Lectures on Molecular Population Genetics & Genomics

Thomas Flatt & Margot Paris

Department of Biology
Ecology & Evolution
University of Fribourg

Three lectures

1. Genetic variability and its measurement, with a focus on DNA sequence variability.

2. Recombination and characterization of population structure with individual-based methods.

3. Detecting selection using sequencing data.

Acknowledgements

We gratefully acknowledge Vitor Sousa,

Brian Charlesworth, Laurent Excoffier and Dmitri Petrov

for sharing slides and/or exercises.

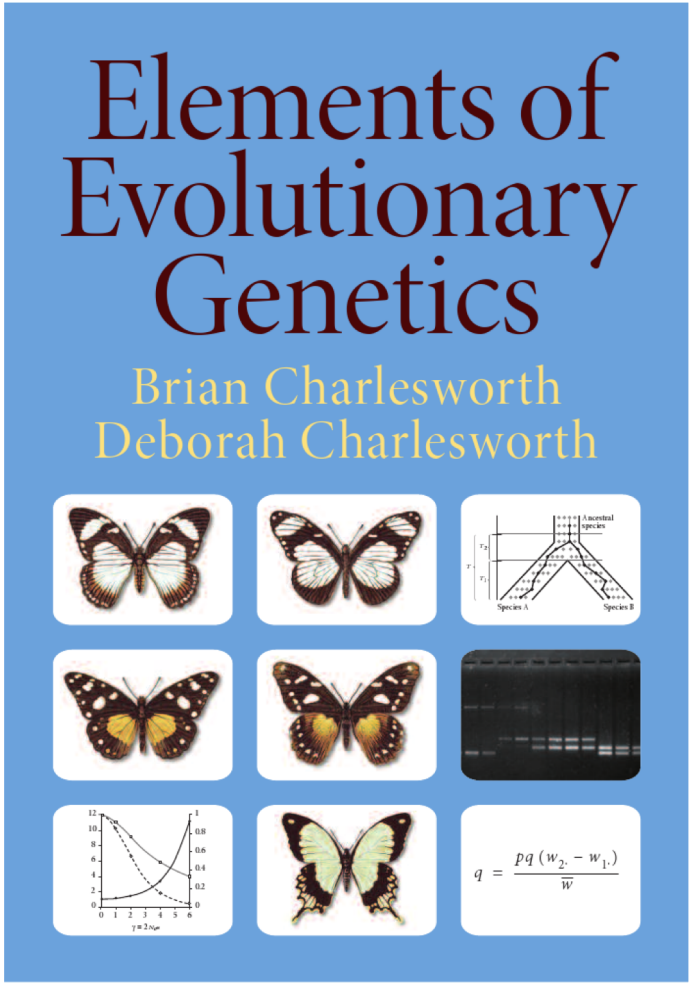
(1) Genetic variability and its measurement

(with a focus on DNA sequence variability)

Thomas Flatt

Department of Biology
Ecology & Evolution
University of Fribourg

Page references for further reading



Page references are often made on the lecture slides to the textbook *Elements of Evolutionary Genetics*, by Brian and Deborah Charlesworth (2010; Roberts & Company, Greenwood Village, CO, USA).

(*The lecture here, on genetic variability, relies heavily on this book and on a lecture by Brian Charlesworth*)

Preview of lecture 1

In this lecture we will take a closer look at some central questions of population genetics, with a focus on DNA sequence variability:

- How much genetic variation is there within and between populations?
- How can we quantify genetic variability and genetic differentiation between populations, especially using genomic (sequencing) data on DNA sequence variability?
- How can we build some simple models to understand variability?

Why is intra-specific variability interesting?

"A high degree of variability is obviously favourable,
as freely giving the materials for selection to work on…".
(Charles Darwin, 1859, *The Origin of Species,* Chapter 1)

Darwin was the first person to recognize clearly that evolutionary change over time is the result of processes acting on genetically controlled variability among individuals within a population, which eventually cause differences between ancestral and descendant populations.

Knowledge of the nature and causes of this variability is crucial for an understanding of the mechanisms of evolution, animal and plant breeding, and human genetic diseases (*cf.* GWAS).

Structure of the lecture

1. A brief discussion of evidence concerning variability derived from classical and quantitative genetics

2. An account of how variability can be detected and quantified using DNA sequencing ("population genomics")

3. An account of how to understand DNA sequence variability in terms of the interaction between mutation and genetic drift

4. How to measure genetic differentiation between populations

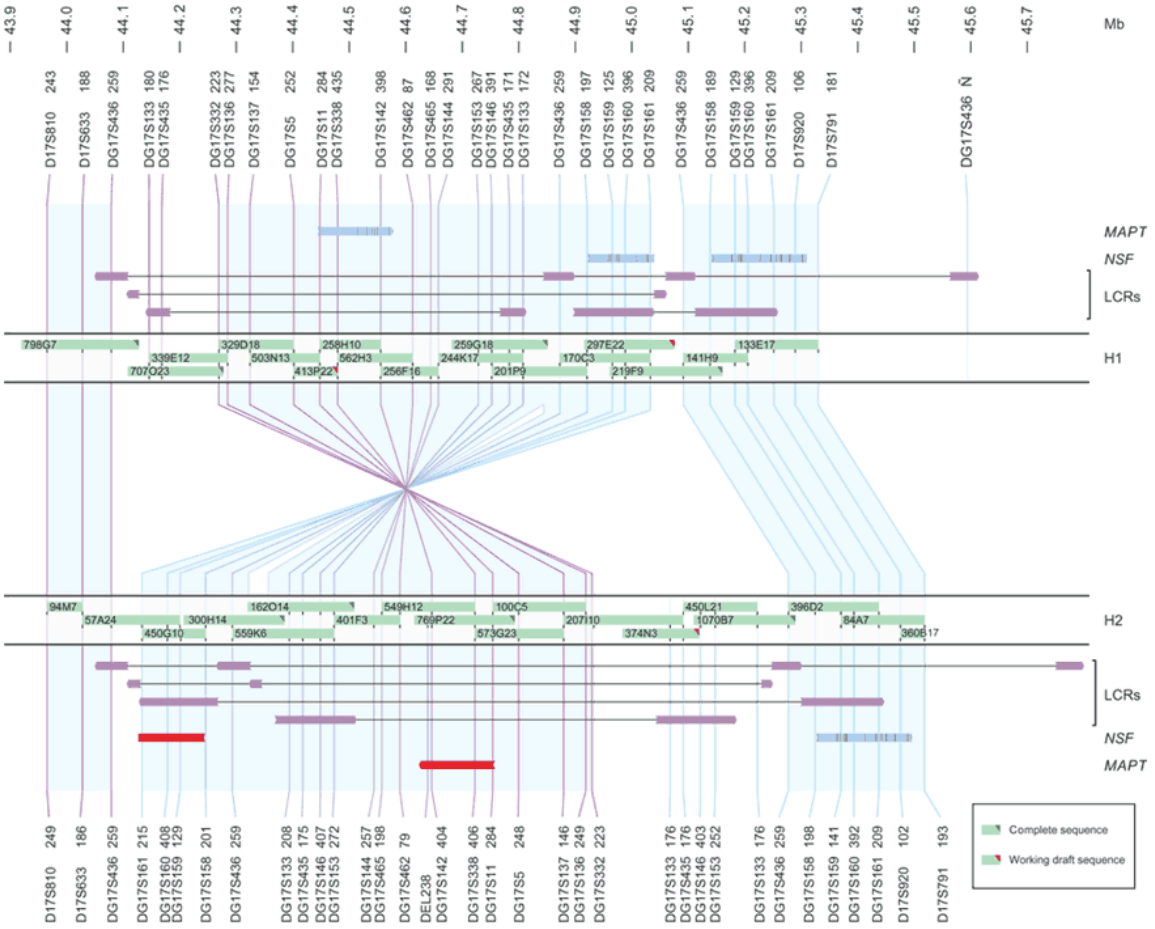
1. Classical and quantitative genetic studies of variation

Classical genetics reveals the existence of discrete genetic polymorphisms in natural populations, but these were necessarily limited either to chromosomal rearrangements such as inversions that can be detected cytologically, or to conspicuous phenotypes such as eye color or body color (e.g., in humans; or in fruit flies carrying certain eye color mutations such as *cardinal* can be found in natural populations).

Within a given species, only a handful of such polymorphisms can easily be detected. Relatively few cases of discrete polymorphisms affecting morphological traits are known.

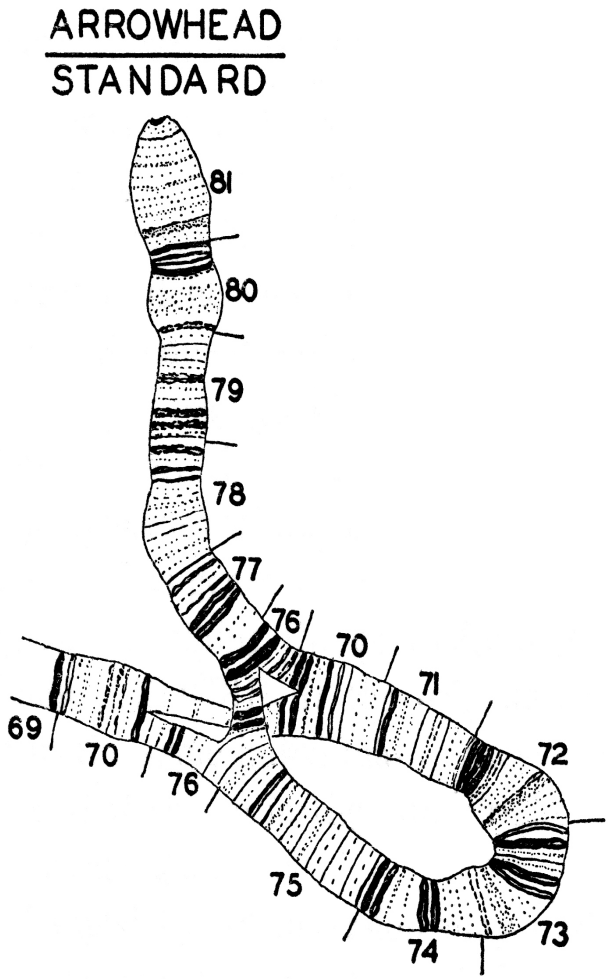
(text book, pp. 3-4)

A 900-kb human inversion polymorphism on chr. 17



Stefansson *et al.* 2005 *Nature Genetics*

The classic inversion polymorphism of *Drosophila pseudoobscura*



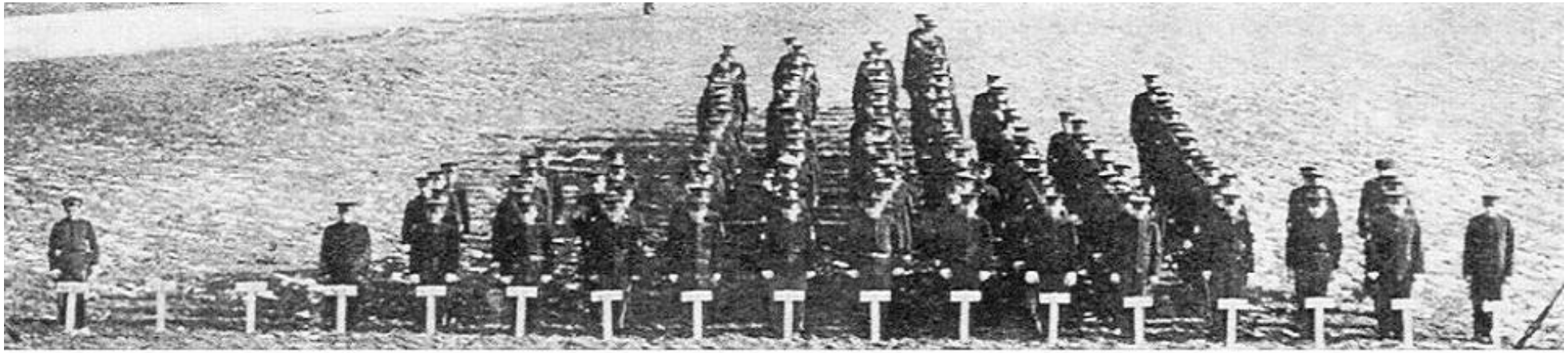
Shell color and banding polymorphism in the grove snail *Cepaea nemoralis*

Quantitative traits

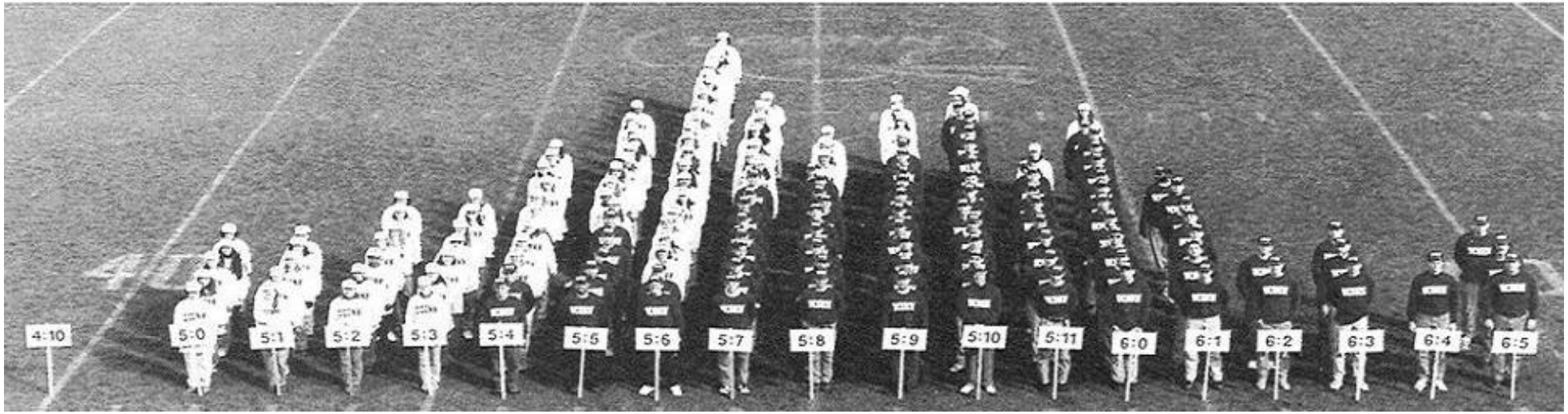
Quantitative genetics reveals the existence of ubiquitous genetic variation in metric (continuously varying) traits, such as human body height (text book, pp. 4-11)

Most metric traits have a coefficient of variation (the ratio of the standard deviation to the mean) of 5-10%.

Normal (Gaussian) distribution of height in humans in 1914 versus 1996



4:10 4:11 5:0 5:1 5:2 5:3 5:4 5:5 5:6 5:7 5:8 5:9 5:10 5:11 6:0 6:1 6:2



Quantitative traits

Measurements of the resemblances between relatives (e.g., Fisher 1918) show that 20%-80% of the variance in such traits is typically due to (additive) genetic factors (heritability = relative proportion of phenotypic variation that is due to genetic variation). For example, for human height heritability is approximately 80%.

This type of quantitative genetic variation is of great evolutionary, medical and economic significance, but measuring it does not tell us anything about the details of its genetic control (i.e., numbers of loci involved, frequencies of variant alleles, etc.).

Evolution as change of allele frequencies

Experimental genetics shows that the genetic factors underlying quantitative trait variability are just like those underlying Mendelian traits controlled by single genes. The only difference is that the effects of individual genes cannot be detected directly, because their contributions are obscured by non-genetic variation and by the effects of other genes (text book, pp. 8-11).

This means that we can think about evolution in terms of changes in allele frequencies of alternative variants at Mendelian loci, located on chromosomes.

There are some exceptions, notably maternally inherited traits controlled by mitochondrial genes or chloroplast genes in plants.

While amply validating Darwin's view that there is plenty of variation available for evolution to utilize, this evidence leaves two important questions unanswered:

1. How much variation within a natural population is there at an average locus? Classical genetics provides no means of sampling loci at random from the genome, without respect to their functional importance or level of natural variability.

2. To what extent does natural selection, as opposed to mutation and/or genetic drift, control the frequencies of allelic variants within populations?

(text book, pp. 13-14)

2. Measuring genetic variation

The solution to question (**a**) is to use the fact that genetic information is encoded in the DNA (or RNA, in the case of some viruses).

If either the protein sequence corresponding to a gene, or its DNA sequence, can be studied directly, then we can look at variation within the population without having to follow visible mutations, i.e. there is no need for prior knowledge of the existence of variation.

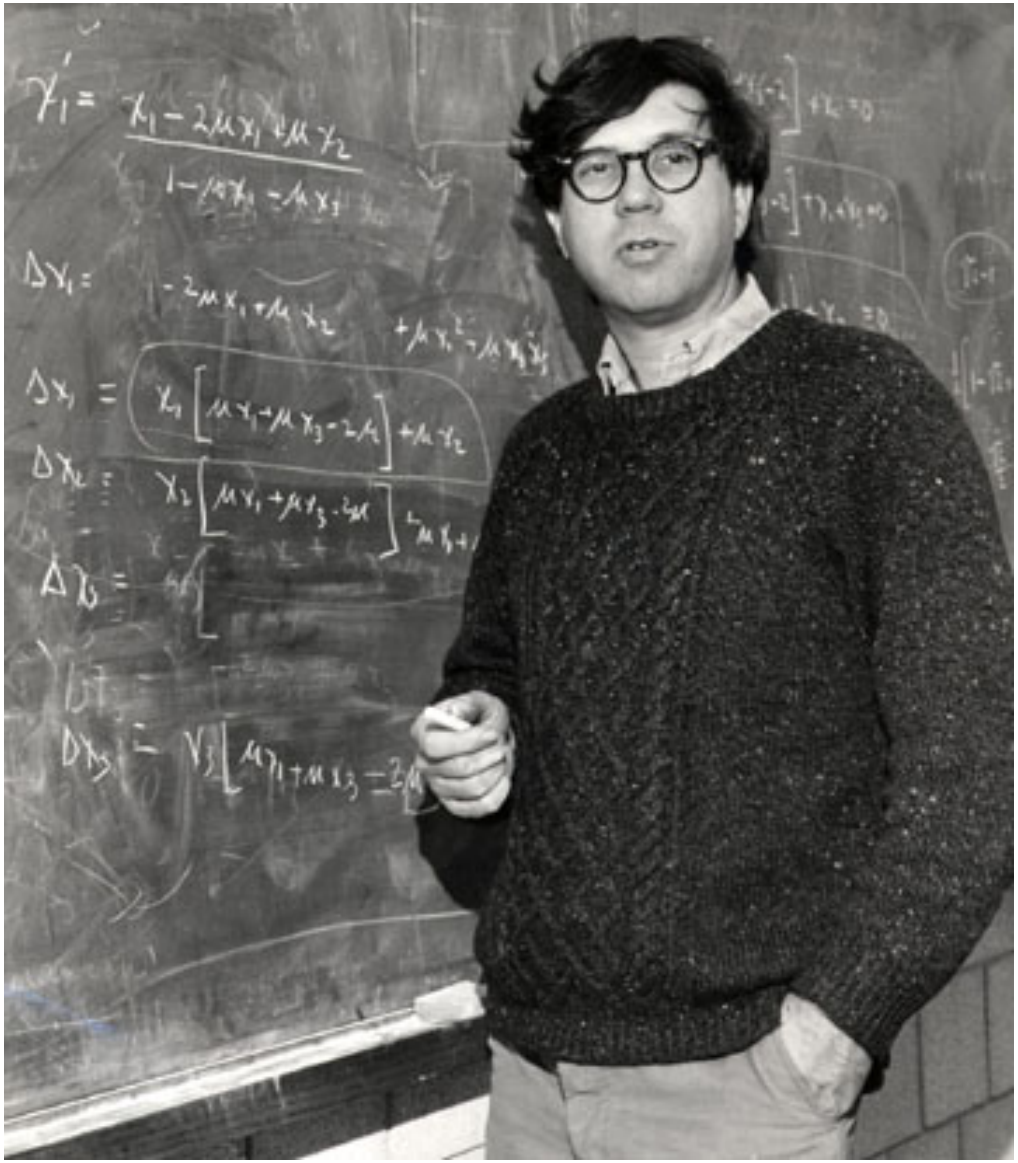
We can also look at variation in non-coding sequences.

Question **(b)** is still an active area of research.

A bit of history: electrophoretic variation

The first steps were taken in the mid-1960s by Richard Lewontin and Jack Hubby, working in Chicago on the fruit fly *Drosophila pseudoobscura*, and by Harry Harris in London, working on humans.

They used the technique of gel electrophoresis of proteins to screen populations for variants in a large number of soluble proteins controlled by independent loci, mostly enzymes with well-established metabolic roles. The proteins were chosen purely because they could be studied easily.



Lewontin (1974) *The Genetic Basis of Evolutionary Change* Columbia University Press

Massive amounts of genetic variation

The results of the early electrophoretic surveys were startling: a large fraction (up to 40%) of loci were found to be polymorphic (i.e., they exhibited one or more minority alleles with frequencies greater than 1%).

An average *D. pseudoobscura* individual was estimated to be heterozygous at 13% of the 24 protein loci that had been studied by 1974, i.e. a random individual sampled from the population would be expected to have distinct maternal and paternal alleles at 13% of its protein-coding loci.

Much lower levels of heterozygosity (or gene diversity: the chance that two randomly chosen copies of a gene are different) were found in mammals, and much higher levels in bacteria.

This work conclusively refuted the view that loci are only rarely polymorphic. It showed that there are massive amounts of genetic variation, and that the “classical view”, pioneered by Muller, i.e. that mainly variation is fueled by rare deleterious *de novo* mutations that are efficiently purged from populations

However, it raised more questions than it answered. In particular, there were several biases in the data. Only soluble proteins could easily be studied, and amino acid changes that do not affect the mobility of proteins on gels are not detected by electrophoresis.

Similarly, any changes in the DNA that do not affect the protein sequence go undetected.

DNA sequence variation

The advent in the late 1970s of methods for cloning and sequencing of DNA meant that studies of natural variation could be carried out at the DNA level. This eliminates virtually all the possible biases in quantifying variability.

This approach can be used to study variation in both coding sequences and non-coding sequences; the latter make up 98% of the human genome, and 75% of the fruit fly (*Drosophila*) genome.

(text book, pp. 21-33)

DNA sequence variation

With the advent of PCR amplification for isolated specific regions of the DNA, and with relatively cheap automated sequencing, this is now the method most commonly used in surveys of variation.

At first, this was done using classical Sanger sequencing. Today, 'next-generation' sequencing methods mean that it is now possible to sequence entire genomes from multiple individuals of a species, without any need to clone specific regions of the genome.

e.g., see The 1000 Genomes Project Consortium 2010 in *Nature* 467: 1061 (humans); see Langley et al. 2012 in *Genetics* 192: 533 (*Drosophila melanogaster*).

Marty Kreitman's pioneering work

The pioneering work on directly comparing homologous DNA sequences sampled within a species was carried out by Martin Kreitman in Lewontin's lab at Harvard in the early 1980s (Kreitman 1983 *Nature* 304: 412). This was the first study of DNA sequence variability (text book, pp. 24-25).

