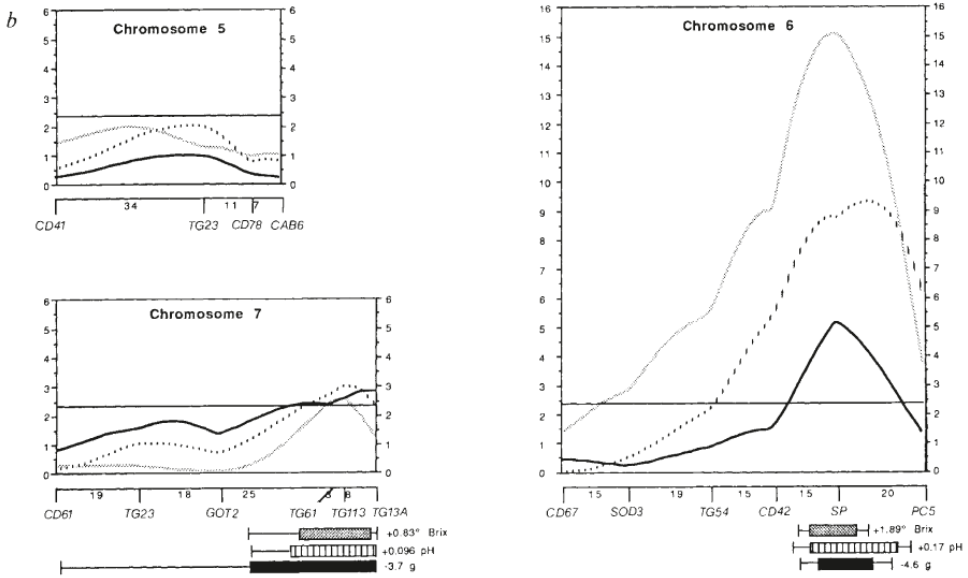
Distribution of traits in the E parental strain and backcrossed progeny

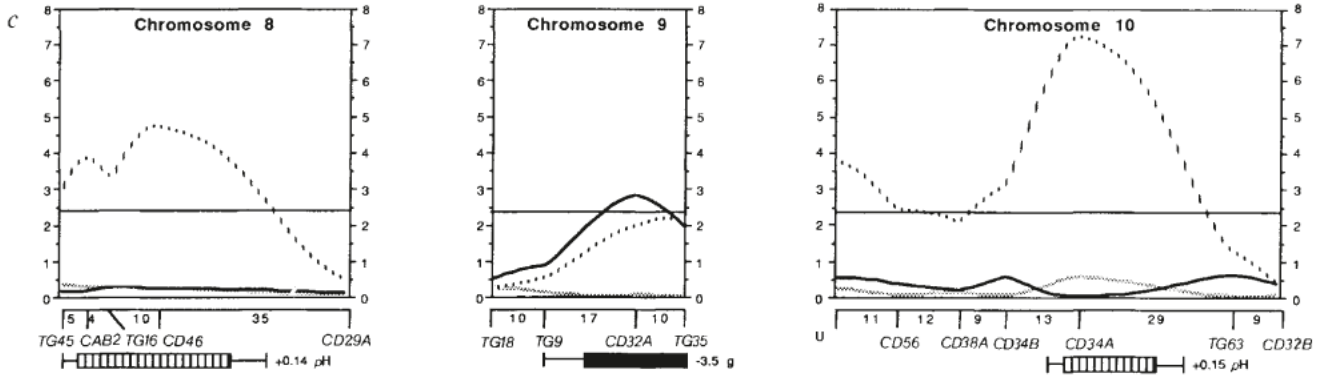


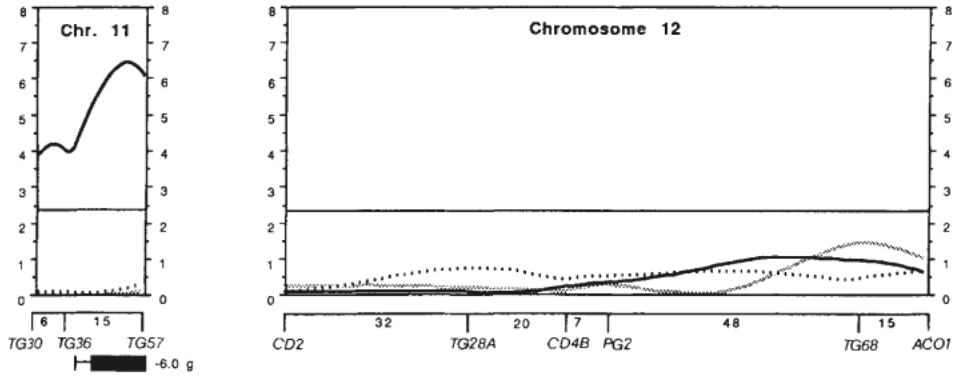
Filled bars represent the E parental strain, open bars represent the backcross (BC) progeny

Loci throughout the genome contribute to trait variation



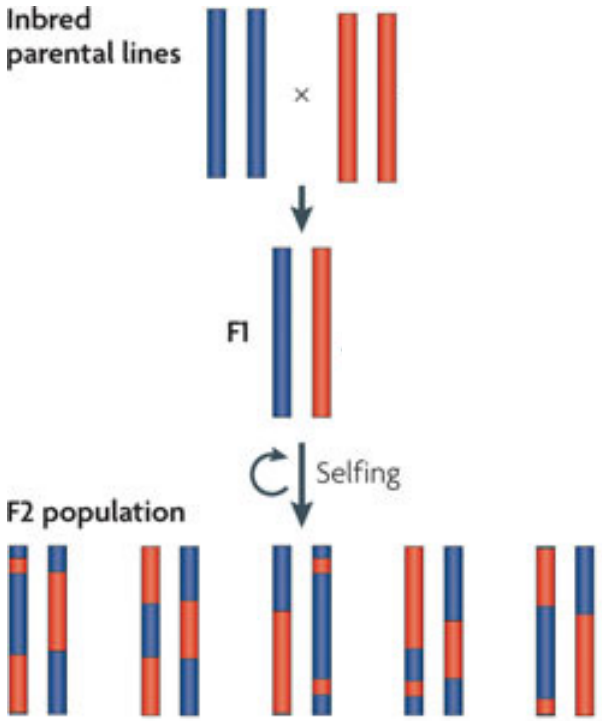






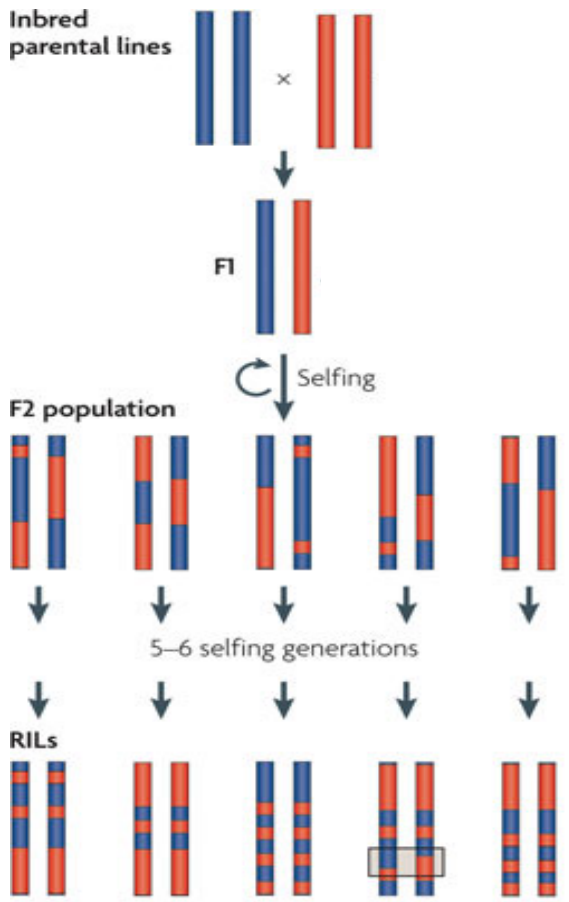
Types of mapping populations in model organisms

Mapping populations: F2



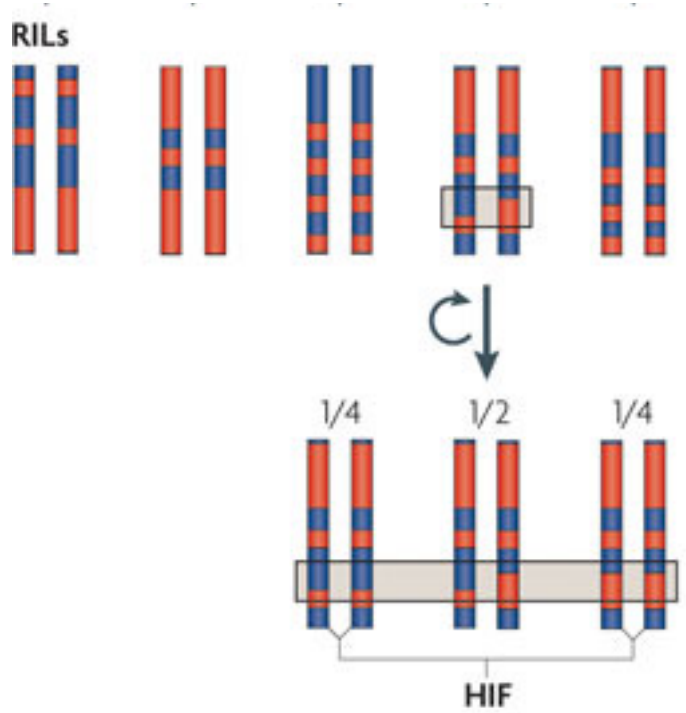
Bergelson and Roux 2010

Recombinant inbred lines (RILs)



