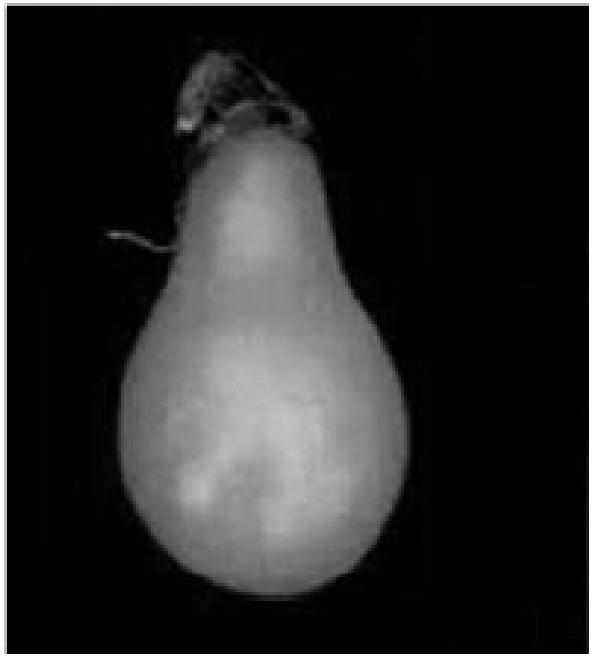
Mapping the genes underlying phenotypic variation

Virginie Courtier-Orgogozo
Institut Jacques Monod, Paris

Tomato shape

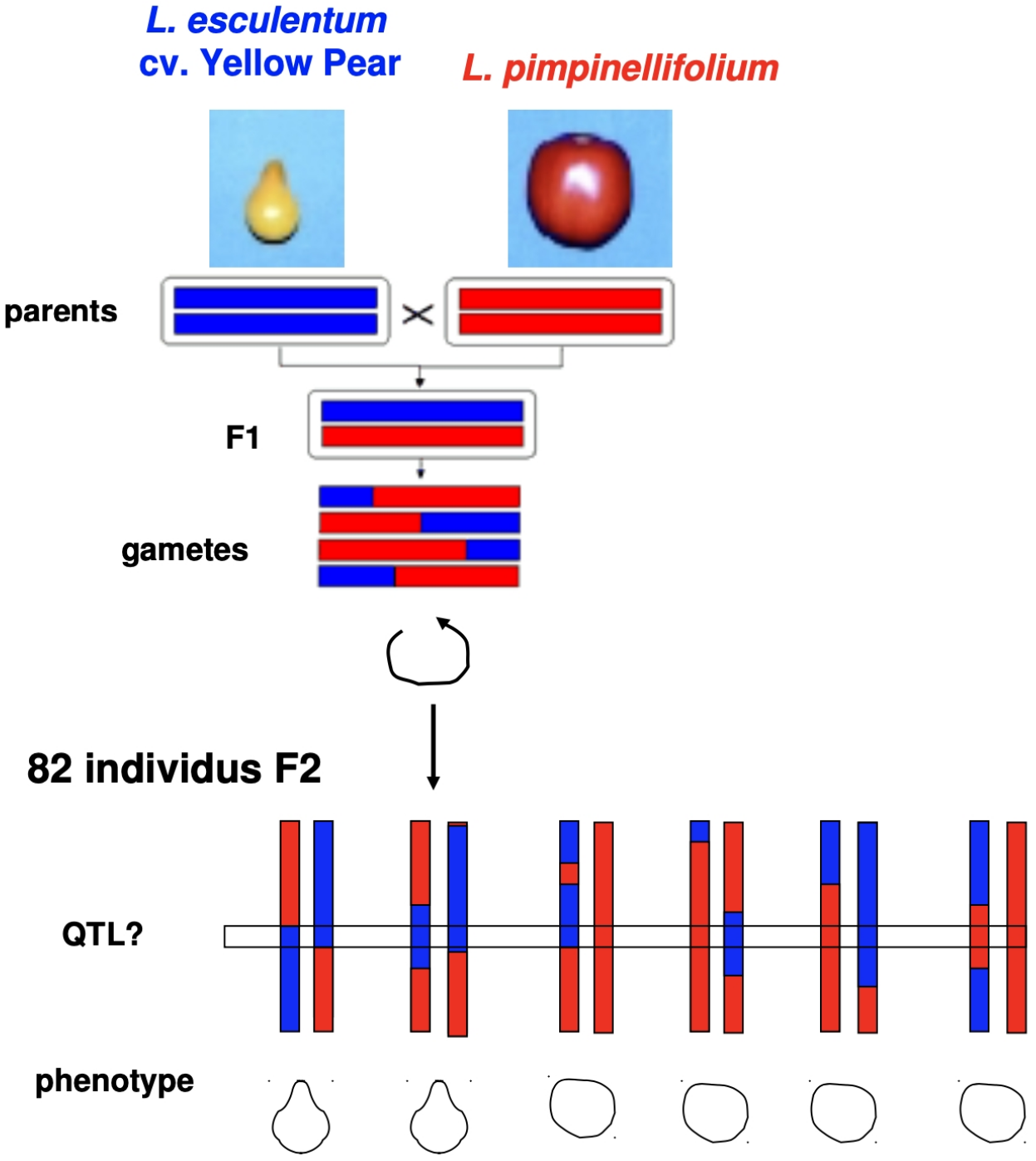


Lycopericon esculentum

Lycopersicon esculentum **cv. Yellow Pear**

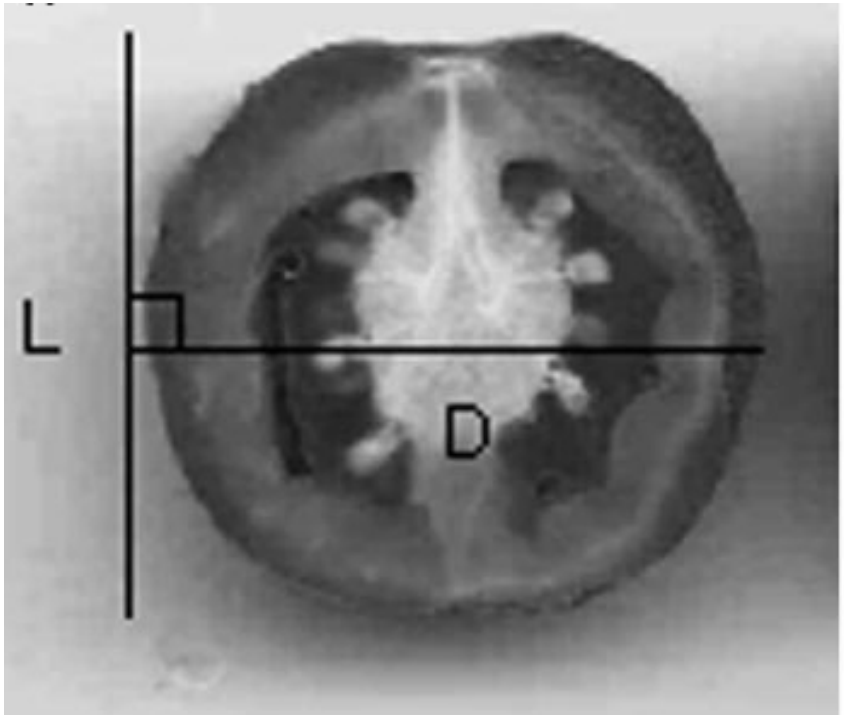
(Ku et al., 1999; Liu et al., 2002)

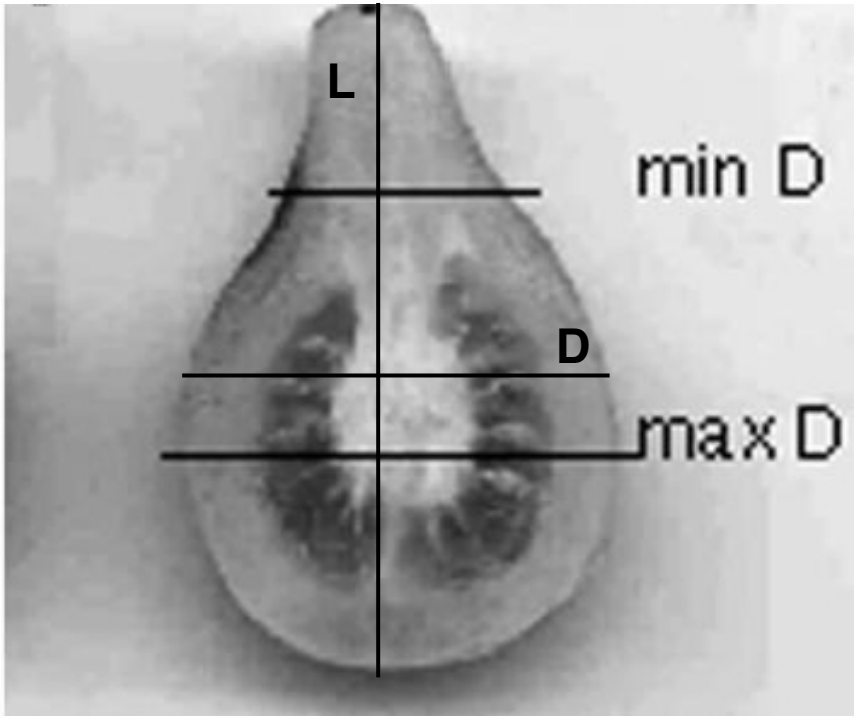
QTL mapping


L. esculentum
cv. Yellow Pear
L. pimpinellifolium
parents
F1
gametes
82 individus F2
QTL?
phenotype

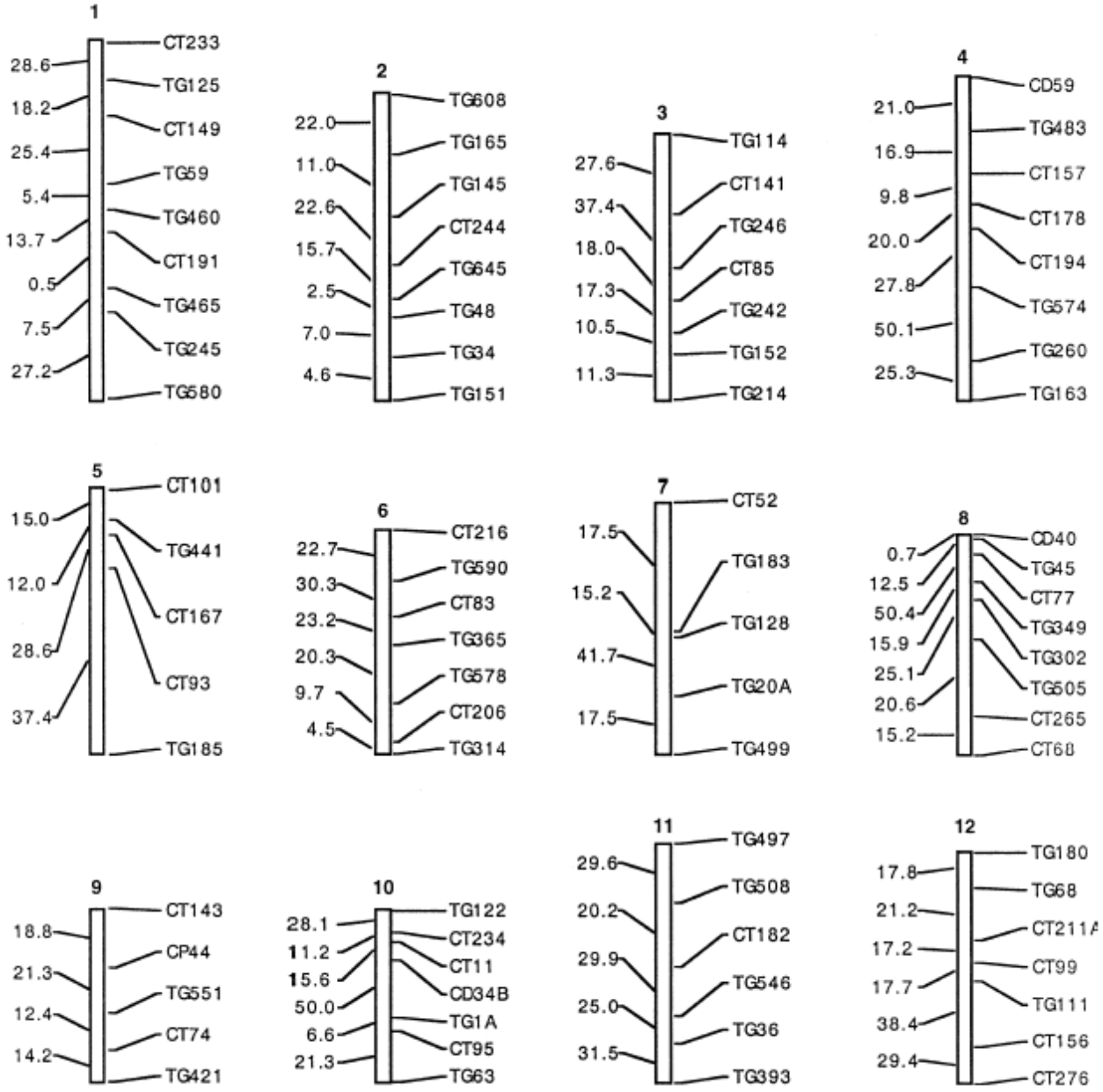
Quantitative measure of the phenotype

Measure of 2 indexes L/D and Dmin/Dmax for 10 fruits per plant
L/D : L= length, D = diameter at equator
Dmin/Dmax



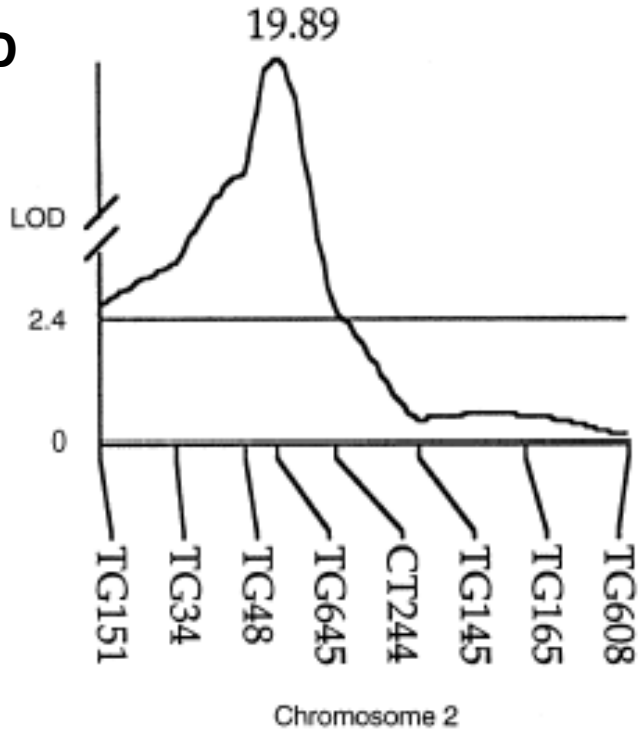


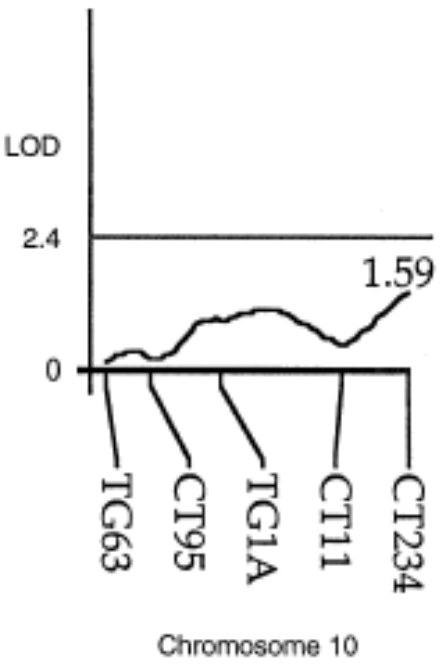
82 molecular markers on the 12 tomato chromosomes



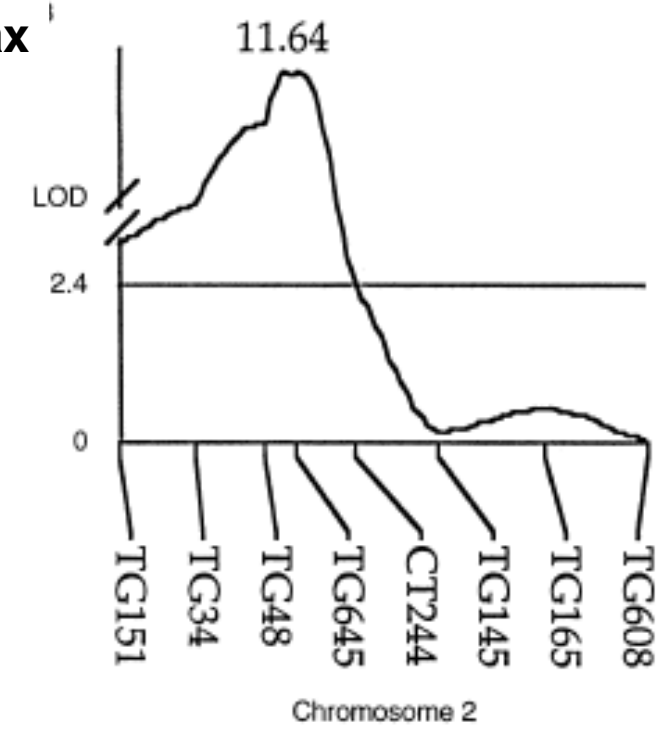
One major locus near marker TG645

L/D





Dmin/Dmax



responsible for 67% of L/D variance

allele YP = recessive

Two main files

Markers file

```
-start
-Chromosome 1
CF5475          0.4
CF5573          24.7
CT7895          41.0
CT8903          59.0
CF5613          67.7
CT7892          76.0
CT890           89.0
CT233           39.0
Telomere        50.0
-Chromosome 2
CF5671          0
CF5675          10.4
CF5673          34.7
CT789           41.0
CT890           89.0
CT567           115.0
Telomere        130.0
…
```