Kreitman sequenced 11 independent copies (alleles) of the *Adh* (alcohol dehydrogenase) gene of *D. melanogaster*, isolated from collections made around the world. He sequenced 2379 bases from each of these alleles, an heroic effort in those days.

Most variants were single nucleotide polymorphisms (SNPs), i.e. alternative base pairs at the same nucleotide site.

Marty Kreitman's pioneering work

Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*

Martin Kreitman

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA

The sequencing of eleven cloned Drosophila melanogaster *alcohol dehydrogenase* (Adh) *genes from five natural populations has revealed a large number of previously hidden polymorphisms. Only one of the 43 polymorphisms results in an amino acid change, the one responsible for the two electrophoretic variants (fast, Adh-f, and slow, Adh-s) found in nearly all natural populations. The implication is that most amino acid changes in Adh would be selectively deleterious.*

Kreitman 1983 *Nature* 304: 412

Kreitman's work succeeded in:

- Demonstrating a high level of variability at the level of individual nucleotide sites, a factor of ten or so higher than would have been expected from the typical level of heterozygosity for protein polymorphisms.
- Showing that nearly all of this variability involved silent changes that did not affect protein sequences, i.e. the changes were either in regions that did not code for amino acids (non-coding sequences) or involved synonymous changes in codons (i.e., they did not change the amino acid sequence of the protein coded for by the gene)
- The only amino acid polymorphism detected was that already known to cause the difference between the fast (*F*) and slow (*S*) electrophoretic alleles of *Adh*.

These results demonstrate that the protein sequence is highly constrained by selection, i.e. most mutations affecting the amino acid sequence of a protein (non-synonymous mutations) cause selectively disadvantageous changes to its functioning, and are eliminated rapidly from the population.

Most variation that is detected in coding sequences (typically over 85% in *Drosophila*) thus involves synonymous variants, which must therefore have much smaller average effects on fitness than non-synonymous mutations.

Non-coding region variation shows a similar level to synonymous variation (on average slightly less).

A similar pattern is observed when between-species comparisons of DNA sequences are made.

These results suggest that most variation and evolution at the DNA level may be due to neutral or nearly neutral mutations, whose fate is controlled by genetic drift (random changes in allele frequencies due to finite population size) rather than selection, especially as much of the genome is made up of non-coding sequences, even in *Drosophila*.

Describing levels of genetic variability

- The allele frequencies provide a useful reduced description of the variation at each locus, but they are still cumbersome if many loci are studied. “Summary statistics” have thus been developed to measure variability in samples.

- The three main measures are:

  1. The proportion of polymorphic loci *(P)*
  2. The number of different alleles at the locus (*A*)
  3. The *gene diversity* (*H*), also called the *heterozygosity.*

- Note that, to measure diversity, all loci surveyed must be included, not only the the polymorphic ones.

(text book, pp. 17-19, 27-29)

The proportion of polymorphic loci *(P)*

- A locus is classified as polymorphic if there is at least 1 minority allele with a frequency greater than some cut-off value, e.g. 0.01 or 0.05 (i.e., the majority allele has a frequency equal to or smaller than 0.99 or 0.95)
- The rationale for this is that rare alleles do not contribute much variation to the population, compared to alleles at intermediate frequencies.
- For example, a survey that finds variants at 7/30 loci, two of which are not polymorphic according to the above cut-off criterion, yields $P = 5/30 = 0.17 = 17\%$.
- This measure is quick and simple, however, it relies on an arbitrary cut-off and it is difficult to compare to other studies. Other measures are therefore preferable.

Allele number *(A)*

- For each different locus, the number of different alleles in a sample is counted and the mean can be taken over the set of loci that are studied.

- This method suffers from the problem that *A* is highly dependent on sample size.

Gene diversity or Heterozygosity (*H*)

- For a given locus, a useful diversity measure is the frequency with which a pair of randomly chosen alleles from the sample differ in state (Nei 1973).
- Let's look at a simple example. Consider an autosomal allozyme locus in a diploid species, with three alleles: *S* (slow), *I* (intermediate) and *F* (fast). There are thus 6 different possible genotypes. Let's assume we get the following numbers:
- SS: 12
- II: 21
- FF: 60
- SI: 34
- SF: 46
- IF: 77
- Total number of individuals $k = 250$.

Gene diversity or Heterozygosity (*H*)

- These genotypic data can be reduced to three allele frequencies. A given allele (e.g., *S*) is always present in 2 copies in the homozygotes (e.g., *SS*) and in 1 copy in the heterozygotes (e.g., *SI*, etc.).

- The allele frequencies *f* are thus:

- $f(S) = (2 \times 12 + 34 + 46) / 500 = 0.208$
  (where $500 = 2k$ = number of gene copies)

- $f(I) = (2 \times 21 + 34 + 77) / 500 = 0.306$

- $f(F) = (2 \times 60 + 46 + 77) / 500 = 0.486$

- $f(S) + f(I) + f(F) = 1.$

Gene diversity or Heterozygosity (*H*)

- Now we can calculate *H* by taking all three pairs of different alleles and finding the sum of twice the product of their frequencies. For our example, this would be:

- $H = 2x ((0.208 \times 0.306) + (0.208 \times 0.486) + (0.306 \times 0.486)) = 0.627$

- The factor 2 arises because the order on which the 2 alleles are chosen is irrelevant / equivalent

- A quicker way of obtaining the same result is to subtract from 1 the frequencies with which a pair of alleles is the same, i.e.:

- $H = 1 - (0.208)^2 - (0.306)^2 - (0.486)^2 = 0.627$

- In other words: $H = 1$ - frequencies of the homozygotes.

Gene diversity or Heterozygosity (*H*)

- This is of course familiar from the Hardy-Weinberg (H-W) expression for 1 single diploid locus with, say, 2 alleles:

- $f(AA) = p^2, f(Aa) = 2pq, f(aa) = q^2$

- where $f(AA) + f(Aa) + f(aa) = p^2 + 2pq + q^2 = 1$, and where

- $1 - f(AA) - f(aa) = f(Aa) = 2pq$ = frequency of heterozygotes = H

- These expressions are, however, only valid if the assumption of random mating is fulfilled (see below).

- If we want to summarize the results for a set of loci, we simply take the mean of the *H* values for each locus.

Gene diversity or Heterozygosity (*H*)

- *H* measures the level of variability in a convenient way, since alleles present at low frequencies contribute little to its value; this measure is also not strongly dependent on sample size.
- For a single locus with 2 alleles, *H* reaches a maximum at 0.5., when each allele has a frequency of 0.5, and is close to zero when either allele is at low frequency.
- e.g., for $p = 0$ and $q = 1$, we have $2pq = 0$, and for $p = q = 0.5$, we have $2pq = 0.5$.
- *(note that for binomial sampling, var = n (pq) so that 2pq can be interpreted as a variance)*
- Under random mating and H-W assumptions, H values will be measured to be the same as those predicted from H-W.
- Therefore *H* is sometimes called *expected heterozygosity,* H_E.

Gene diversity or Heterozygosity (*H*)

- The terminology *expected heterozygosity,* H_E, is not recommended because in a non-randomly mating population (violating H-W), *H* as calculated above is obviously NOT the expected heterozygote frequency.

- *H*, as derived from allele frequencies, is thus best called *gene diversity*, and represents the probability with which a pair of randomly chosen alleles from the sample differ in state (Nei 1973).

- It is thus advisable to use "*heterozygosity*" to mean the *observed frequency* of heterozygotes (*observed heterozygosity,* H_O) and to distinguish it from *H*.

Measuring DNA sequence variability

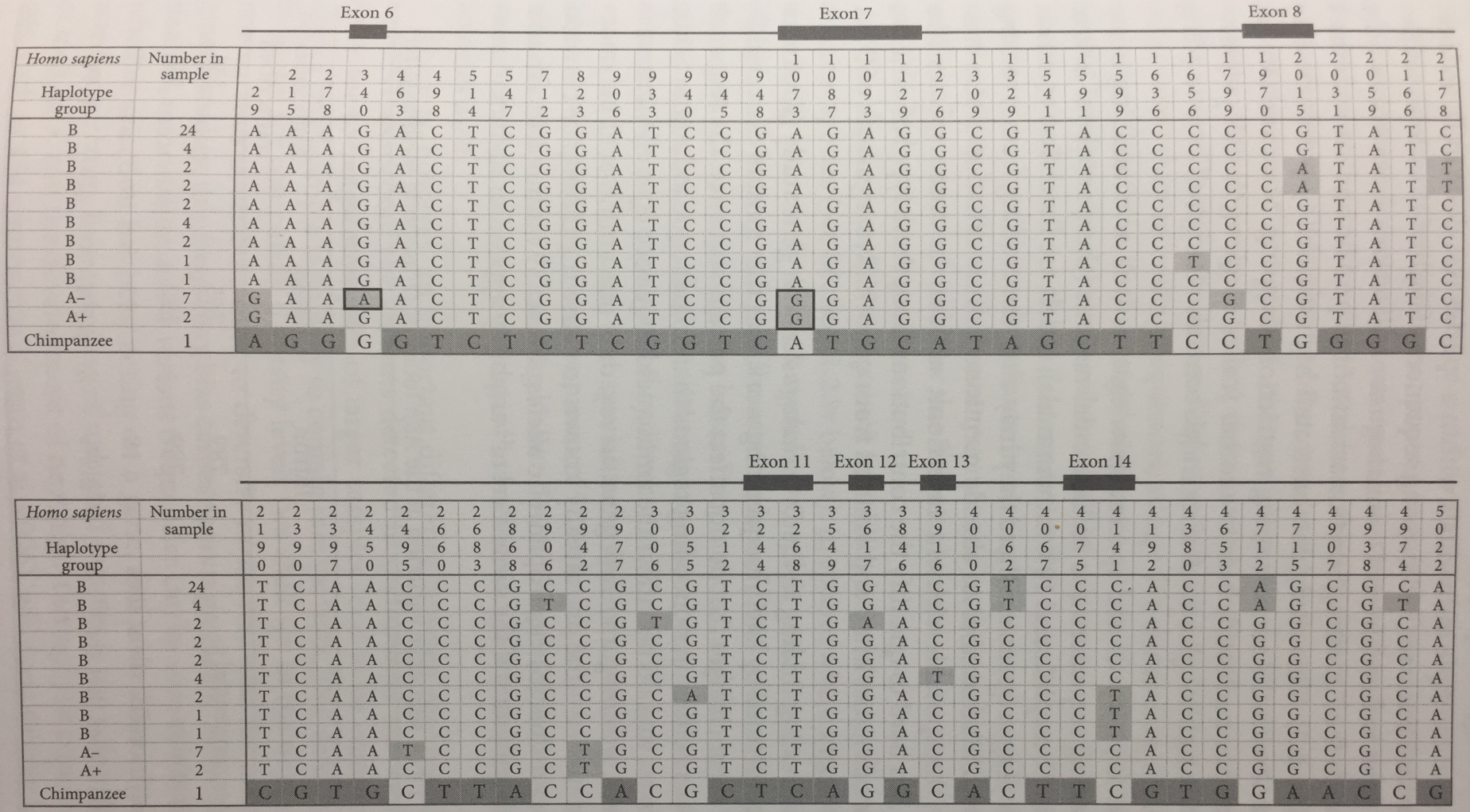
- Now let's look at how to obtain quantitative measures of sequence variation (text book, pp. 27- 33).
- In studying DNA sequence variation, we often refer to a stretch of DNA from a region of the genome as a *locus* (even if only a part of a gene was sequenced, or even if the sequenced region is non-coding), and the independently sampled sequences as *alleles.*
- This is very different from the terminology used for allozyme variability, where the locus was used to refer to a gene coding for a polypeptide identified in an electrophoretic gel, and alleles mean variants at such a locus that segregate from each other in a Mendelian fashion.

Measuring DNA sequence variability

- The basic information for quantifying DNA sequence variability comes from haplotypes, i.e. the above-mentioned stretches of sequence from one of the two homologous chromosomes of an individual, in the case of diploid genomes / organisms.

- It is not necessarily easy to assemble haplotypes from diploid individuals, but we we will not consider this problem here (see p. 369 of the text book for references about this).

- The next slide shows an example of a set of allelic DNA sequences at the *glucose-6-phosphate dehydrogenase* (G6PD) locus in a worldwide sample of humans (Saunders *et al.* 2002, *Genetics* 162:1849), which is subject to malarial selection.



Homo sapiens	Number in sample	2	2	2	3	4	4	5	5	7	8	9	9	9	9	9	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
Haplotype group		2	1	7	4	6	9	1	4	1	2	0	3	4	4	4	0	0	0	1	2	3	3	5	5	5	6	6	7	9	0	0	0	1	1
		9	5	8	0	3	8	4	7	2	3	6	3	0	5	8	3	7	3	9	6	9	9	1	1	9	6	6	9	0	5	1	9	6	8
B	24	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	G	T	A	T	C
B	4	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	G	T	A	T	C
B	2	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	A	T	A	T	T
B	2	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	A	T	A	T	T
B	2	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	G	T	A	T	C
B	4	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	G	T	A	T	C
B	2	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	G	T	A	T	C
B	1	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	T	C	C	G	T	A	T	C
B	1	A	A	A	G	A	C	T	C	G	G	A	T	C	C	G	A	G	A	G	G	C	G	T	A	C	C	C	C	C	G	T	A	T	C
A–	7	G	A	A	A	A	C	T	C	G	G	A	T	C	C	G	G	G	A	G	G	C	G	T	A	C	C	C	G	C	G	T	A	T	C
A+	2	G	A	A	G	A	C	T	C	G	G	A	T	C	C	G	G	G	A	G	G	C	G	T	A	C	C	C	G	C	G	T	A	T	C
Chimpanzee	1	A	G	G	G	G	T	C	T	C	T	C	G	G	T	C	A	T	G	C	A	T	A	G	C	T	T	C	C	T	G	G	G	G	C

Homo sapiens	Number in sample	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	5
		1	3	3	4	4	6	6	8	9	9	9	0	0	2	2	2	5	6	8	9	0	0	0	0	1	1	3	6	7	7	9	9	9	0
Haplotype group		9	9	9	5	9	6	8	6	0	4	7	0	5	1	4	6	4	1	4	1	1	6	6	7	4	9	8	5	1	1	0	3	7	2
		0	0	7	0	5	0	3	8	6	2	7	6	5	2	4	8	9	7	6	6	0	2	7	5	1	2	0	3	2	5	7	8	4	2
B	24	T	C	A	A	C	C	C	G	C	C	G	C	G	T	C	T	G	G	A	C	G	T	C	C	C	A	C	C	A	G	C	G	C	A
B	4	T	C	A	A	C	C	C	G	T	C	G	C	G	T	C	T	G	G	A	C	G	T	C	C	C	A	C	C	A	G	C	G	T	A
B	2	T	C	A	A	C	C	C	G	C	C	G	T	G	T	C	T	G	A	A	C	G	C	C	C	C	A	C	C	G	G	C	G	C	A
B	2	T	C	A	A	C	C	C	G	C	C	G	C	G	T	C	T	G	G	A	C	G	C	C	C	C	A	C	C	G	G	C	G	C	A
B	2	T	C	A	A	C	C	C	G	C	C	G	C	G	T	C	T	G	G	A	C	G	C	C	C	C	A	C	C	G	G	C	G	C	A
B	4	T	C	A	A	C	C	C	G	C	C	G	C	G	T	C	T	G	G	A	T	G	C	C	C	C	A	C	C	G	G	C	G	C	A
B	2	T	C	A	A	C	C	C	G	C	C	G	C	A	T	C	T	G	G	A	C	G	C	C	C	T	A	C	C	G	G	C	G	C	A
B	1	T	C	A	A	C	C	C	G	C	C	G	C	G	T	C	T	G	G	A	C	G	C	C	C	T	A	C	C	G	G	C	G	C	A
B	1	T	C	A	A	C	C	C	G	C	C	G	C	G	T	C	T	G	G	A	C	G	C	C	C	T	A	C	C	G	G	C	G	C	A
A–	7	T	C	A	A	T	C	C	G	C	T	G	C	G	T	C	T	G	G	A	C	G	C	C	C	C	A	C	C	G	G	C	G	C	A
A+	2	T	C	A	A	C	C	C	G	C	T	G	C	G	T	C	T	G	G	A	C	G	C	C	C	C	A	C	C	G	G	C	G	C	A
Chimpanzee	1	C	G	T	G	C	T	T	A	C	C	A	C	G	C	T	C	A	G	G	C	A	C	T	T	C	G	T	G	G	A	A	C	C	G

FIGURE 1.8 Sequences of part of the gene encoding the glucose-6-phosphate dehydrogenase enzyme in a worldwide sample of humans and a single chimpanzee. A region of just over 5 kilobases (kb) was sequenced in a sample of 51 humans, including three types of allele (B, A–, and A+); these differ at two amino acids. B alleles encode a protein with high enzyme activity. A– alleles are quite common and encode an allele with only 12% of normal activity (a deficiency allele), whose presence confers malaria resistance, with a 50% reduction in the risk of severe malaria in heterozygous females and in males. A+ is a mild deficiency allele (85% of normal activity and no malaria resistance), which is rarer than A– in human populations. Exons are marked with solid boxes. The diagram shows only sites with variants, either within humans (18 sites), or between the human and chimpanzee sequences. Their positions from the start of the region sequenced are given at the top of each column. SNPs are highlighted and the two amino acid variants are boxed. [Data from Saunders et al. (2002).]

How to measure DNA sequence variation

- A sample of 3 haploid sequences = 3 alleles
- Sequence length: 30 bp
- 4 nucleotide site variants = 4 polymorphisms

Allele 1

ATGCTTAGCGTTGGCATCCTAGCGATCGAG

Allele 2

ATGCTT**G**GCGTTGGCATCCTAGCGATCGAG

Allele 3

AT**A**CTTAGCGTTGGCATCCT**C**GCGAT**T**GAG

Summary statistics for sequence data

- The nucleotide diversity (π, sometimes also called θ_π) for a given set of alleles sampled from a population is the frequency with which a randomly chosen pair of alleles differ at a given site.
- It can be calculated from data on a sample of homologous DNA sequences, by determining the sum of the numbers of differences between all possible pairs of sequences.
- The result is divided by the product of the number of sequences that were compared (equals $k(k\text{-}1)/2$, if there are k independent alleles) and the number of bases studied (to normalize the measure and get an "average" per nucleotide measure).

- In our above example, $k = 3$, so that $k(k\text{-}1)/2 = 3(3\text{-}1)/2 = 3$ (*i.e.*, 3 pairwise comparisons, excluding all comparisons to *itself* along the diagonal of a 3 x 3 matrix).

- The total number of pairwise differences between all 3 combinations of sequences is 1 + 3 + 4 = 8.

- To get the pairwise diversity per nucleotide site, we divide this by 3 times the number of sites, so that:

- $\pi = 8/(3 \text{ x } 30) = 0.089$ = probability per site that 2 randomly chosen sequences differ from each other .

- This is analogous to the heterozygosity of gene diversity *H* for the case of single nucleotides.

An alternative method of measuring variation is simply to count the number of sites that are segregating in the sample, S_k (k = sample size).

By dividing the number of segregating sites S_k by the product of the number of bases in the sequence and the following sum (correction factor)

$$a_k = \text{sum over } 1/i \text{ (for } i = 1 \text{ to } i = k\text{-}1)$$

we obtain a per nucleotide diversity estimate called Watterson's θ or θ_w.

- If the population is at equilibrium (mutation-drift equilibrium), and if there is no selection, then θ_w is expected to be same in value as π.
- In the example, we have $S_k = 4$, and for $k = 3$, we have:

$$a_k = 1 + 0.5 = 1.5.$$

- Hence, for a sequence of 30 bp length, we have:

$$\theta_w = 4/(30 \text{ x } 1.5) = 0.089.$$

- Note that this is exactly the same value as we obtained for π.
- Under non-equilibrium conditions (deviations due to demography, selection, etc.) the two measures can differ: while Watterson's method only takes the number of SNPs into account, π also takes into account the frequency of the polymorphisms.

Neutral theory and mutation-drift equilibrium

Under the neutral theory of evolution, variability in DNA sequences reflects the balance between the input of new variants by mutation and their loss by random fluctuations in frequencies caused by finite population size (genetic drift).

Under this model the two measures of nucleotide variability or diversity are expected to be the same.

Under this model, variant frequencies at a locus are always shifting around, but a statistical equilibrium will eventually be reached if population size stays constant.

The expected value of the pairwise diversity in the population is then given by (for diploids):

$$\theta = 4N_e\mu = \text{“population-scaled mutation rate”},$$

where μ is the neutral mutation rate per site, and N_e is the effective population size, which controls the rate of genetic drift. (For haploids, this would be $\theta = 2N_e\mu$.).

The expected values of both π and θ_w are equal to θ, so that, in other words:

$$\mathrm{E}(\pi) = \mathrm{E}(\theta_w).$$

Some empirical estimates

Estimates of θ_w or π have now been obtained from many different kinds of organisms, by sampling sets of homologous genes from natural populations and sequencing them.

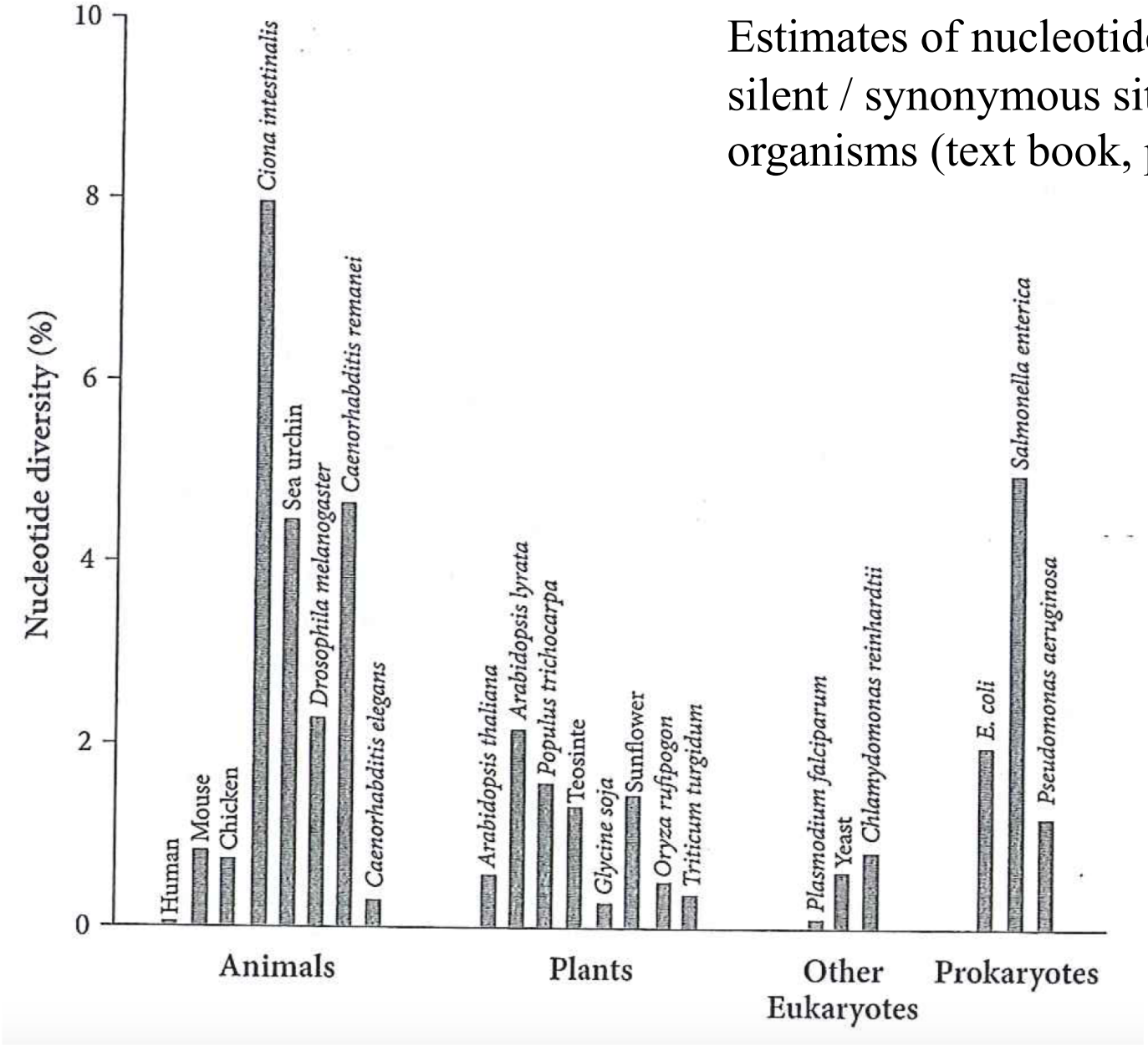
Rough average values over many genes for SNPs that do not affect protein sequences (silent site changes) are as follows:

- *Escherichia coli* (bacterium): 0.05
- *Drosophila melanogaster* (African) 0.02
- *Homo sapiens* 0.001

A value of θ_w or π = 0.01 (= 1%) implies that there is 1 SNP every 100 base pairs.