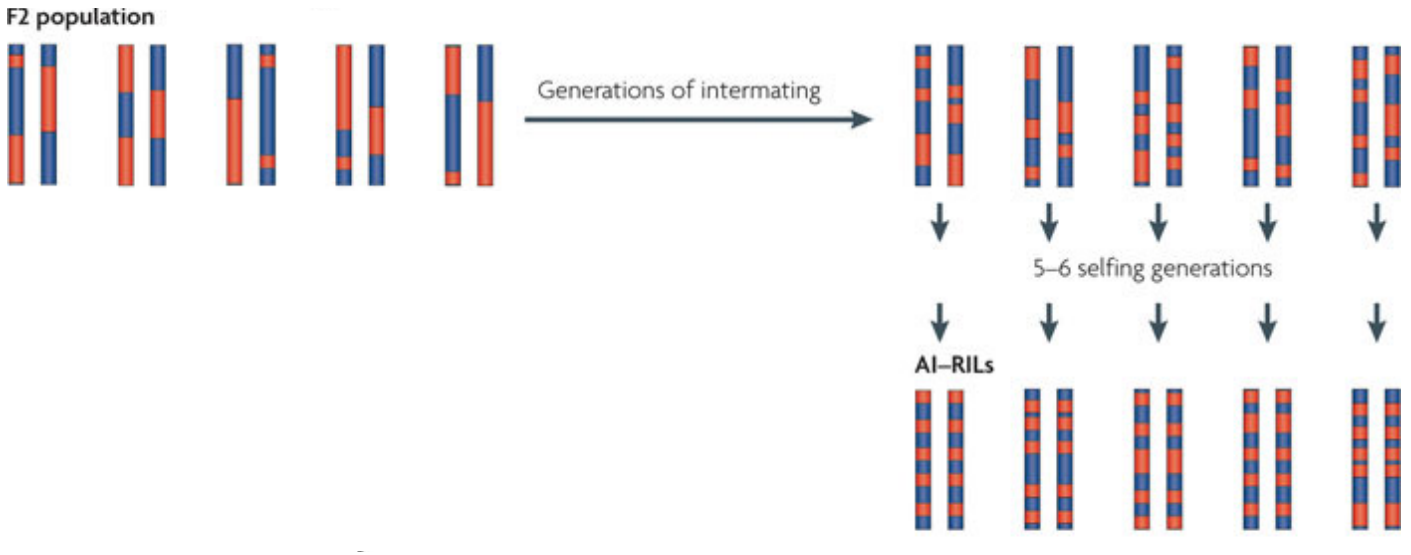
Bergelson and Roux 2010

Heterogeneous inbred families

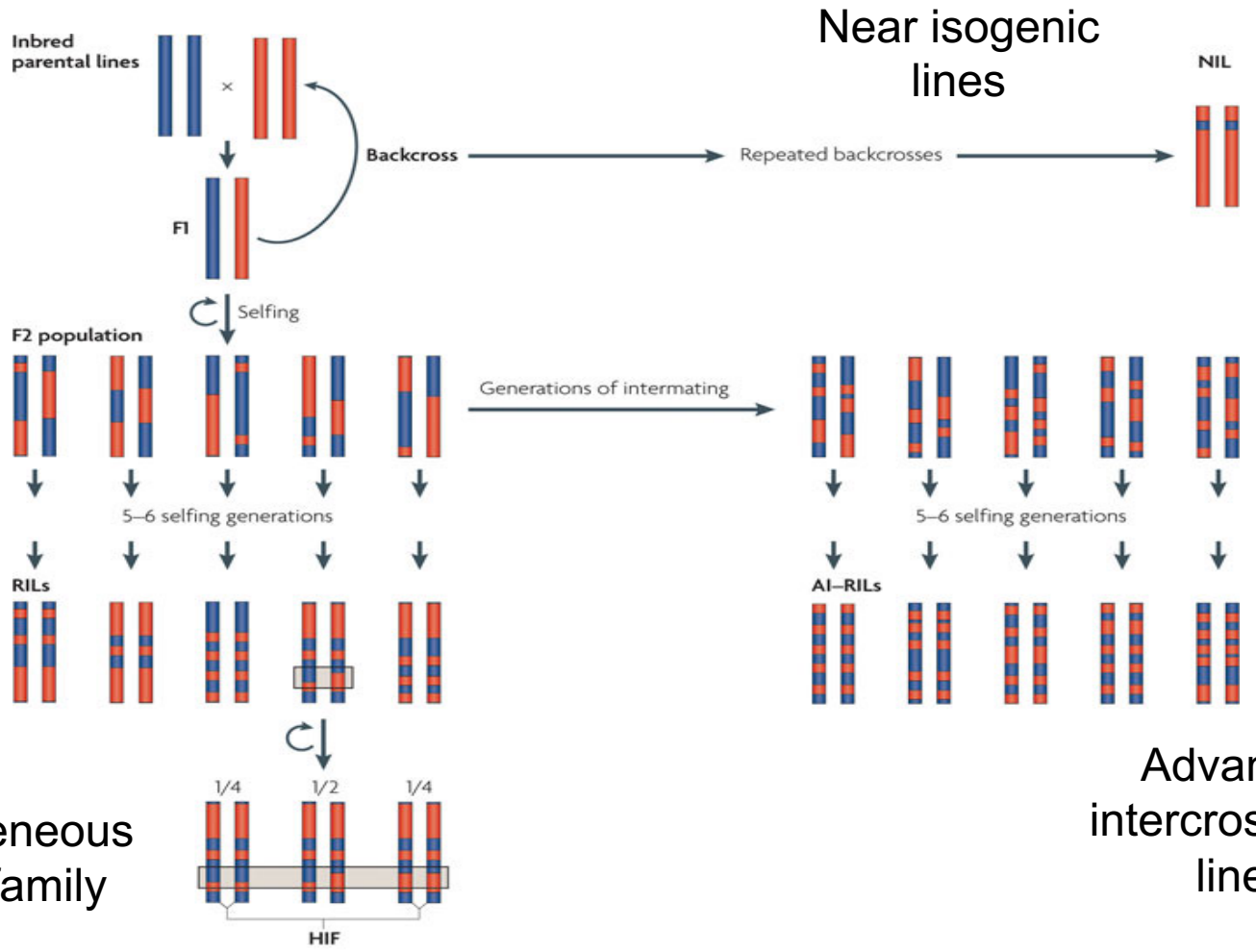


Backcross to get a population that has heterogeneity only in a small region

Advanced intercross RILs



Overview of mapping populations



Near isogenic lines

F2s

Recombinant inbred lines

Heterogeneous inbred family

Advanced intercross RILs lines

Bergelson and Roux 2010

Intercross populations

1) Start with a set of diverse strains

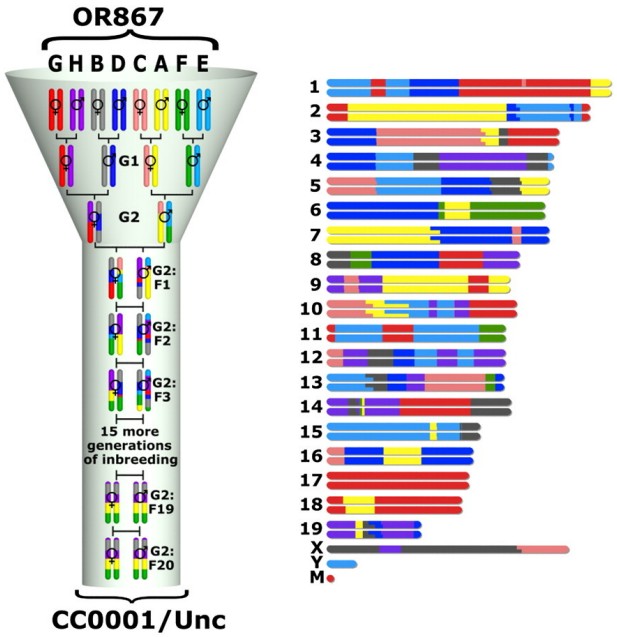
2) Cross them

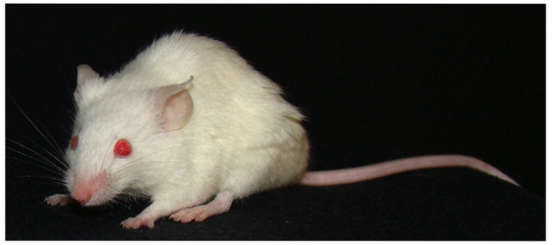
3) Inbreed for several generations

Benefits:

- increase genetic diversity
- Increase recombination events

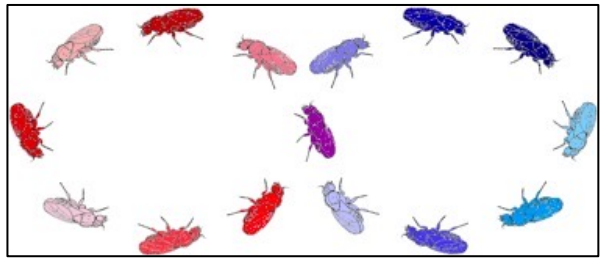
Collaborative cross (mice)