Genotypes and phenotype(s) file

```
-start individuals markers
Ind_1 0 0 1 1 0 0 0 0 0 1 2 2 2 2
Ind_2 0 0 0 1 0 1 0 0 1 1 1 1 0 0
Ind_3 2 2 2 2 2 1 0 1 1 1 1 0 0 0
Ind_4 0 1 0 0 0 0 1 1 1 2 2 1 1 1
Ind_5 0 1 0 0 0 0 1 1 1 1 2 2 2 2
Ind_6 1 1 1 1 1 1 1 1 1 0 0 0 0 0
Ind_7 1 1 1 1 1 1 1 0 1 n n 1 1 1
Ind_8 2 2 2 1 1 1 1 0 1 1 1 1 1 0
Ind_9 1 1 1 1 1 1 1 0 0 1 1 1 1 1
Ind_1 0 2 2 1 1 1 1 1 0 0 0 1 1 2
-stop individuals markers

-start individuals traits 1 LoverD named
Ind_1    5.5
Ind_2    3.0
Ind_3    4.0
Ind_4    7.0
Ind_5    6.5
Ind_6    5.0
Ind_7    3.5
Ind_8    6.0
```

Simple linear regression for each marker

L/D of individual i = a + b.xi + ε

xi = 0 if Le/Le, = 1 if Le/Lp, = 2 if Lp/Lp

a,b = best fit parameters (least square regression)

ε assumed to have a normal distribution

Test Ho: b = 0 *versus* H1: b = estimated b

Likelihood ratio test statistic

$$D = -2(\ln(\text{likelihood for null model}) - \ln(\text{likelihood for alternative model}))$$
$$= -2\ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right).$$

The probability distribution of the test statistic can be approximated by a chi-square distribution with (df1 − df2) degrees of freedom, where df1 and df2 are the degrees of freedom of models 1 and 2 respectively

Interval mapping

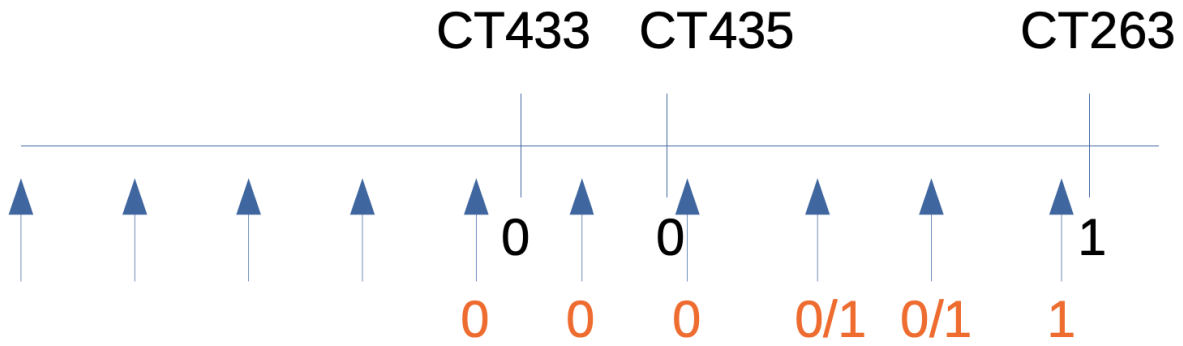
L/D of individual i = a + b.xi + e

xi = indicator variable specifying the probabilities of an individual being in different genotypes for the tested position, constructed by flanking makers

xi = 0 if Le/Le, = 1 if Le/Lp, = 2 if Lp/Lp

a,b = best fit parameters (maximum likelihood)

Test Ho: b=0 versus H1: b=estimated b



Interval mapping

L/D of individual i = a + b.xi + e

xi = indicator variable specifying the probabilities of an individual being in different genotypes for the tested position, constructed by flanking makers

xi = 0 if Le/Le, = 1 if Le/Lp, = 2 if Lp/Lp

a,b = best fit parameters (maximum likelihood)

Test Ho: b=0 versus H1: b=estimated b

Composite Interval mapping

L/D of individual i = a + b.xi + c.xi + e

xi = indicator variable specifying the probabilities of an individual being in different genotypes for the tested position, constructed by flanking makers

xi = 0 if Le/Le, = 1 if Le/Lp, = 2 if Lp/Lp

yi = 0 if Le/Le, = 1 if Le/Lp, = 2 if Lp/Lp at marker y

LOD score

L/D of individual i = a + b.xi + e

Test Ho: b = 0 *versus* H1: b = estimated b

Lo = pr (data | no QTL) – phenotypes assumed to follow a normal distribution
L1 = pr (data | QTL at tested position)

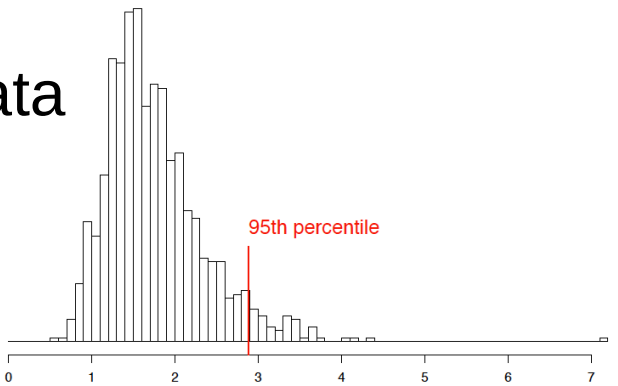
$$LOD = -\log \frac{L_0}{L_1}$$

Significance threshold

10,000 permutations of phenotype/genotype data

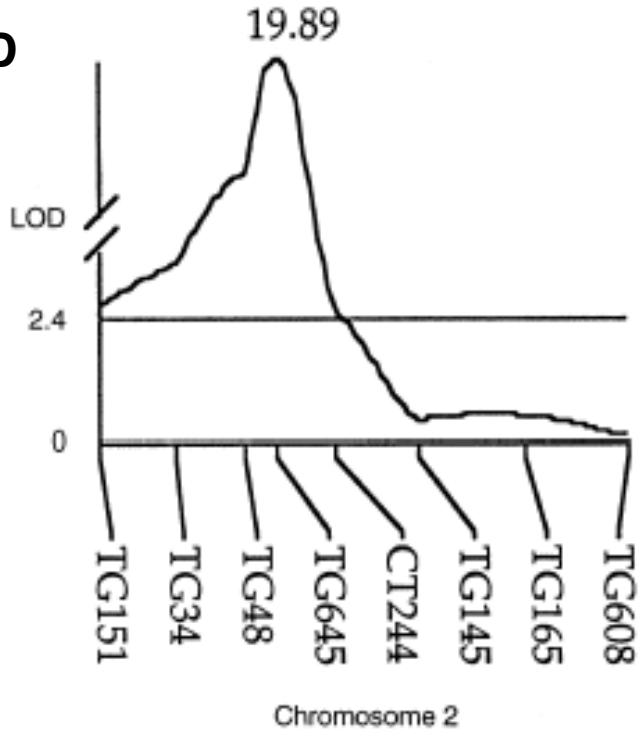
→ random distribution of LOD scores

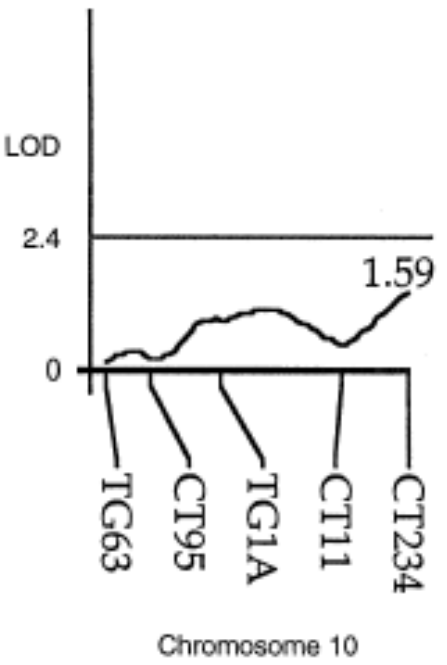
→ 1% or 5% significance threshold



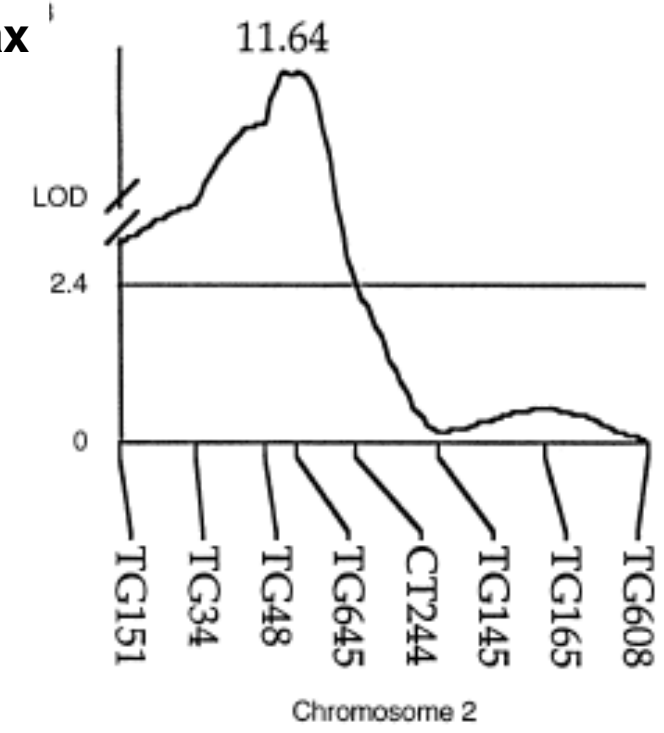
One major locus near marker TG645

L/D





Dmin/Dmax



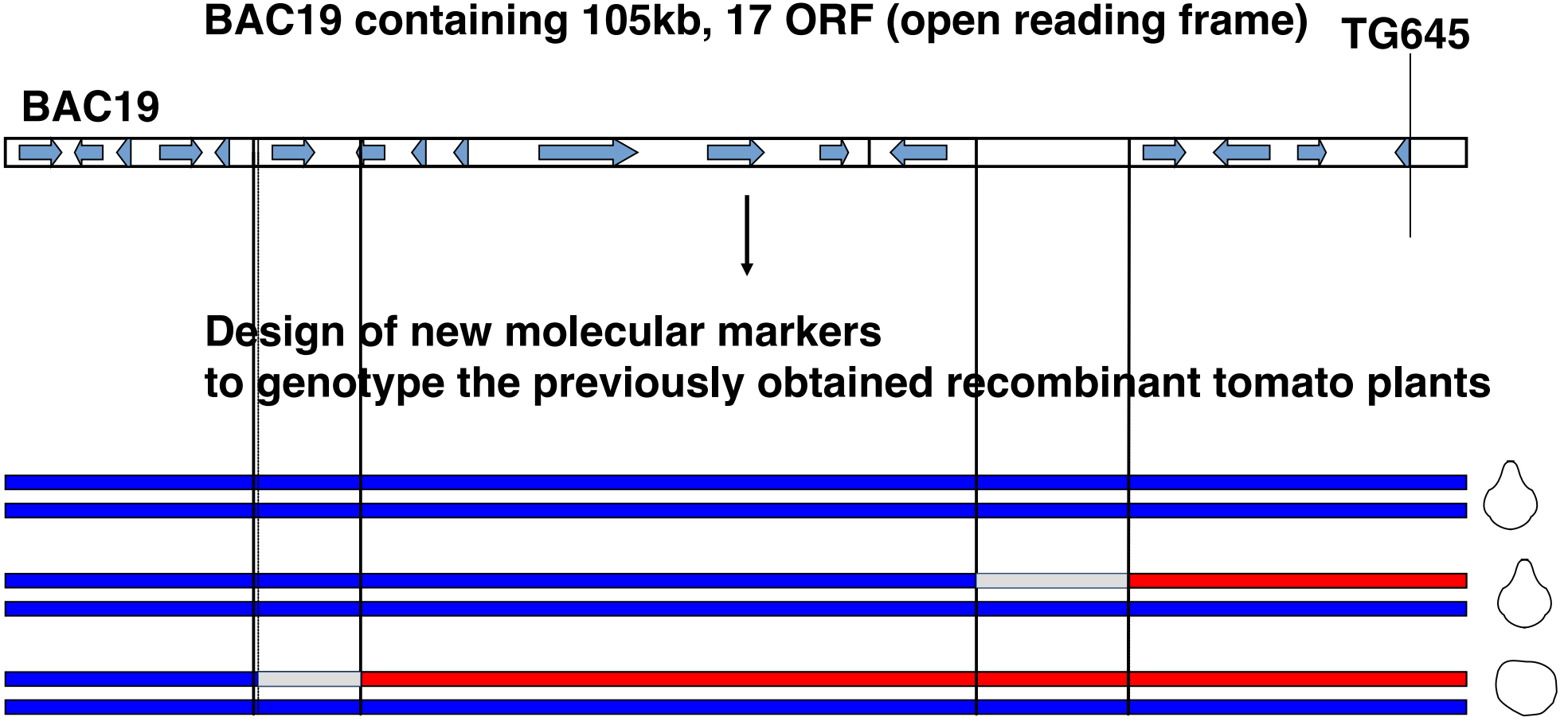
responsible for 67% of L/D variance

allele YP = recessive

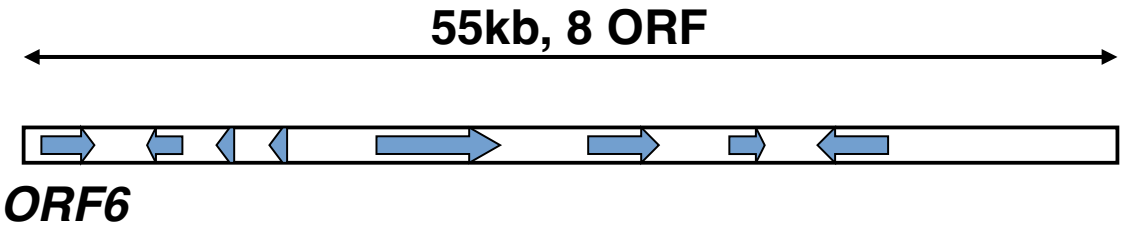
BAC library (Bacterial Artificial Chromosomes)
contains genomic DNA fragments of 100-350kb

Screen of the library with marker TG645

BAC19 containing 105kb, 17 ORF (open reading frame)



Design of new molecular markers
to genotype the previously obtained recombinant tomato plants



L. esculentum

Sequencing of the region in the 2 tomato varieties

1 SNP (single nucleotide polymorphism)
et 1 indel (insertion-deletion) of 2bp
in non-coding regions

L. esculentum
cv. Yellow Pear

X

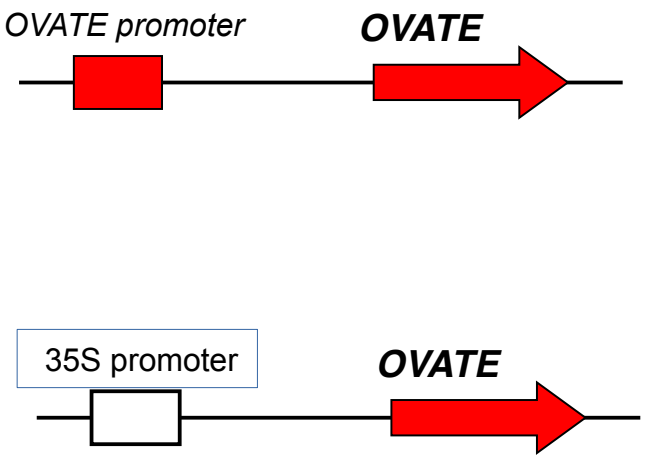
1 SNP in *ORF6* : G496T, stop codon stop,
truncated protein with last 75 amino acids
missing

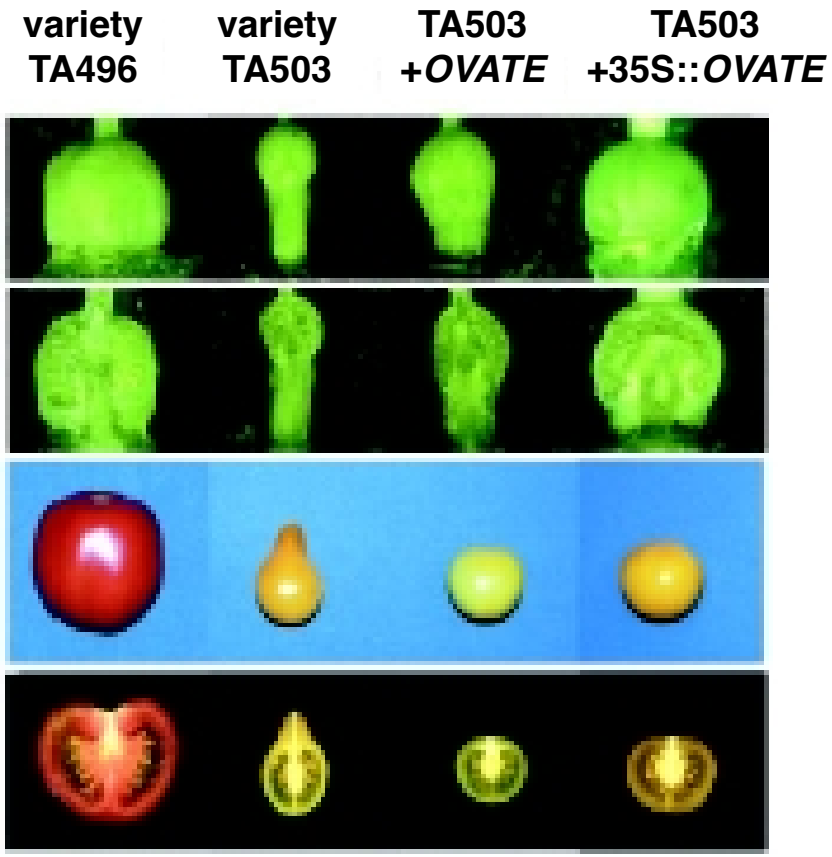
Hypothesis: the causing gene is *ORF6* = *OVATE*