If we take account of the genome size, these values imply that there can be several millions of SNPs within a species such as humans.

Some empirical estimates



Estimates of nucleotide diversity for silent / synonymous sites from different organisms (text book, p. 30)

Estimating N_e from θ and μ

- Under neutrality, knowledge of the mutation rate μ enables us to estimate N_e from θ or π because of the relation $\theta = \pi = 4N_e\mu$

- For example, using estimates from *Drosophila*, $\mu = 4 \times 10^{-9}$, and $\theta = 0.02$, we obtain $N_e = 1.25 \times 10^6$

- *Drosophila* effective population sizes are therefore very large.

- For humans, we have $\pi = 0.001$ and $\mu = 1 \times 10^{-8}$, so we obtain $N_e = 0.001 / (4 \times 10^{-8}) = 25{,}000$.

- The human effective population size is thus quite small.

Other types of molecular variability

- Several other types of DNA sequence variants can be detected in surveys of variability, but these do not occur as frequently as SNPs. Two important categories are:

1. Insertions of transposable elements (TEs). These are almost invariably found outside coding regions, indicating selection against such insertions.

2. Another, usually much more abundant category consists of short insertions or deletions (indels), usually a few bases in length, but sometimes much larger, and usually also found outside coding regions. Microsatellites are an especially important example; these consist of repeated short sequences that vary in the numbers of repeats.

Where does all this variability come from?

- We can first ask: what happens to variability in the absence of any evolutionary forces?

- i.e., if there is no mutation, selection, migration (= gene flow), and if the population is so large that random fluctuations in the frequencies of variants can be ignored.

- The answer is that, contrary to what was believed by Darwin and his contemporaries, NOTHING CHANGES with respect to variability at a single nucleotide site.

(text book, pp. 33-35)

- This is because Mendelian inheritance is particulate rather than blending in nature; i.e. at a given locus, the contribution from either parent is transmitted intact to the gametes.

- The 19th century view of inheritance as a process of blending of material from the two parents implied that variability is constantly being lost (i.e., halved in each generation) from sexual populations. This turned out to be wrong.

- In fact, if variability is present in a population, it will be maintained indefinitely in the absence of any evolutionary forces.

(text book, pp. 33-35)

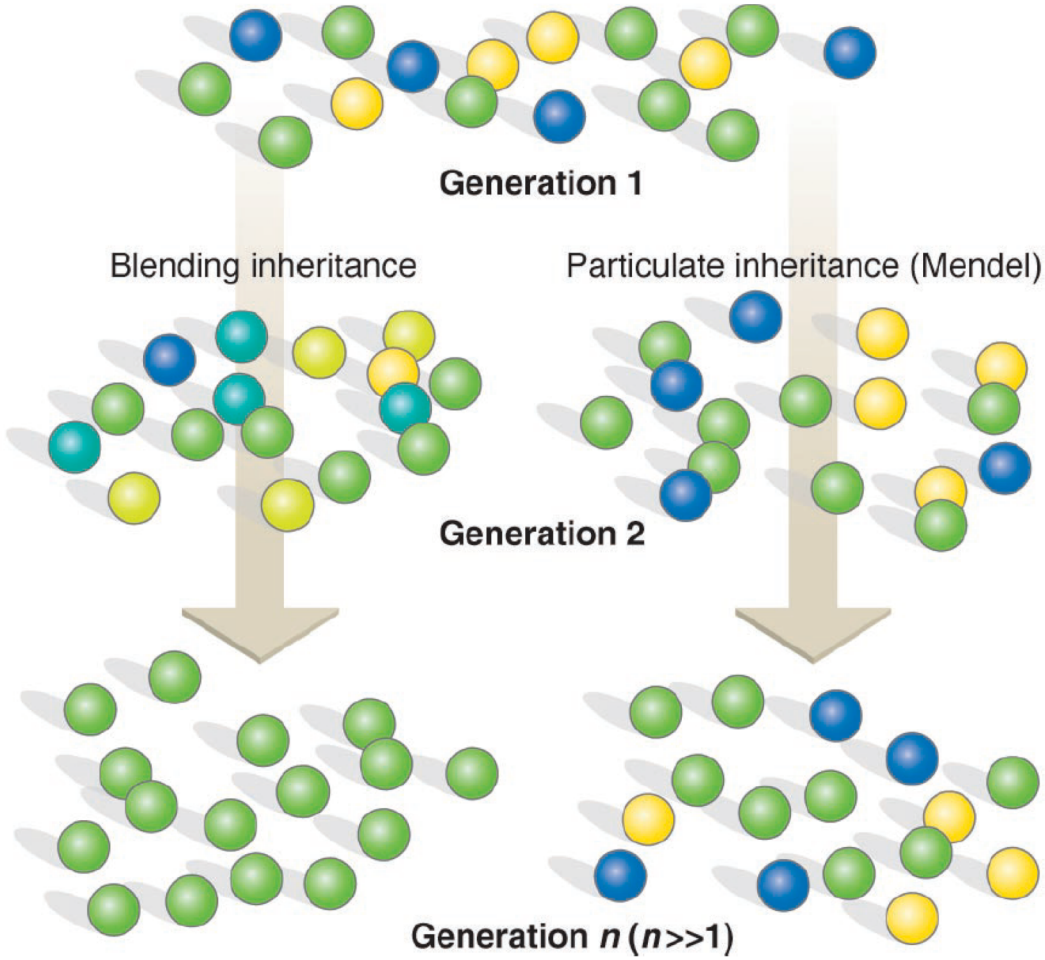


Fig. 1. Difference between the outcomes from blending and from particulate inheritance. In post-Mendelian terms, we assume a single diallelic locus, and hence three diploid genotypes (AA, blue; Aa, green; aa, yellow). Under particulate inheritance, the population's variability is preserved from generation to generation. In contrast, the conventional wisdom of Darwin's day saw offspring inherit a blend of parents' characteristics, here represented as the average of the two parental shadings. The result is that the variability diminishes in successive generations (the variance is halved each generation if mating is at random).

The Hardy-Weinberg equilibrium

(see Vitor Sousa's 1st lecture)

- The mathematical statement of this fact is known as the Hardy-Weinberg (H-W) equilibrium:
- In an infinitely large, diploid, randomly mating population with no mutation, selection, migration or drift, the frequencies of the three genotypes at an autosomal locus with 2 alleles, A and a, with frequencies p and $q = 1\text{-}p$, remain constant over time at:
- $p^2 : 2pq : q^2$
- This can be easily generalized to an arbitrary number of alternative alleles.

(text book, pp. 33-35)

- But we still need to know where the variants present in a population come from and what has caused them to attain the frequencies at which they are now present.

- Let’s start to explore this question.

Mutation as an evolutionary force

- Mutation is the ultimate source of all genetic variation.
- For our purposes, a mutation is any type of change in the genetic material that is transmitted from parent to offspring.
- It is now possible to measure at the DNA sequence level the rates at which new mutations arise, either in laboratory experiments on model organisms, or by sequencing parents and offspring from populations.

(text book, pp. 43-45; Kong et al. 2012 *Nature* 488:471; Schrider *et al.* 2013 *Genetics* 194:937)

Mutation as an evolutionary force

- The most common type of stably inherited mutations are single nucleotide substitutions.

- In organisms other than RNA virues, they arise at an extremely low rare per generation: around 1×10^{-8} per site per generation in humans but closer to 5×10^{-9} in *Drosophila* and 1×10^{-9} in bacteria (text book, pp. 41-43).

- This means that mutation is not very effective at changing the composition of a population and is thus usually not considered to be a major directional factor in evolution.

A simple model of mutation pressure

- This result can be obtained more formally as follows, giving you a gentle introduction to the methods of mathematical analysis in population genetics (text book, pp. 43-45)

- Consider the simple case of two alternatives, A_1 and A_2, at a particular site (e.g., a GC versus an AT base pair).

- We assume discrete generations (= non-overlapping generations), i.e. all parents in the population produce their offspring at the same time and then die. The offspring then grow up to repeat the cycle.

Change in frequency under mutation pressure

- Let A_1 mutate to A_2 at a rate u and let A_2 mutate back to A_1 at a rate v. Let p be the frequency of A_1 and $(1\text{-}p) = q$ the frequency of A_2 .

- If the population is so large that random changes in frequencies can be ignored, the frequency of A_1 in the next generation, p', is given by:

- $p' = p\,(1\text{-}u) + (1\text{-}p)\,v = p\,(1\text{-}u) + q\,v$

- The change in frequency, Δp, is obtained by subtracting p:

- $\Delta p = (1\text{-}p)\,v - up = qv - up$

Change in frequency under mutation pressure

- This shows, not surprisingly, that the rate of change in allele frequencies due to mutation pressure is of the same order of magnitude as the mutation rate: e.g., if A_1 is initially very rare (i.e., p is close to 0), it will spread at a rate of approximately v. This is because, when A_1 is very rare, this allele will be mainly generated and spread by mutation of A_2 into A_1.

- The above model has a stable equilibrium in allele frequency which satisfies $\Delta p = 0$: $p^* = v / (u + v)$.

- Thus mutation pressure could in principle produce DNA sequence variants that are present at intermediate frequencies.

- But the time involved for a given type of variant to increase substantially in frequency is hundreds of millions of generations (e.g., $\mu = 1 \times 10^{-8}$ in humans implies 100 Mio generations).

- This means evolutionary forces other than mutation, such as selection and genetic drift, which act more quickly, are likely to be much more important in controlling the frequencies of variants that contribute to variability (see R. A. Fisher 1930, *The Genetical Theory of Natural Selection,* Oxford at the Clarendon Press).

3. The maintenance of neutral variability by mutation and drift

- For the last part of the lecture, we'll examine the basic theory that enables us to interpret data on sequence variability.

- This is based on the simplest possible assumption – that the variants concerned have no effects on Darwinian fitness, i.e. they are selectively neutral.

- This is the assumption that underlies the neutral theory of molecular evolution, famously advocated by Motoo Kimura

(Kimura 1968 *Nature* 217:624; Kimura 1983 *The Neutral Theory of Molecular Evolution,* Cambridge University Press)

- If neutrality is the case, then mutations can only rise to intermediate frequencies in a closed population as a result of genetic drift, the process of random sampling of variants in a finite population (*cf.* the first lecture by Vitor Sousa)

- The discovery that random sampling of variant frequencies in finite populations may be a significant factor in evolution is a major contribution of population genetics. This process has two aspects:

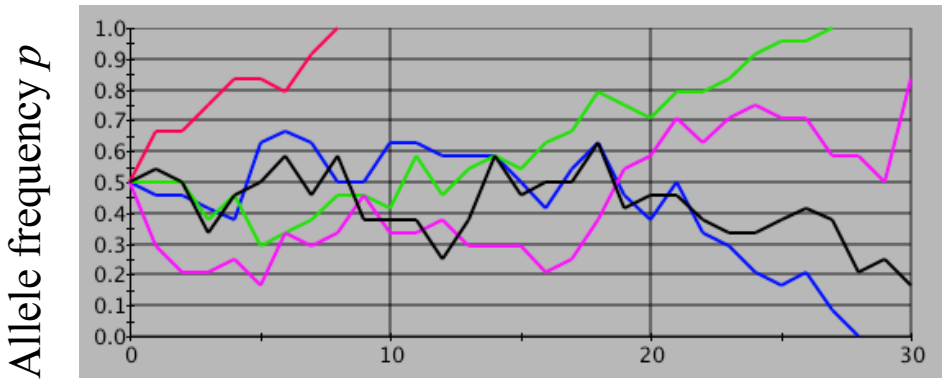
- **a.** The tendency of a finite population to become genetically uniform, owing to the fact that there is an increasing tendency as time goes on for all the copies of the gene (i.e., all alleles at a locus) to be descended from a single ancestral allele.

- **b.** The tendency of isolated populations to diverge in variant frequencies over time, since independent trials of a population with the same initial state will arrive at different allele frequencies by chance (population differentiation).

- See Vitor's lecture: the underlying model is the Wright-Fisher model, first introduced by Fisher (1922): random binomial sampling of allele frequencies from one generation to the next.
- The process is analogous to the tossing of a coin.

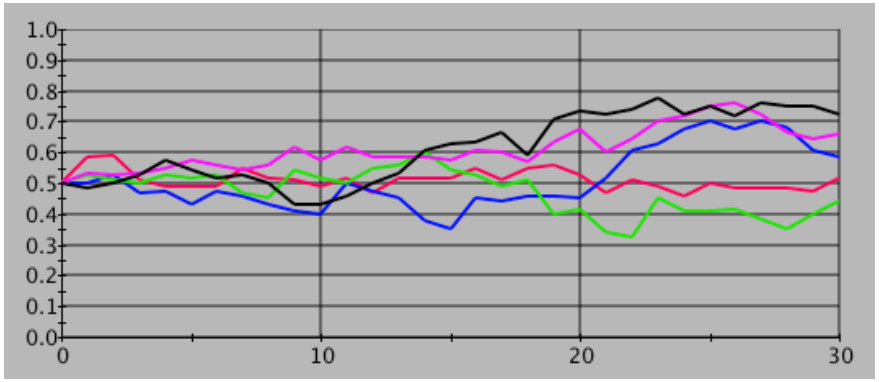
Computer simulation of genetic drift at 1 locus with 2 alleles;
each curve represents an independent population (statistical trial)

Small population (18 individuals



Time (generations)

Larger population (60 individuals



Time (generations)

- The first effect of drift (genetic uniformity of closed populations) is closely related to the decrease in heterozygosity H ($H = 1 - G$ where G is homozygosity), i.e. the increase in homozygosity G, that accompanies the inbreeding of close relatives.

- Both are conveniently studied by means of the concept of identity by descent (i.b.d.) (text book, pp. 35-36; *cf.* coalescence theory).

- Two alleles at the same locus drawn from a population are said to be identical by descent if they trace their ancestry back to a single ancestral allele. The probability of i.b.d. is $1/2N$ (also called the inbreeding coefficient).

- This also corresponds to the probability that a neutral allele becomes fixed; under the Wright-Fisher model $1/2N$ represents the rate of genetic drift.

Decline in heterozygosity over time due to drift

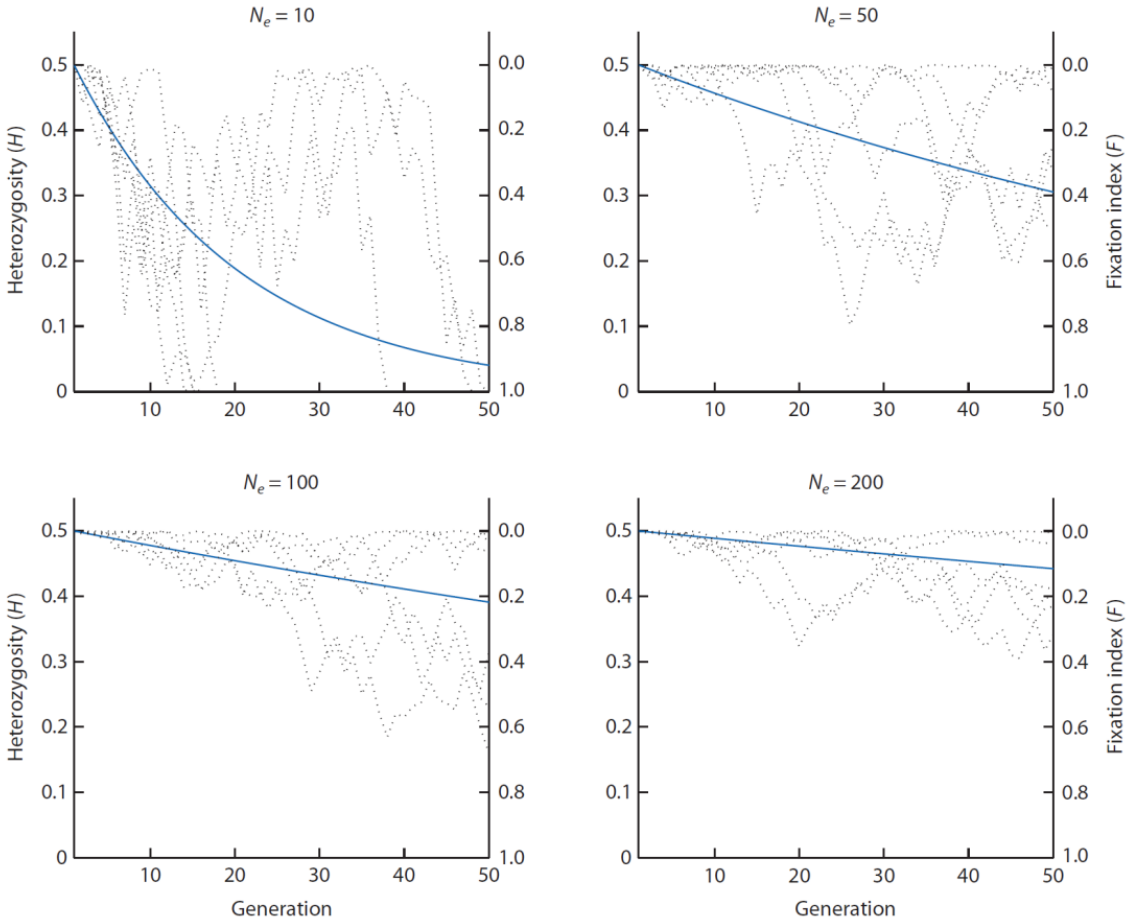


Figure 3.19 The decline in heterozygosity as a consequence of genetic drift in finite populations. The solid lines show expected heterozygosity over time according to $H_t = \left(1 - \frac{1}{2N_e}\right) H_{t-1}$. The decrease in heterozygosity can also be thought of as an increase in autozygosity or the fixation index (F) through time under genetic drift. The dotted lines in each panel are levels of heterozygosity ($2pq$) in six replicate finite populations experiencing genetic drift. There is substantial random fluctuation around the expected value for any individual population.

4. How to measure differentiation between populations?

- As we have seen, genetic drift can lead to genetic divergence or differentiation between 2 or more populations. Selection and migration as well as spatial distance (isolation by distance) can also contribute to among-population differentiation.

- How can we measure genetic differentiation? (*cf.* first lecture by Vitor Sousa).

- To quantify genetic among-population differentiation we use the fixation index, F_{ST}, developed by Sewall Wright (e.g., Wright 1943; also see Wright 1950 *Nature* 166:247) (text book, pp. 305-308).

Among-population differentiation F_{ST}

F_{ST} is the relative proportion of genetic variation that occurs among (sub-)populations as compared to the total combined variation in the total population.

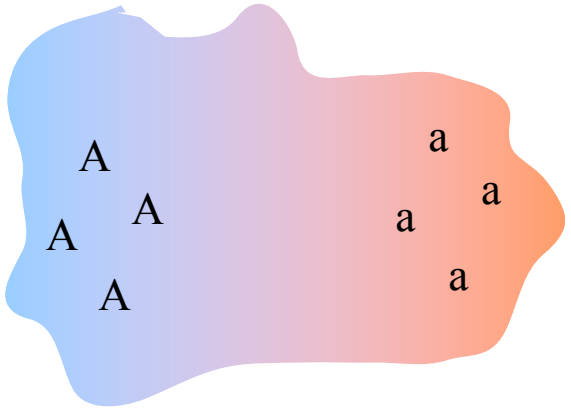
F_{ST} is a summary statistic that measures allele frequency similarity

$$0 \leq F_{ST} \leq 1$$

identical allele frequencies

fixation of different alleles

$F_{ST} = 1$



Subpopulation 1 Subpopulation 2

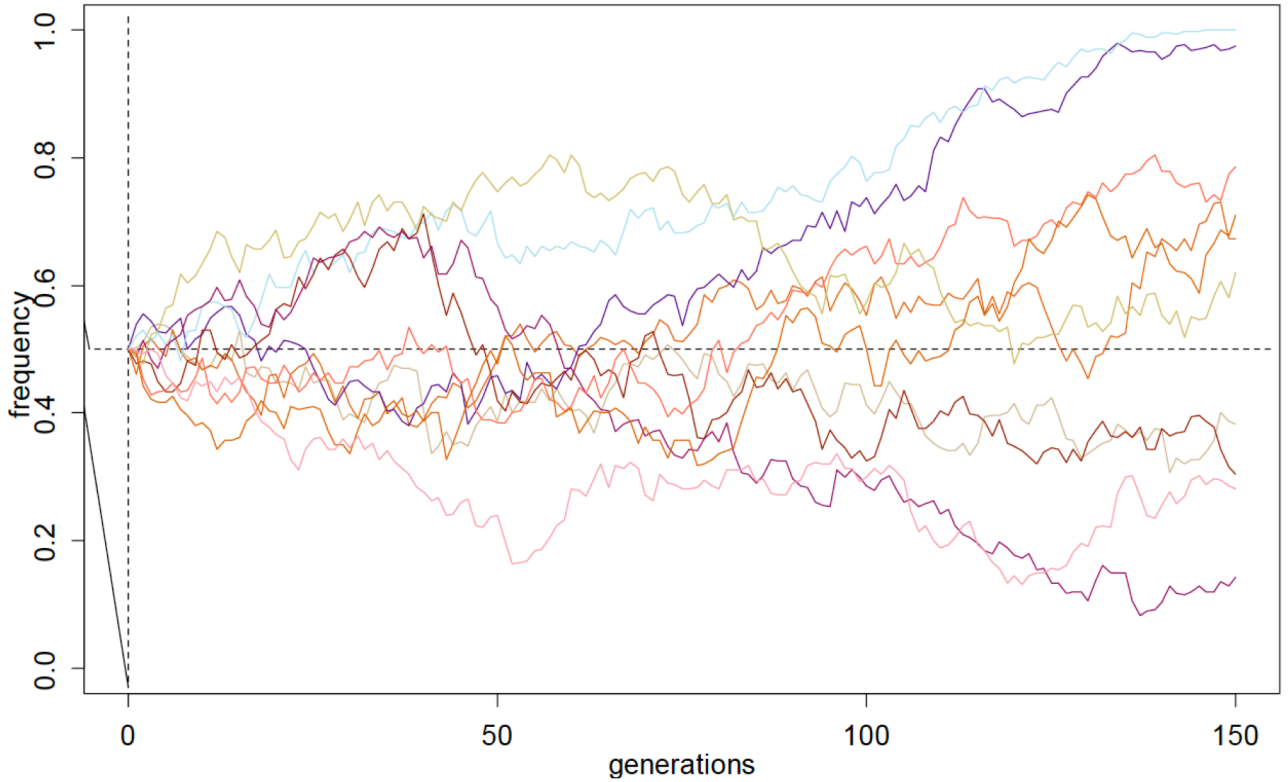
Courtesy: Vitor Sousa

Technical definition of fixation index, F_{ST}

- F_{ST} is a measure of genetic differentiation among populations (or subpopulations, demes), originally defined by Wright (1943) as:
- F_{ST} = (var (p) $_{\text{among subpopulations}}$) / (var ($p$) $_{\text{within total population}}$)
- This is equivalent to (var (p) $_{\text{sub}}$) / (p_{average} q_{average}), where the denominator is the maximum variance possible in the total population (note that for binomial sampling, var = n (p q)).
- F_{ST} can therefore be viewed as measuring the between-population variance relative to the maximum possible total variance for the total, combined population.
- It can thus range from 0 to 1 (100%). F_{ST} = 0 implies that two or more populations have identical allele frequencies (no differentiation), whereas F_{ST} = 1 implies that they have fixed different alleles.