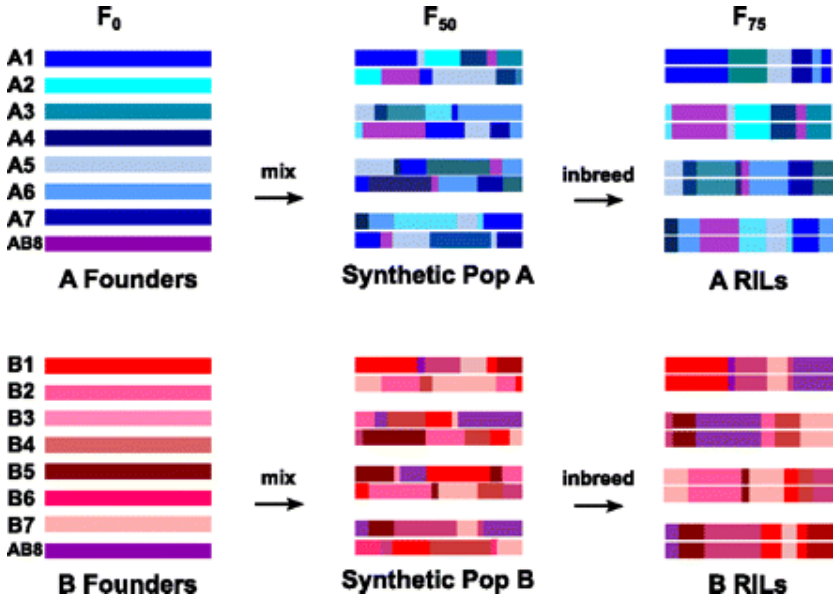




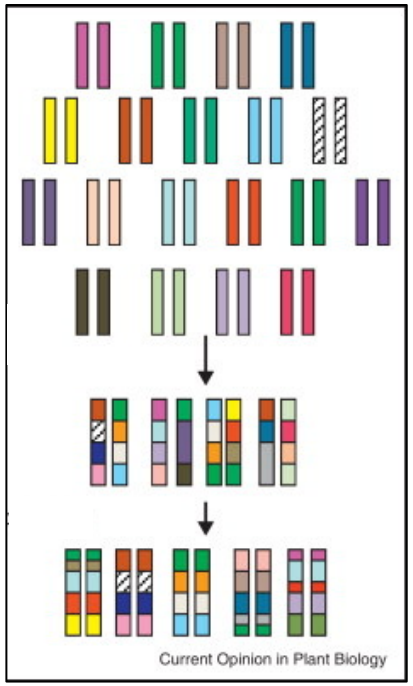
Other model intercross populations

Drosophila Synthetic Population Resource





Multiparent Advanced Genetic InterCross (MAGIC)



Kover and Mott 2012

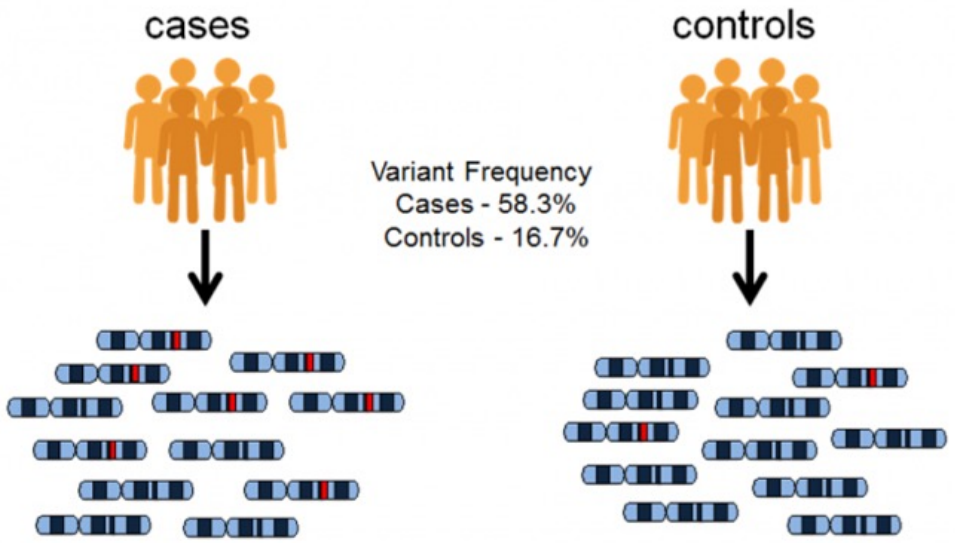
Comparison of recombinant mapping populations

Mapping material	Advantages	Drawbacks	Time (generations)	Refs
Backcross	• Detecting genetic basis of heterosis*	• Low mapping resolution • Limited genetic diversity	2	149,150
F2 population	• Estimation of QTL dominance	• Genotyping individuals for each phenotyping experiment • Limited genetic diversity	2	52,151
RILs	• Genotyped once • Unlimited replicates	• Limited genetic diversity	7–8	33,58
AI-RILs	• High-resolution mapping • Genotyped once • Unlimited replicates	• Limited genetic diversity	10	152,153
MAGIC lines	• High-resolution mapping (up to 300 kb) • Increased genetic diversity • Genotyped once • Unlimited replicates	• Genetic and allelic heterogeneity	10	31
NILs	• Single introgression segment in homogeneous genetic background • Increased power to detect small-effect QTL • Unlimited replicates	• Time consuming: size of the introgression segment will depend on the number of backcross generations • Limited genetic diversity	>6	36
HIFs	• Single introgression segment in heterogeneous genetic background • Increased power to detect small-effect QTLs • Increased power to detect epistasis • Unlimited replicates • The same genomic region covered by independent HIFs	• Limited genetic diversity	9–10	35,154

*Heterosis is the equivalent to hybrid vigour; superiority in one or more phenotypes of the hybrid individual over the parents. AI-RILs, advanced intercross–recombinant inbred lines; HIF, heterogeneous inbred family; MAGIC lines, multiparent advanced generation intercross lines; NIL, near-isogenic line; QTL, quantitative trait locus; RILs, recombinant inbred lines.

Bergelson and Roux 2010

Mapping in natural populations



- Variation in natural populations can be used for identifying loci statistically associated with a trait
- A common study design in humans is to compare allele frequencies in cases and controls

Advantages

- Can contain high phenotypic variation
- Reduced LD relative to families because lots of time for recombination to occur

Disadvantages

- Unbalanced design
- Confounding with environment and population structure

Confounding effects

- Effects of environment
- Effects of population history/population structure

Uncontrolled environmental variance can impact trait mapping results

- In model systems and breeding populations, researchers aim to control the environment and measure traits in a common environment, but it can be difficult to perfectly control environmental variance
- Statistical techniques (replicates and blocking) can help to measure environmental variance and control for it
- In human populations, environmental variance is a much bigger problem. Here, family-based studies and questionaires can help.

Population stratification

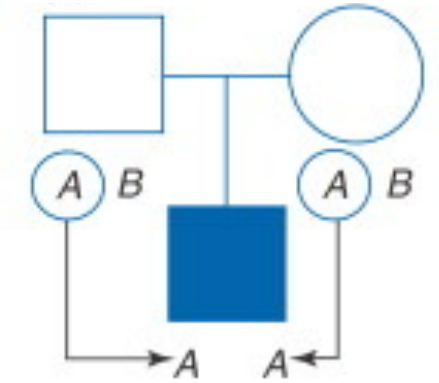
- Confounding due to population stratification can cause false positives in trait mapping
- For an extreme example, you can imagine how this problem could arise for a trait with no genetic basis (e.g., a learned cultural behavior) that co-segregates with other traits (and genetic loci)
- Or a set of traits that are genetic that co-segregate (e.g., multiple traits that are responses to the same selection pressure)
- When traits are correlated due to genetic or non-genetic causes, population stratification can be an issue
- Or when traits are correlated with population history then many loci across the genome may appear to be linked to the trait

Transmission disequilibrium test (TDT): an approach to deal with confounders in humans

- The TDT test is a family-based association test
- It tests for the presence of linkage between a marker and a trait in the presence of genetic association
- The test uses a set of parent-child trios to test for over-transmission of an allele from heterozygous parents to affected offspring
- The results can be tested using a Chi-square test

Transmission disequilibrium test (TDT): an approach to deal with confounders in humans

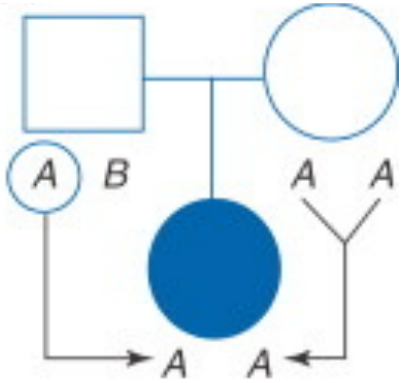
Here, the A allele is transmitted and the B allele is untransmitted from both parents



b=2, and a, c and d are 0

		Untransmitted allele: A	Untransmitted allele: B
Transmitted allele	A	a	b
Transmitted allele	B	c	d

Transmission disequilibrium test (TDT): an approach to deal with confounders in humans