The causing gene is *OVATE/ORF6*

Same mutation in 3 other pear tomato varieties

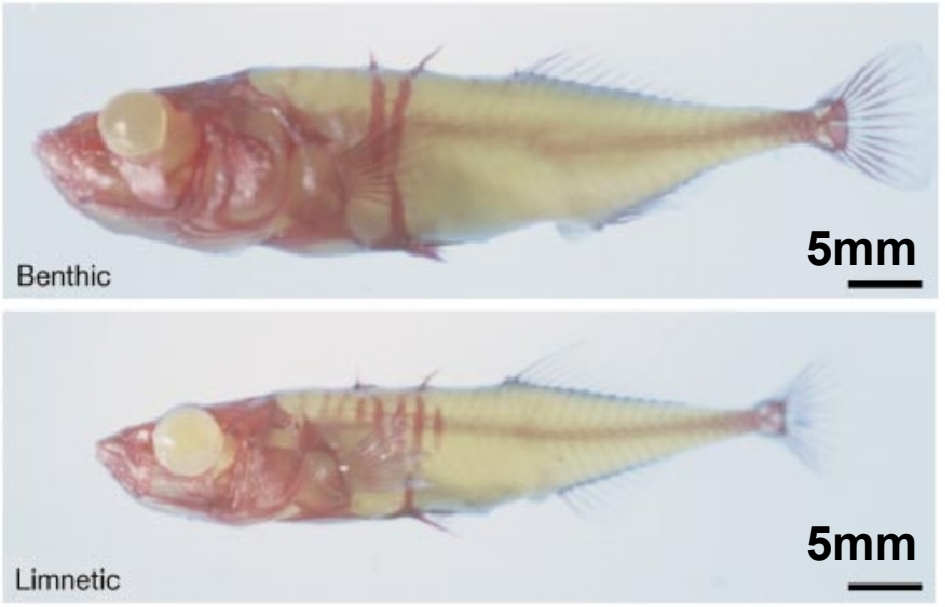
Complementation of the mutation by transgenesis





OVATE = protein with NLS (nuclear localization signal), unknown function, expressed in developing fruits

Evolution of morphology in threespine sticklebacks



marine

Paxton Lake, Canada

Gasterosteus aculeatus

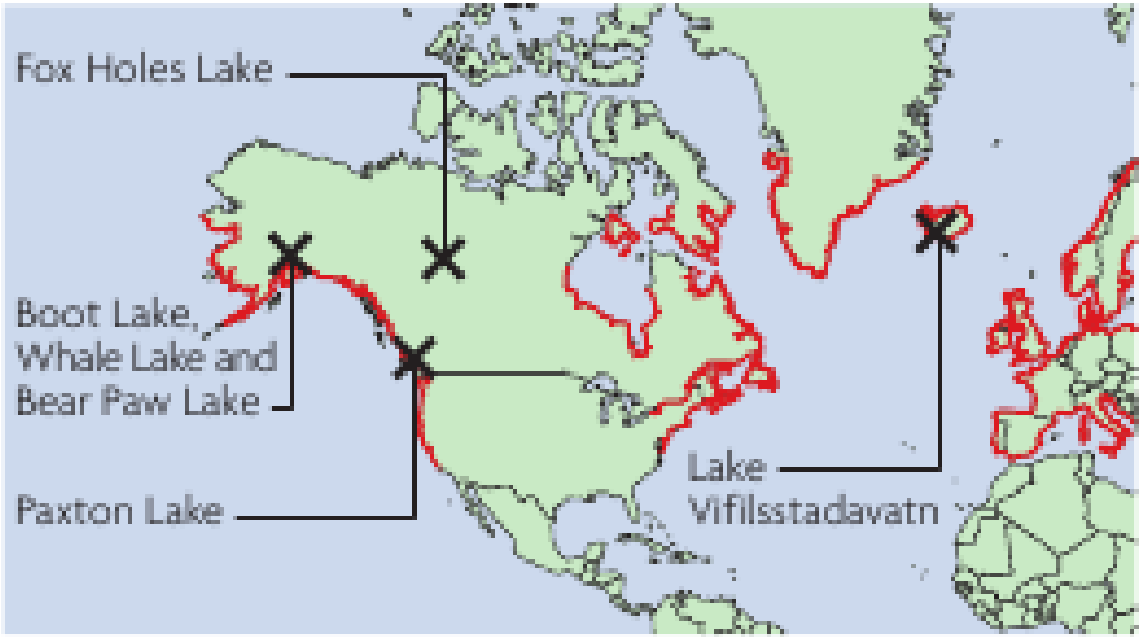
(Peichel et al., 2001 ; Shapiro et al, 2004 ; Chan et al. 2010)

Marine fishes with robust pelvis = ancestral

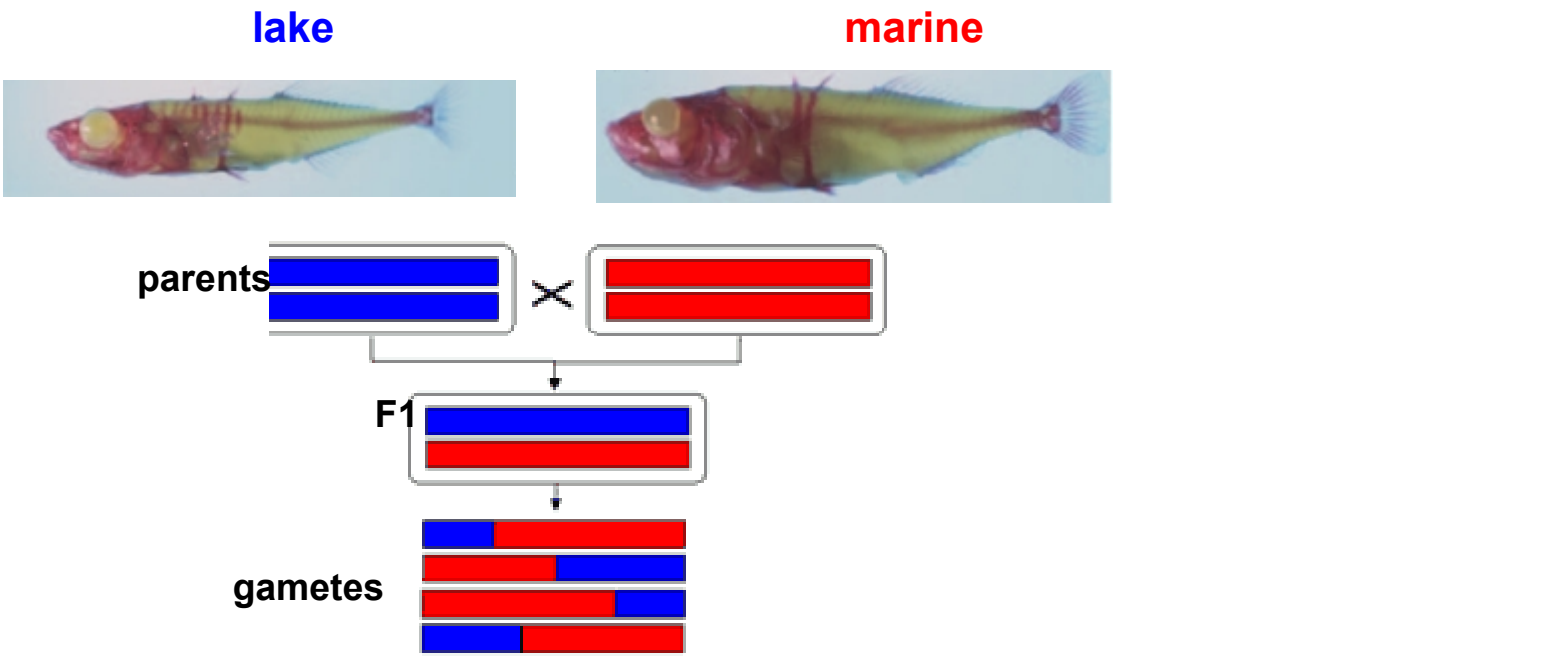
Freshwater fishes with reduced pelvic structures = derived, independently at least 20 times

- **limited calcium availability**
- **absence of gape-limited predatory fishes**
- **predation by grasping insects**

Last glacier retreat = 10 000 – 20 000 years ago



QTL mapping

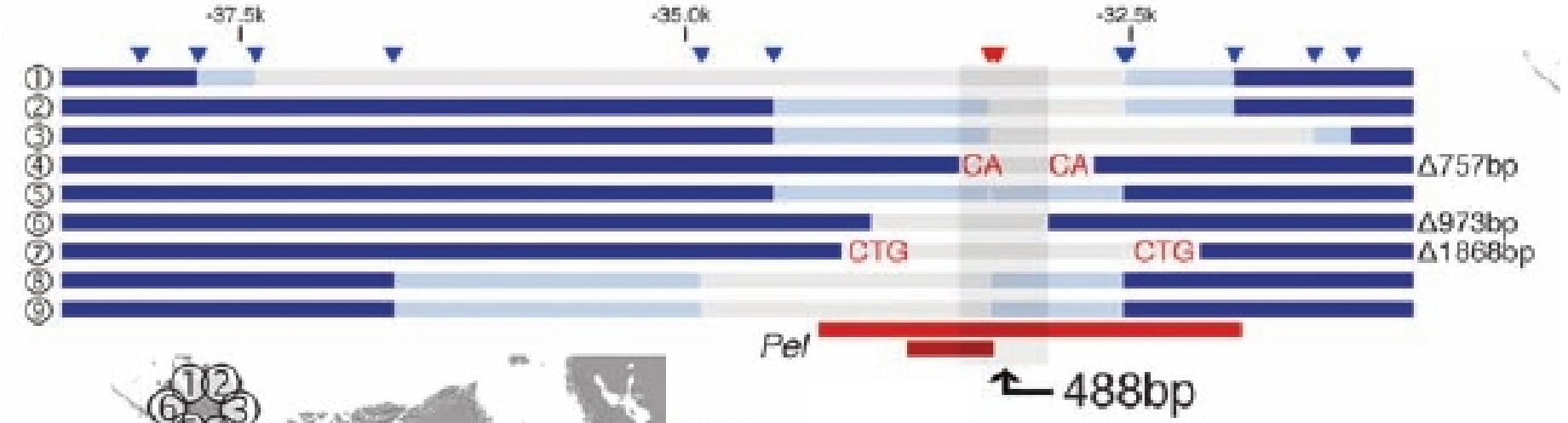


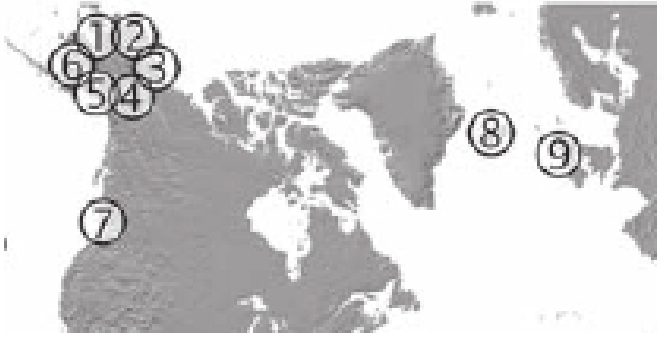
(Shapiro et al., 2004)

Several independent deletions in the cis-regulatory region of *Pitx1*

Region sequenced in two lake populations: a 2-kb deletion in one and a 757-bp deletion in the other one

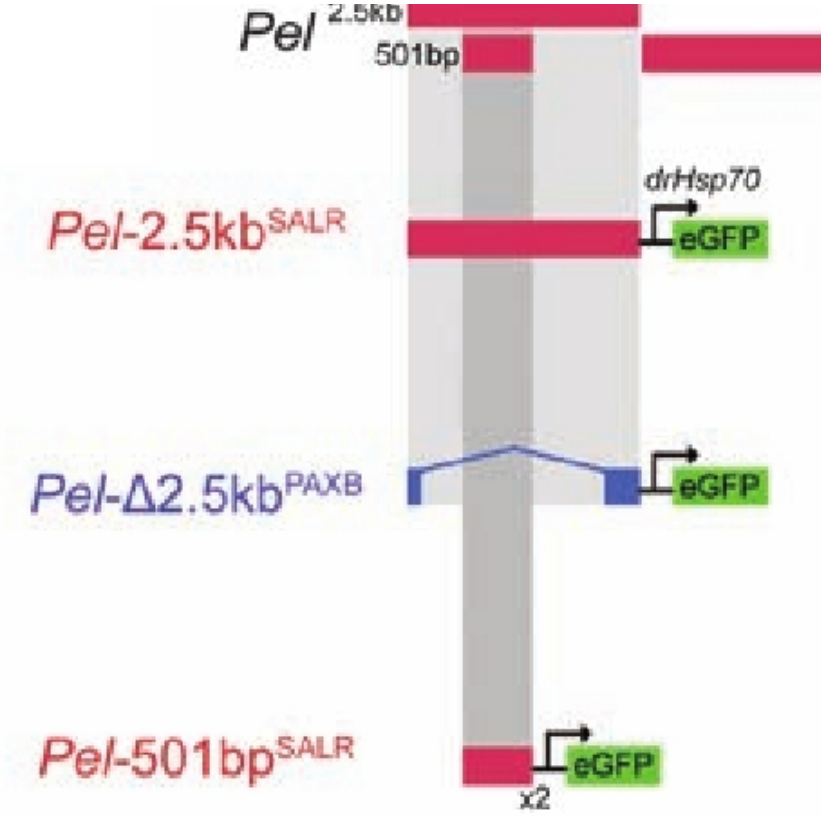
SNP genotyping in 13 populations with reduced pelvis and in 21 populations with complete pelvis