Factors affecting F_{ST}

- See Vitor Sousa's first lecture
- When measured between a pair of populations separated some *t* generations ago, F_{ST} increases over time due to genetic drift – the variance in allele frequencies between the diverged populations will increase over time.



Courtesy: Vitor Sousa & Laurent Excoffier

Factors affecting F_{ST}

- In general, we can say that migration and mutation are analogous processes that increase diversity, whereas drift and selection reduce diversity.
- This means that genetic drift and isolation (no migration; e.g. isolation by distance) increase genetic differentiation, whereas migration decreases genetic differentiation.
- An equilibrium can be reached between drift (which increases differentiation) and migration (which reduces differentiation).

$$F_{ST} \approx \frac{1}{1+4Nm}$$

(2) Recombination & characterization of population structure with individual-based methods

Margot Paris

Department of Biology
Ecology & Evolution
University of Fribourg

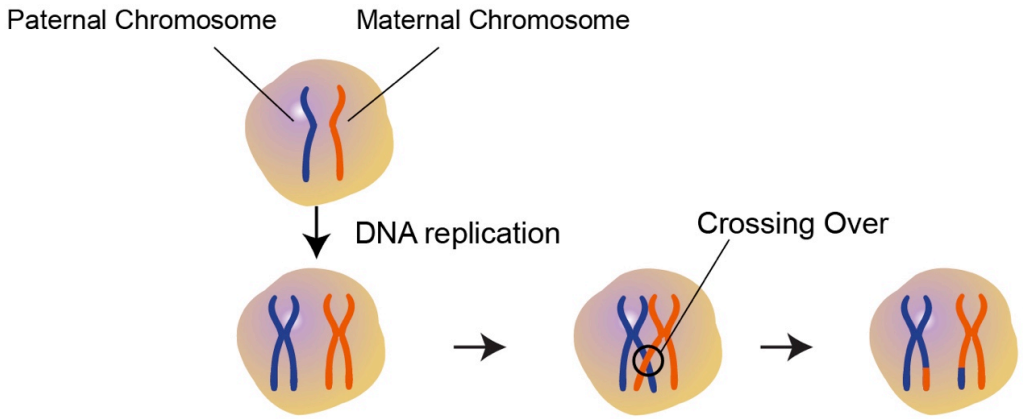
Structure of the lecture 2

1. A description of the factors affecting patterns of linkage disequilibrium (LD) between loci.

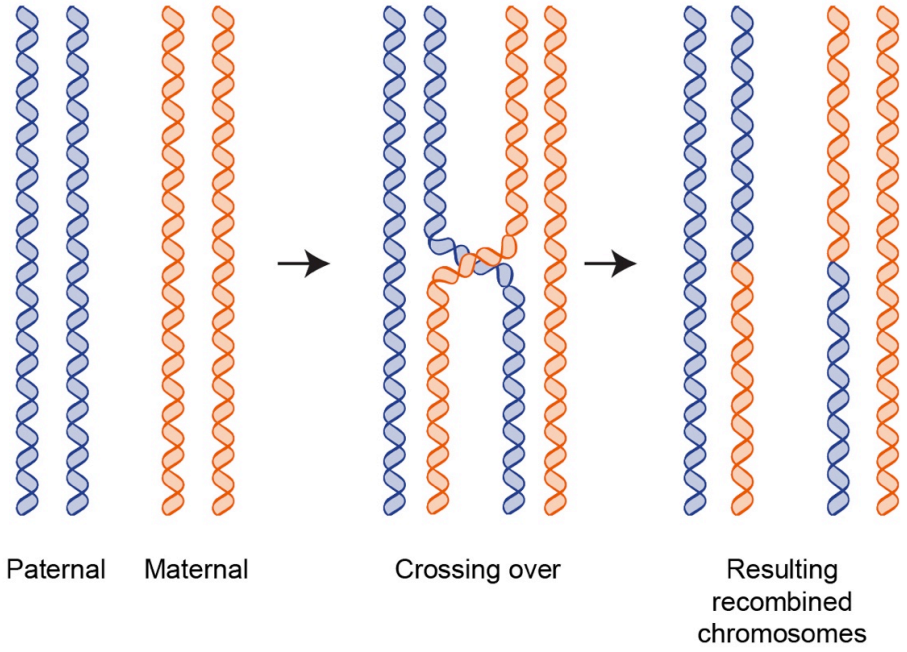
2. An account of how to study LD using linkage coefficients and local ancestry inference.

3. An account of how to characterize population structure with individual-based methods: Principal Component Analysis (PCA) and Structure-like clustering programs.

Recombination during meiosis



Recombination between 2 homologous chromosomes



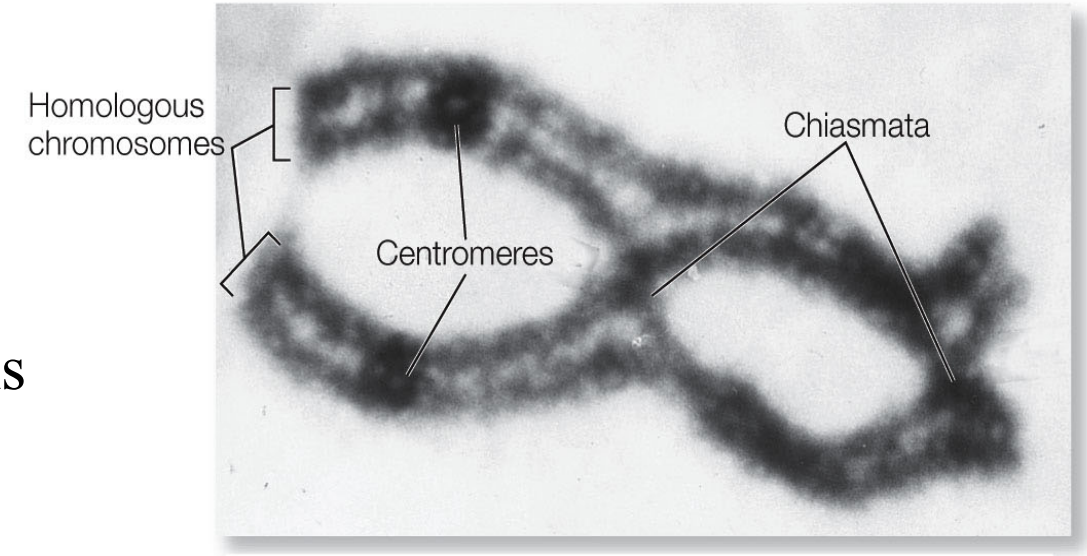
www.genome.gouv

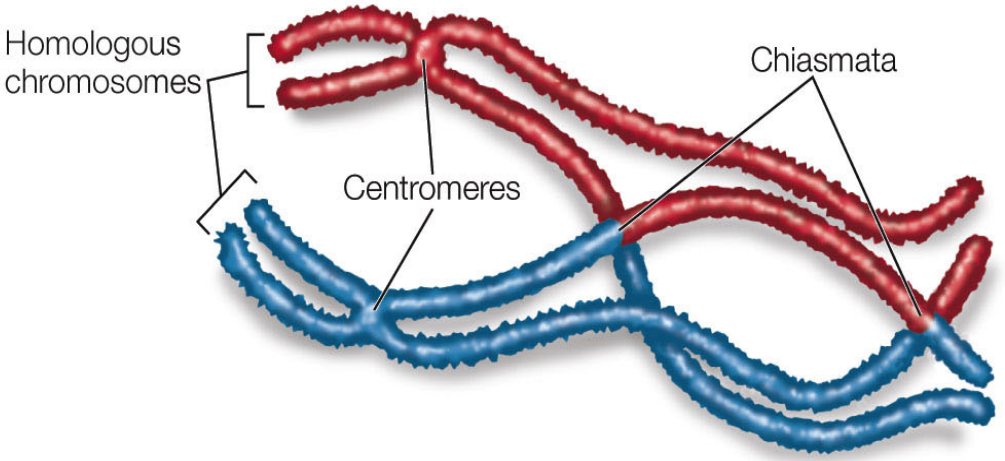
Thomas Morgan's work (1909)

- Thomas Morgan's work (1909) on drosophila leads to the discovery of linked genes and recombination due to crossing-over.

- He proposed that the chiasmata visible on chromosomes were regions of crossing-over.

- It mainly occurs between non-sister chromatids of homologous chromosomes.





Recombination varies among species/populations/individuals

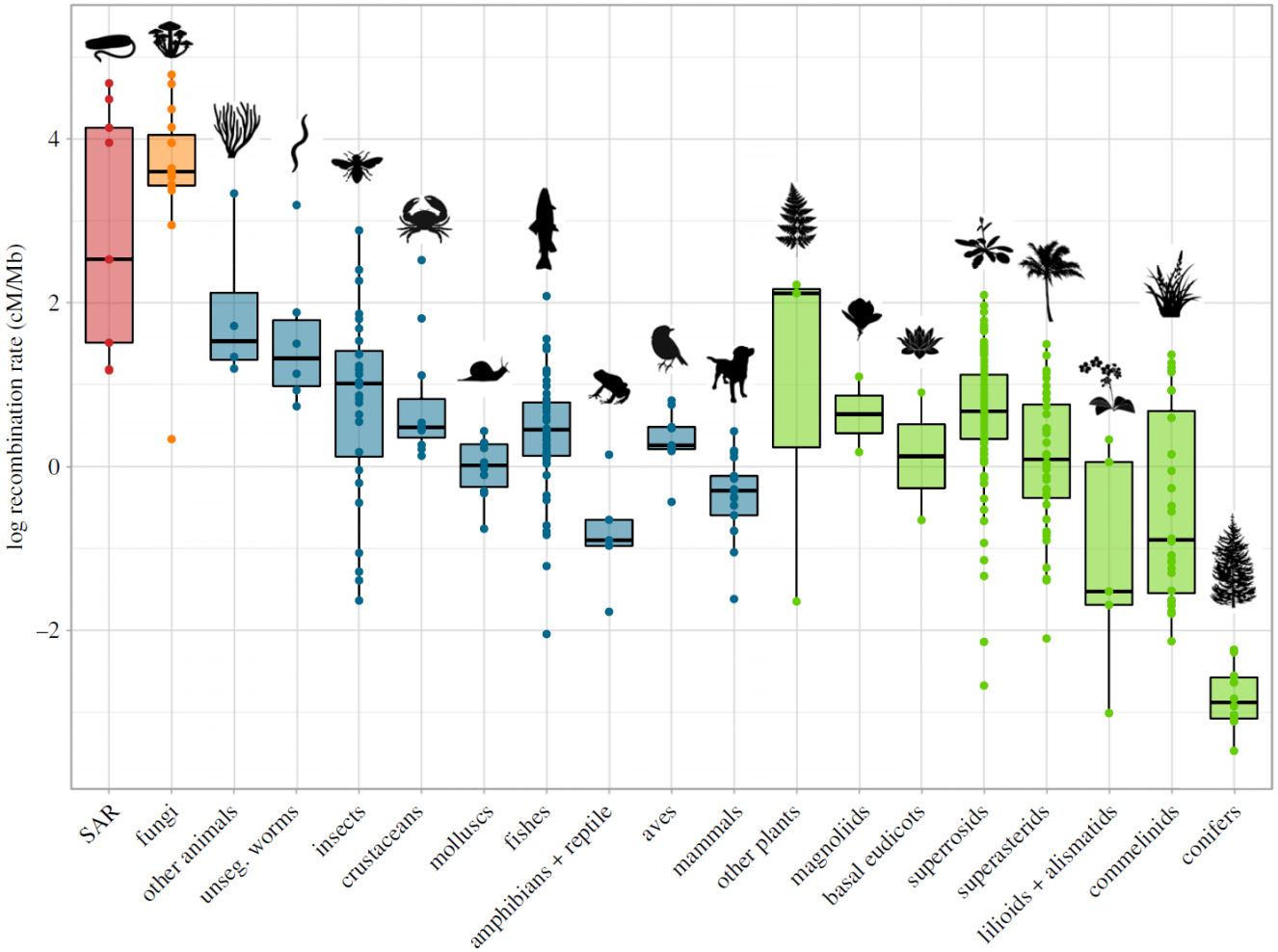


Figure 2. Variation in the log of recombination rate, estimated by dividing linkage map length in centimorgans (cM) by genome size (Mb) across eukaryotic taxa. Other plants: Pteridophyta, Chlorophyta, Bryophyta. Other animals: Anthzoa, Holothuriodea, Ascidacae. unseg, unsegmented.

Courtesy: Vitor Sousa; Stapley *et al.* (2017) *Phil. Trans. Roy. Soc.*

Because of recombination different regions of the genome have different gene trees

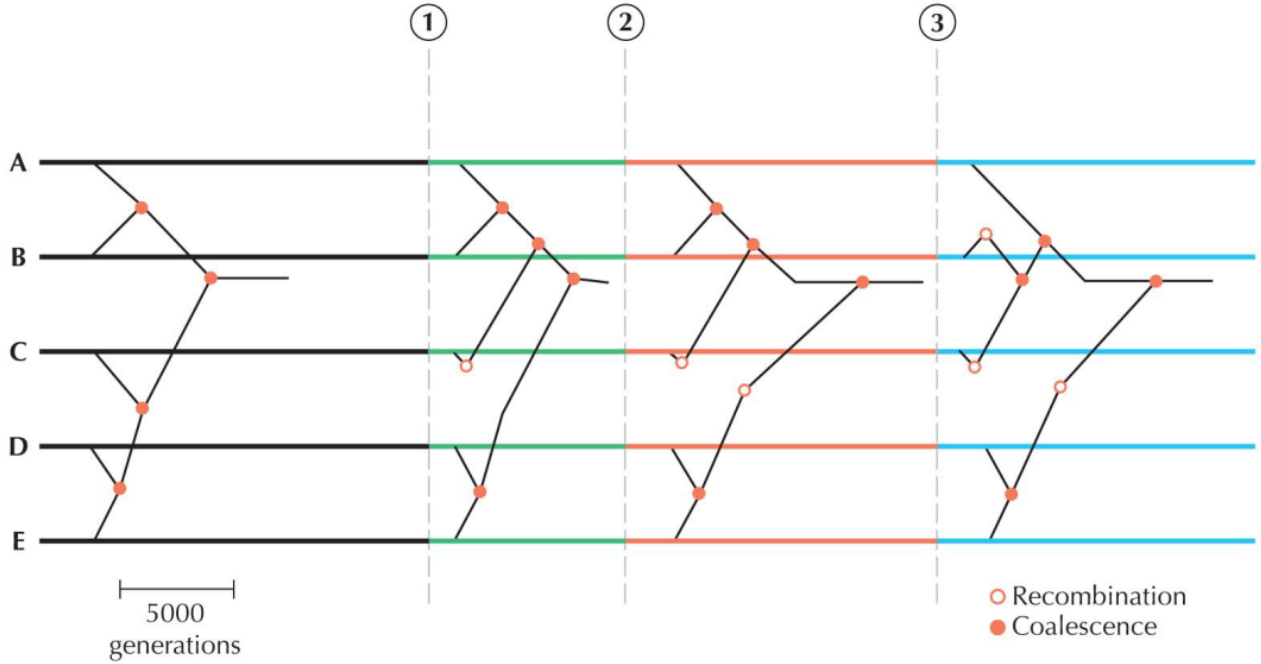


FIGURE 15.11. In a sexual population, different parts of the genome have different ancestry. This example shows a small region from five genomes (labeled A, B, C, D, and E). In the *leftmost section*, A and B share a common ancestor 2000 generations back; D and E share an ancestor 1000 generations back, and lineage (D, E) traces back to a shared ancestor with C 2000 generations into the past. The whole sample shares a common ancestor at 5000 generations. The genome C descended from an ancestral genome that underwent a recombination event 500 generations back; thus, the section to the *right* of position ① in the genetic map has a slightly different ancestry: C is more closely related to A and B than to D or E. Moving to the *right*, the next event is a recombination event at ② in the genetic map, which occurred 3000 generations back in the ancestor of D, E. This event did not change the qualitative relationships, but now the five genomes share a common ancestor 8000 generations into the past. Finally, moving to the *rightmost section*, a recombination event at 1000 generations makes B more closely related to C than to A.

Courtesy: Vitor Sousa

Measuring linkage disequilibrium (LD) between 2 loci with 2 alleles

- Lets consider two loci with alleles A_1, A_2 and B_1, B_2.

- If the two loci are independent: *freq*. (A1B2) = $P_{A1B2} = P_{A1}P_{B2}$.

- Thus, one measure of this association is:

$$D_{12} = P_{A1B2} - P_{A1}P_{B2},$$

- When $D = 0$, the two alleles are in linkage equilibrium

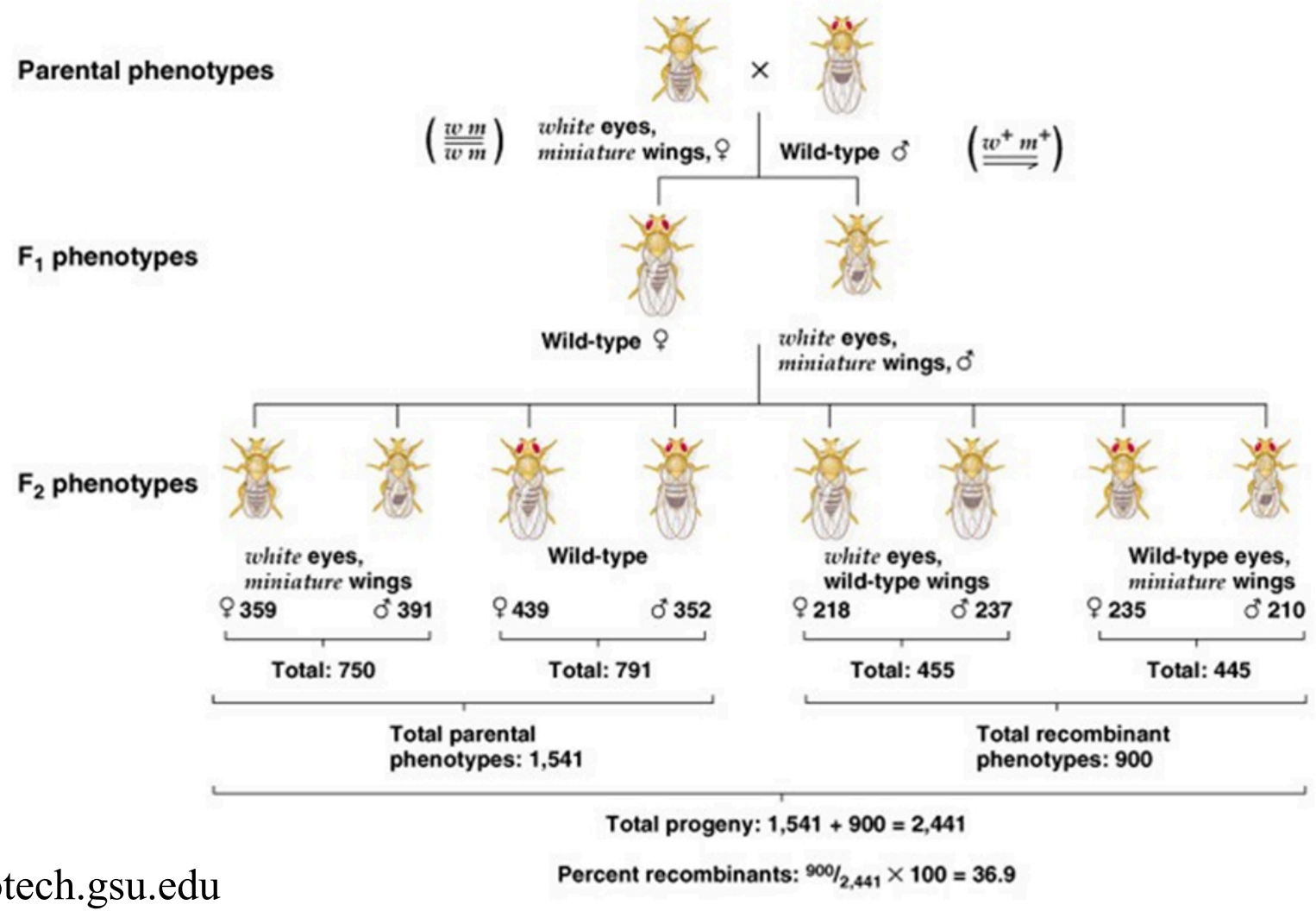
- LD can be generated by several factors:
  - physical linkage
  - natural selection
  - demographic effects: population structure, founders effects, admixture,…

Courtesy: Vitor Sousa

Thomas Morgan's work (1909)

- Example of crosses output for two linked genes located on X crhomosomes

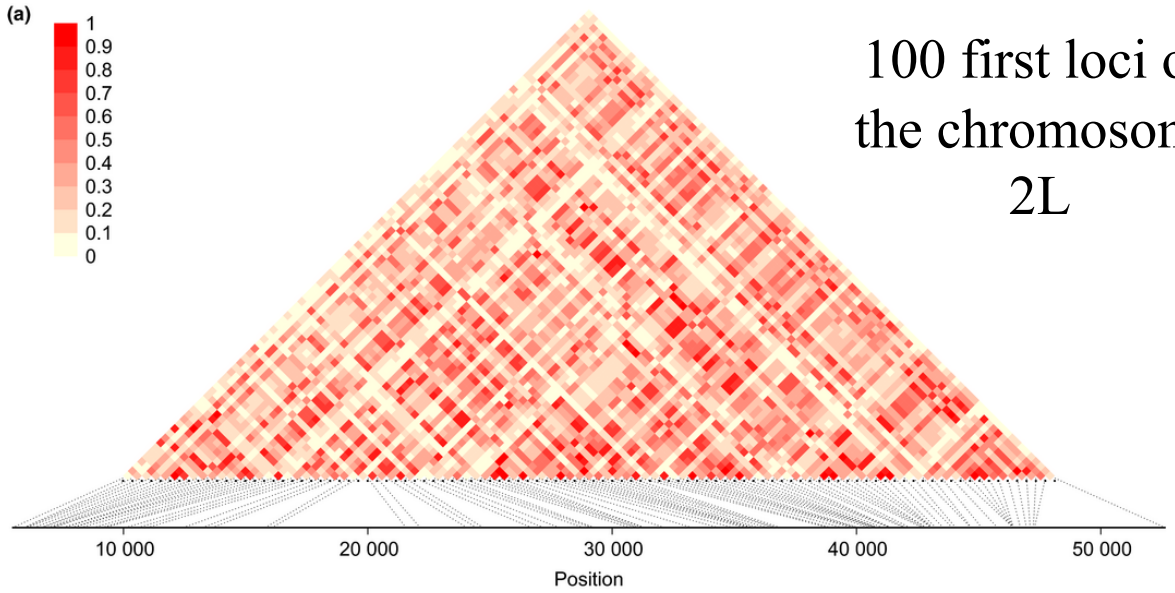
Fig. 14.1, Morgan's experimental cross of white-eyes and miniature wings.