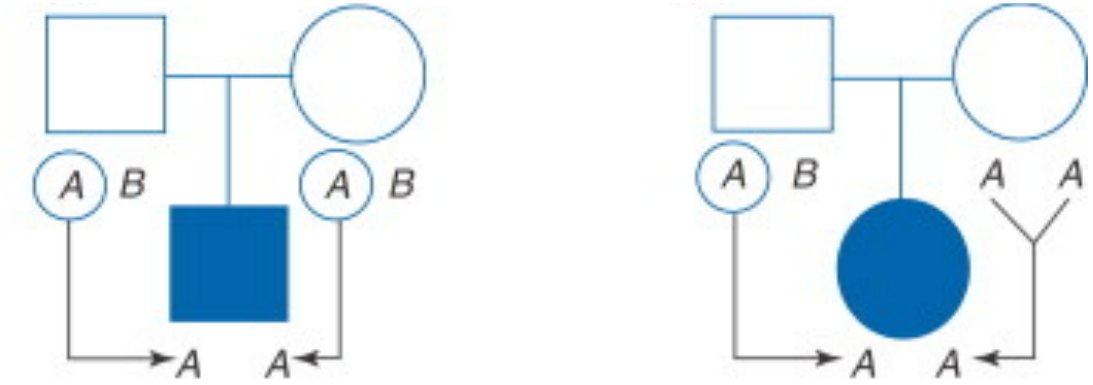


Here, the A allele is transmitted from the heterozygous father

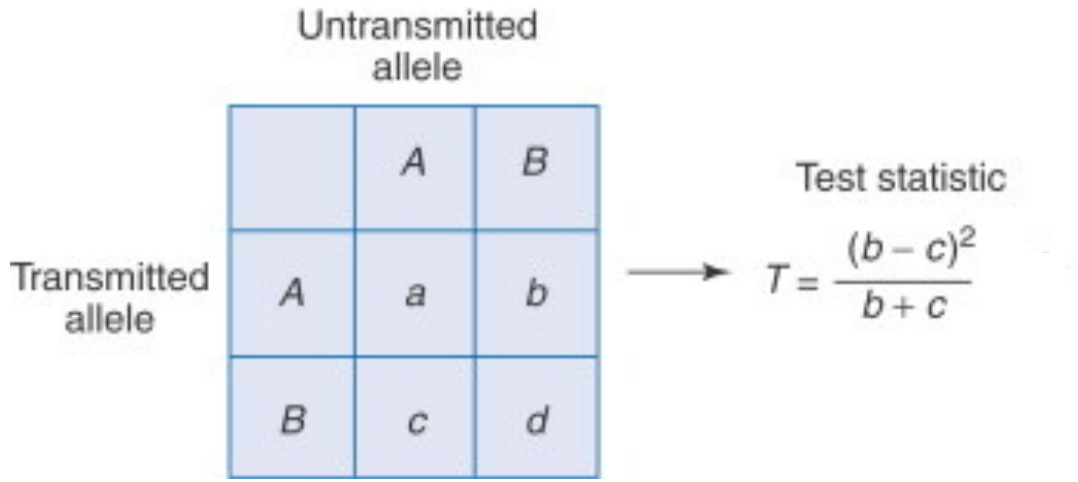
Transmitted allele \ Untransmitted allele		A	B
	A	a	b
	B	c	d

Here, a = 1, b=1, and c and d are 0

Transmission disequilibrium test (TDT): an approach to deal with confounders in humans



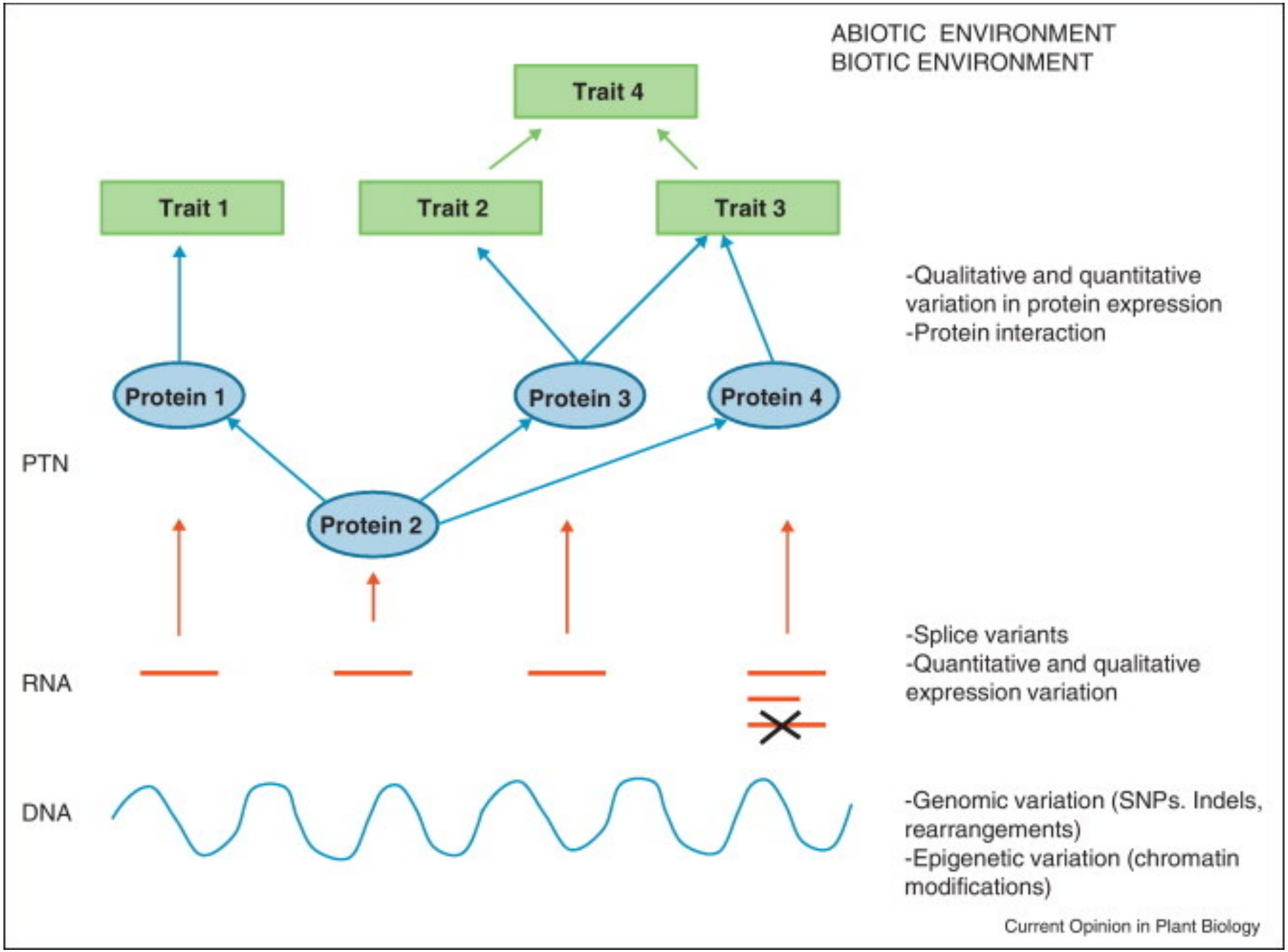
The same procedure is conducted over many families and then the final contingency table is tested using a Chi-square test



Some complications

Defining traits, pleiotropy and epistasis

What is a phenotype?



What is a phenotype?

Community-based (e.g., pathogens, microbes, neighbors)

Organismal, readily observable

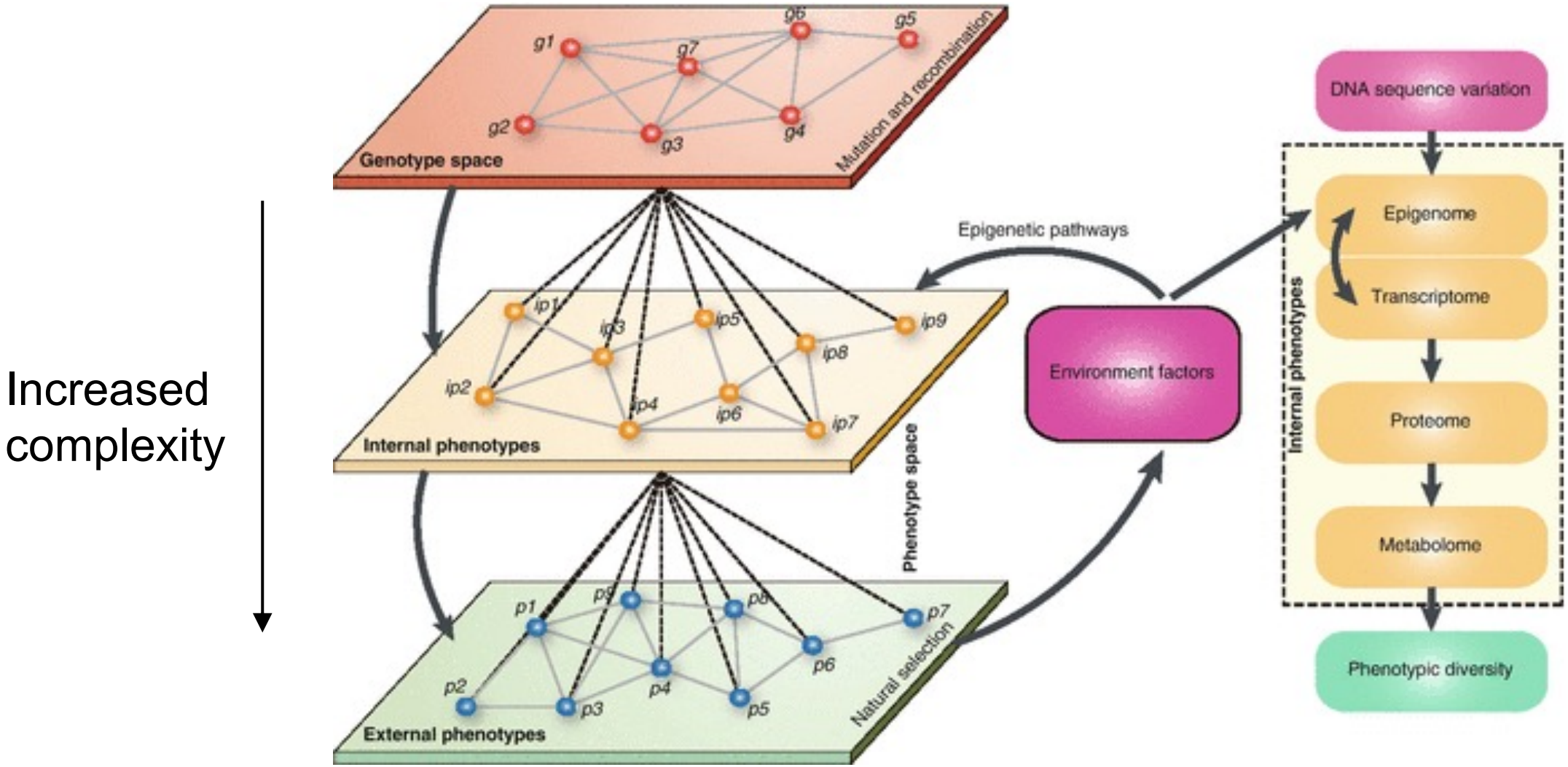
Epigenetic (e.g., silencing)