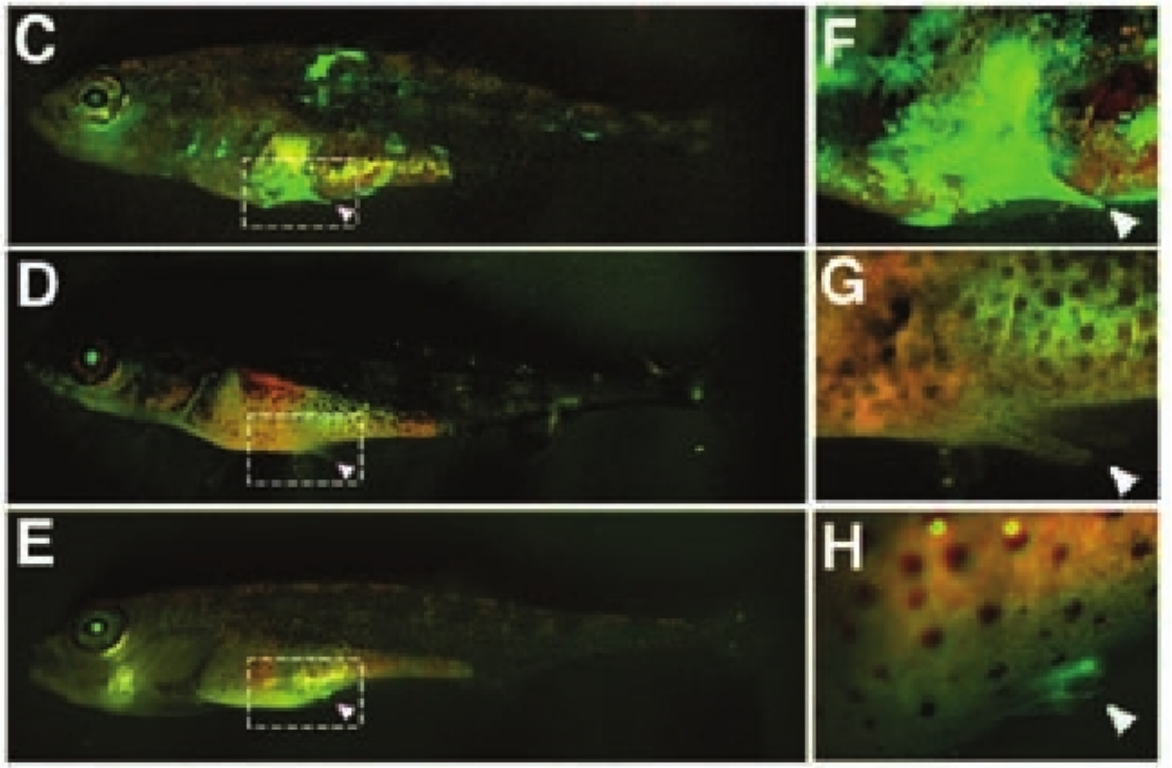




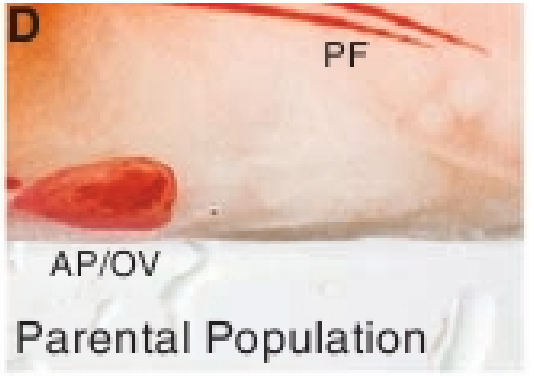
9 different deletions

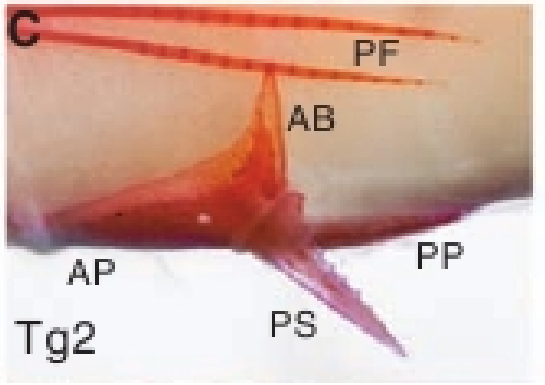
Test of *Pitx1* cis-regulatory regions



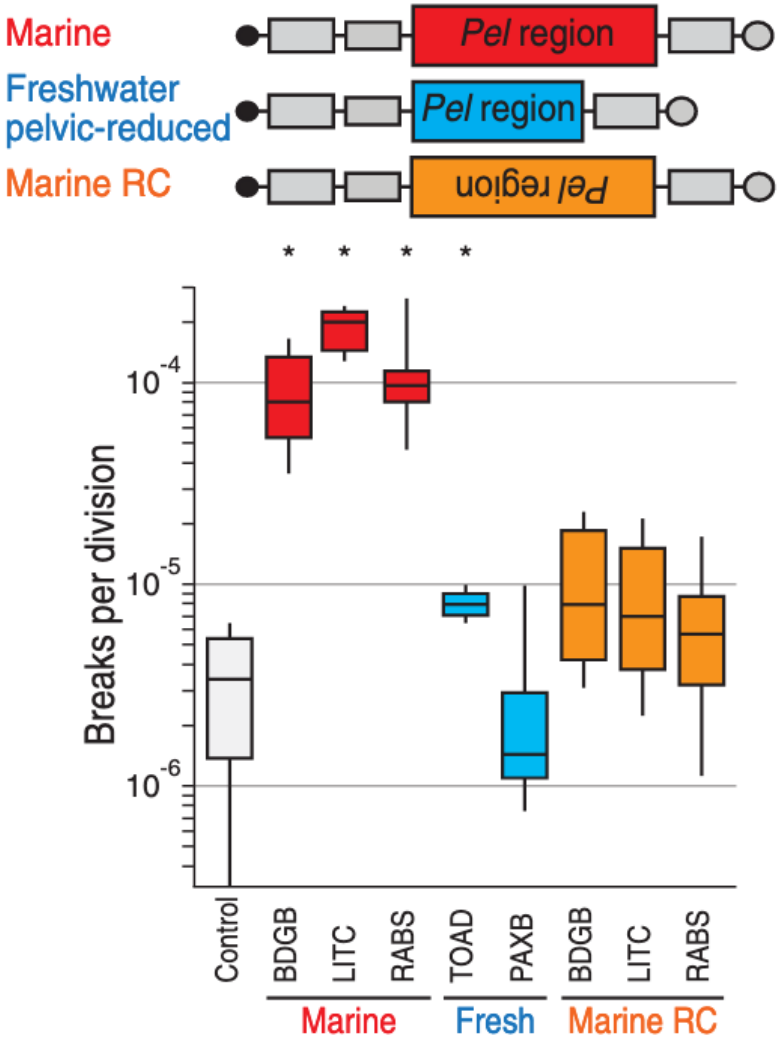


Rescue of a pelvis in freshwater individuals

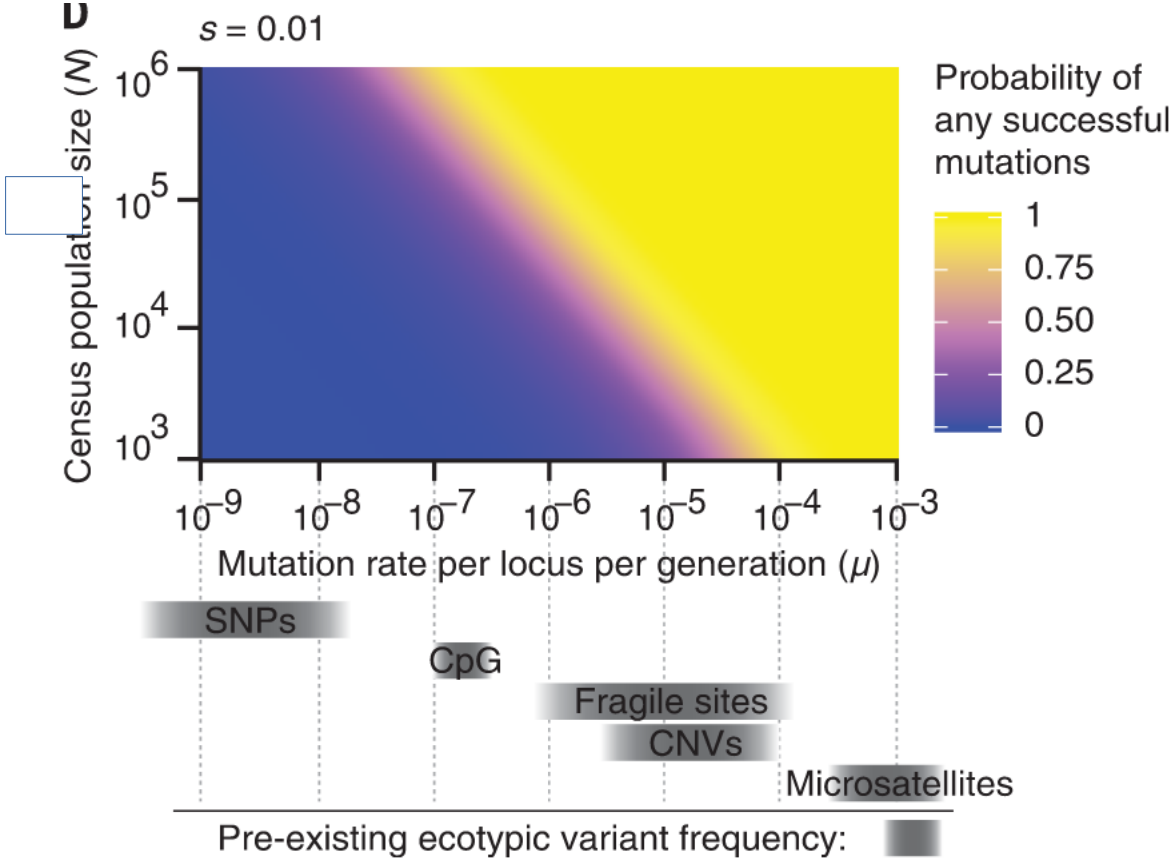




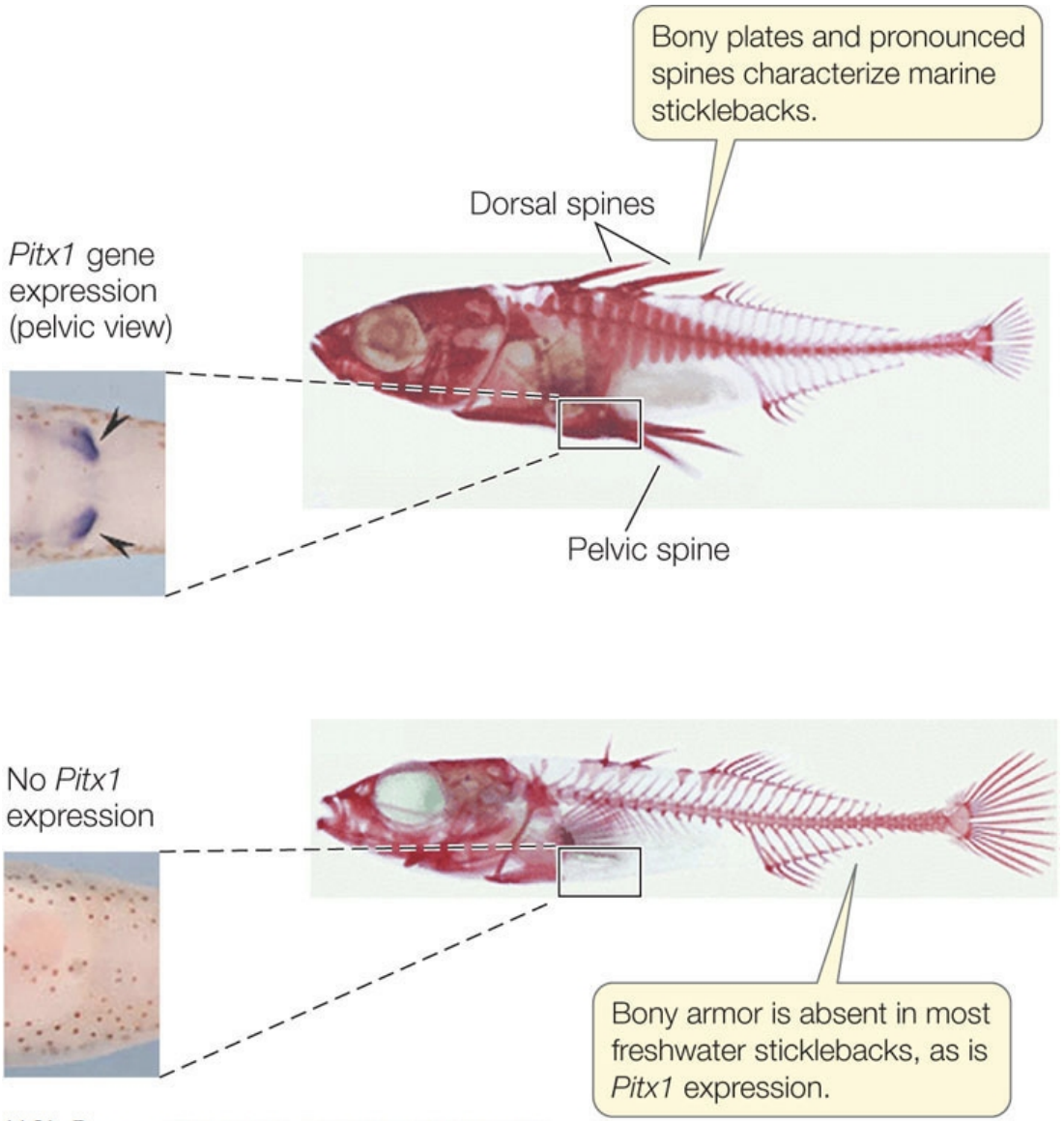
Pitx1 is in a fragile DNA region



Control = artificial chromosome without test region



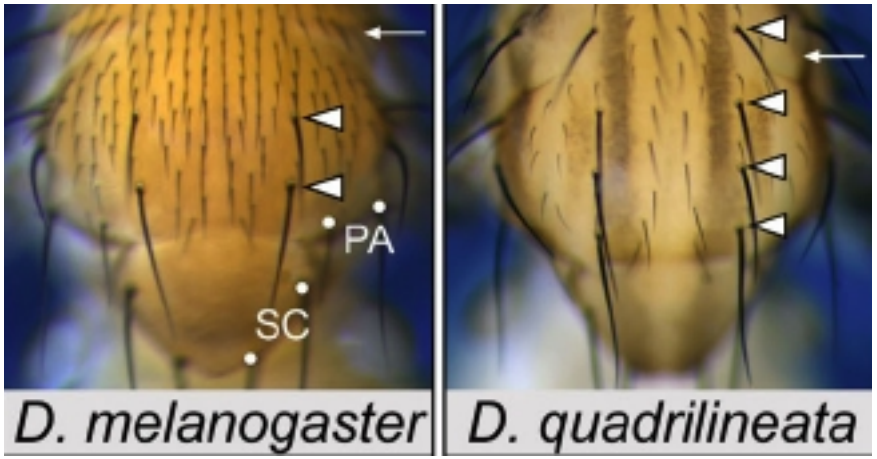
(Xie et al., 2019)



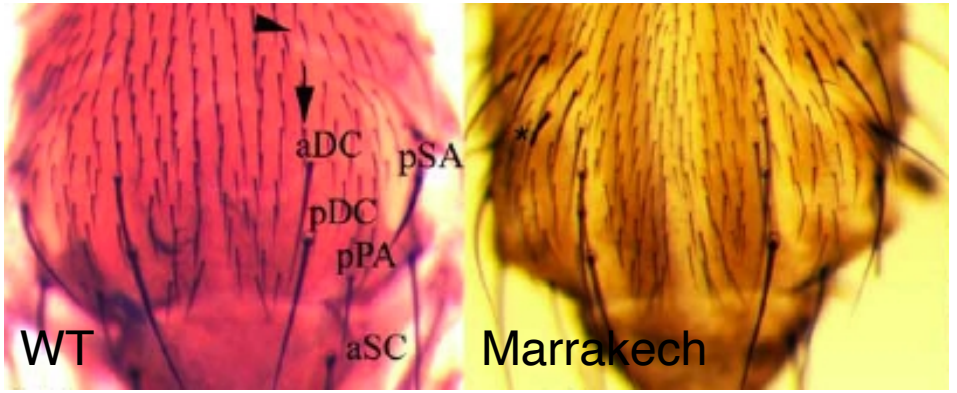
14.21: Courtesy of Mike Shapiro and David Kingsley.

Evolution of extra bristles

Interspecific change in *D. quadrilineata*

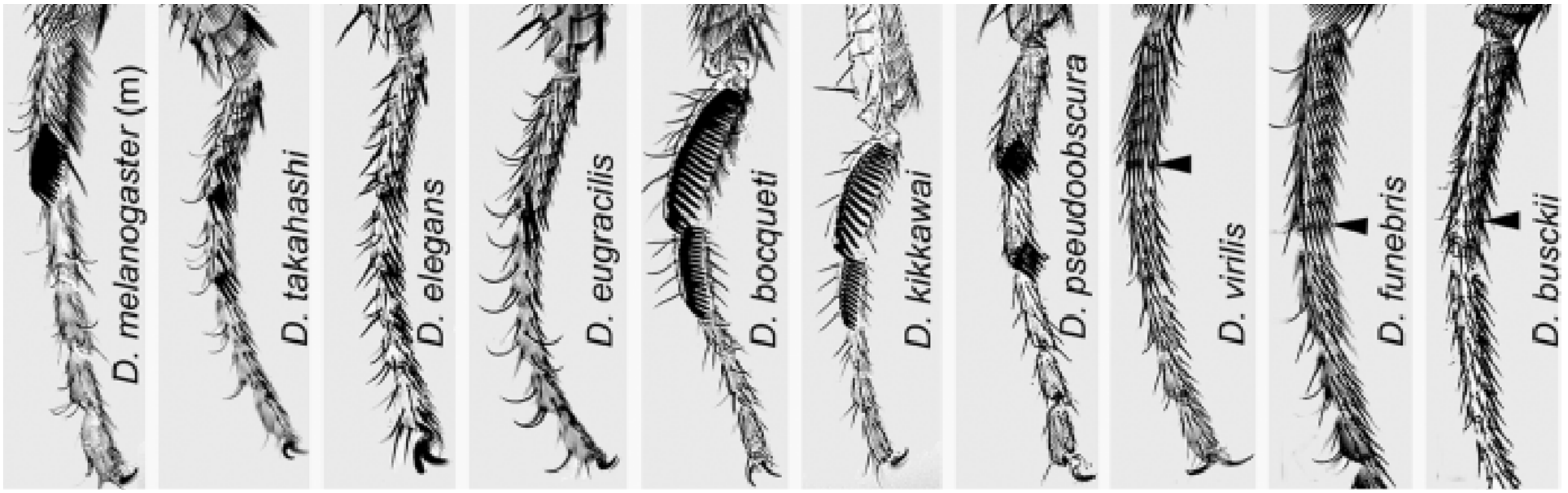


Intraspecific change in *D. melanogaster*

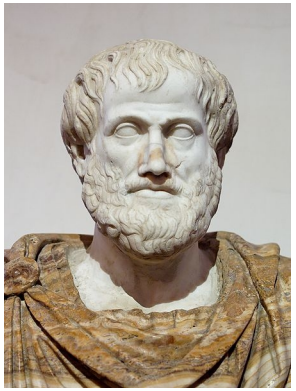


Marcellini and Simpson 2005, Gilbert et al. 2005

Finding genetic rules on bristle evolution



Randsholt and Santamaria 2008



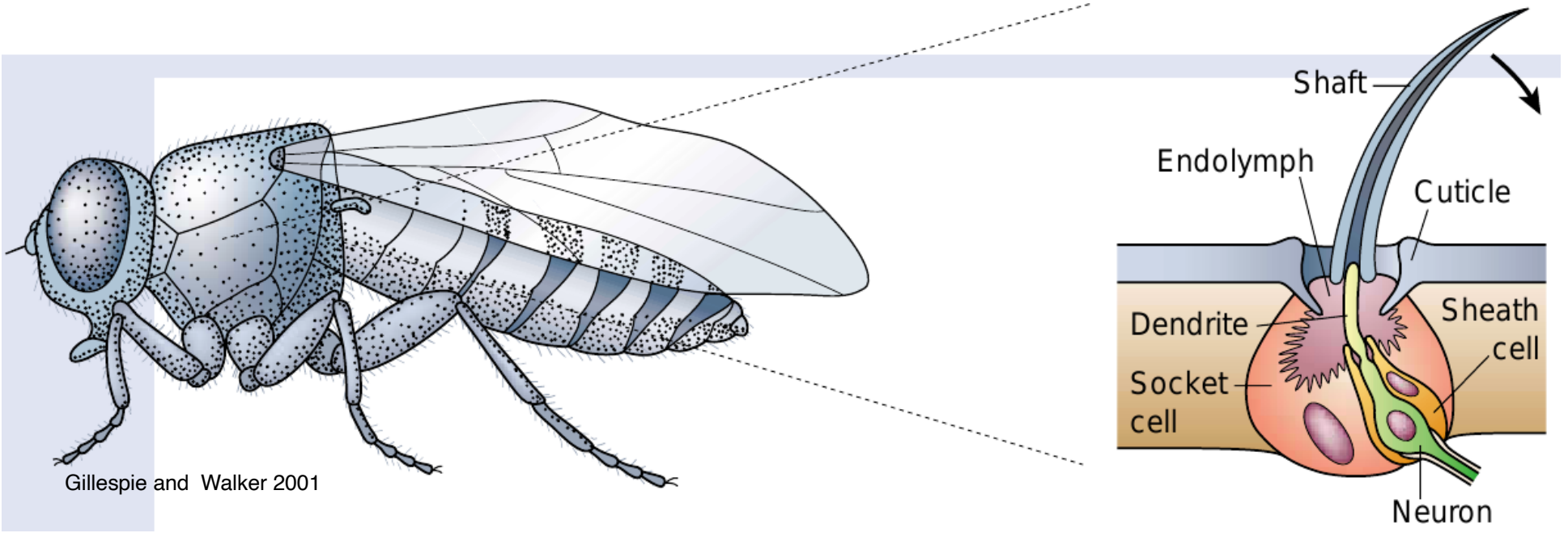
Aristotle, Historia animalium, book I, 2, 300BC