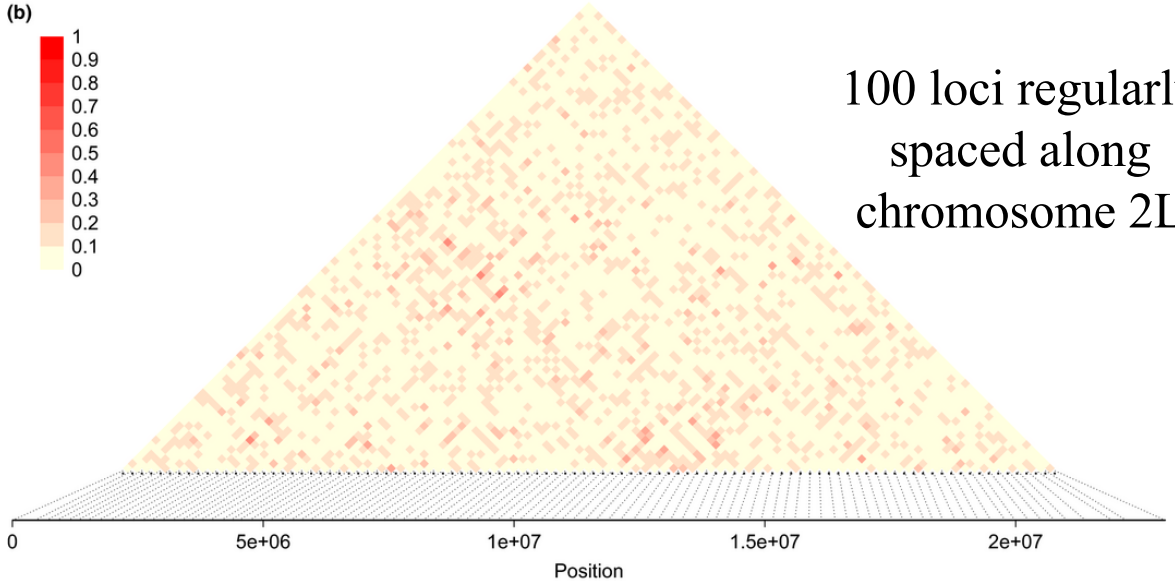
http://biotech.gsu.edu

LD decay with physical distance

- LD maps at two different genomic scales for the fruit fly data.
- The horizontal axis indicates the position of the loci on the chromosome.
- The linkage coefficients (r^2) between each pair of loci are indicated as coloured squares: the squares at the bottom of the triangle are for nearby loci, whereas the square at the top is for the two most distant loci.



100 first loci on the chromosome 2L



100 loci regularly spaced along chromosome 2L

Courtesy: Vitor Sousa; Paradis *et al.* (2016) *Molecular Ecology*

LD decay with physical distance

- Because recombination can occur with small probability at any location along chromosome, the frequency of recombination between two locations depends on the distance separating them.

- Therefore, for genes sufficiently distant on the same chromosome, the amount of crossover is high enough to destroy the correlation between alleles.

Wikipedia.org

LD decay though time

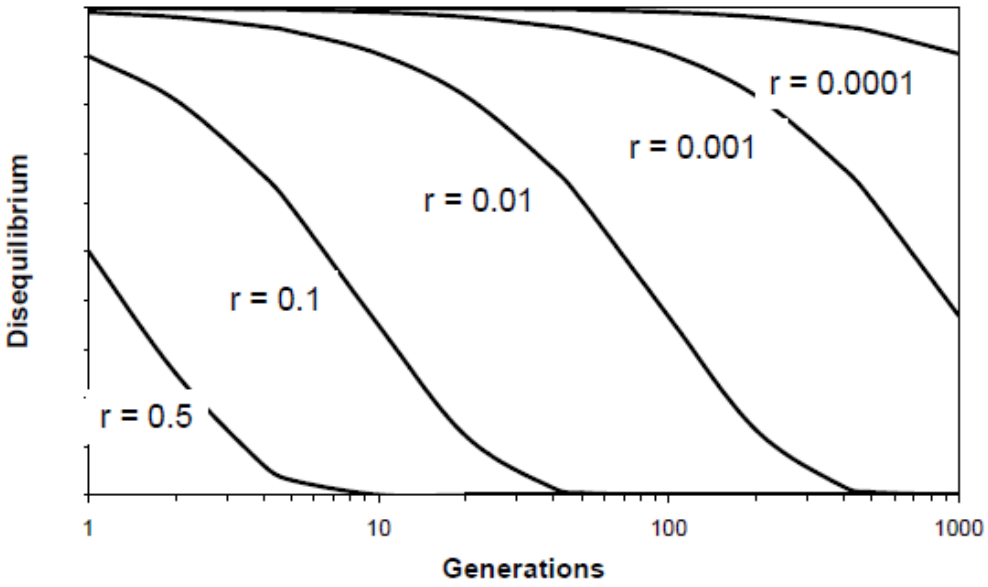
- Under the assumptions of a large effective population size, given r recombination rate between loci A and B:

$$p_{AiBj}(t) = (1-\boldsymbol{r})p_{AiBj}(t\text{-}1) + \boldsymbol{r}\, p_{Ai}\, p_{Bj},$$

- Which can be written as :

$$D_t = (1 - \boldsymbol{r})\, D_{t-1} = (1 - \boldsymbol{r})^t\, D_0$$ with D_0 = LD at time zero

- LD decreases between two loci geometrically as a function of r.
- For two independent loci, r = 0.5 and LD is reduced by half at each generation.
- The observed LD patterns are a function of demographic history and selection.



Courtesy: Vitor Sousa

Recombination hotspots in human genes

- In humans and other mammals recombination hotspots seem related with the binding of PRDM9 protein.

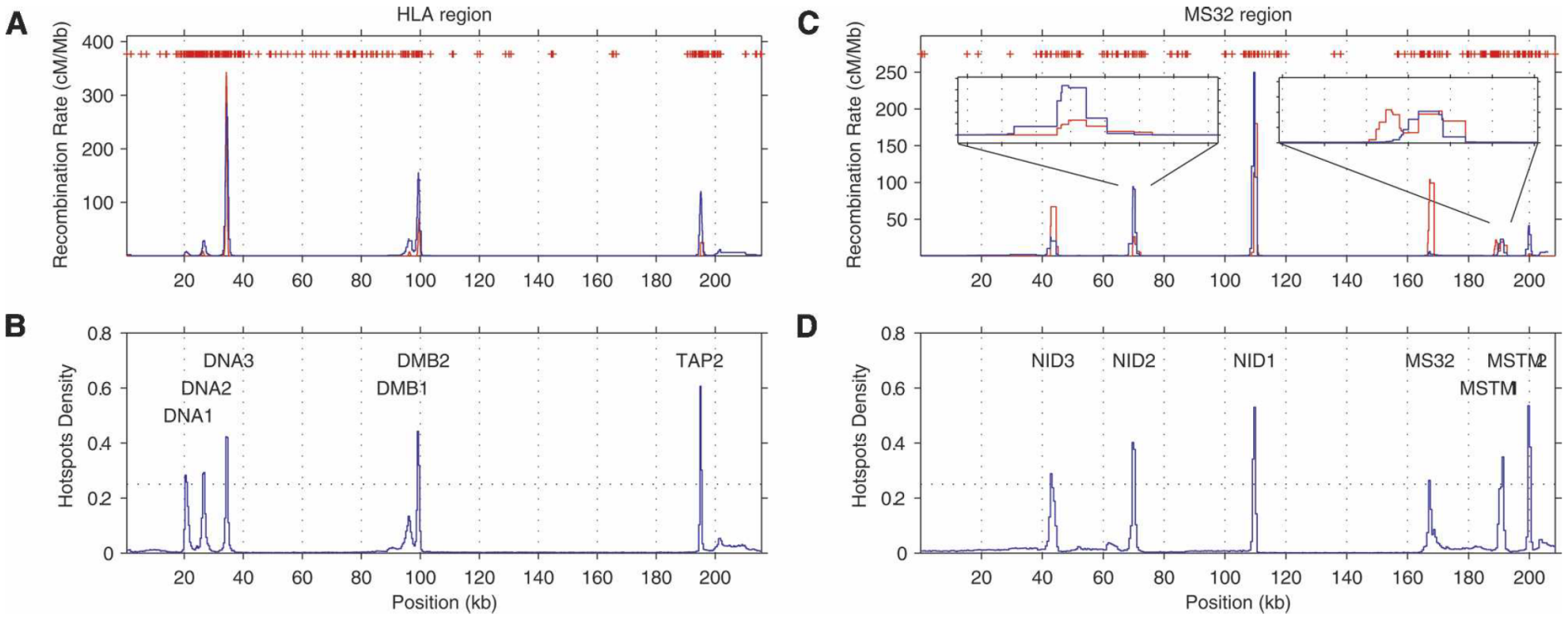
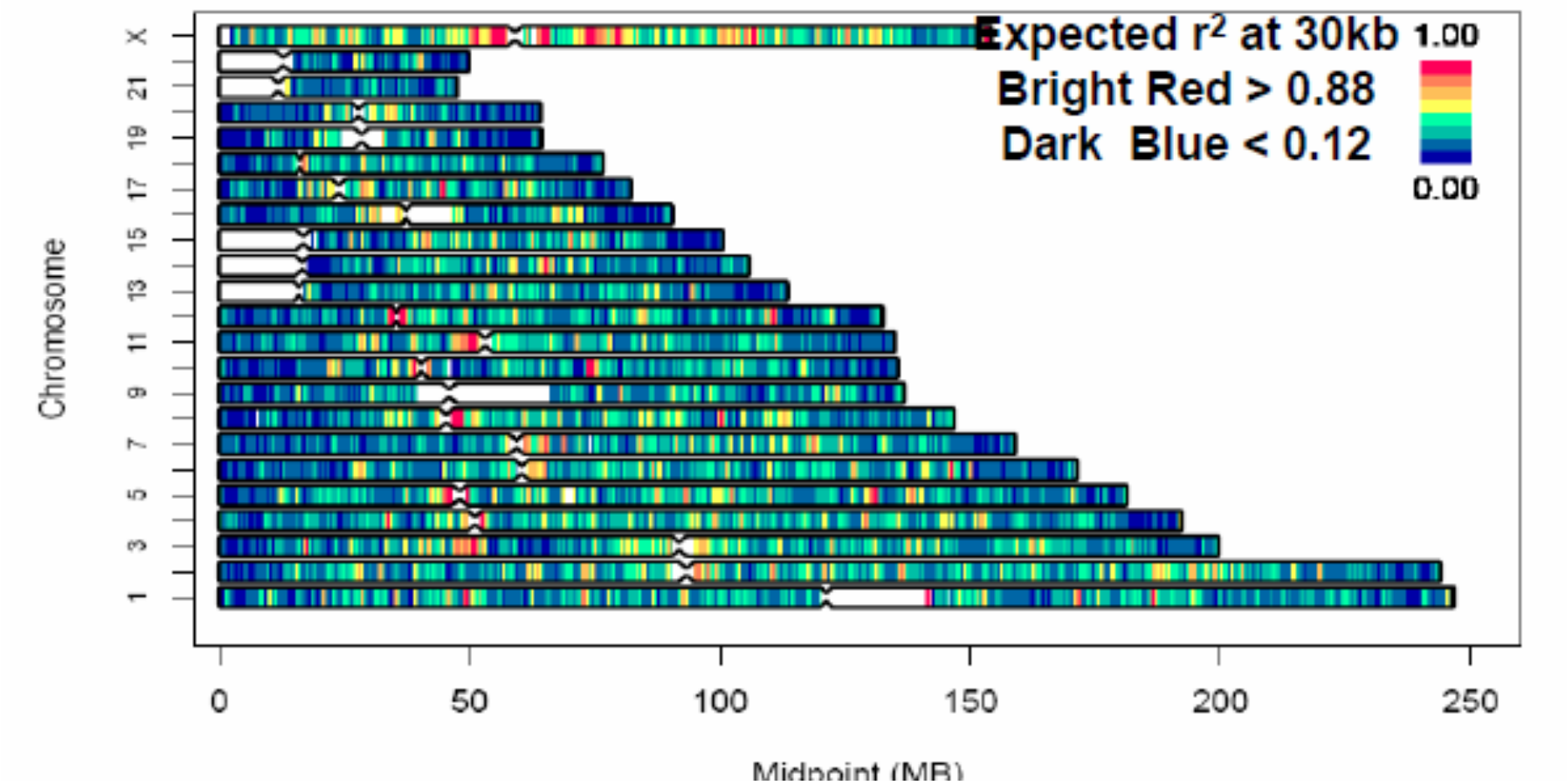


Figure 6. Output of *rhomap* for the HLA and MS32 regions. Plots *A* and *C* show the recombination rate estimates of the HLA and MS32 regions respectively, with the estimated rate in blue, and (sex-averaged) sperm typing rate in red. SNP locations are shown as red marks. Estimates from *rhomap* were converted to cM/Mb by assuming N_e = 10,000. Also shown in plot *C* is the detail of the NID2a/b and MSTM1a/b estimates. Plots *B* and *D* show the average number of hotspots per sample per kb for the same regions.

Courtesy: Vitor Sousa; Austen & McVean (2007) *Genome Research*

Patterns of LD vary among chromosomes

Genomic Variation in LD (CEPH)



Courtesy: Vitor Sousa

The coalescent with recombination

- Recombination happens at rate :

$$\rho = 4N_e r$$

where *Ne* is the effective size and *r* is the recombination rate between 2 sites per generation.

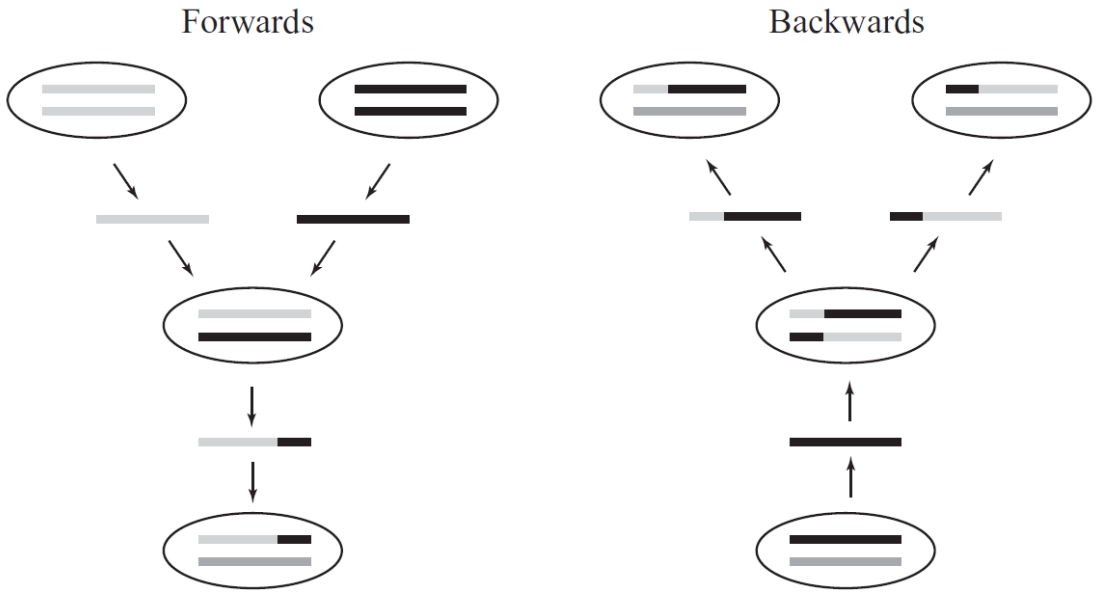
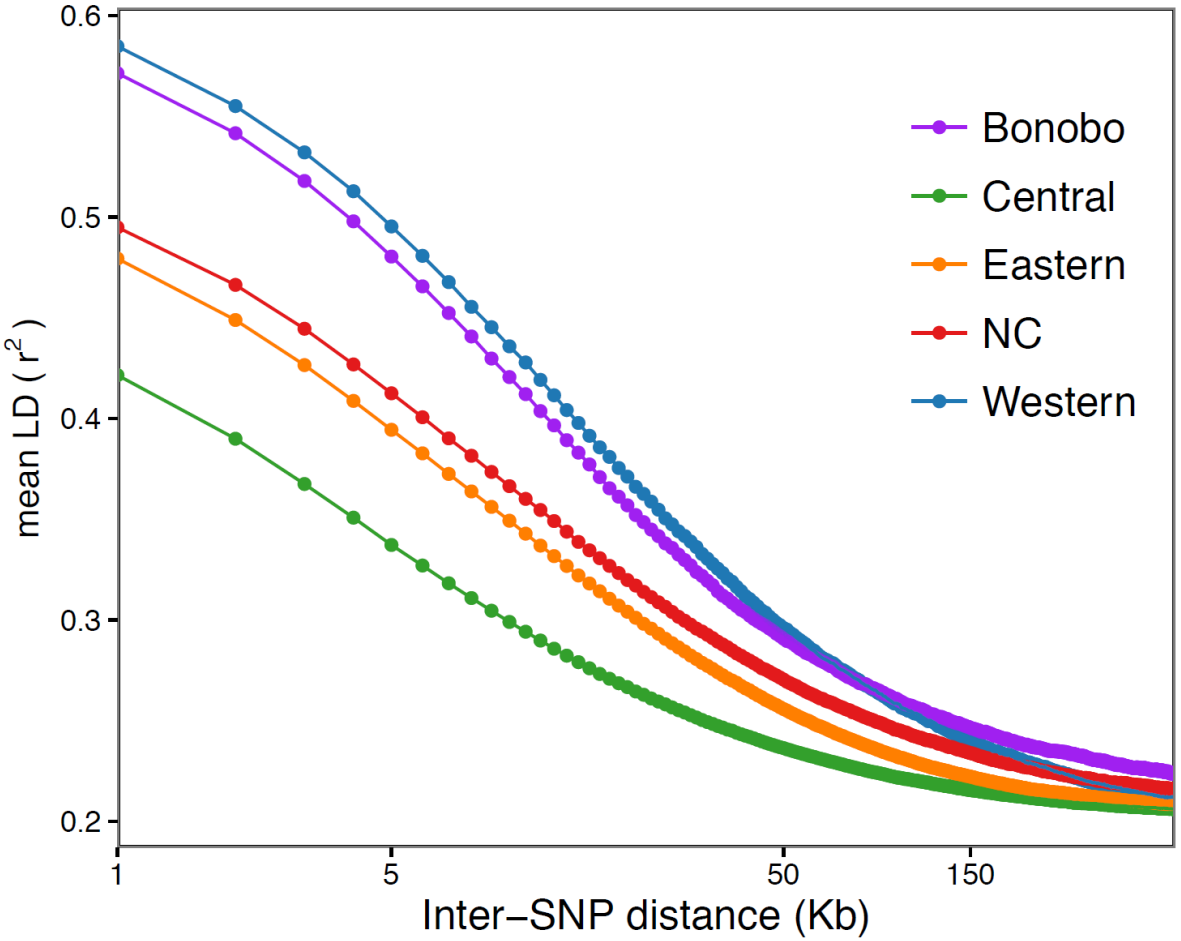


Figure 5.5 Hudson's recombination model on a continuous representation of a sequence. A sequence is made by recombination when an individual creates a haploid genome (sperm cell or egg). Looking forwards, two sequences are recombined into one recombinant sequence. Knowing the allelic states of the grandparent's chromosomes determines one of the child's two chromosomes; the other, the dark grey, originates from the second set of grandparents. Looking backwards, an individual chooses a chromosome from a parent. This chromosome is split onto two grandparental chromosomes. The child's dark grey chromosome is inherited through the other parent and the dark grey chromosomes in grandparents have unknown allelic states.