Transcriptional (e.g., eQTL associations)

Process-based (e.g., metabolites)

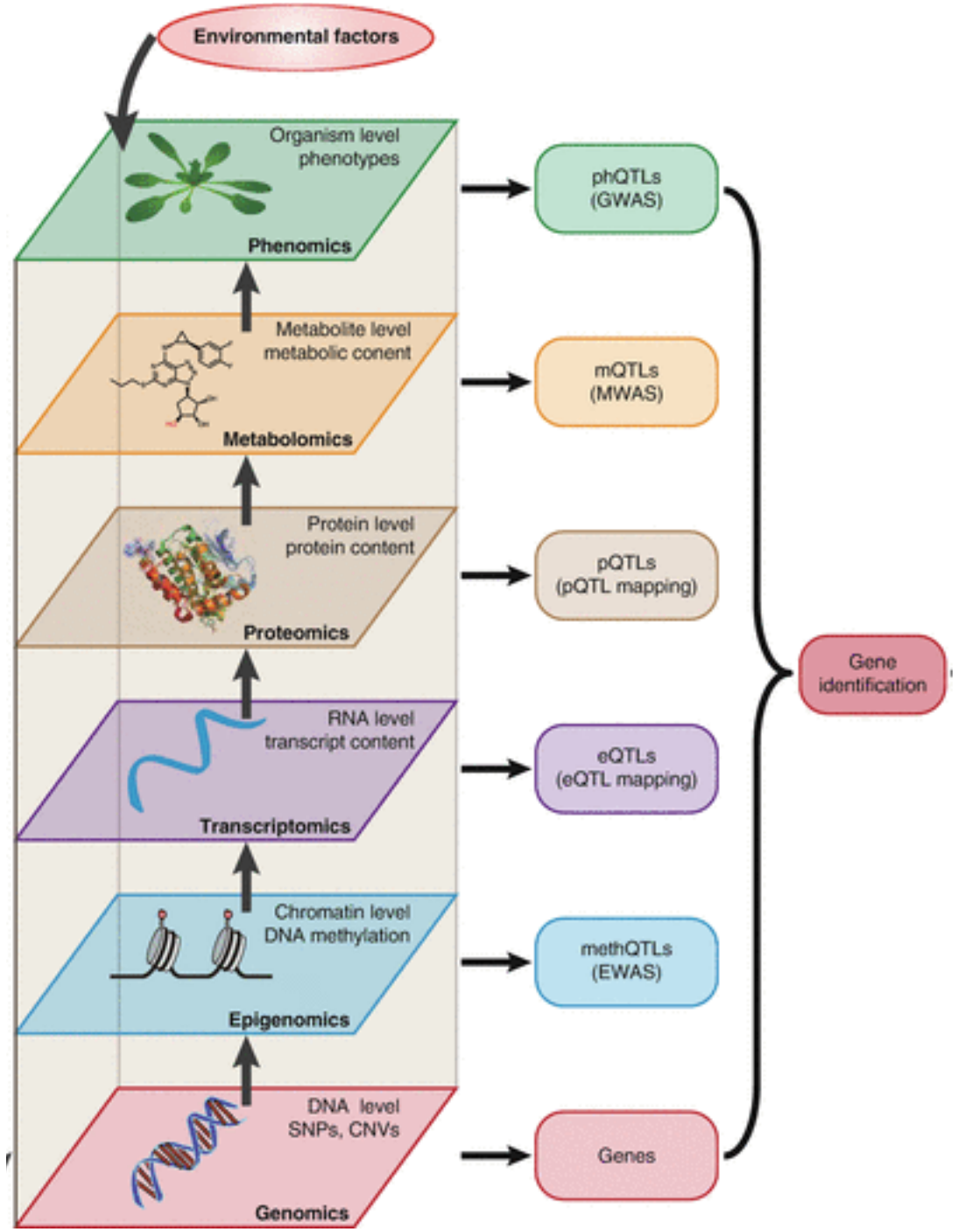
Traits can be defined at multiple levels

Increased complexity



https://link.springer.com/chapter/10.1007/978-3-642-41281-3_11/figures/1

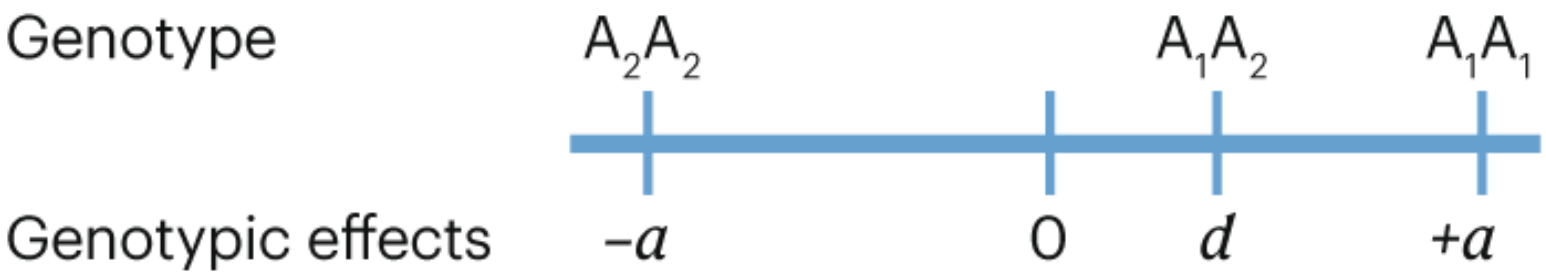
Traits can be defined at multiple levels



Pleiotropy and complex traits

- Pleiotropy is defined as the case where **variants in a single gene affect multiple traits**
- The extent of pleiotropy can be measured from mutant analysis or in natural populations
- In forward genetic screens (mutant analyses) effects of gene knockouts are often used
- In natural populations or recombinant mapping populations
  - Effects of variants are often more subtle than knock-outs
  - Linkage disequilibrium can result in apparent pleiotropy

Measuring pleiotropy

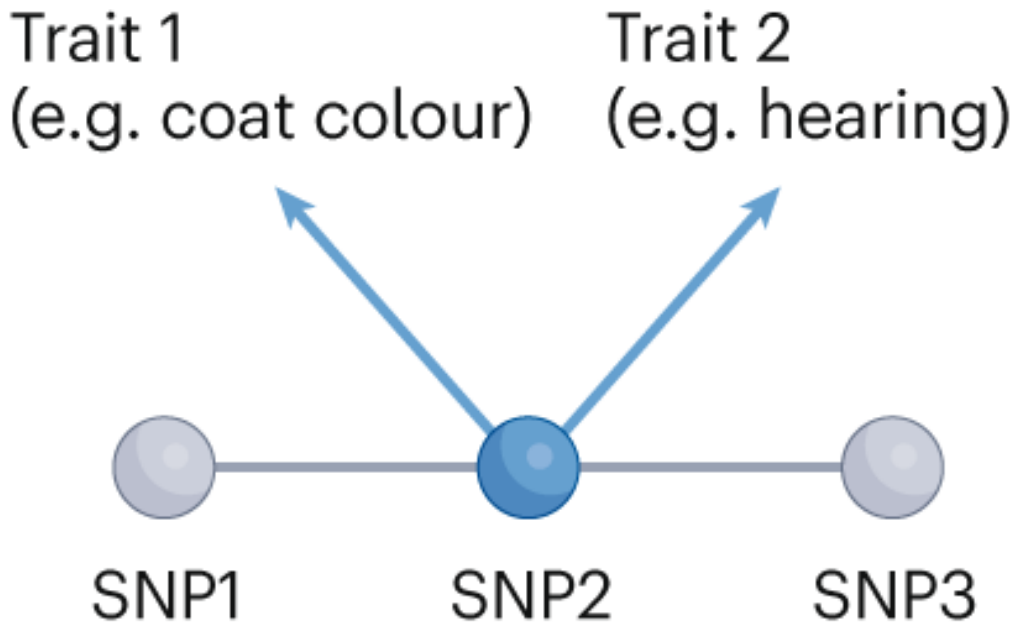


	μ_{A1A1}	μ_{A2A2}	μ_{A1A2}	a	d
Trait 1	5	5	5	0	0
Trait 2	4	6	8	2	0
Trait 3	8	14	16	4	2