color, type, orientation
shape and size
presence/absence
position

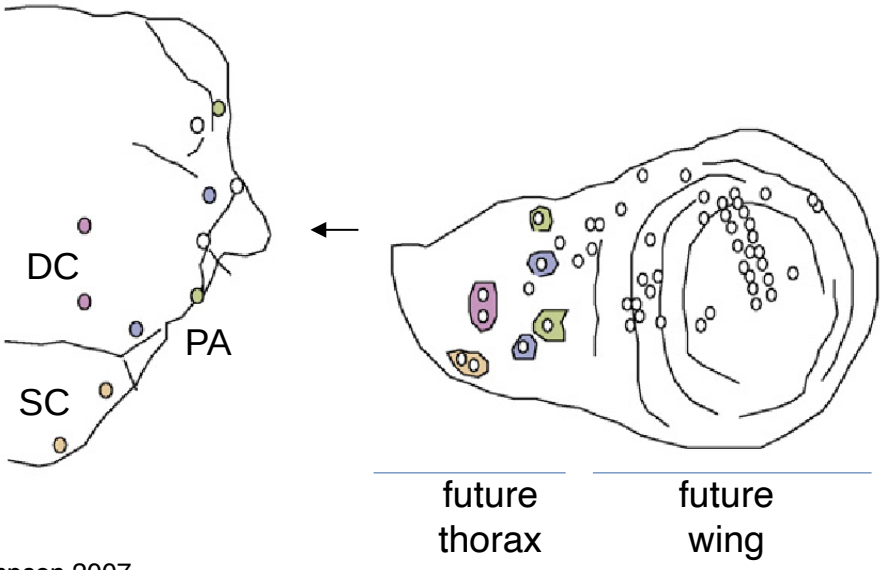
} CRE mutations in *achaete-scute*

Stern and Orgogozo 2009

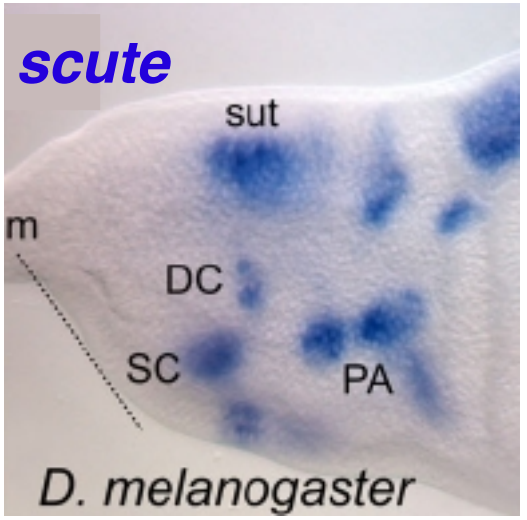
Bristle development



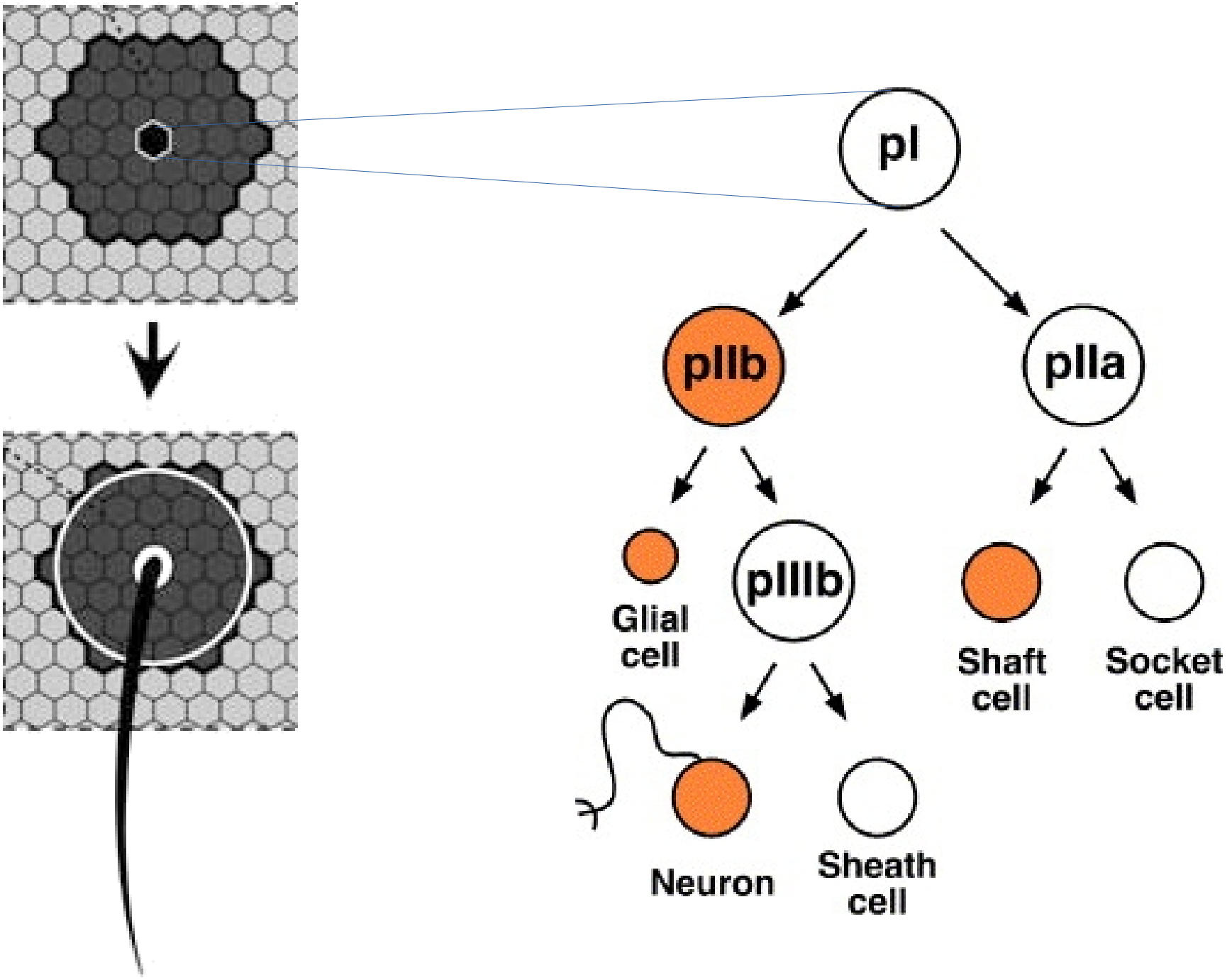
Gillespie and Walker 2001



Simpson 2007

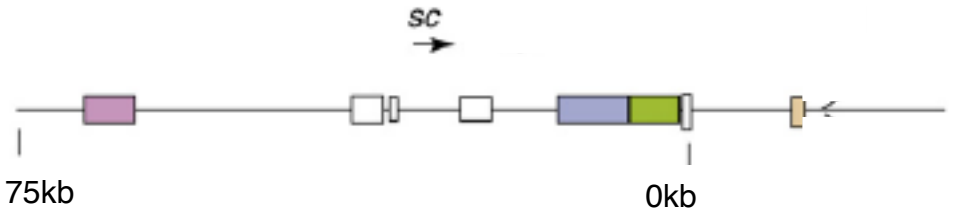


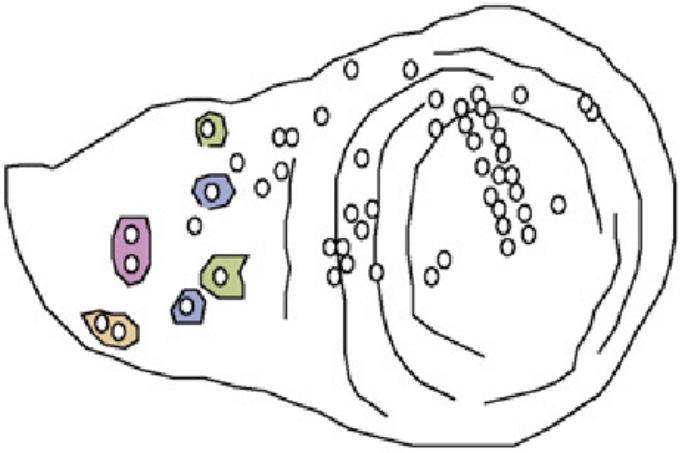
Marcellini 2006 - PloS Biology



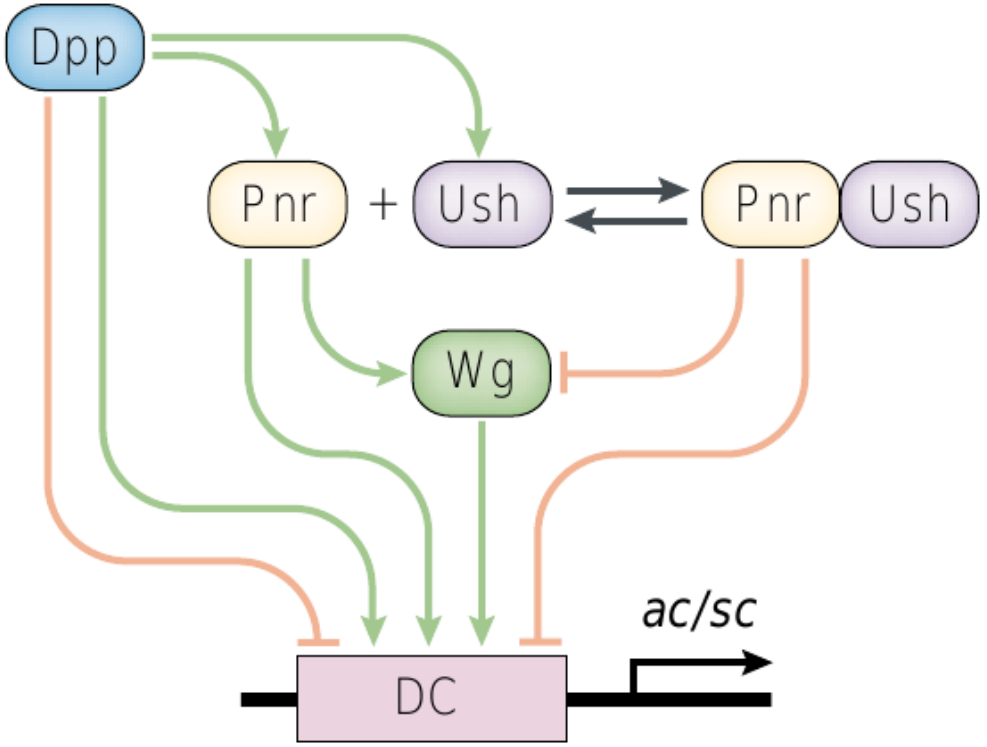
Lewis - Imaginal discs

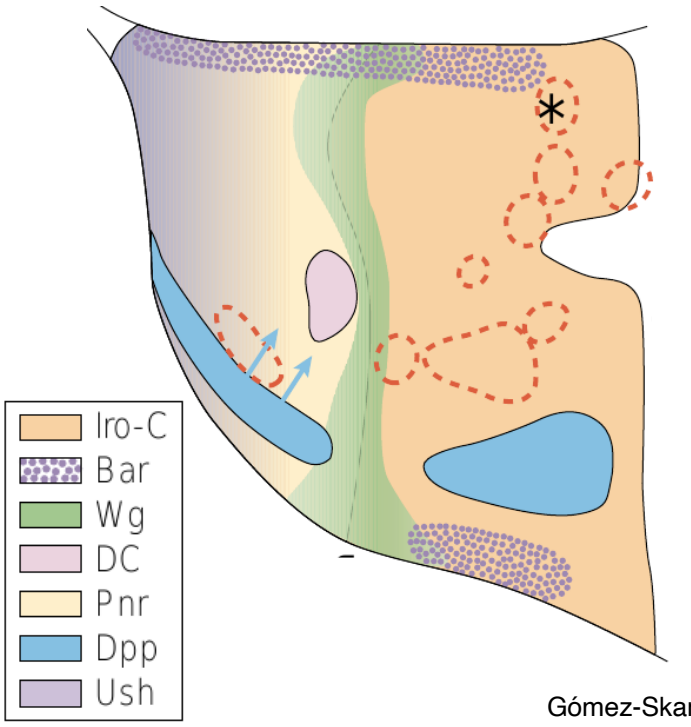
scute cis-regulatory elements are “master switches”





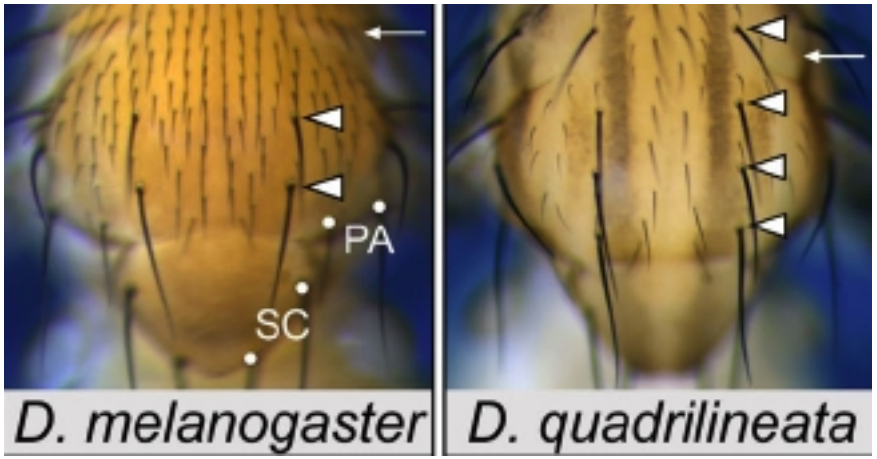
Simpson 2007

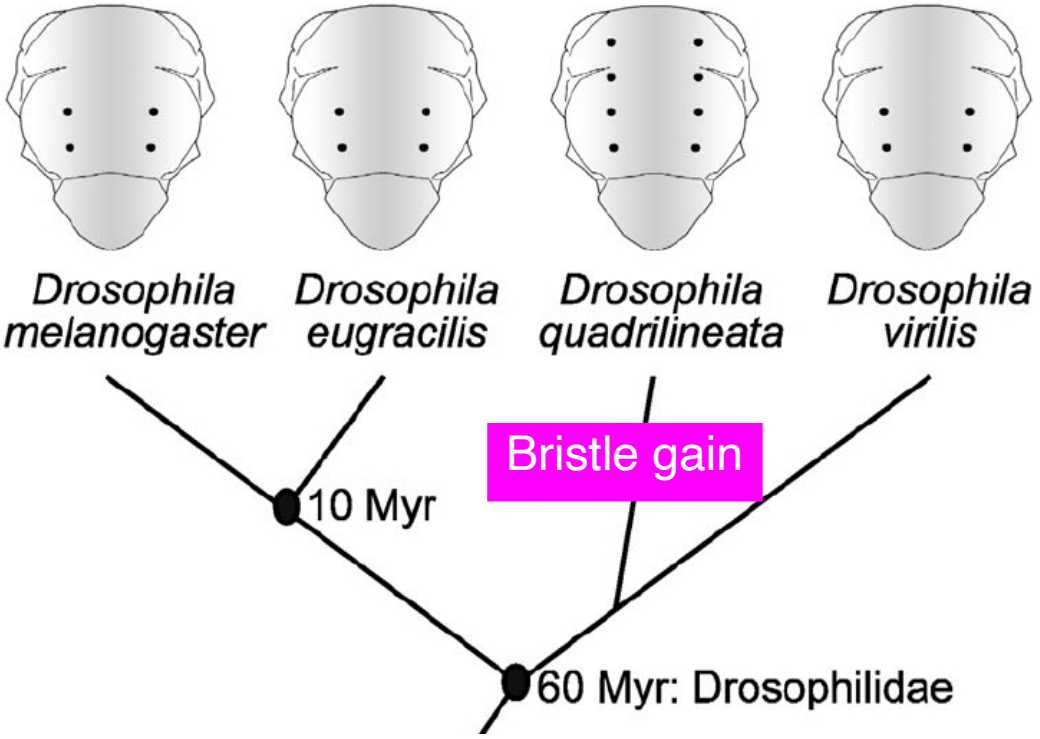




Gómez-Skarmeta 2003

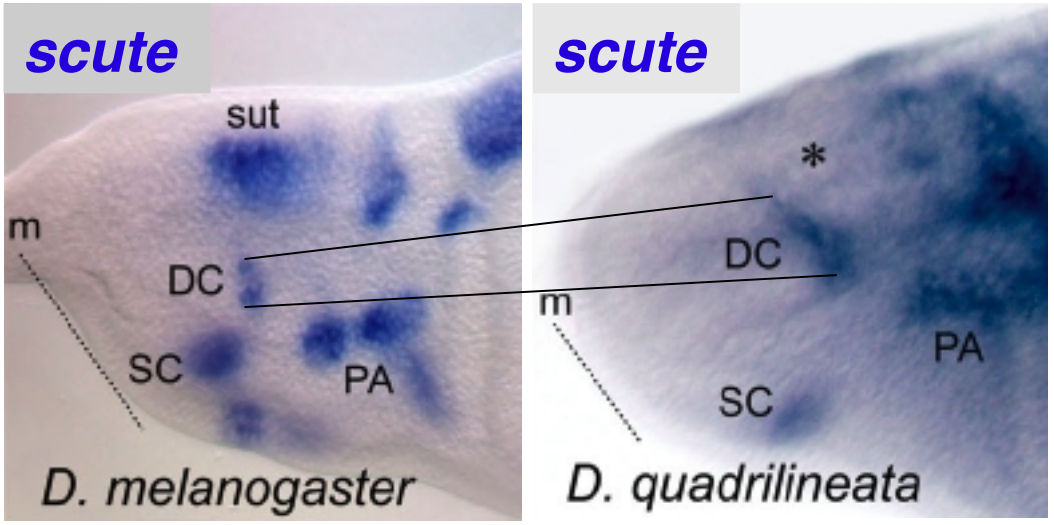
Extra bristles in *D. quadrilineata*





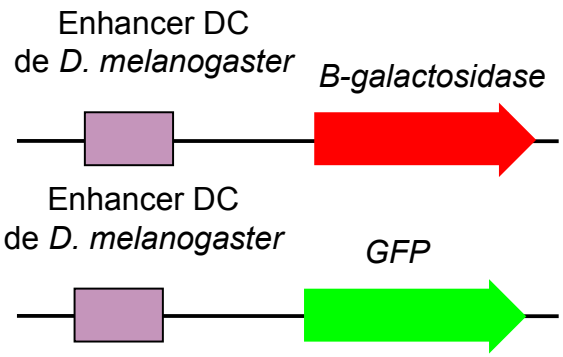
Extra bristles in *D. quadrilineata* correlate with larger *scute* expression domain