Courtesy: Vitor Sousa

LD patterns differ among populations



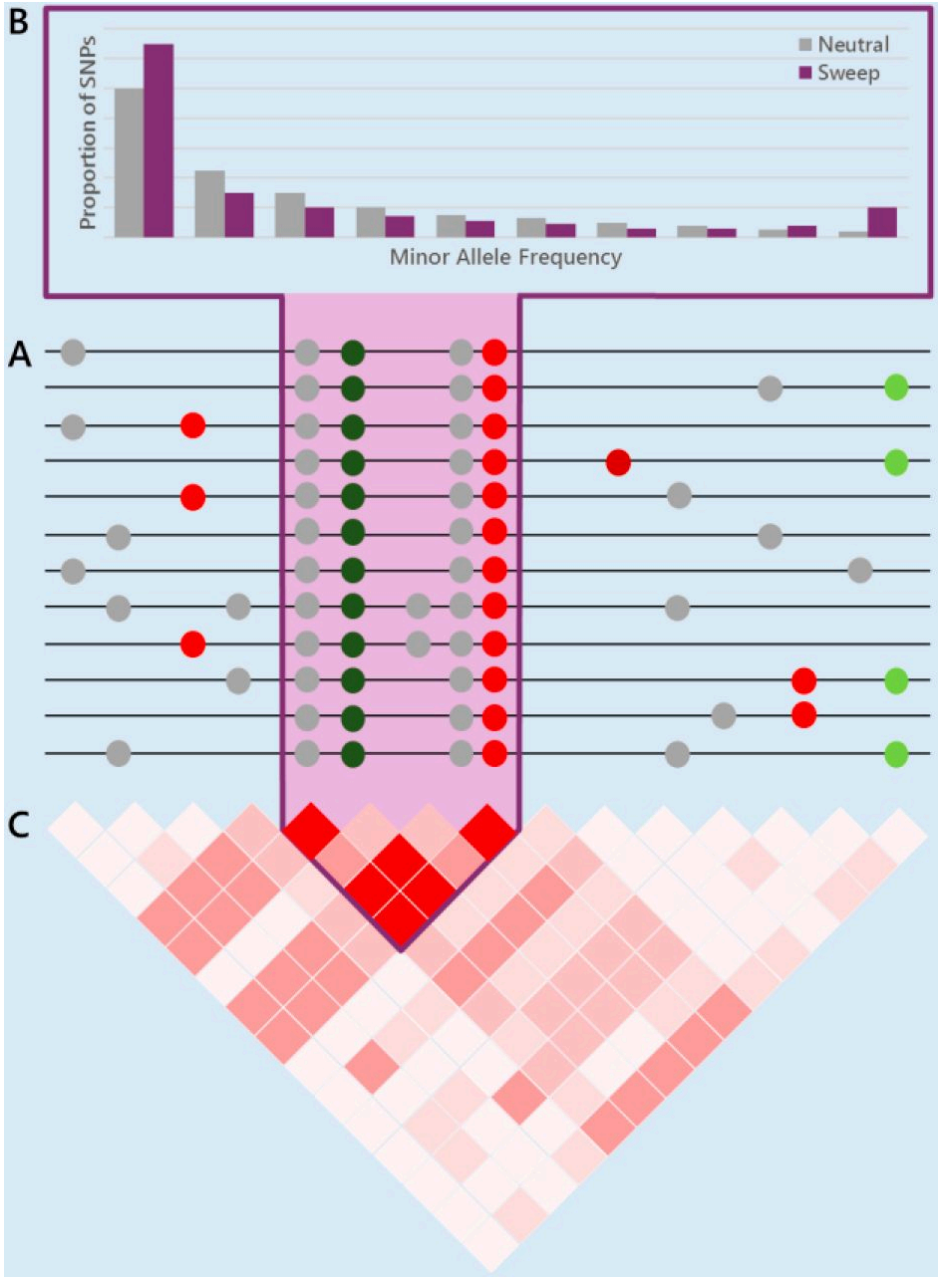
- Example: Bonobos and chimpanzees
- Higher LD in populations with smaller N_e, as the rate of recombination is lower

Decay of linkage disequilibrium (LD) in the *Pan* populations. Note that X-axis is in log10 scale. Colours represent: Purple; bonobo. Green; central chimpanzee. Orange; eastern chimpanzee. Red; Nigeria-Cameroon chimpanzee. Blue; western chimpanzee.

Courtesy: Vitor Sousa; De Manuel et al. (2016) Science

Selective sweeps (selection) can locally affect LD

- Higher LD is observed in genomic regions affected by selective sweeps, as N_e is strongly reduced locally due to selection.



Signatures of a selective sweep in the genome (A) A reduction in genetic diversity, (B) a skew toward rare derived alleles, and (C) an increase in LD (see text for details). Colored ● reflects different classes of mutations according to their fitness effects: maroon, strongly deleterious (very infrequent, in their way to elimination by natural selection); red, slightly deleterious; gray, neutral; light green, slightly advantageous; dark green, advantageous. Note that in the region of the selective sweep (purple), an advantageous mutation has been driven to fixation together with its linked neutral and nearly neutral variants. In this region, genetic diversity is reduced, most polymorphisms are shared among different chromosomes (high LD), while recently arisen mutations are still at low frequency (gray ● present in two chromosomes).

Casillas & Barbadilla (2017) *Genetics*

Patterns of LD and migration/introgression

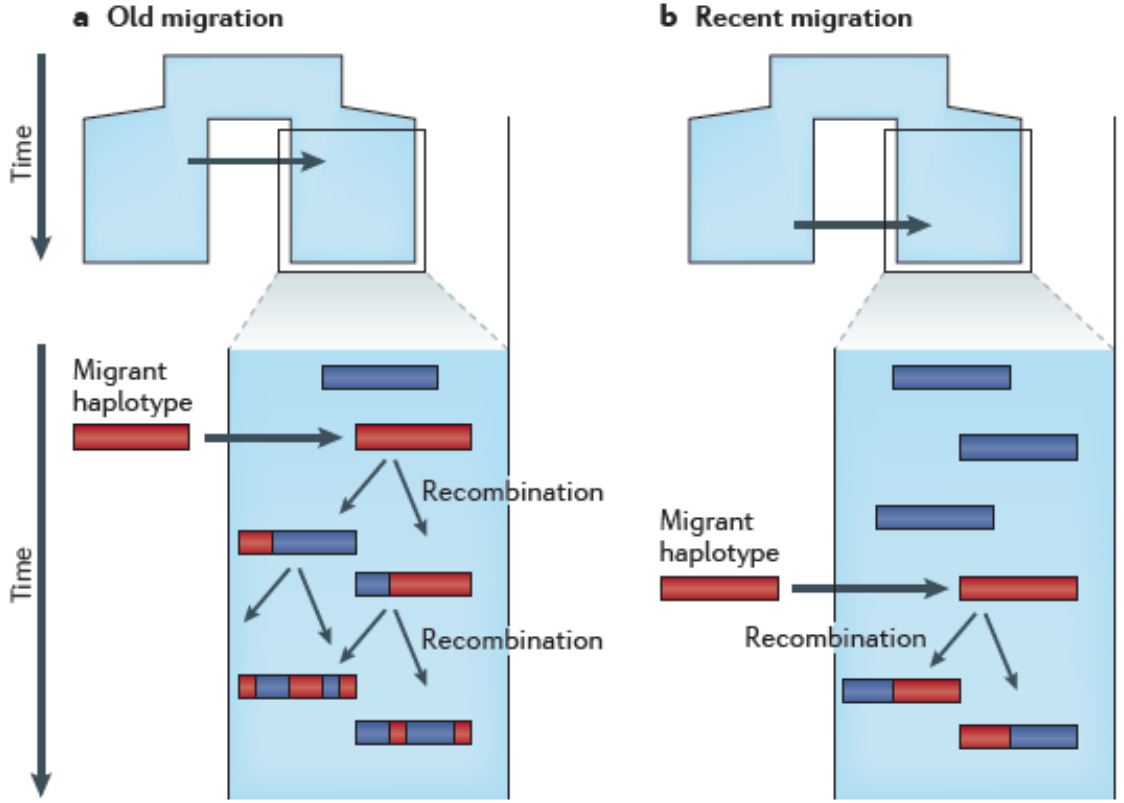


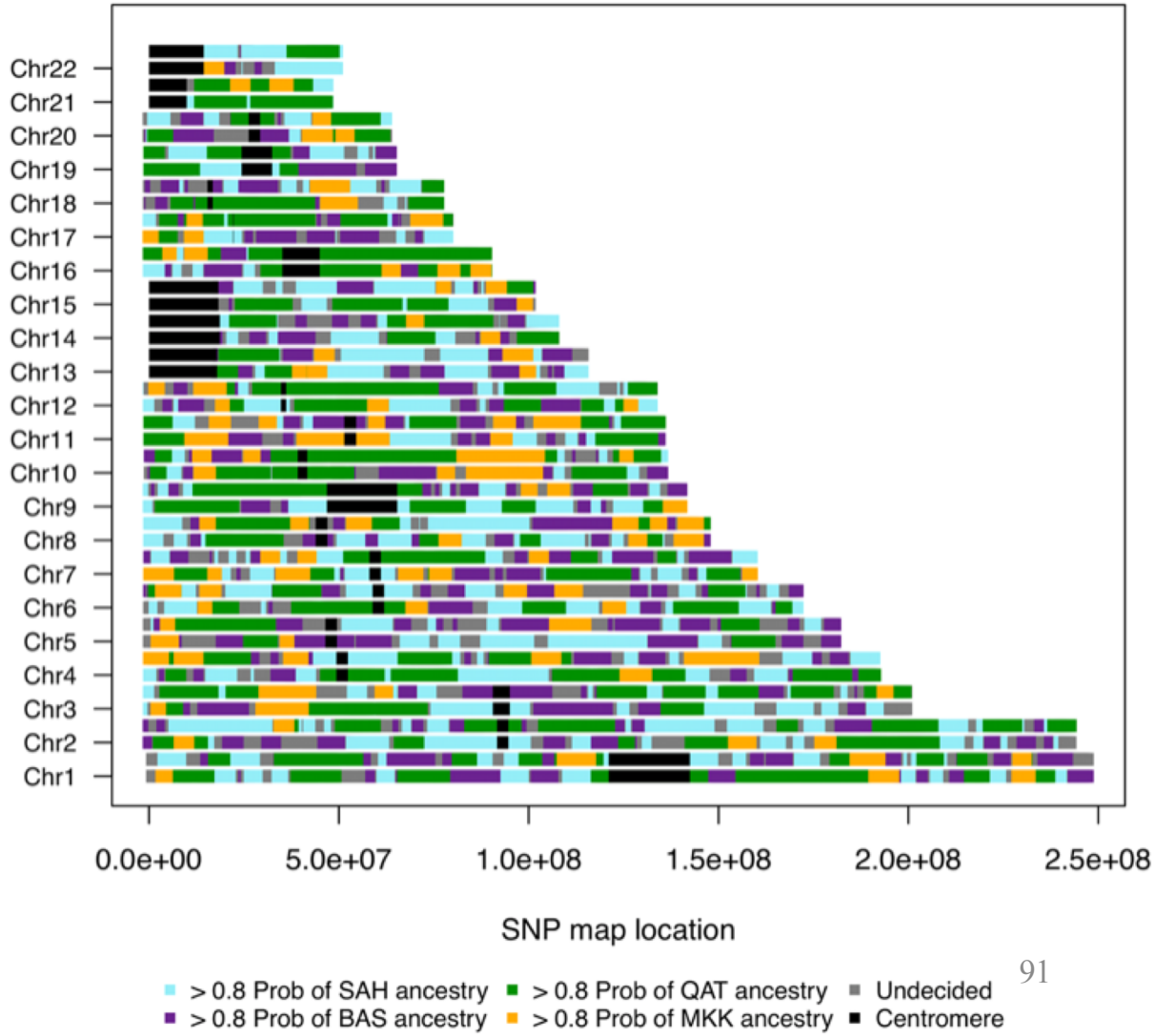
Figure 4 | **Distinguishing migration events based on linkage disequilibrium block structure.** Schematic representation of the expected distribution of the haplotype block lengths for an old migration event (**a**) and a recent migration event (**b**). The diagram shows two diverging populations that experience migration at some time in the past after the split and a zoom-in of what happens at the population that receives immigrant haplotypes. For simplicity, we assumed that all individuals share the same haplotype in the destination population (blue haplotype in the figure): that is, this haplotype has reached fixation. When a migrant haplotype (shown in red in the figure) enters a population, as times goes by, recombination breaks it into smaller fragments. Thus, blocks are expected to be shorter following an old migration event (**a**) than directly after a recent migration event (**b**), for which blocks are expected to be larger.

Sousa & Hey (2013) *Nature Reviews Genetics*

Local ancestry inference

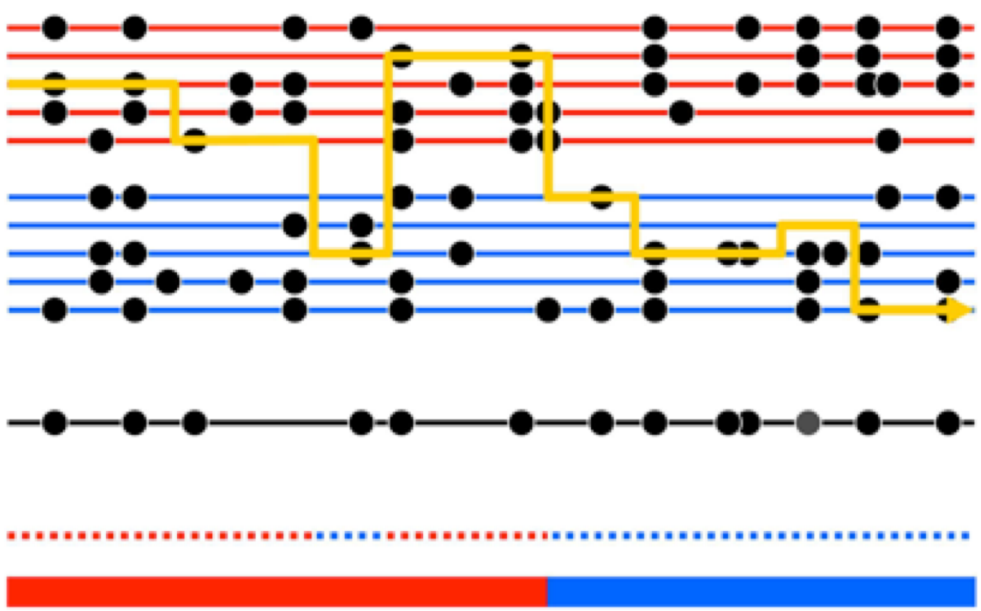
- Find population of origin of each SNP
- For example, chromosome painting of a single Egyptian individual:

- Breakpoints reflect recombination events
- The time at which the ancestor lived vary from segment to segment
- Longer segments are associated with more recent shared ancestor



Henn *et al.* (2012) *PLoS Genetics*

Ancestry inference using Hidden Markov Model (HMM)



2 parental populations

Observed genotype

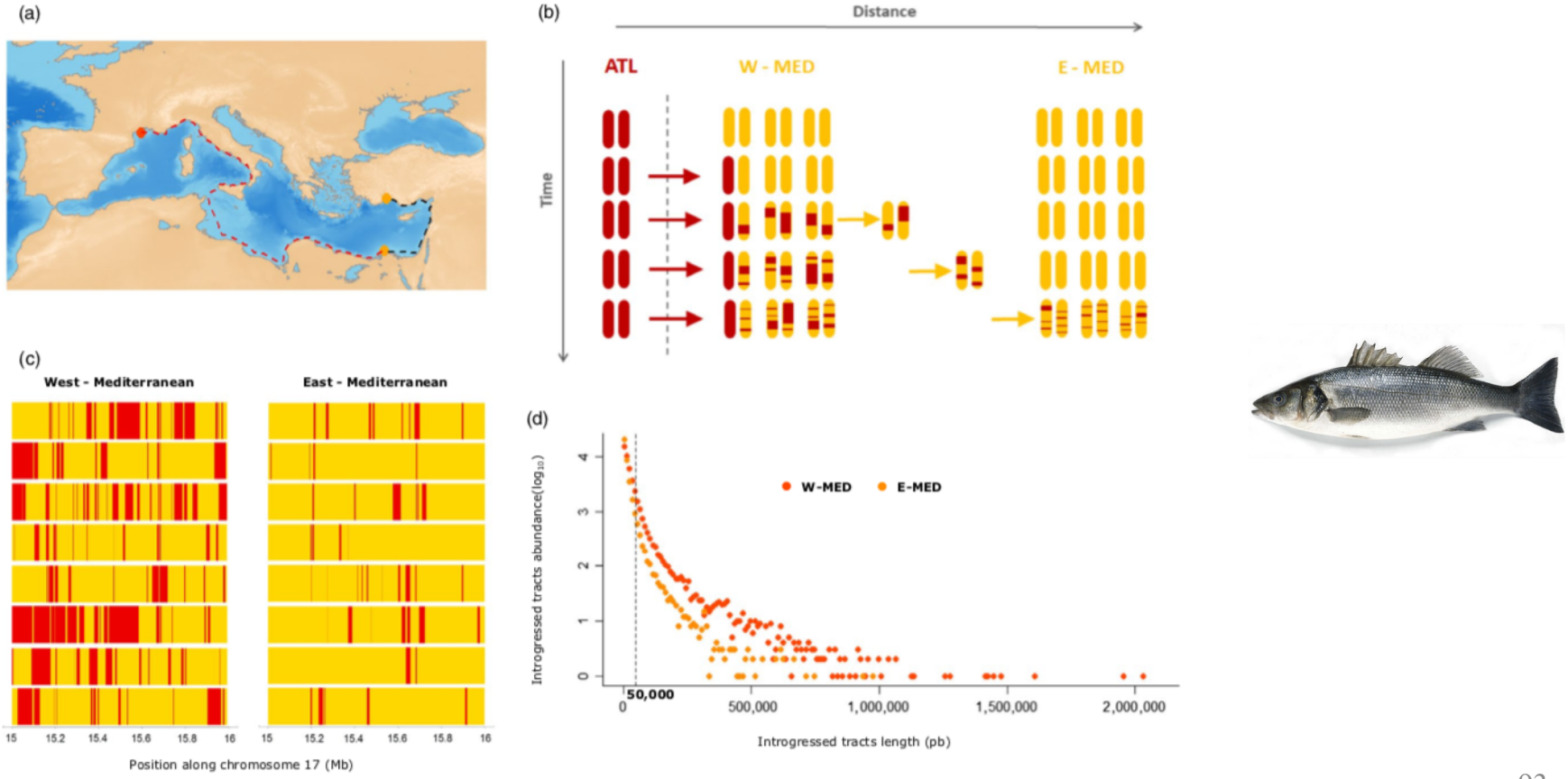
Ancestry (hidden state)

- At each position in the genome, the model estimates the likelihood that a haplotype from an admixed individual is a better statistical match to one reference population or the other.
- HMM is used to combine these likelihoods with information from neighboring loci, to provide a probabilistic estimate of ancestry at each locus.

Price *et al*. (2009) *PLoS Genetics*

Why local ancestry inference?

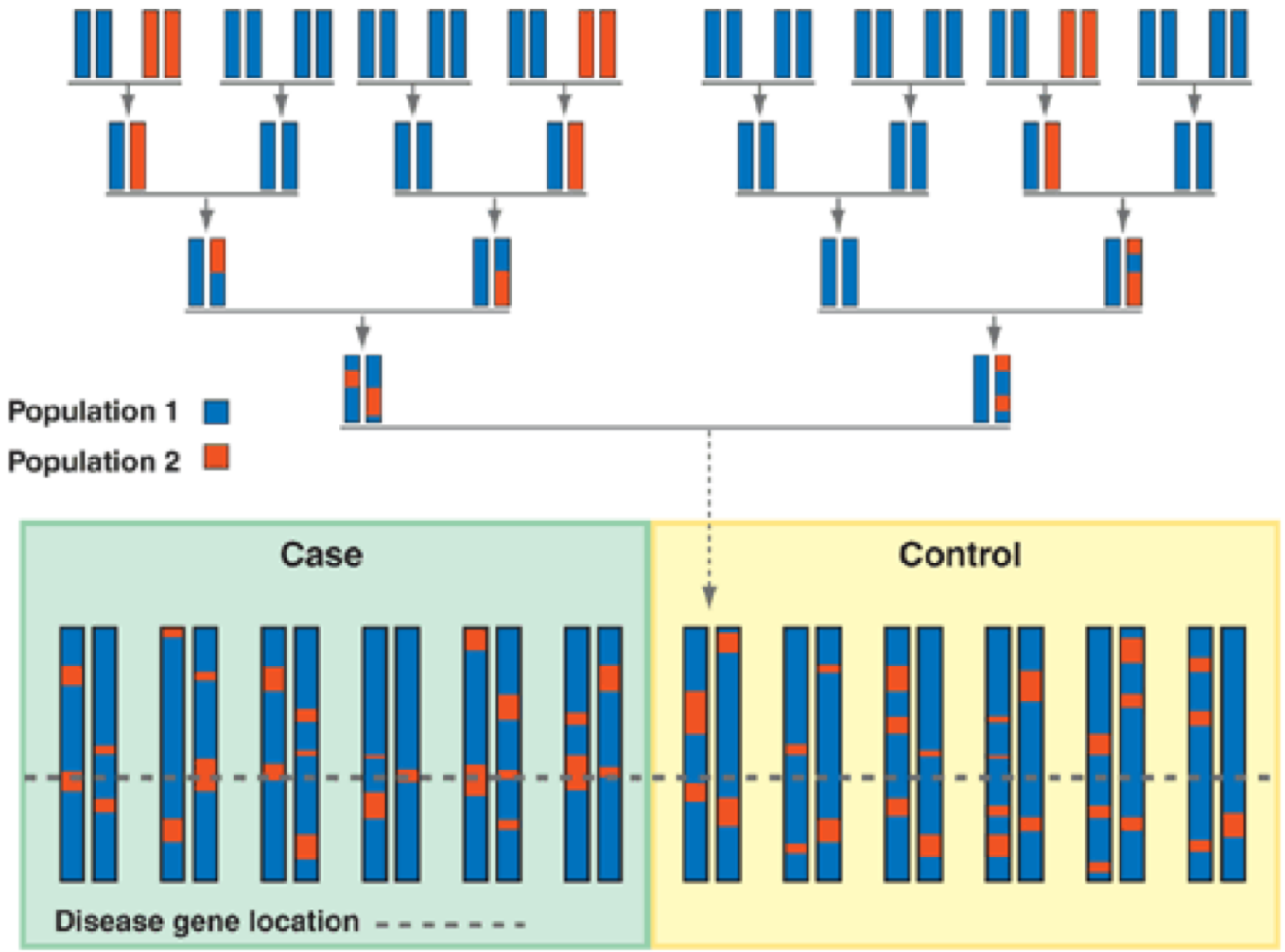
- Characterize admixture events (time, proportion) during a species history
- Example: introgression and diffusion of Atlantic tracts within the Mediterranean genetic background in see bass



Duraton *et al*. (2018) *Evol. App.*

Why local ancestry inference?

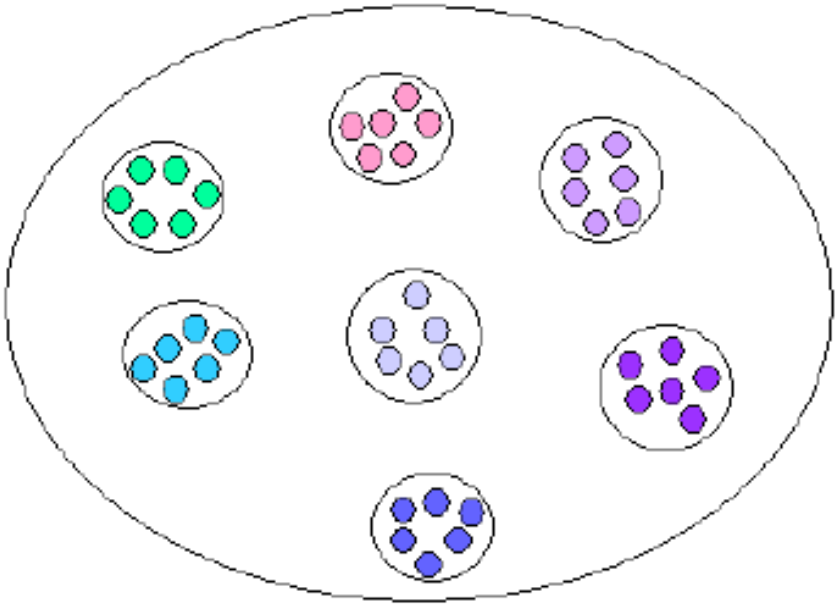
- Explore population-specific disease predisposition



Darvasi & Shifman (2005) *Nature Genetics*

Population structure

- Natural populations are usually subdivided
  - Discontinuous habitat (ponds, lakes, trees)
  - resource limitation
  - Seasonality
  - Behavioral (social or mating system)



- F_{ST} and related measures depend on a priori definition of sampling units

- However, in many cases we do not know if individuals sampled in a given location correspond to a population.