Pleiotropy

Mackay and Anholt, NRG, 2024

Horizontal pleiotropy

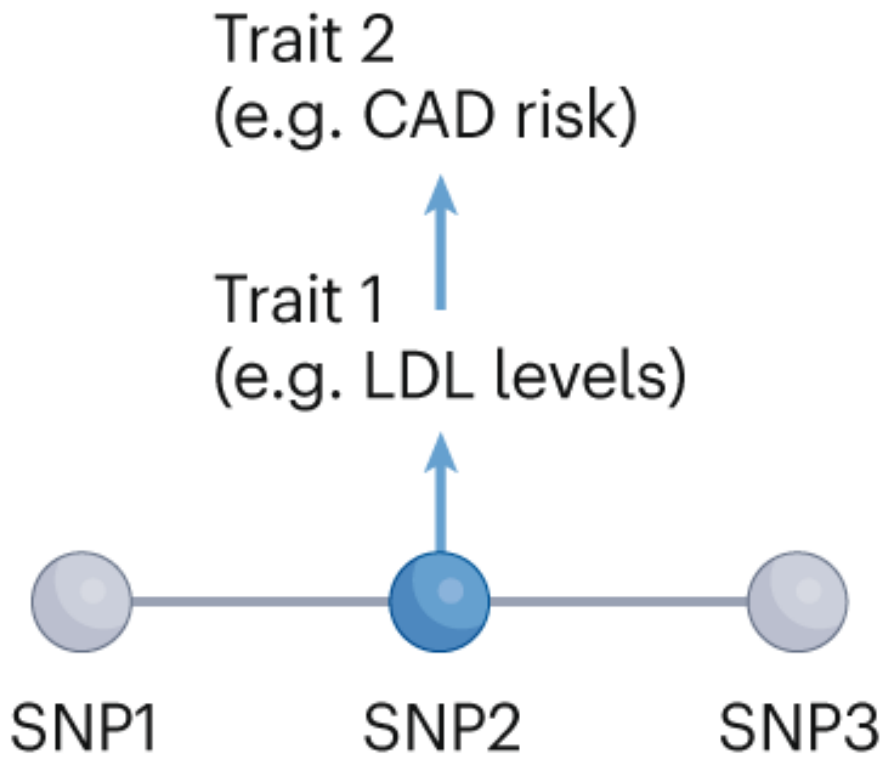


The same SNP independently affects two or more quantitative traits

Horizontal pleiotropy is a form of true pleiotropy

Mackay and Anholt, NRG, 2024

Mediating pleiotropy

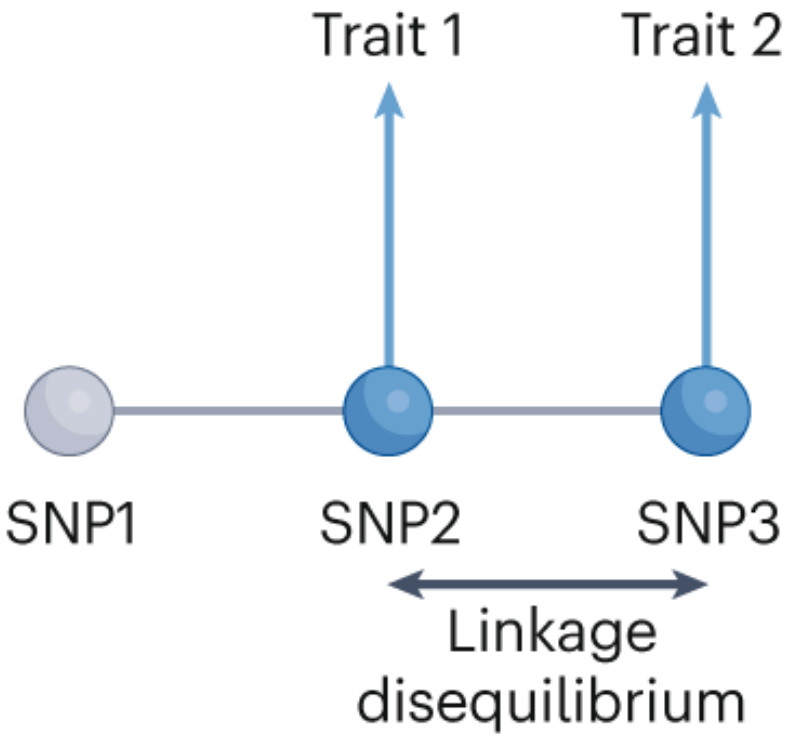


A SNP affects one trait, which in turn affects a second trait

Mediating pleiotropy is a form of true pleiotropy

Mackay and Anholt, NRG, 2024

Apparent pleiotropy



A locus maps to multiple traits but the causal locus is only in LD with the true variant for some traits

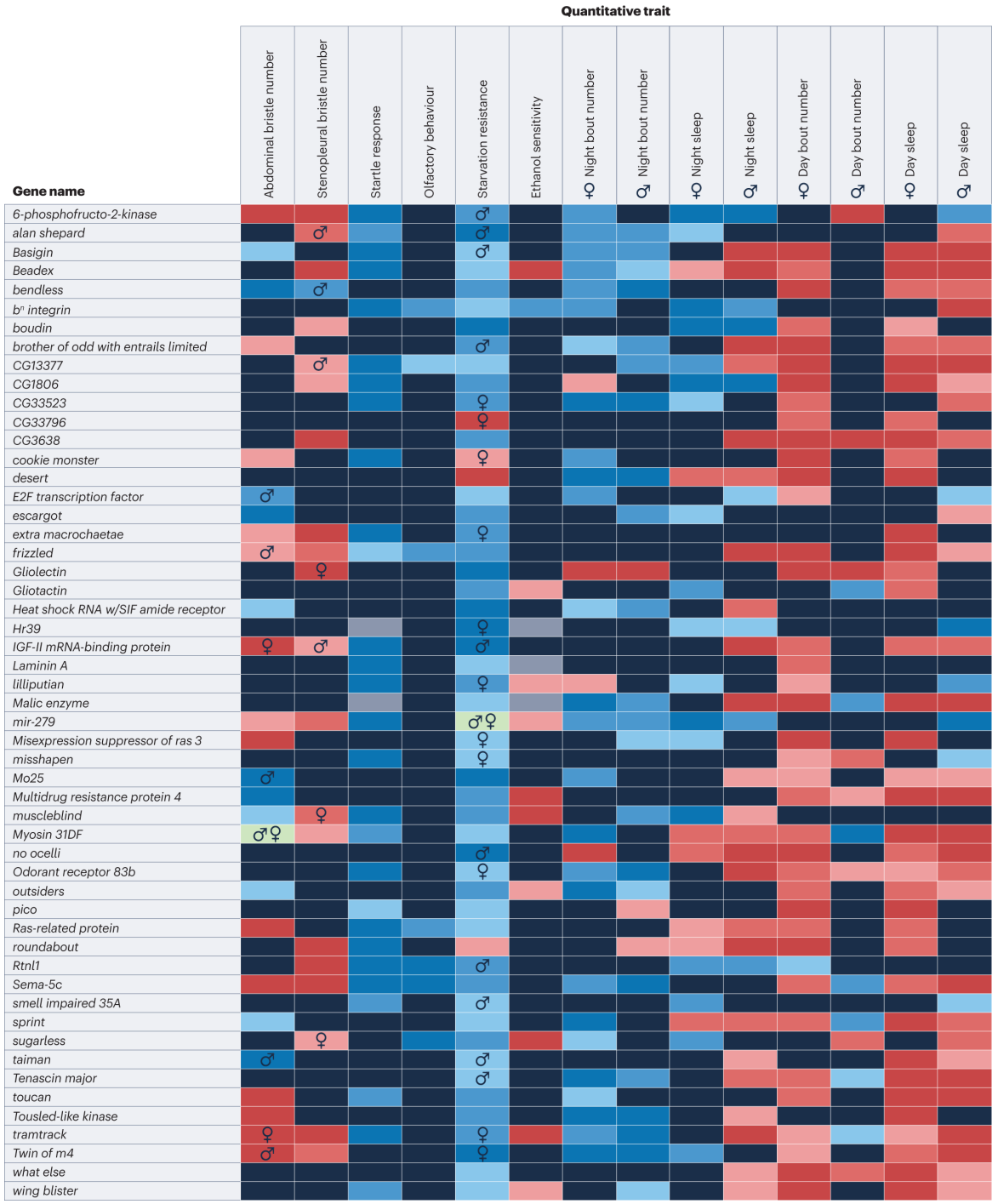
Here, SNP2 affects trait 1 and SNP3 affects SNP3, but because they are in LD in the mapping population, their effects cannot be differentiated

Mackay and Anholt, NRG, 2024

Effects of gene disruptions on a variety of traits

Phenotyping 10 quantitative traits across 53 Drosophila P-element insertions

The intensity of the colour (dark, medium and light) denotes deviations from the mean of co-isogenic controls exceeding the 95%, 99% and 99.9% confidence intervals, respectively. Light green cells indicate decreased trait values in males and increased trait values in females for the P-element insertion compared with the control. Black cells are not significantly different from the control, and grey cells indicate traits for which the effect of the P-element insertion was not measured.