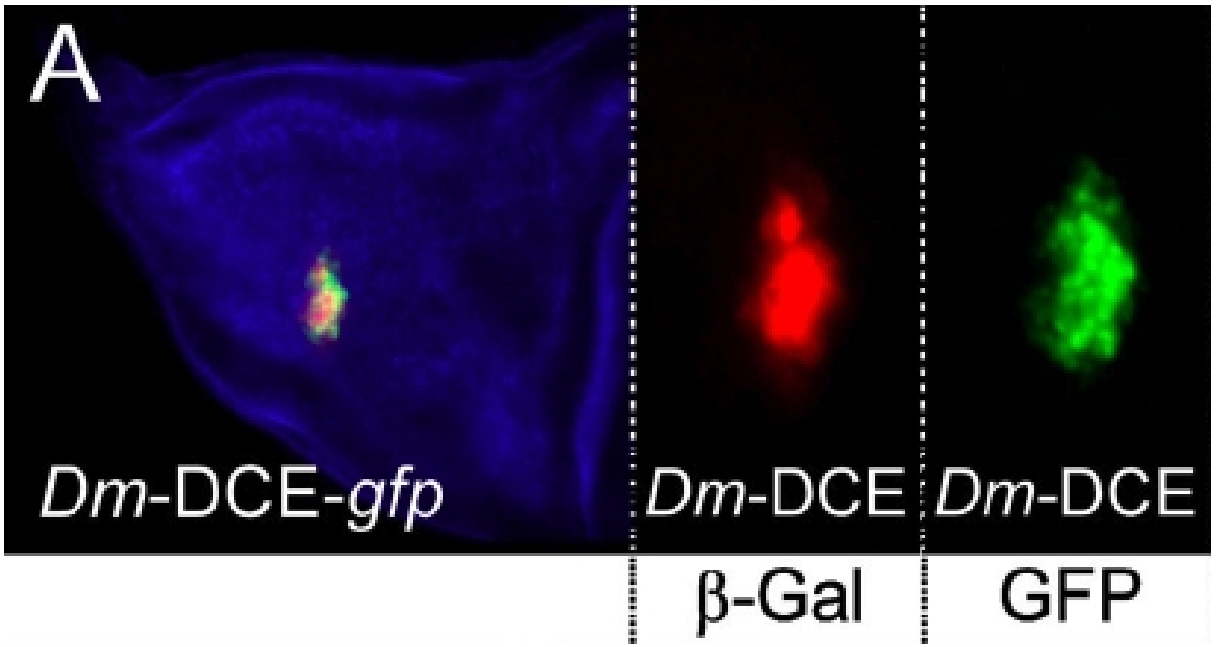
In situ hybridization

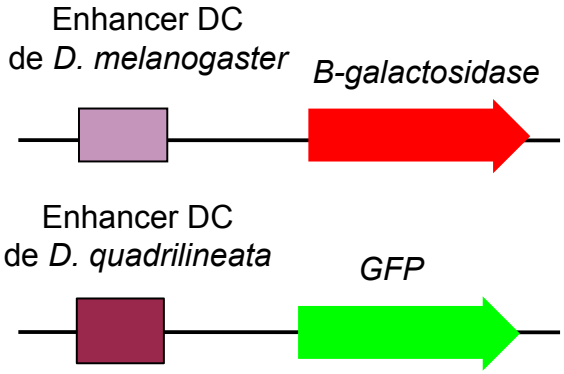


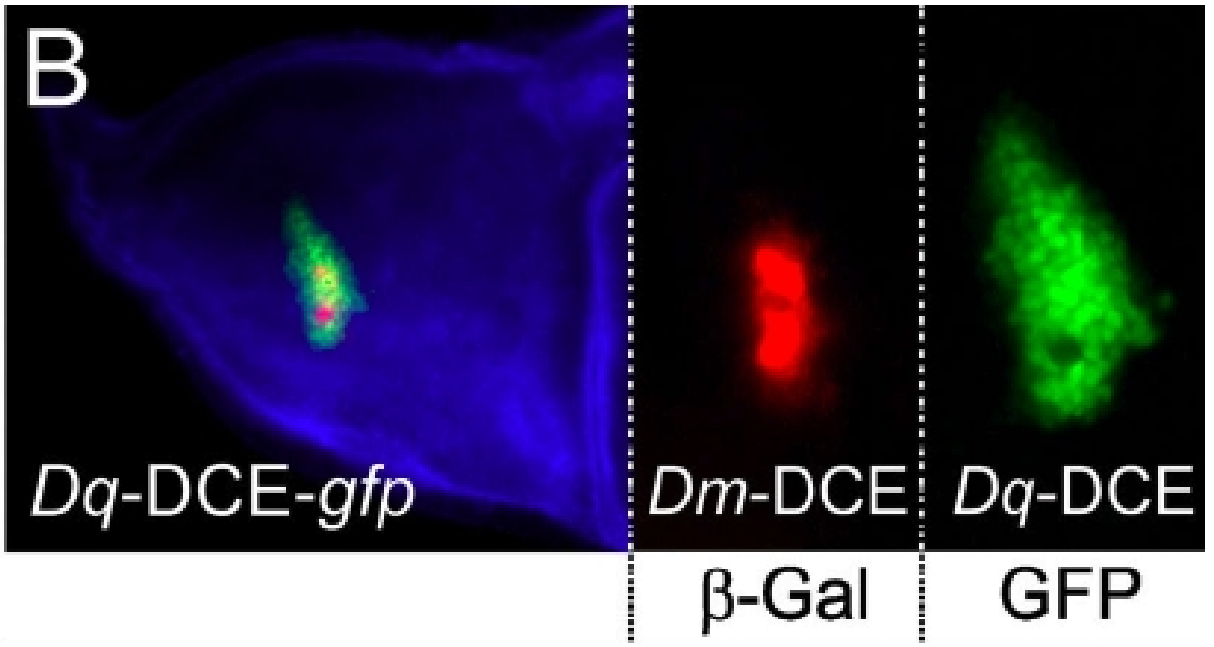
Test for a cis-regulatory change (1)

D.melanogaster
transgenics







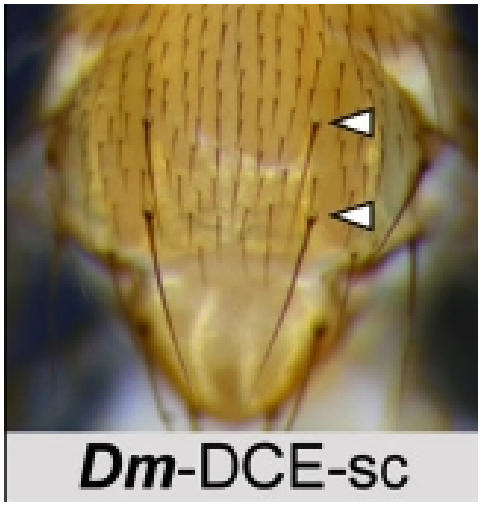


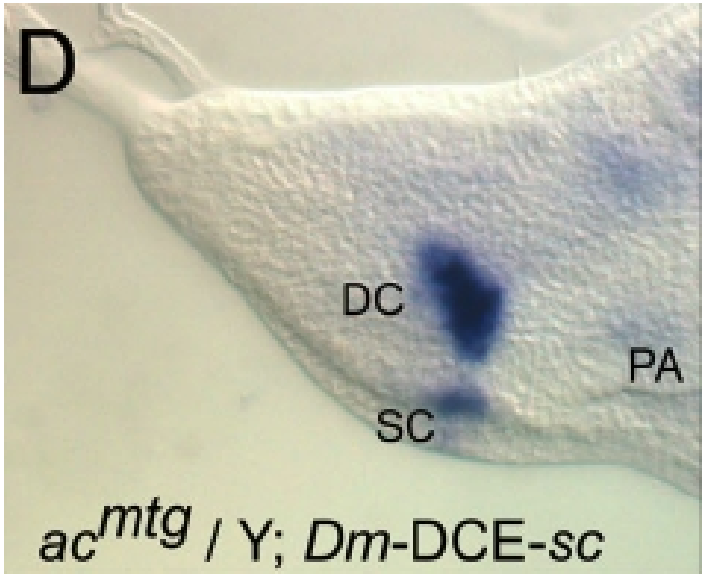
Test for a cis-regulatory change (2)