Courtesy: Vitor Sousa

Characterization of population structure with individual-based methods

- Consider each SNP to be independent.
- Look at patterns of allele frequencies across all the SNPs along the genome.
- Information about the linkage disequilibrium among SNPs is discarded. Two SNPs nearby in the same chromosome are treated as SNPs in two different chromosomes that segregate independently.

Courtesy: Vitor Sousa

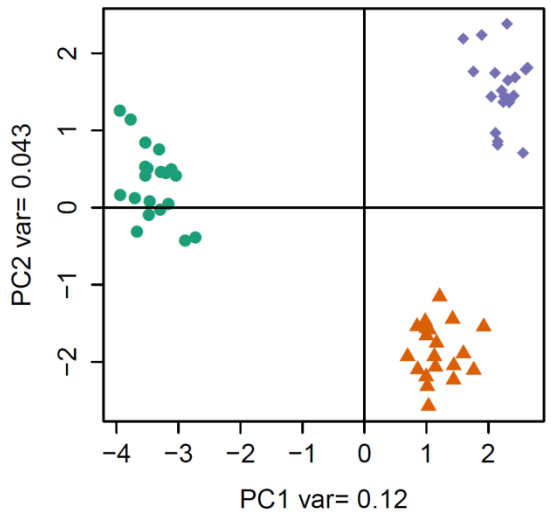
Principal Component Analysis (PCA)

- Principal Component Analysis (PCA) of most likely genotypes is an efficient approach to deal with large amounts of data (Patterson *et al.* 2006)

Alternative allele frequencies

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Ind 1	0.0	1.0	0.0	0.5	0.0	1.0
Ind 2	0.5	0.0	0.0	1.0	NA	1.0
Ind 3	NA	0.5	0.5	1.0	0.0	0.0
Ind 4	0.0	0.5	NA	1.0	0.0	1.0
Ind 5	0.0	0.5	0.5	1.0	0.5	0.5

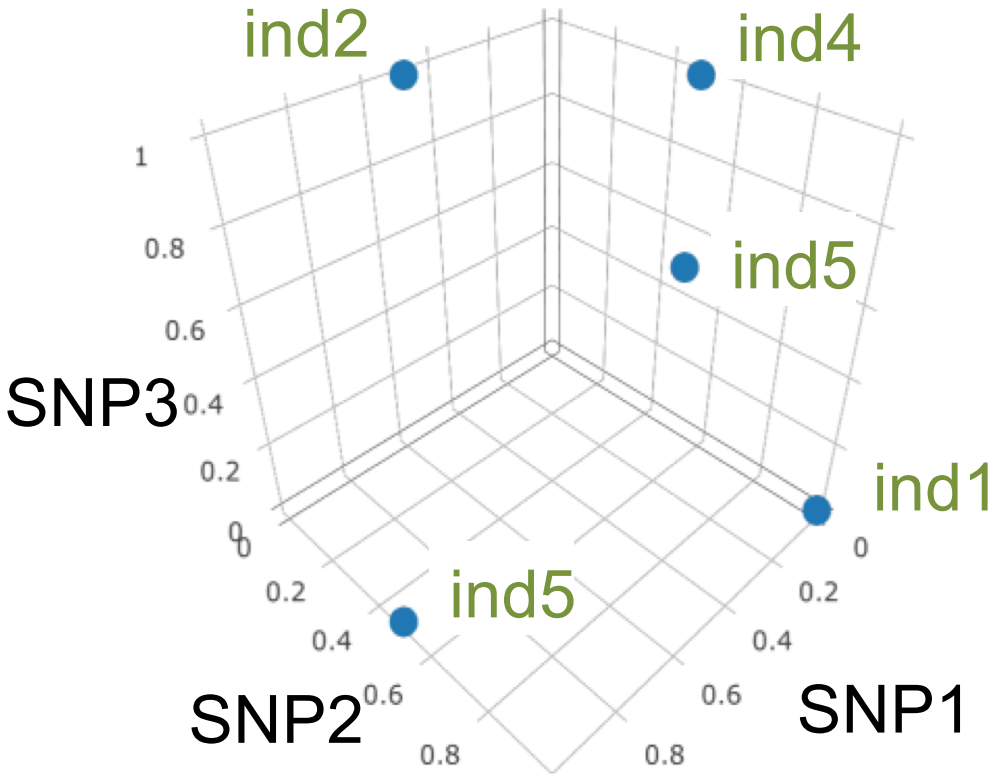
Reduce dimensionality



Courtesy: Vitor Sousa; Patterson *et al.* (2006) *PLoS Genetics*

Principal Component Analysis (PCA)

- Each SNP is treated as a dimension.
- The aim is to explain distance between individuals with few dimensions.

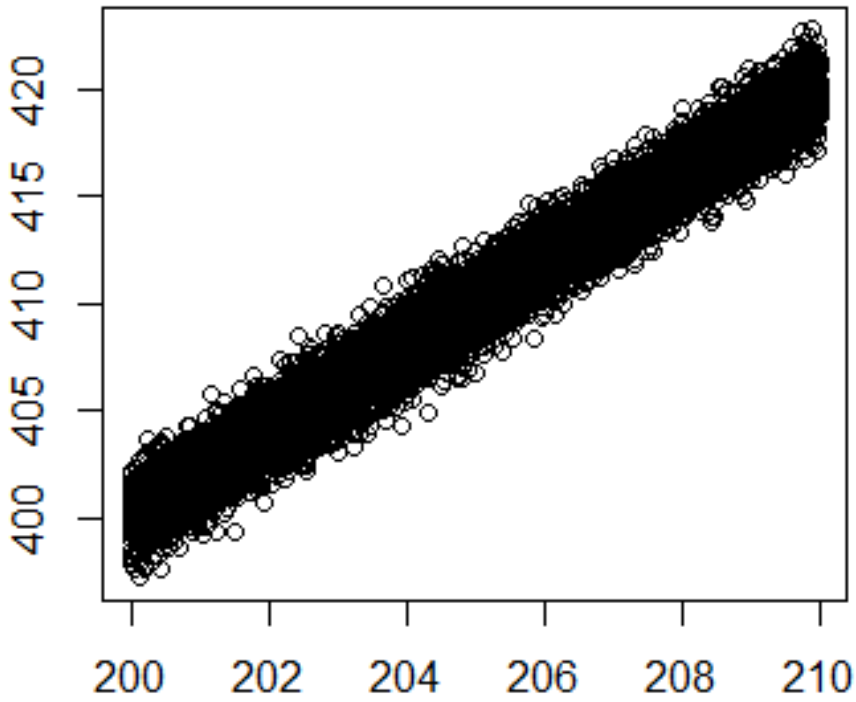


- Now imagine we have millions of SNPs…

Courtesy: Vitor Sousa

Example in toy example data: two variables

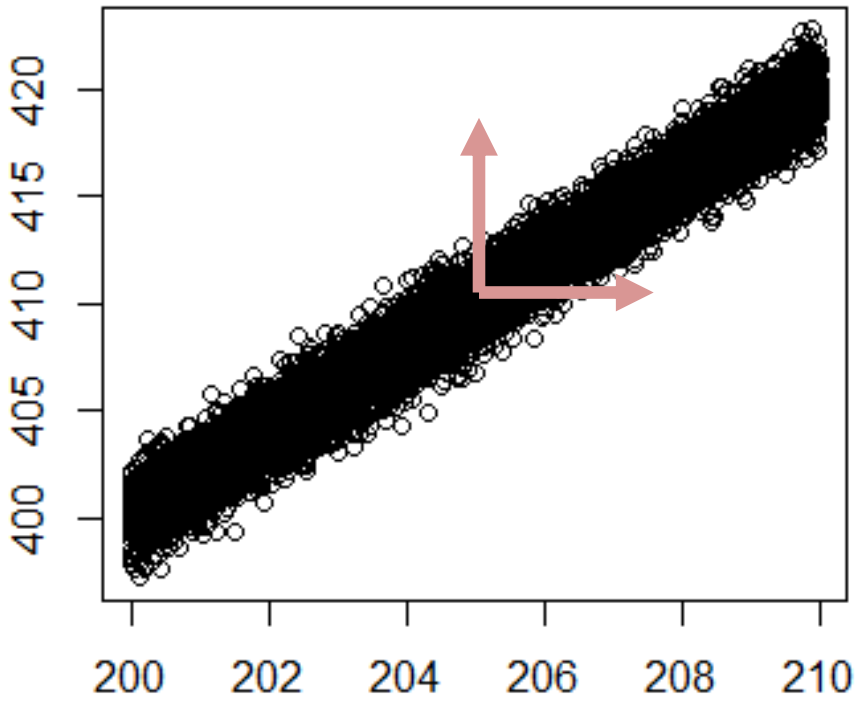
- Do we really need two dimensions to explain this data?



Courtesy: Vitor Sousa

Example in toy example data two variables:

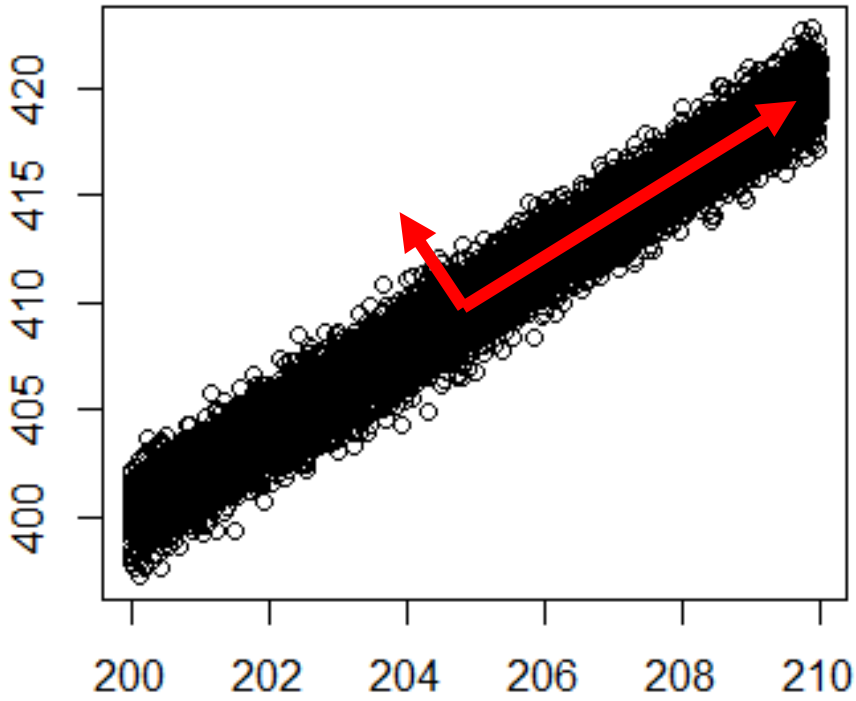
- We can imagine that in this original description both variables are independent and have the same importance, which is described by two perpendicular vectors with the same size.



Courtesy: Vitor Sousa

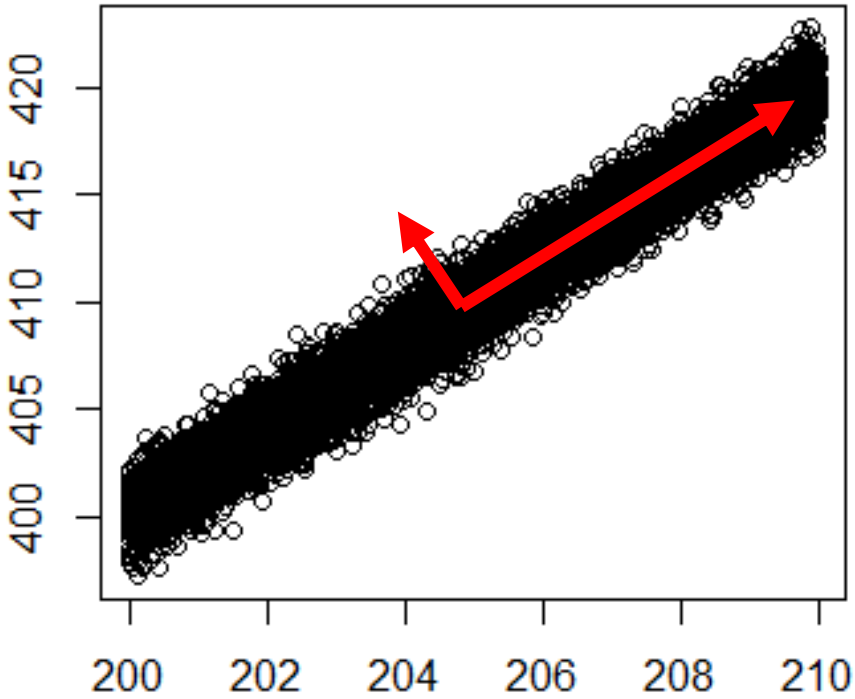
Example in toy example data: two variables

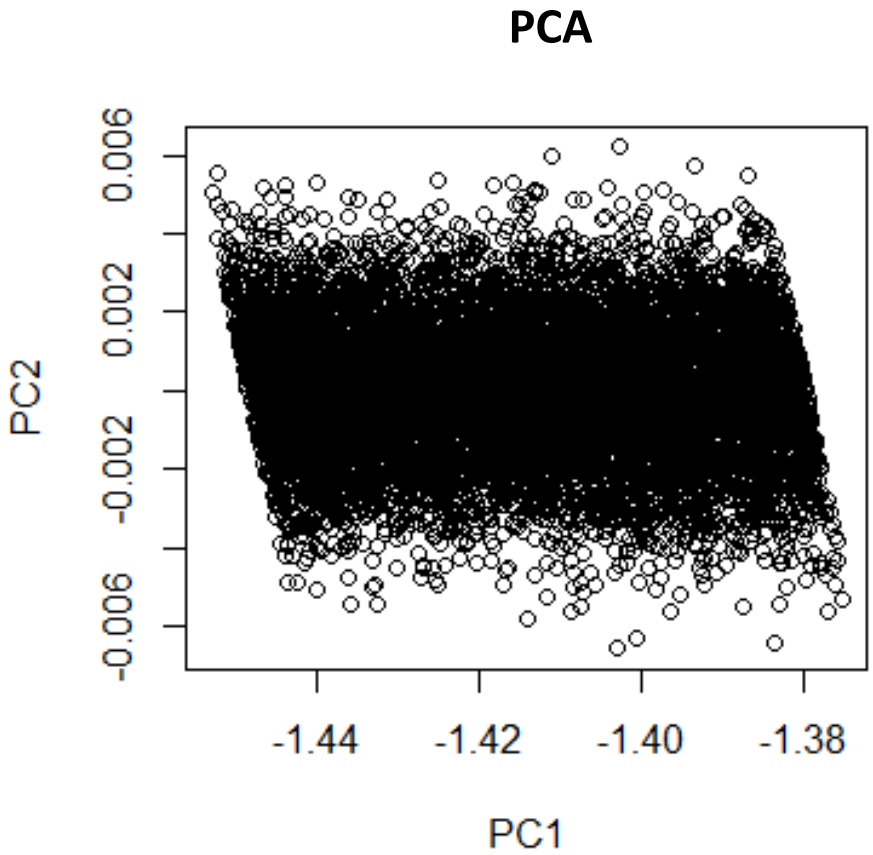
- PCA rotates the data by finding principal components described by eigenvectors that are independent (perpendicular). The size of the vector describe the relative contribution of each component.



Courtesy: Vitor Sousa

Example in toy example data: two variables





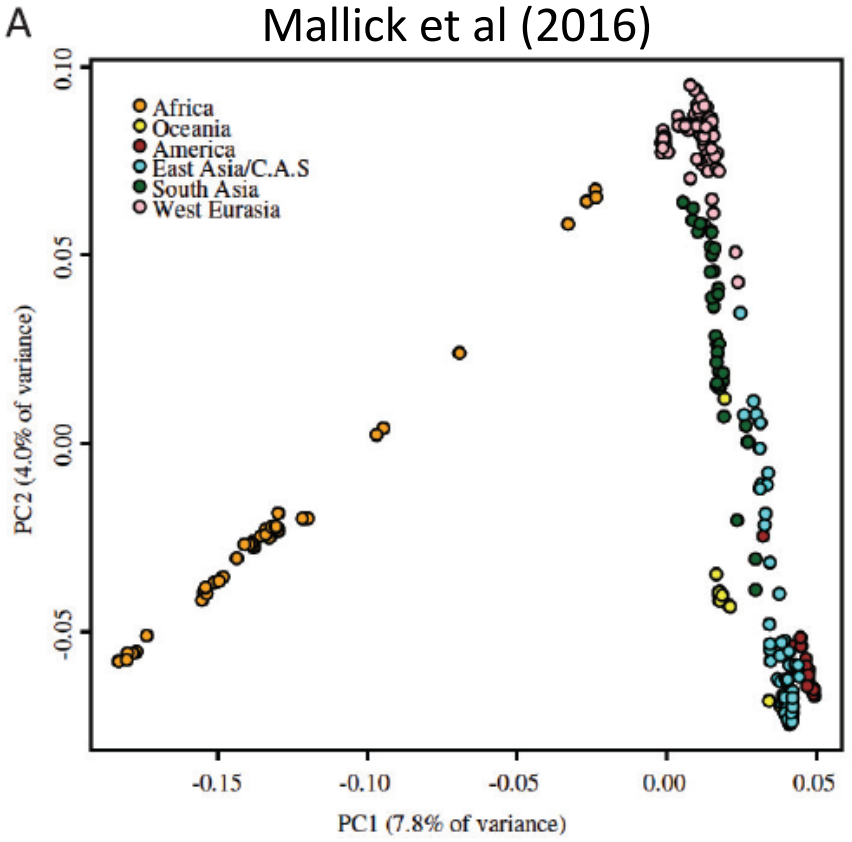
Importance of components:	PC1	PC2
Proportion of Variance	0.9927	0.00727

Courtesy: Vitor Sousa

Results from recent studies in humans

- PC1 separates Africans from non-Africans
- PC2 separates Europeans from Asians

Mallick et al (2016)



Pagani et al (2016)

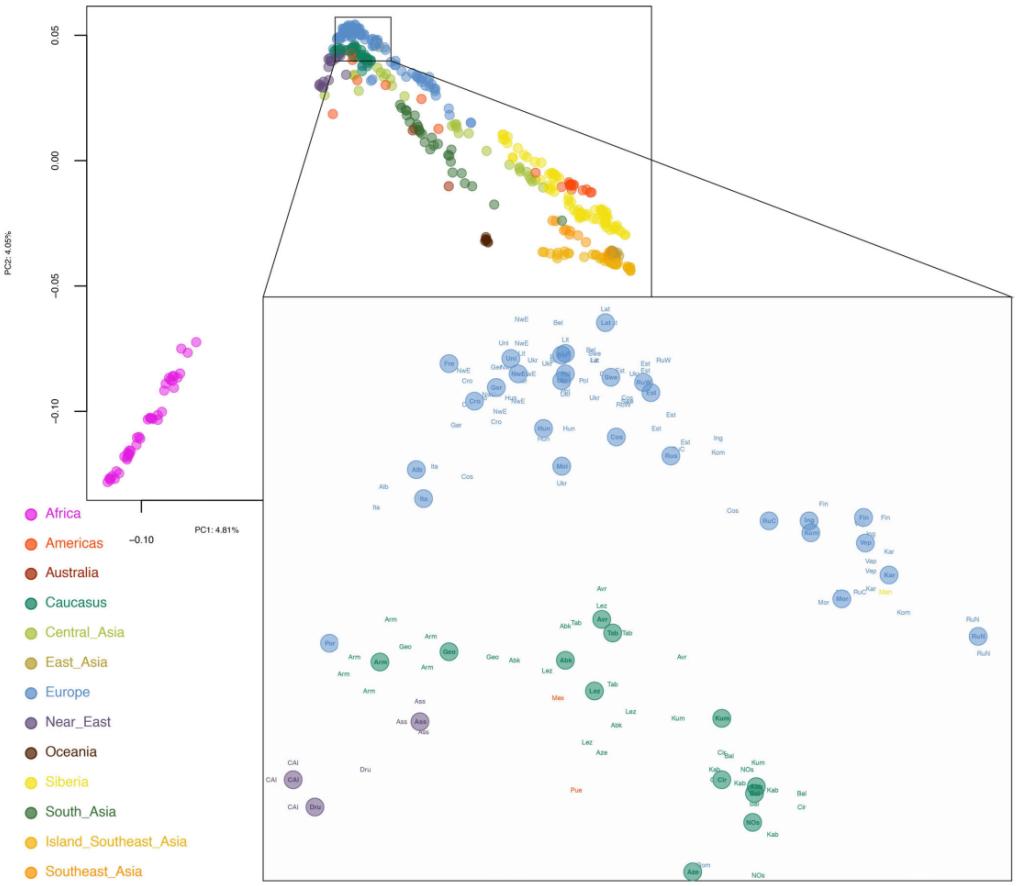


Figure S2.1.1-I PCA on global dataset PC 1 vs PC2

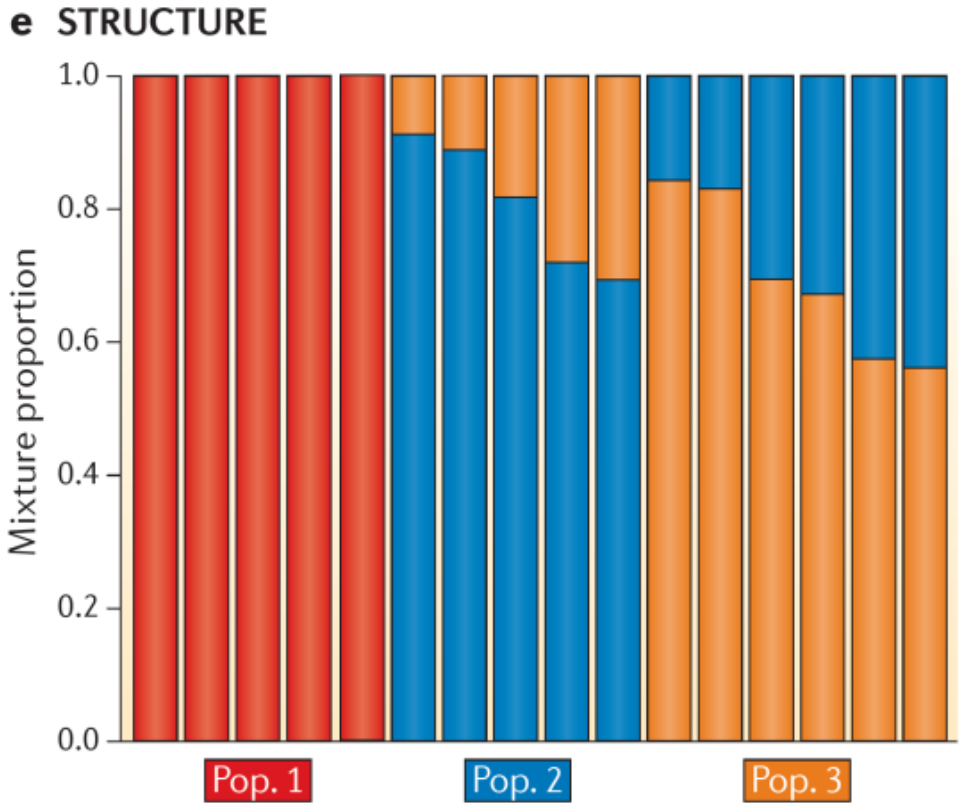
Courtesy: Vitor Sousa

Structure-like clustering programs

- Clustering of individuals based on allele frequencies at many SNPs.