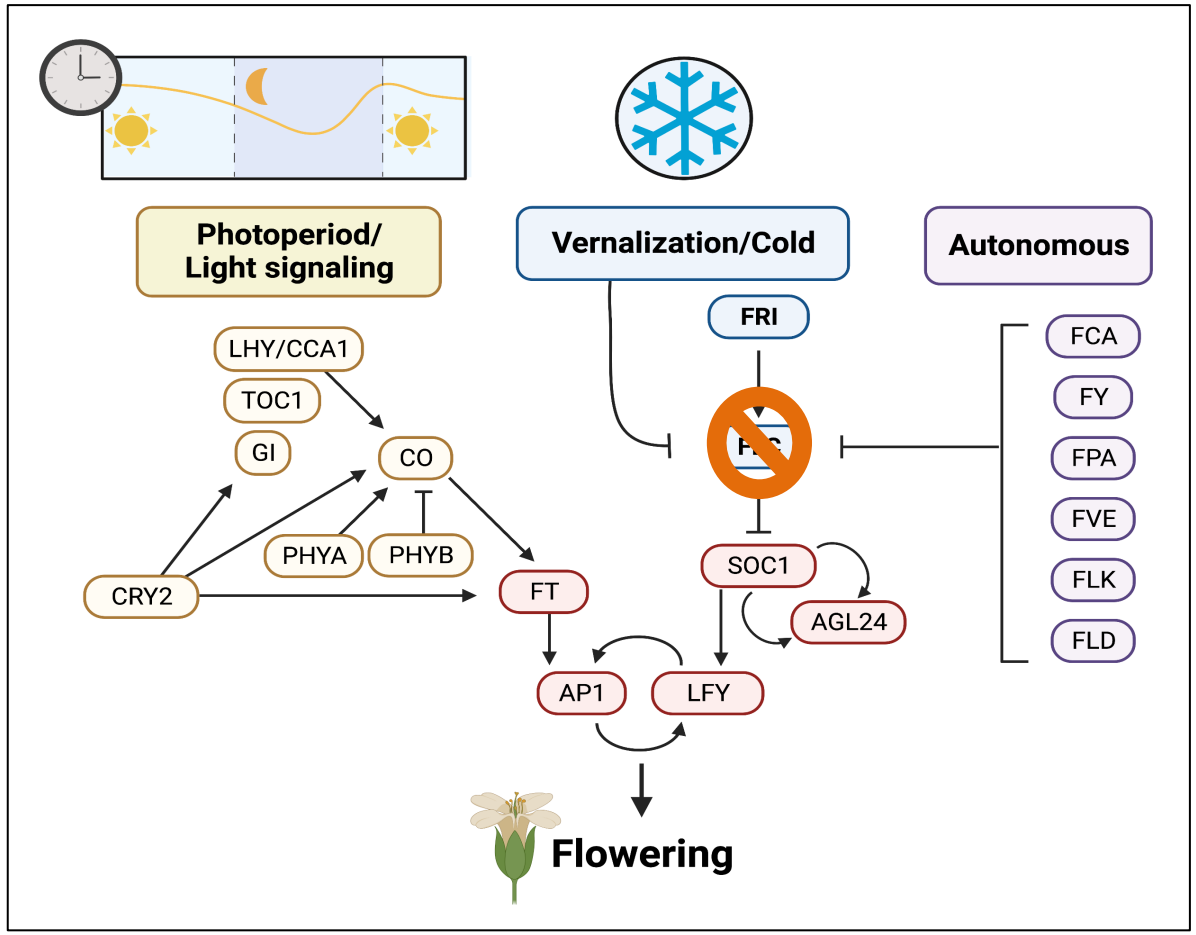
Mackay and Anholt, NRG, 2024

Genotype by environment (GxE) interactions

- Interaction with the environment can differentially affect a genetic response
- These effects can be subtle or strong, acting through ‘genetic switches’
- For example, exposure to a particular environment can cause silencing of a developmental gene through epigenetic modifications

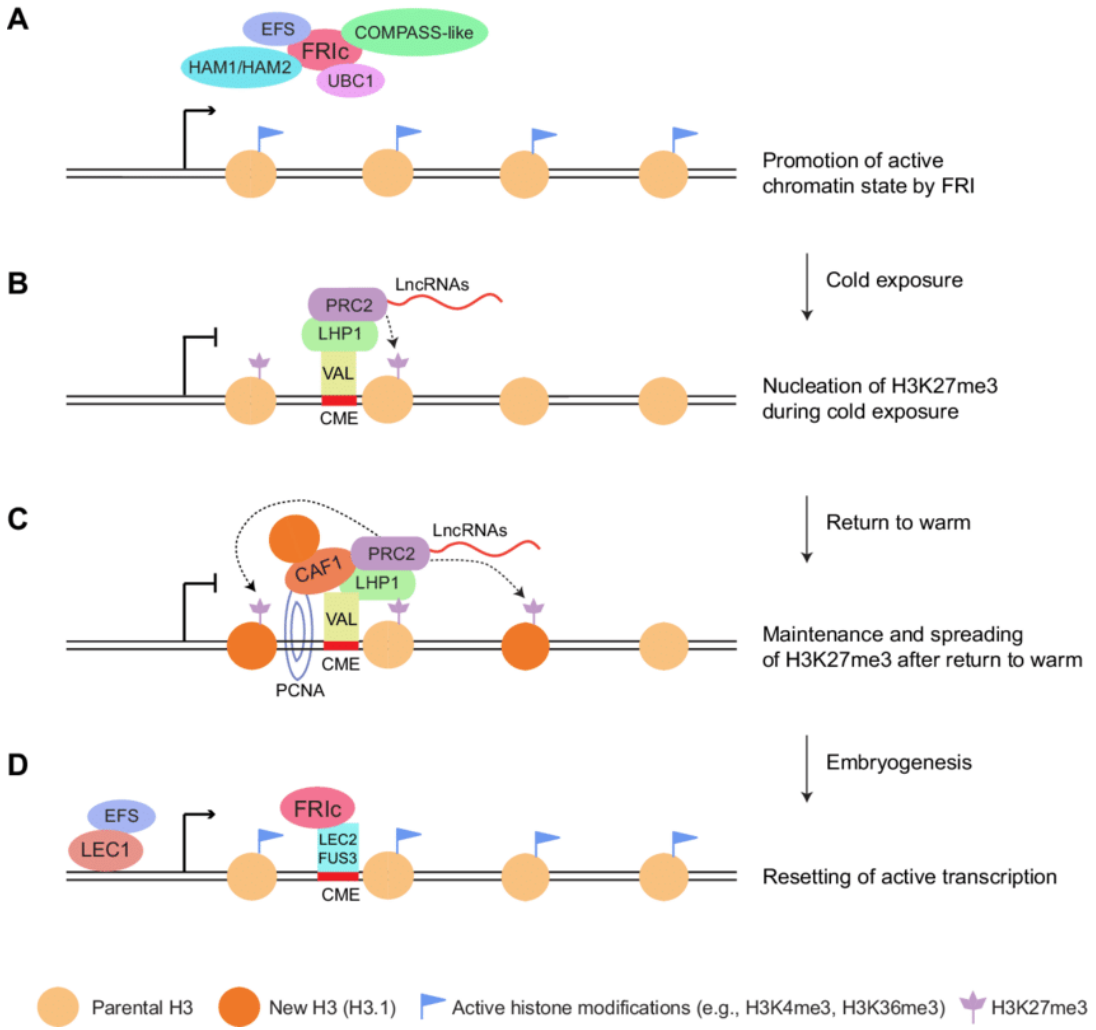
An example of a mechanism that produces a GxE interaction from plants

- In plants, silencing of FLC causes the transition to flowering after cold treatment.
- This is called the ‘vernalization’ response