D.melanogaster
transgenics

Enhancer DC
de *D. melanogaster*

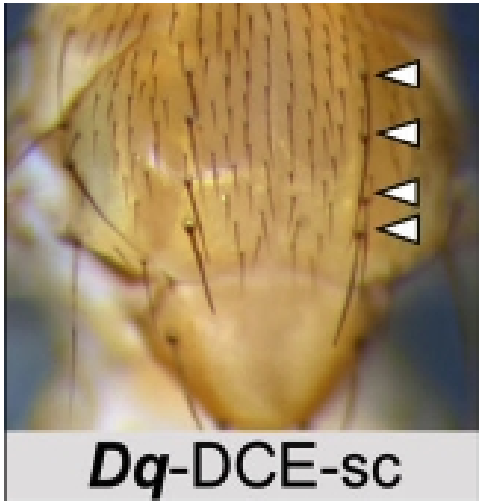
scute

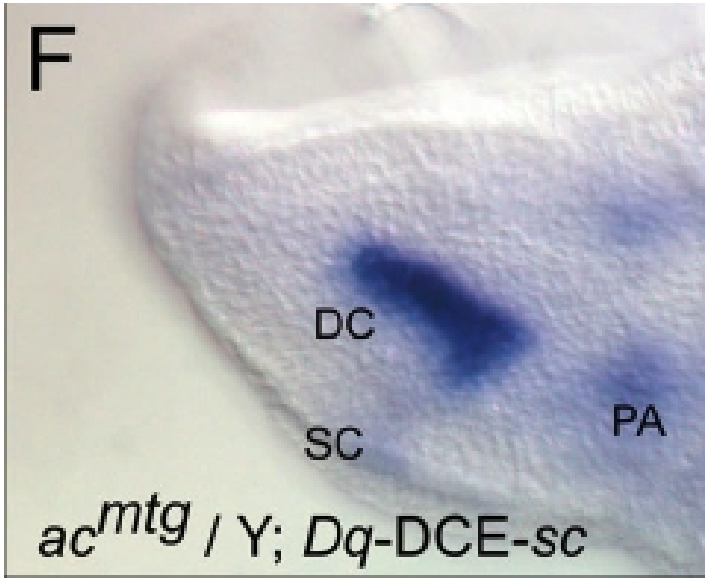




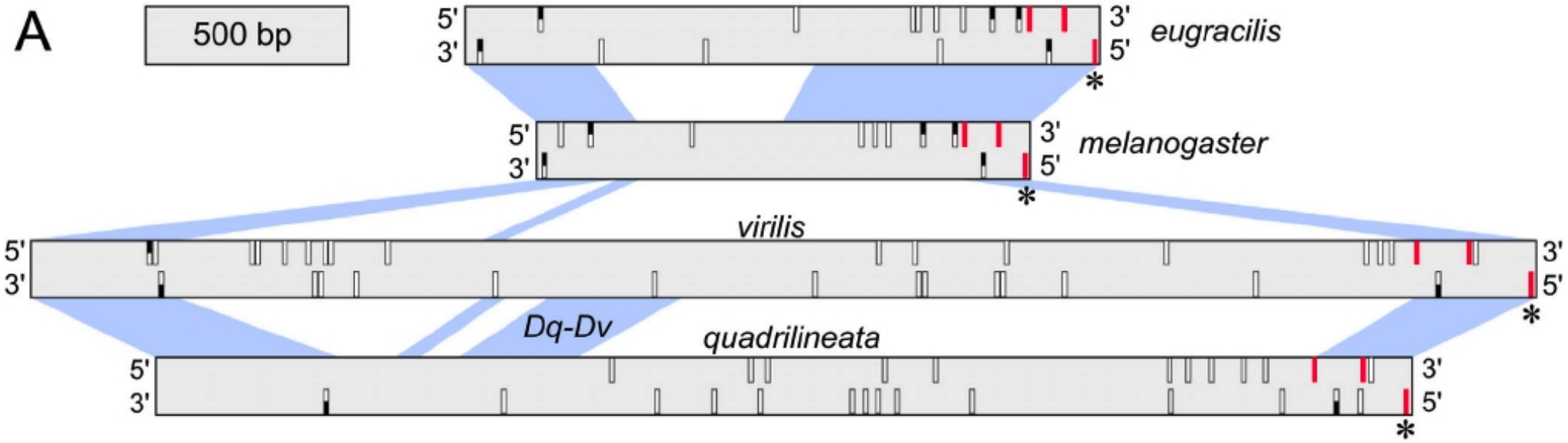
Enhancer DC
de *D. quadrilineata*

scute

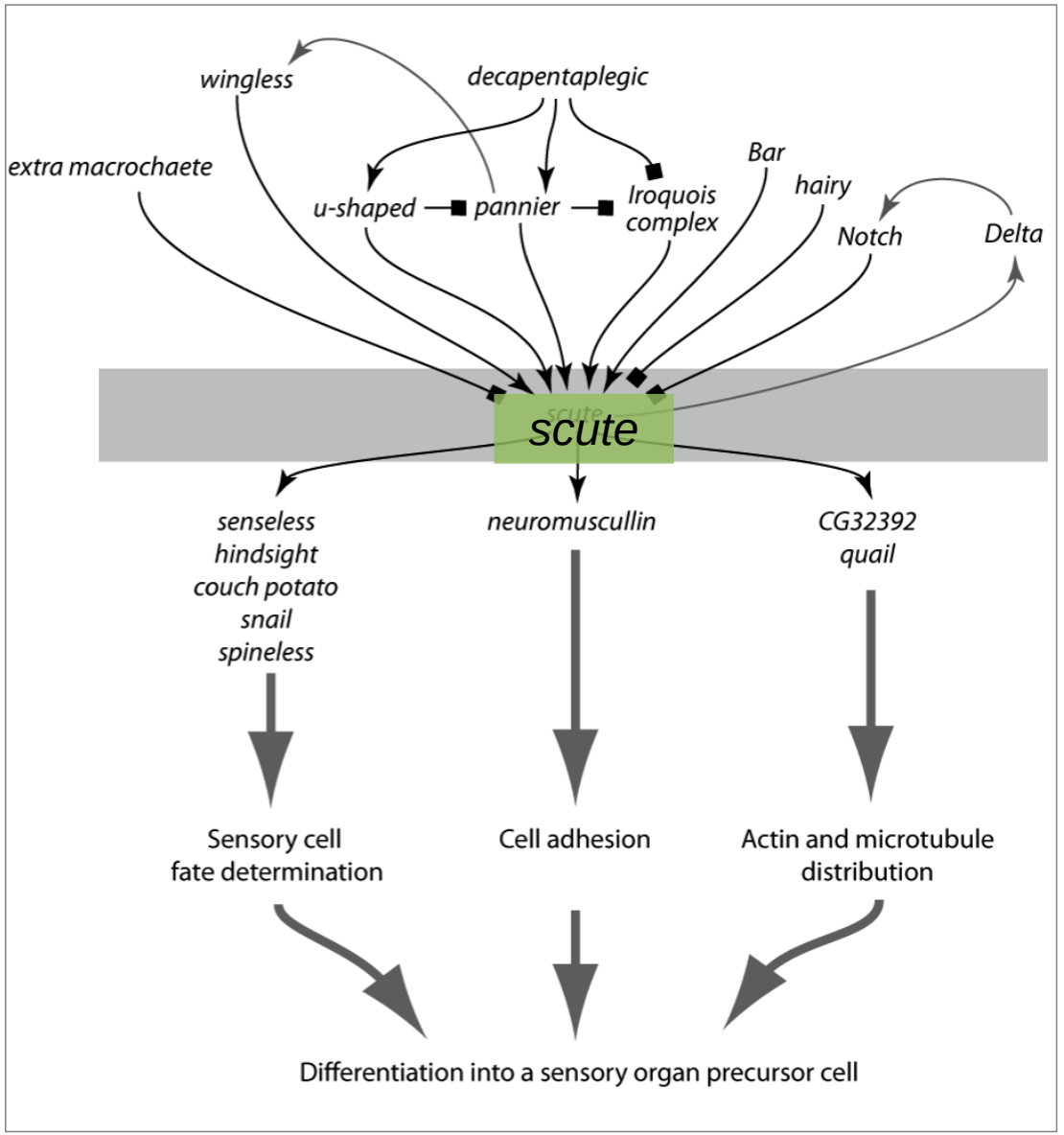


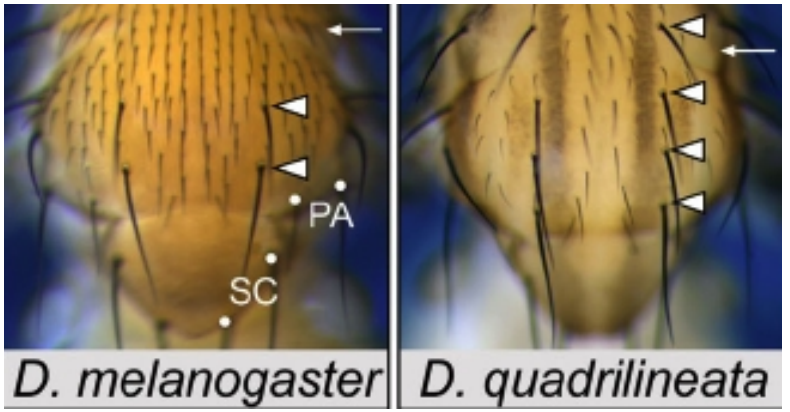


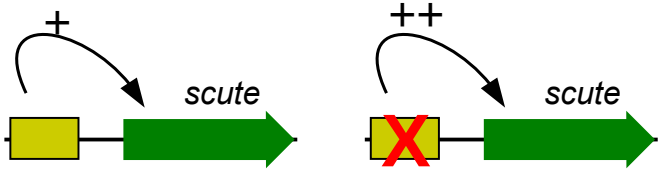
Alignment of the DC region



Genetic evolution is partly predictable

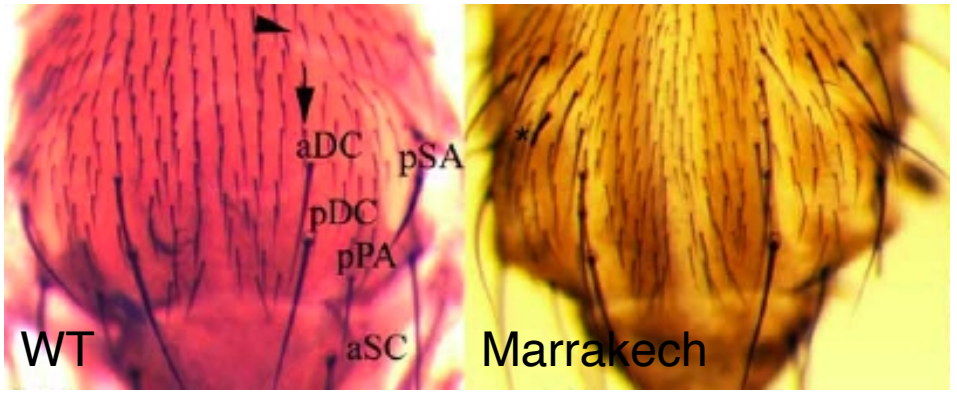




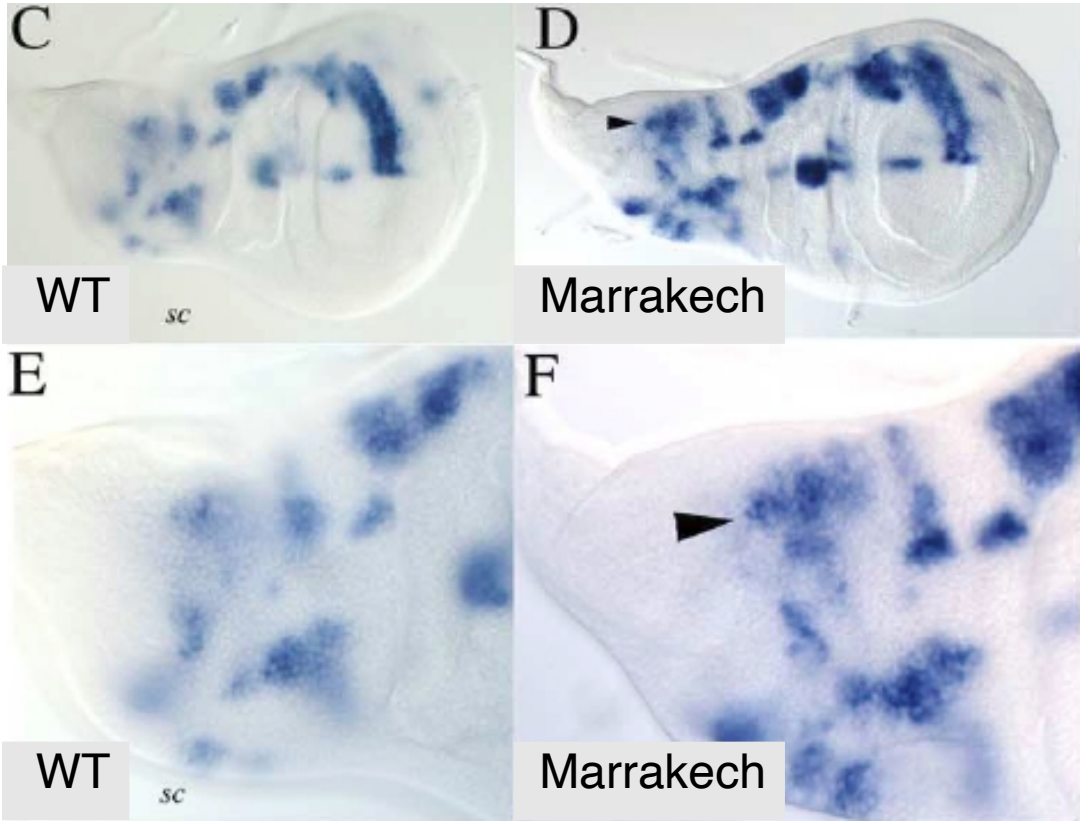


Stern and Orgogozo 2009 Science

Extra bristles in *D. melanogaster*-Marrakech correlate with larger *scute* expression domain

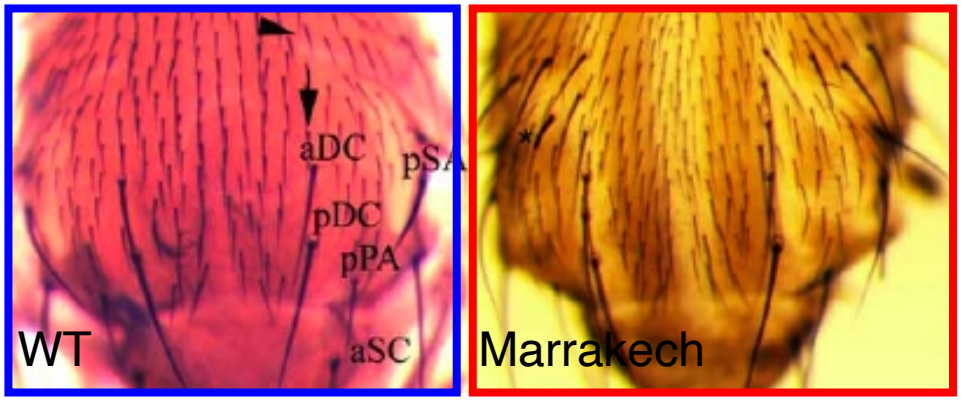


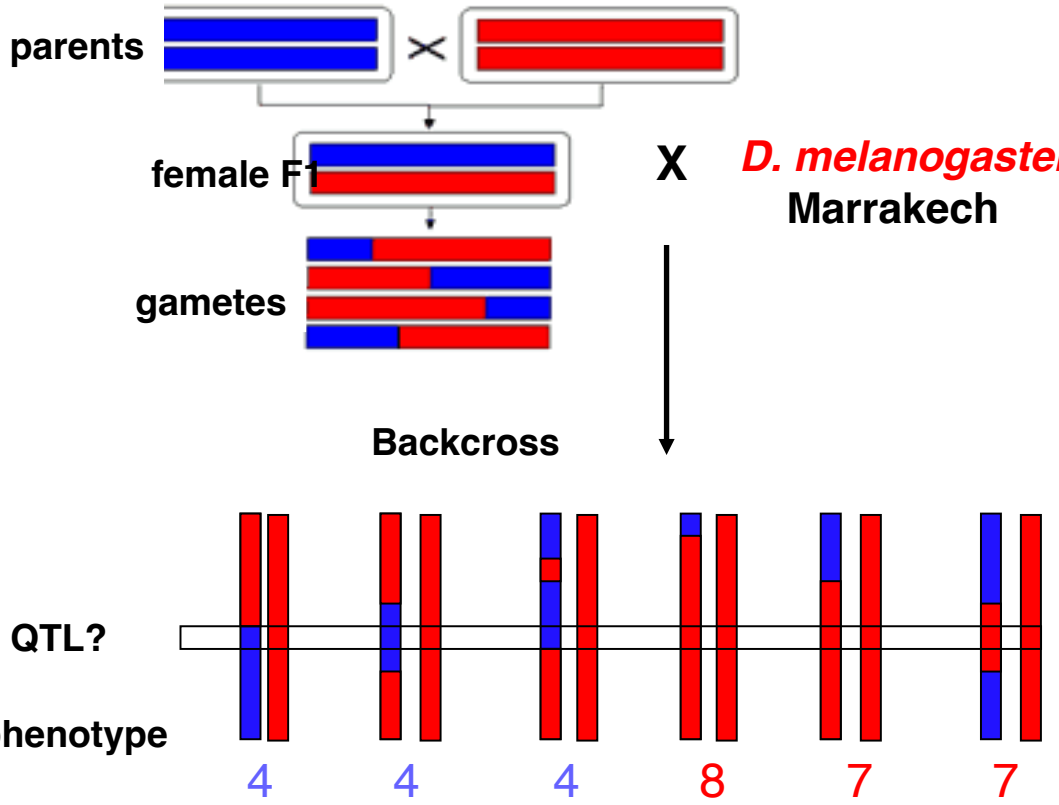
scute

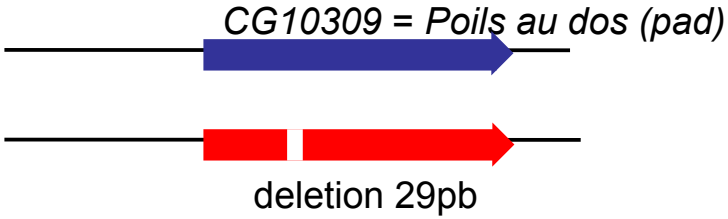


Gibert et al 2005

Extra bristles in *D. melanogaster*-Marrakech due to mutation(s) in *poils-au-dos*

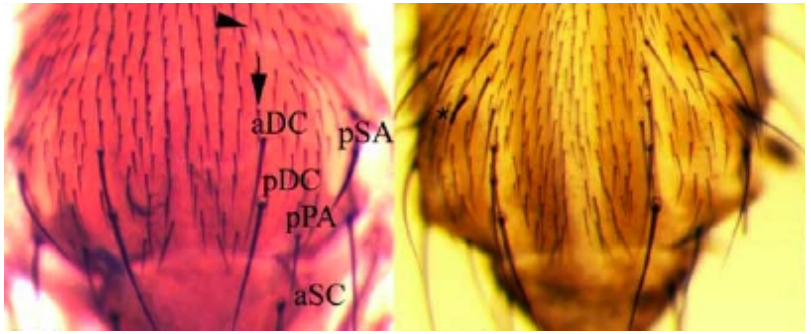




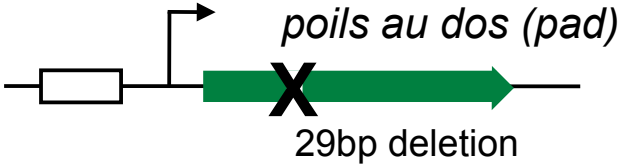


Gibert et al 2005

Short-term evolution…



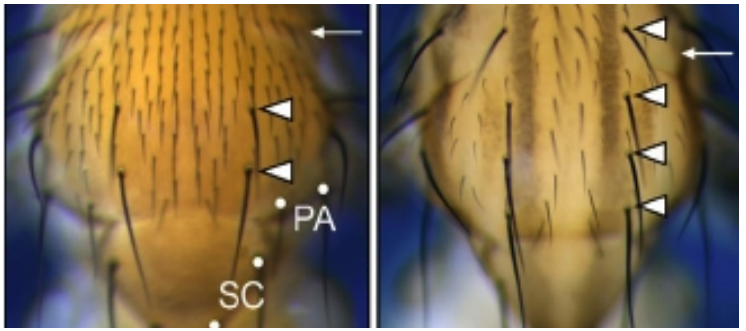
D. melanogaster ***D. melanogaster* variant**



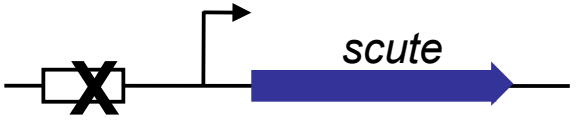
null mutation in coding region
change in thorax and wing

(Gibert et al., 2005)

…versus long-term evolution



D. melanogaster ***D. quadrilineata***



cis-regulatory mutation
change in the thorax only

(Marcellini et al. 2006)

Methods to identify the genes and the mutations responsible for phenotypic evolution

Various methods

Genetic
which chromosome (ex: autosomal versus sex)
QTL mapping
Genetic association studies
Complementation tests

General biology
General knowledge of the genes involved in the phenotype
Similarity with a known phenotype
Correlation with a change in gene expression level/pattern

Final test of protein activity
in vitro in *E. coli*, by transgenesis in the studied species or the closest model organism (ex: *beta-defensin* of dogs tested in mouse)

Final test of cis-regulatory regions
- with reporter constructs, transgenesis, comparison of both regions
- comparison of allele expression levels in hybrids (pyrosequencing)

Two types of approaches

Genetic mapping

Candidate gene

no a priori, fewer bias
long and tedious
rarely ends with identification of the gene

only with strains/species which produce fertile hybrids

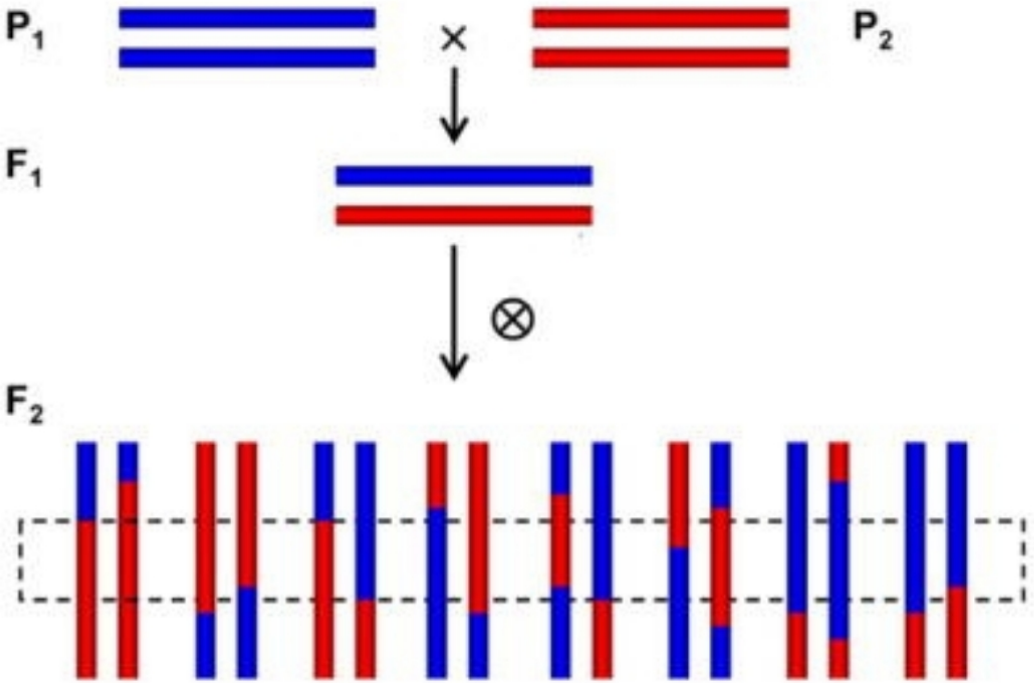
Based on an a priori idea
can be fast and efficient