An example of a mechanism that produces a GxE interaction from plants

Cold exposure results in accumulation of histones over the promoter of FLC

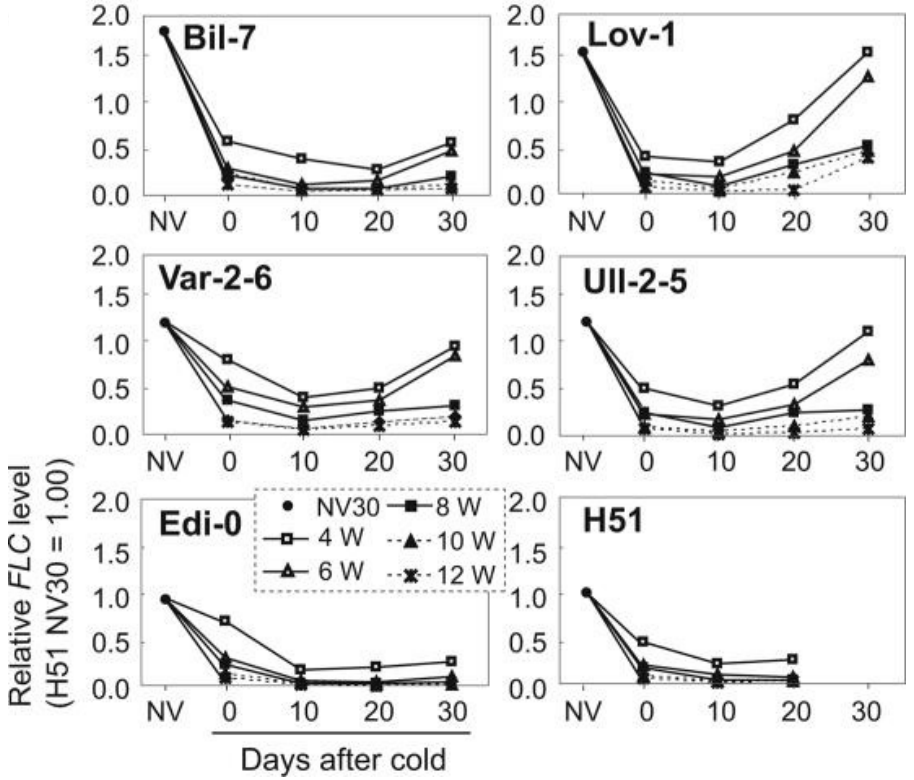


Zhao et al., 2019

An example of a mechanism that produces a GxE interaction from plants

Genetic variation across wild-collected *Arabidopsis* lines in the promoter of FLC results in a GxE interaction

Expression during cold treatment in 6 wild lines of *Arabidopsis*

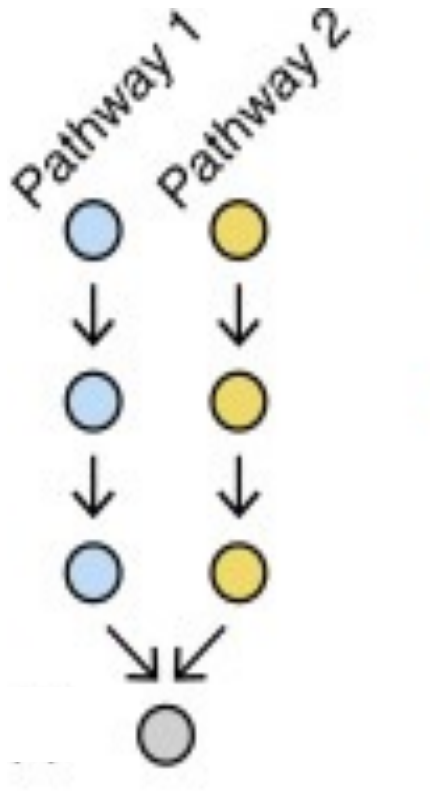


Zhao et al., 2019

Genotype by genotype (GxG) interactions (epistasis)

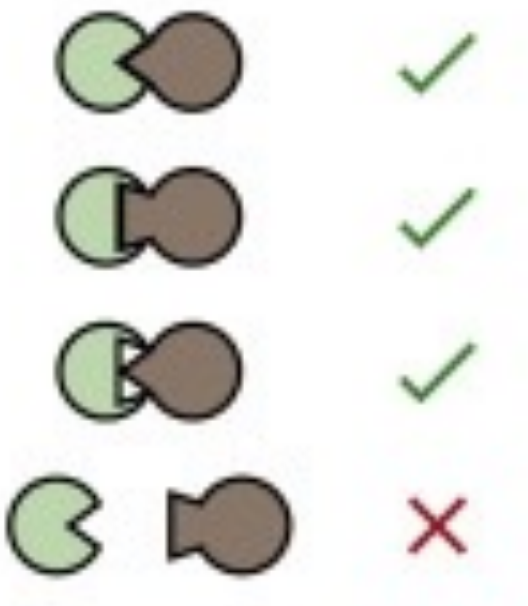
- Epistasis is the circumstance where the effect of one gene is masked, inhibited or suppressed by the expression of one or more other genes
- Epistasis is also sometimes referred to as a ‘background effect’, meaning that some factor in the genetic background effects the the trait
- Epistasis implies a direct or indirect interaction in the context of a molecular pathway
- Epistasis can reduce power in trait mapping

Epistatic effects from loss of function events in two pathways that produce a protein



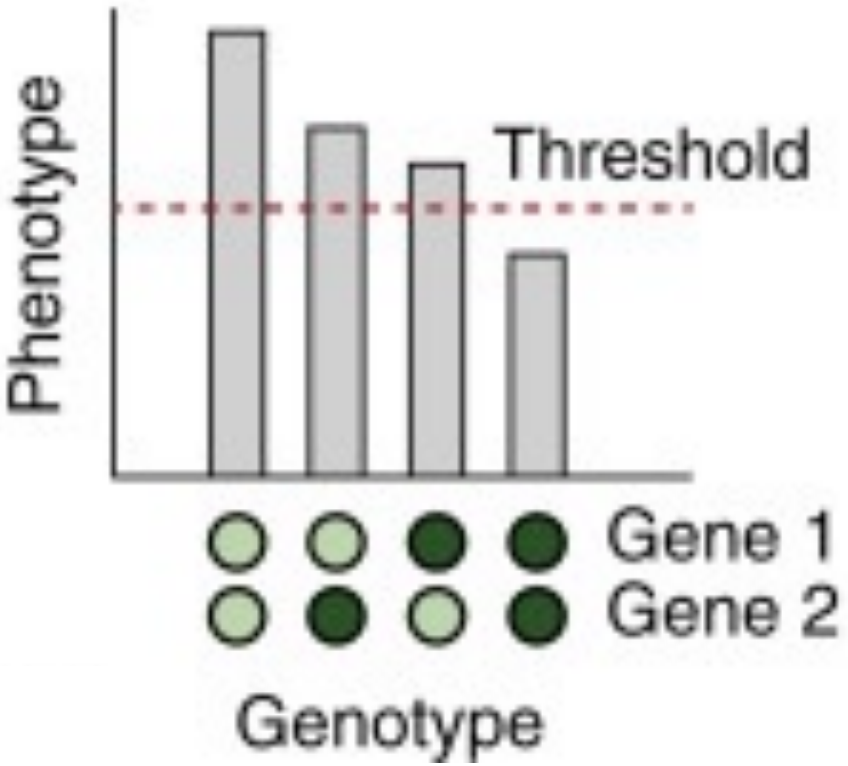
Loss-of-function quantitative trait nucleotides (QTNs) in parallel pathways that converge on the same metabolite or downstream gene can result in a genetic interaction

Interaction in a protein complex



If either of two proteins that interact to form a complex are modified, it can lead to loss of function of the complex

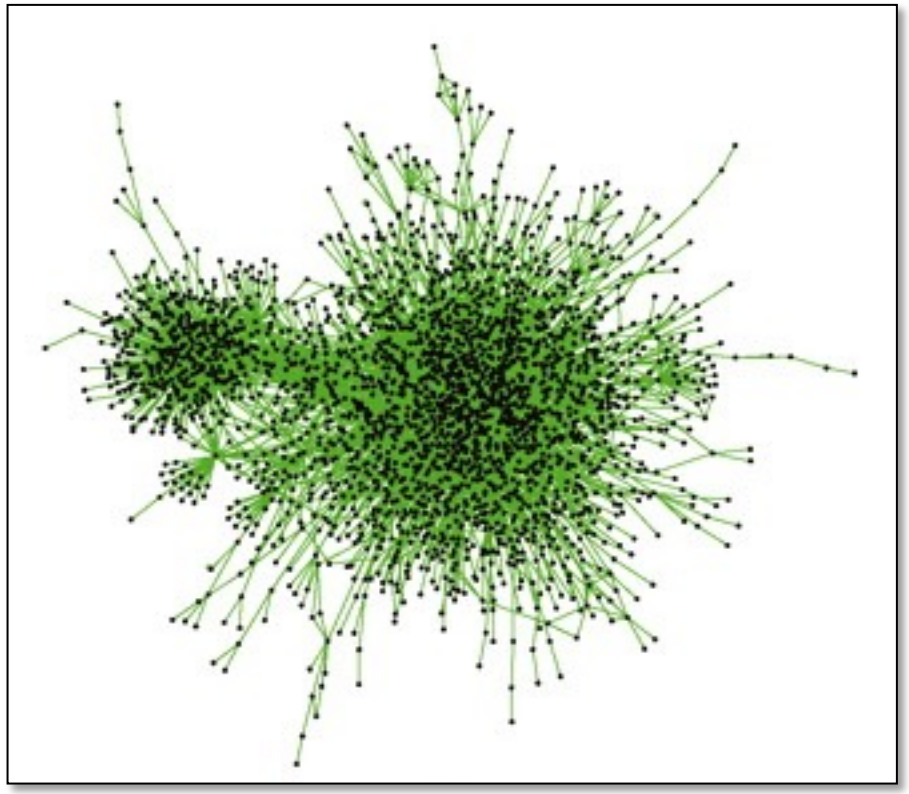
Threshold traits



Sometimes a threshold level of protein is needed to noticeably impact a trait

In that case, multiple genes can contribute additively to the phenotype

CHALLENGE: Interactions among loci may influence trait values



- HIGH COMPUTATIONAL REQUIREMENTS

The number of comparisons scales exponentially with the number of interactions, so the testing burden is great, e.g., for 10K loci, there are 495,000 first-order interactions

- POWER

Enormous sample sizes are therefore be required to find sufficient individuals of each genotype combination to measure small effects accurately

For next time:

Mapping by genome-wide association