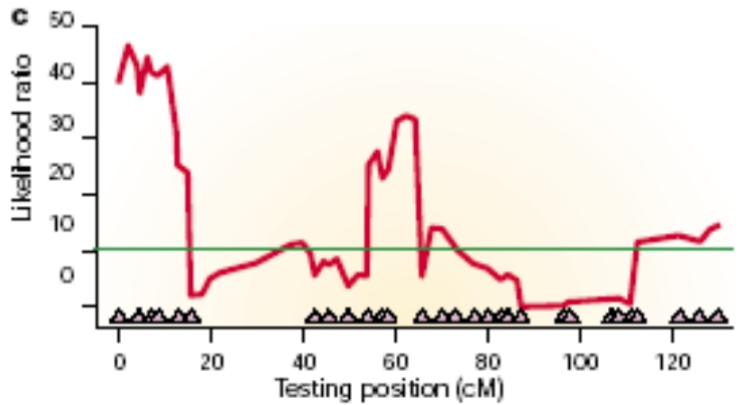
will only find known genes

<u>In both cases, genes with small effect are more difficult to identify</u>

Linkage Mapping

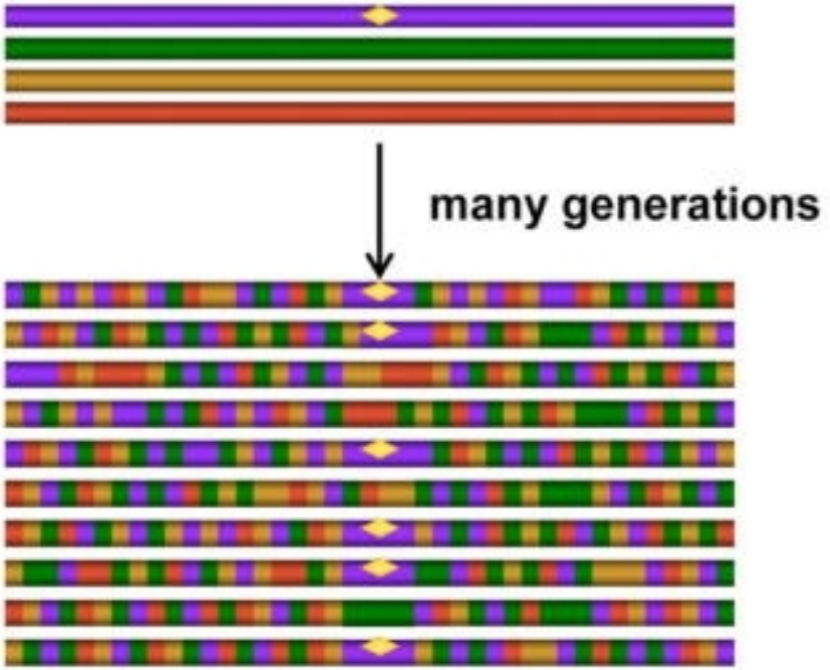
Crosses in the lab

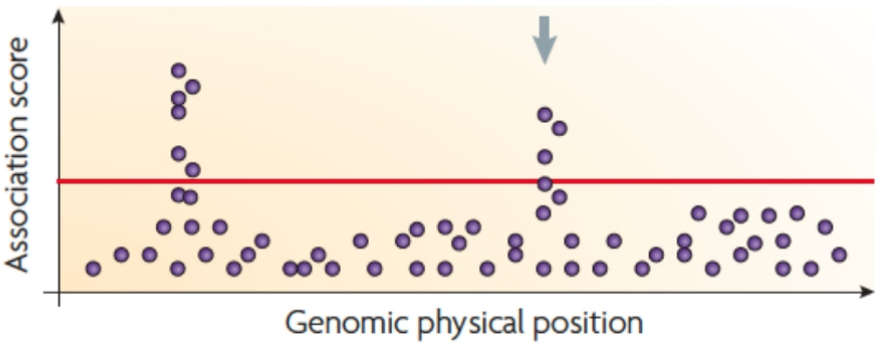




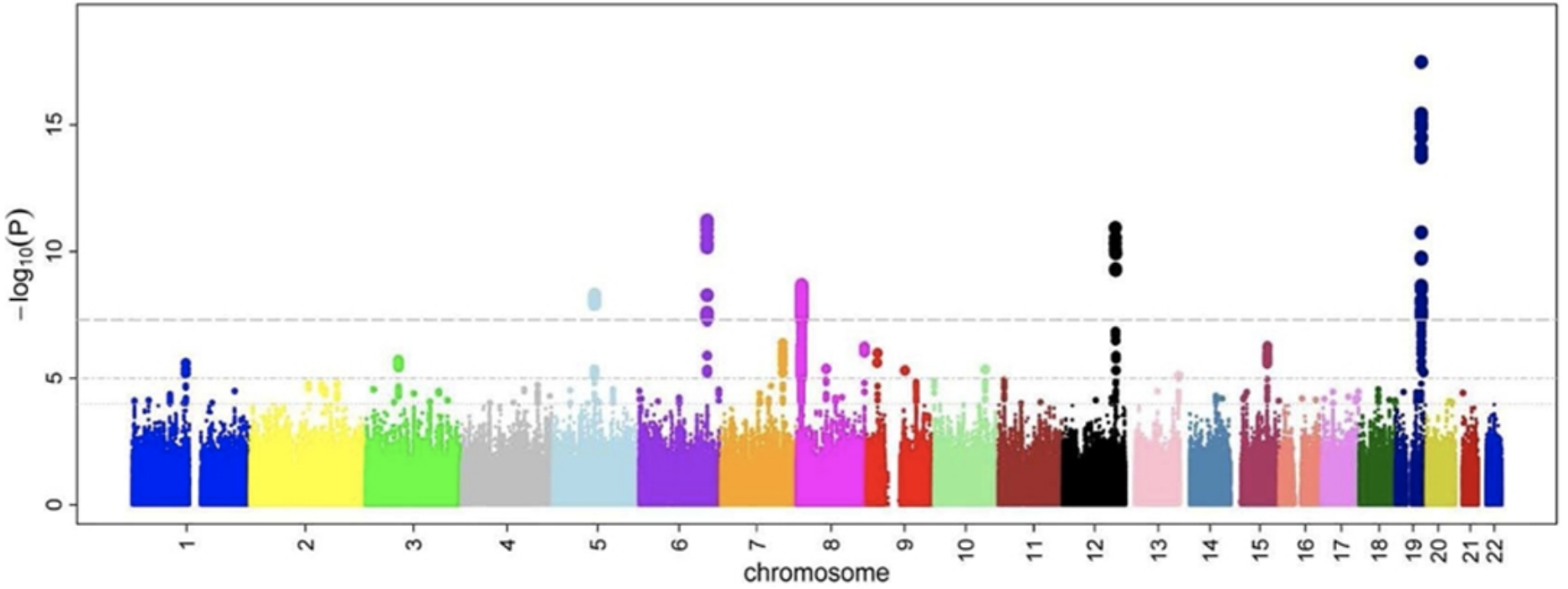
Association Mapping

Past crosses in natural populations





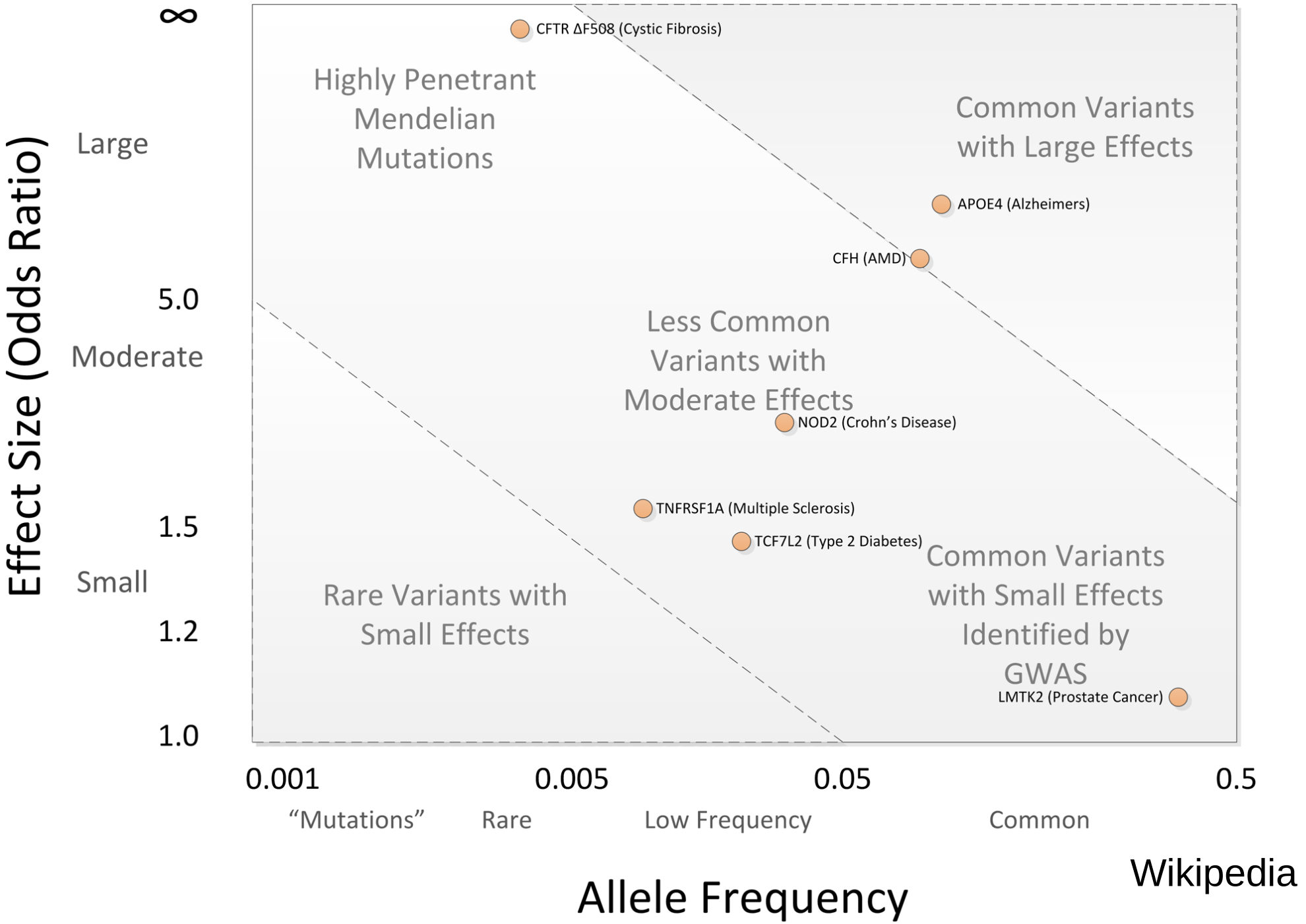
Genome-wide association study (GWAS)



Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level.

Wikipedia

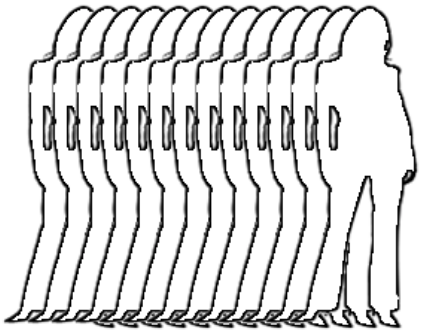
GWAS typically identify common alleles



Wikipedia

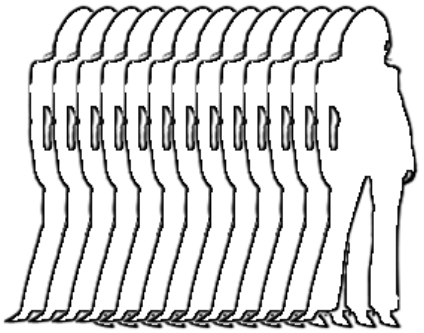
Methodology of a case-control GWA study

Coronary artery disease



GC CC GG GC CC GC GC
GG CC GC GG GC GG

No disease



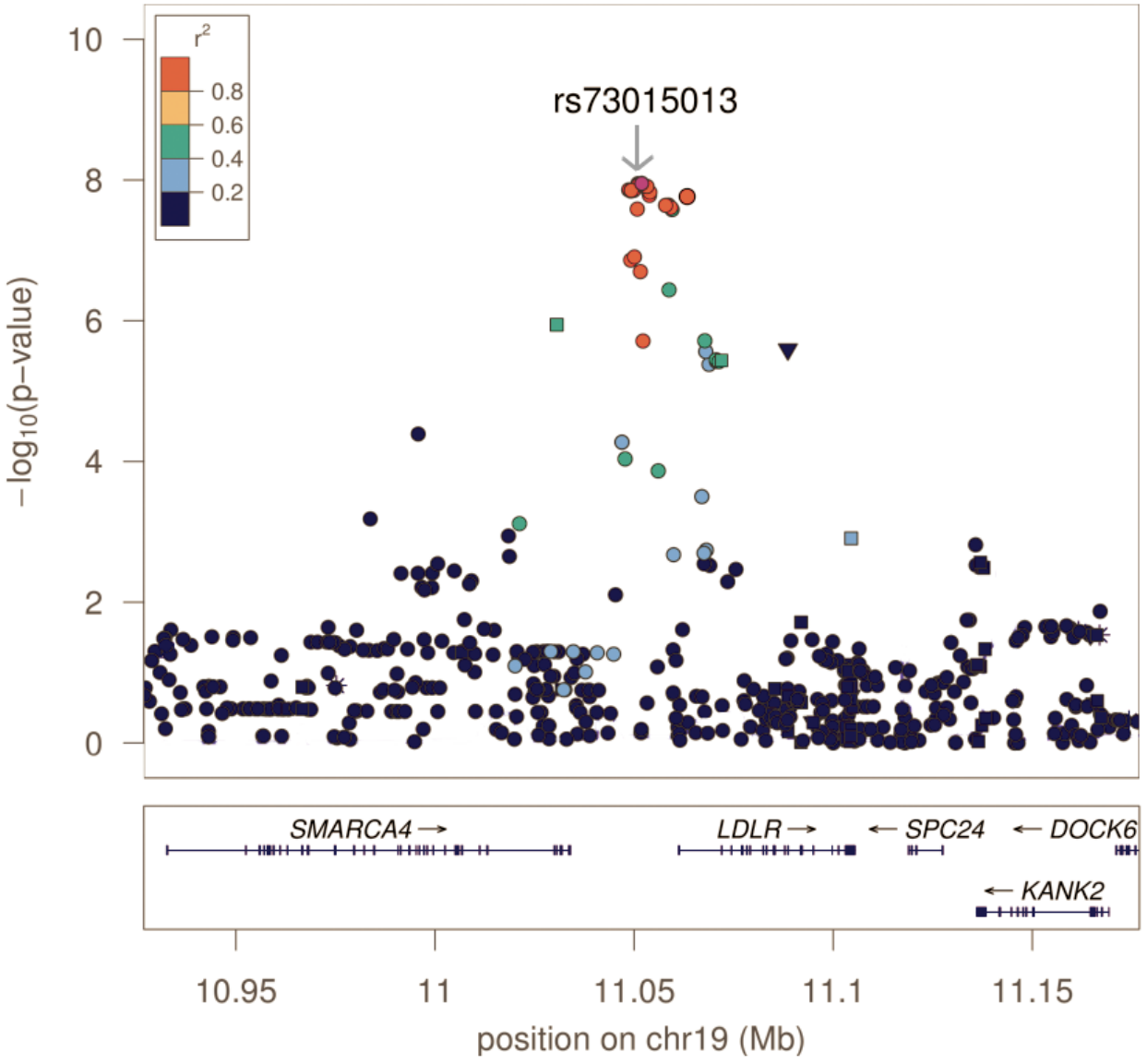
GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1	*SNP2*	*SNP …*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

The allele count of each measured SNP is evaluated, in this case with a chi-squared test, to identify variants associated with the trait in question.

Wikipedia

Regional association plot

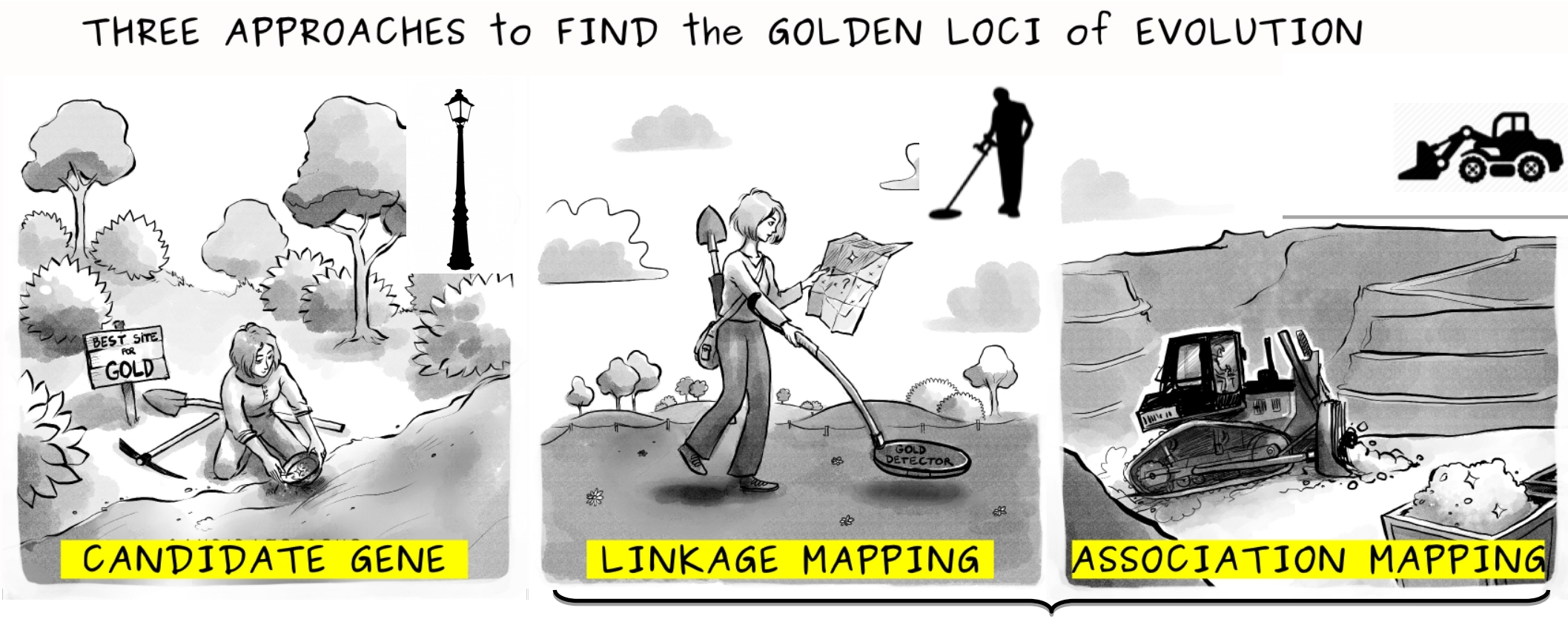


Association to LDL-cholesterol levels.

The haploblock structure is visualized with colour scale and the association level is given by the left Y-axis.

The dot representing the rs73015013 SNP (in the top-middle) has a high Y-axis location because this SNP explains some of the variation in LDL-cholesterol.

Wikipedia



REVERSE GENETICS

From genes to traits

FORWARD GENETICS

From traits to genes

Little Ascertainment Bias, but Micro-Evolution only

Courtier-Orgogozo et al. 2020 NAR

Experimental Evidence

3 categories, each with biases

	Candidate Gene	Linkage Mapping	Association Mapping
Experimental Principle	**Reverse Genetics**: looking for sequence differences and trait effects based on previous studies of a given gene	**Forward Genetics**: trait mapping in hybrids obtained from laboratory crosses, using recombination over a few generations	**Forward Genetics**: statistical SNP/character state association in large cohorts, using recombination over many generations
Example	66 cases of color variation associated to *MC1R* coding mutations in vertebrates	F2 crosses between melanic and amelanic phenotypes in cavefish : identification of *MC1R* and *Oca2* alleles in distinct cave populations	GWAS of human pigmentation (skin, hair, eyes): identification and confirmation of causal variants at >15 genes including *Oca2* p.His615Arg in Eastern Asia
Ascertainment Bias on Locus Identification	**High**	**Low to Intermediate** (depending on resolution / cross size)	**Low**
Molecular Type Bias	Favors identification of **coding mutations**	Little molecular bias	Can miss structural variants (short read genotyping)
Trait Type Bias	Favors traits with small molecular targets, large-effect size	Amenable to dissection of complex traits with small-effect size (large crosses, multiparental families)	Most common approach for complex traits with small-effect size
Taxonomic Breadth	**Large**	**Narrow**, limited to interfertile lineages (populations or sister species)	**Very narrow**, limited to polymorphic or intermixing populations

QTL Mapping

4 steps: crosses, genotyping, phenotyping, statistical analysis

Crosses

Backcross with one line
Backcross in both directions
F2
Crosses for several generations
Introgression lines
Recombinant Inbred Lines
…

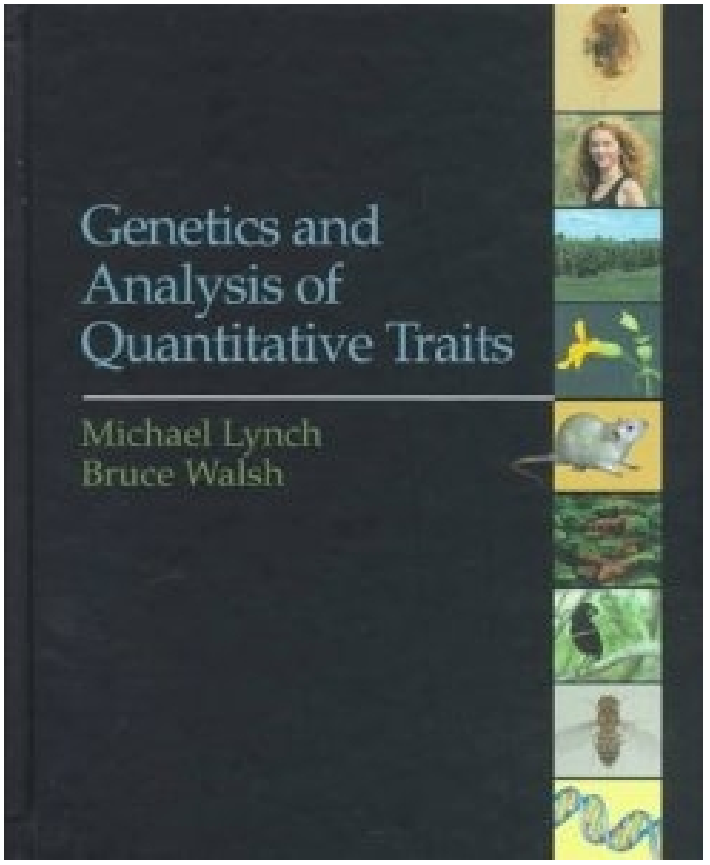
Always try to maximize the number of recombination events

Markers

yes-no PCR
PCR length polymorphism
Pyrosequencing
Probe hybridization
Microarray
RADseq
High-throughput sequencing

How many markers?

theory



practice