- Each column is a given individual.

- Proportion of each color corresponds to the proportion from each ancestral population.



Courtesy: Vitor Sousa; Schraiber & Akey (2015) *Nature Reviews Genetics*

Estimating individual ancestry

- Bayesian method (Structure)
  - Sample values from the posterior using MCMC with Gibbs sampling
  - Slow and difficult to apply to large genomic datasets (Pritchard *et al*. 2001; Falush *et al.* 2003)
- Maximum likelihood estimator (ADMIXTURE)
  - Same likelihood function as Structure, but using a complex optimization algorithm to find the maximum likelihood
  - Much faster than Structure, applicable to large genomic datasets (Alexander *et al*. 2009)
- Non-parametric method (sNMF)
  - Using a sparse non-negative matrix factorization
  - Much faster than ADMIXTURE and reaching similar estimates
  - Uses a cross-validation approach to approximate estimator uncertainty (Frichot *et al*. 2014)

Courtesy: Vitor Sousa

Population structure and model-based clustering methods

- Detect the structure of a genetic sample without prior information on the geographical origin of individuals.
- Do not depend on the units defined by our sampling strategy and try to recover any hidden partition in the data.
- Many methods use LD and Hardy-Weinberg equilibrium to detect clusters (*e.g.*, STRUCTURE, GENELAND).
- Other methods can use spatial information as well (e.g. GENELAND, TESS).

Courtesy: Vitor Sousa

Population structure and Hardy-Weinberg equilibrium: the Wahlund effect

Consider two isolated populations that have fixed different alleles:

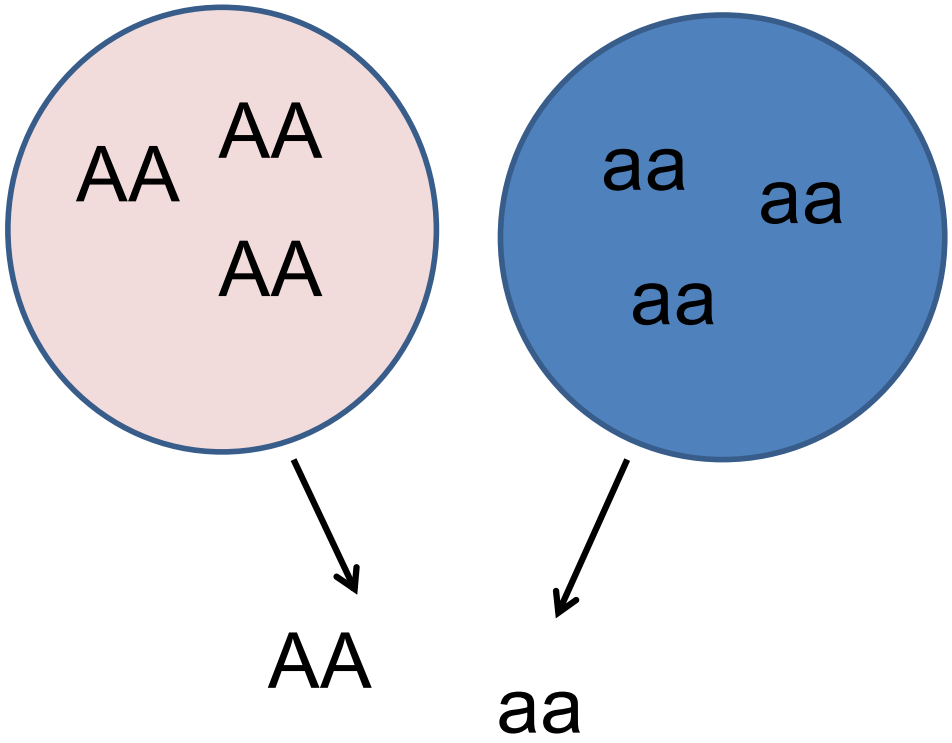
Population 1
Locus 1: p(A)=1.0, p(a)=0.0

Population 2
Locus 1: p(A)=0.0, p(a)=1.0

If we have samples of equal size from a mixture of these two populations, we would observe:

$H_{exp} = 2p_A p_a = 2*0.5*0.5=0.5$

In our example, we do not observe any heterozygotes: $H_{obs} = 0$



This apparent heterozygote deficit due to population structure is called the **Wahlund effect**.

Courtesy: Vitor Sousa

Population structure and LD patterns

Consider two isolated populations that have fixed different alleles:

Population 1

Locus 1: $p(A) = 1.0$, $p(a) = 0.0$

Locus 2: $p(B) = 1.0$, $p(b) = 0.0$

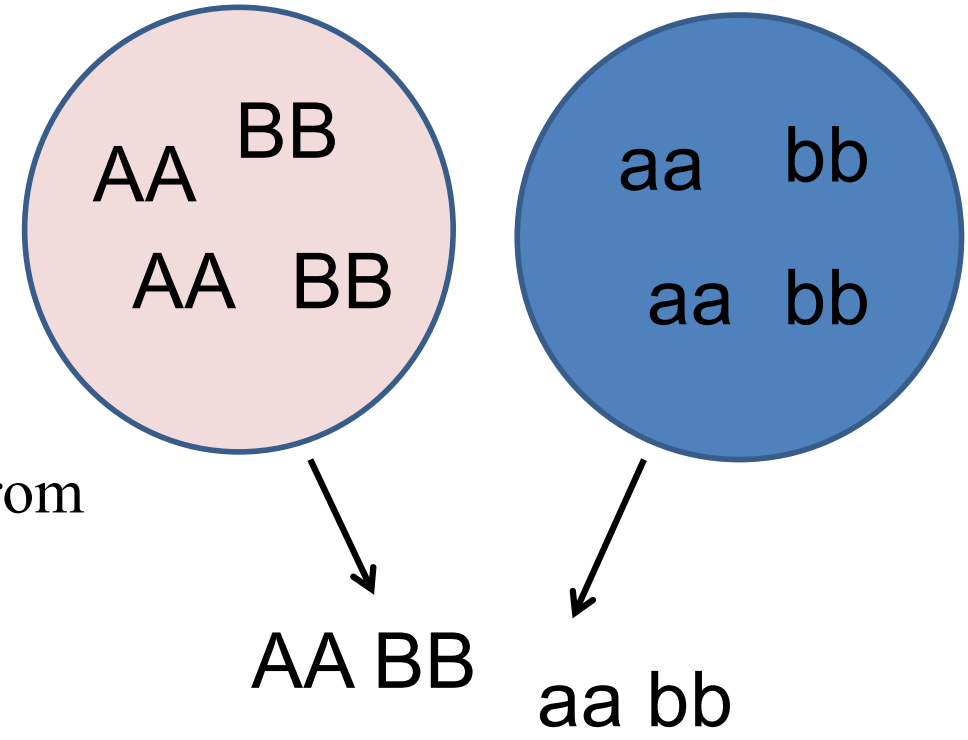
Population 2

Locus 1: $p(A) = 0.0$, $p(a) = 1.0$

Locus 2: $p(B) = 0.0$, $p(b) = 1.0$

If we have a sample of equal sizes from a mixture of these two populations, we would observe:

$$D = p_{AB} - p_A p_B$$
$$= 0.5 - (0.5*0.5) = 0.5 - 0.25$$
$$= 0.25$$



Population structure creates LD

Courtesy: Vitor Sousa

The program STRUCTURE

- Detect the structure of a genetic sample without prior information on the geographical origin of individuals.

- Do not depend on the units defined by our sampling strategy and try to recover any hidden partition in the data.

- Bayesian model-based clustering algorithm:
  - Two ancestry models: *admixture* and *no admixture*;
  - Two allele frequency models: *independent* and *correlated*.

Courtesy: Vitor Sousa

The program STRUCTURE

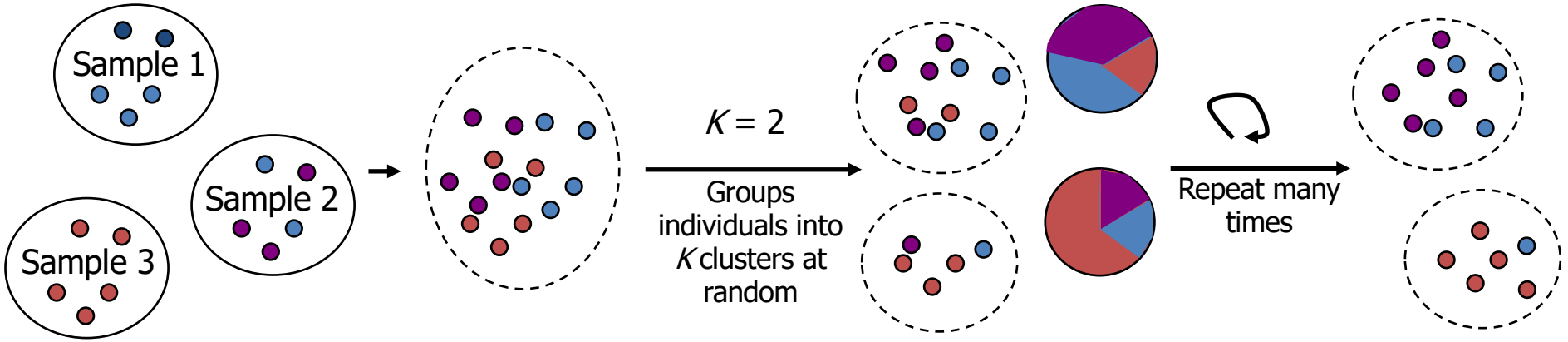
Estimate the posterior distribution Pr(**z, p**| ***X***);

Algorithm *Markov Chain Monte Carlo with Gibbs sampler*

For a given pre-specified *K* value:

Start with initial random $\mathbf{z}^{(0)}$;

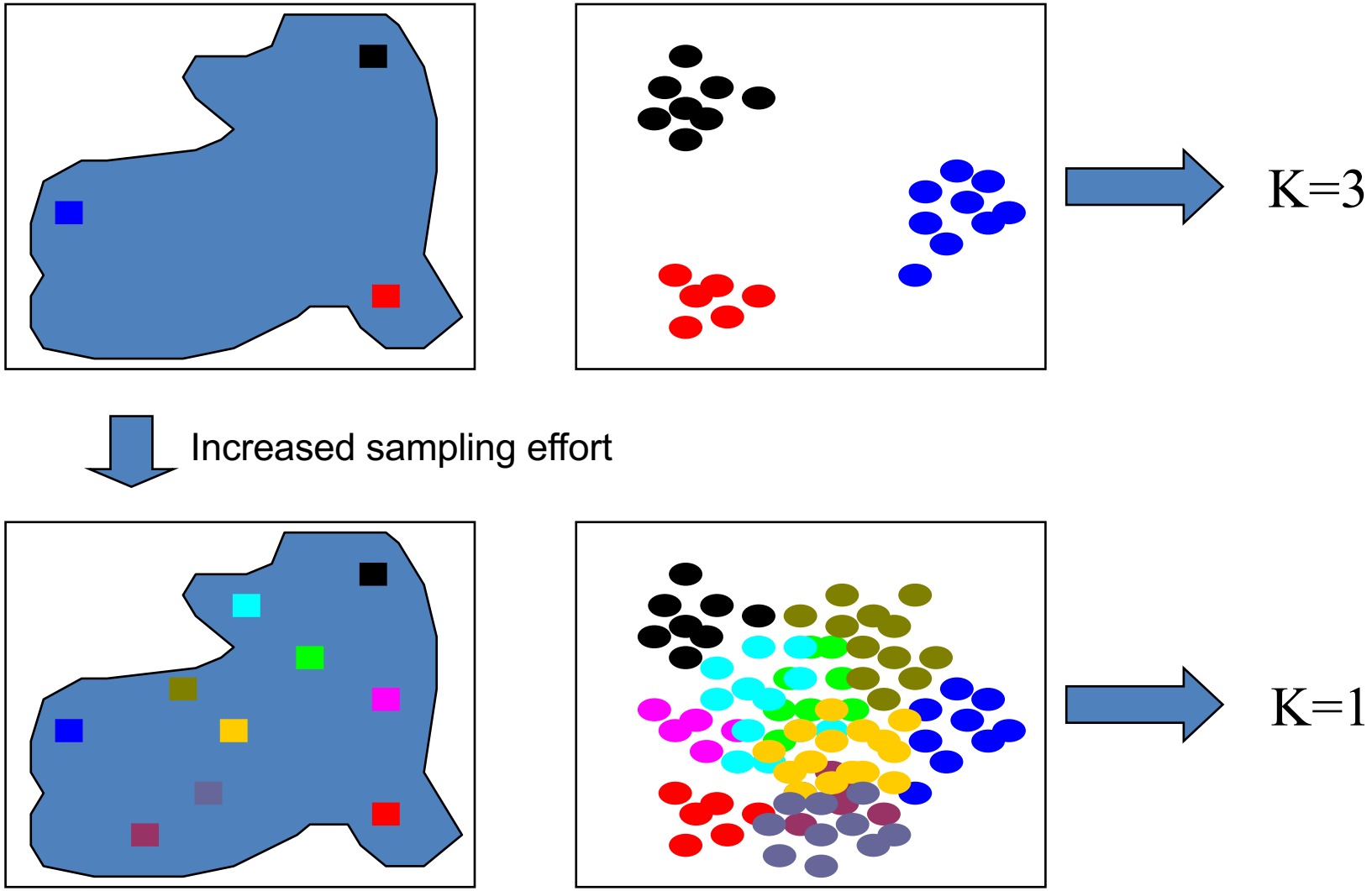
1. Sample $\mathbf{p}^{(m)}$ from Pr($\mathbf{p}$|X, $\mathbf{z}^{(m-1)}$);
2. Sample $\mathbf{z}^{(m)}$ from Pr($\mathbf{z}$|X, $\mathbf{p}^{(m)}$);



Aims at finding the groups that are at HW equilibrium and minimize LD

Courtesy: Vitor Sousa

STRUCTURE: the meaning of K ...

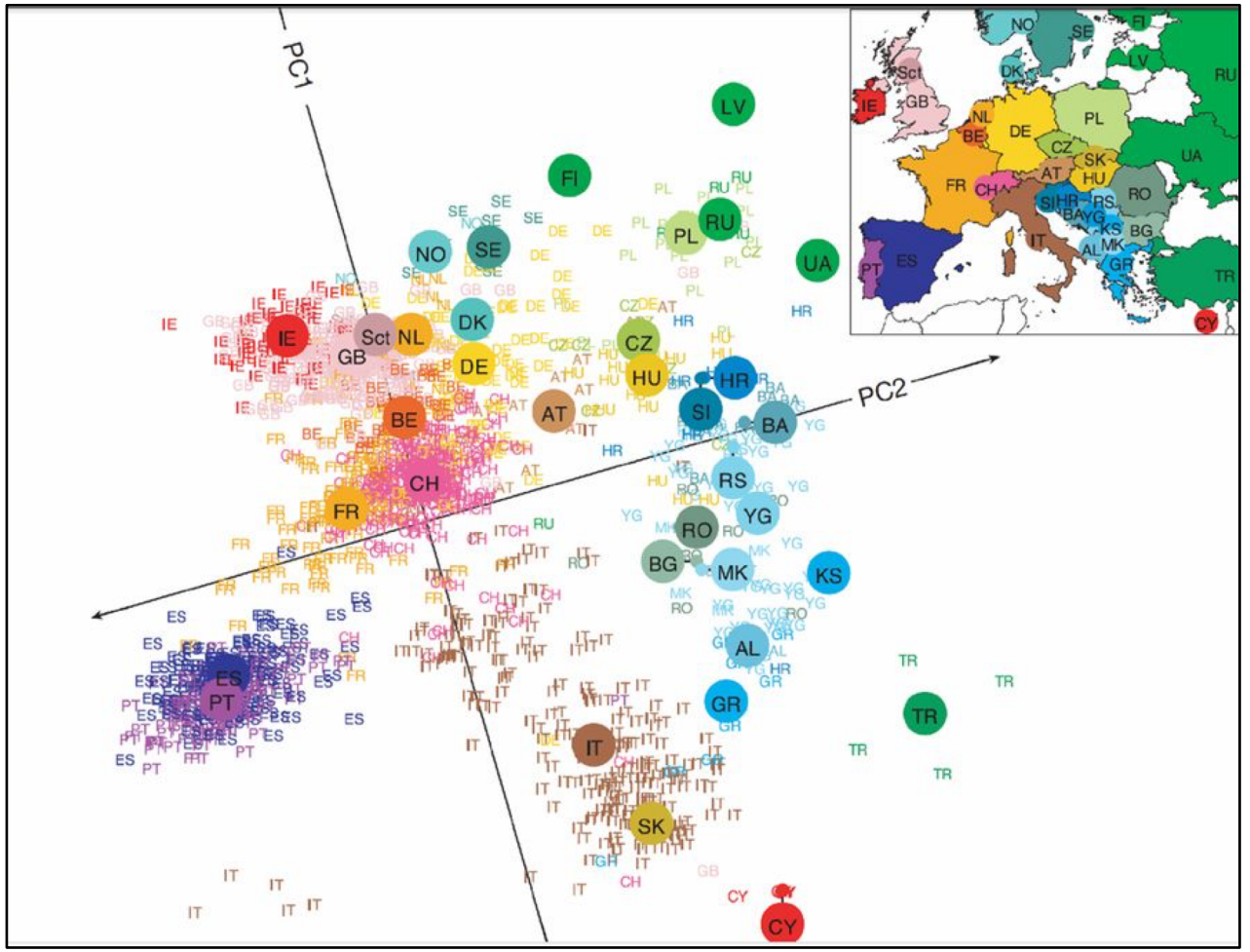


Courtesy: Vitor Sousa

Isolation by distance

- The geographically closer two individuals are, the more genetically similar they will be

- Each dot is an European individual



Courtesy: Vitor Sousa; Novembre *et al.* (2008) *Nature*

Isolation by distance

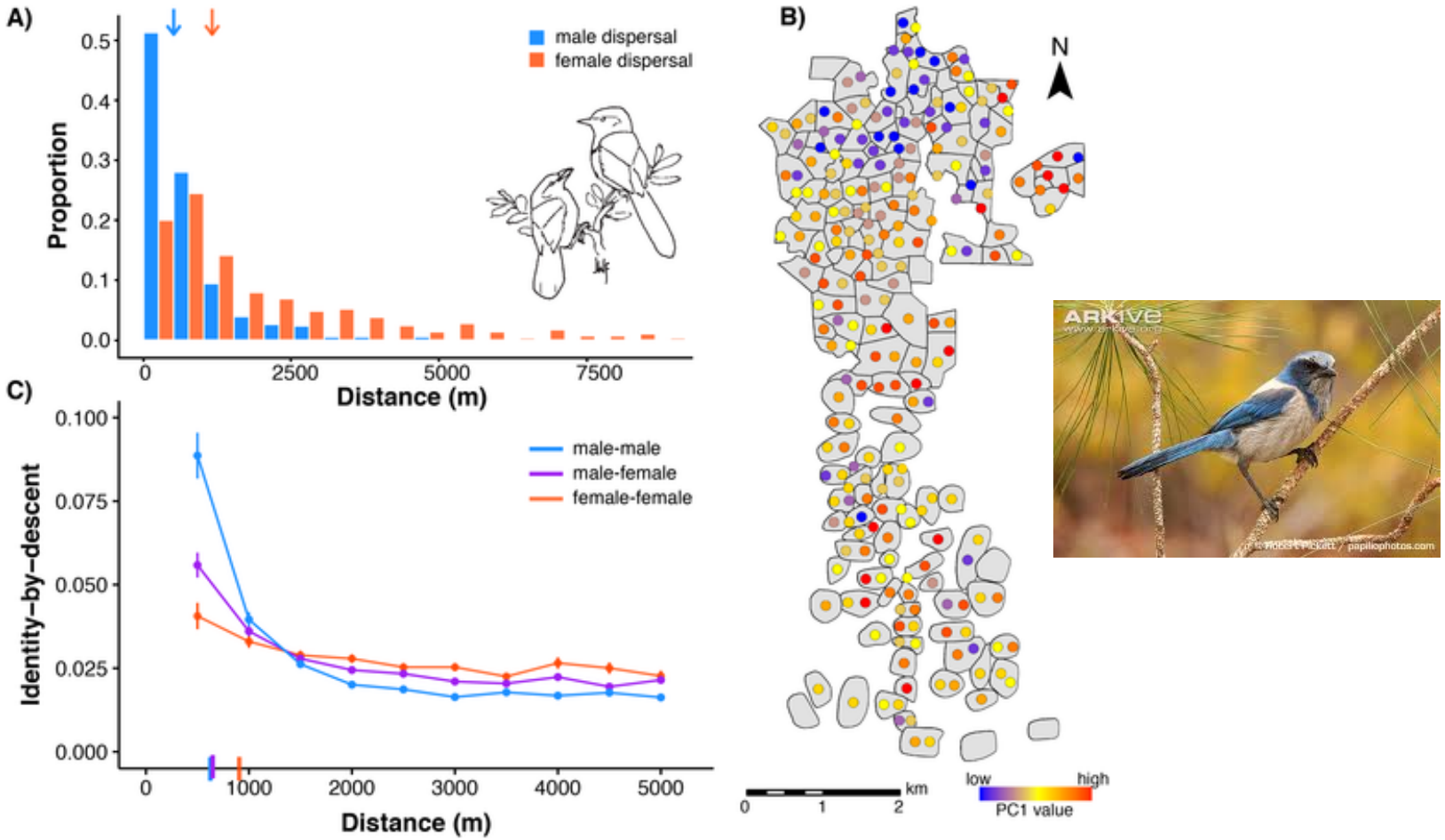
- Isolation by distance (IBD) is a simple consequence of limited dispersal across space, which Sewall Wright described almost seventy years ago: pairs of populations close to each other will be more genetically similar to each other than populations farther away from each other, not because of any selective need for those genetic similarities, but just because individual critters, or their seeds, or pollen, or larvae are less likely to travel longer distances.

Wright, S. 1943. Isolation by distance. *Genetics* **28:** 114-138. PMCID: PMC1209196.

Courtesy: Vitor Sousa

Isolation by distance in nature

- Dispersal curves and isolation-by-distance patterns in the Florida Scrub ay (*Aphelocoma coerulescens*)



Courtesy: Vitor Sousa; Aguillon *et al.* (2017) *PLoS Genetics*

fastSTRUCTURE, ADMIXTURE and sNMF

- For large number of SNPs and individuals, likelihood based methods like Structure become very slow and inefficient.
- Frichot et al (2014) showed that it is possible to infer ancestry coefficients with sparse non-negative matrix factorization (sNMF).

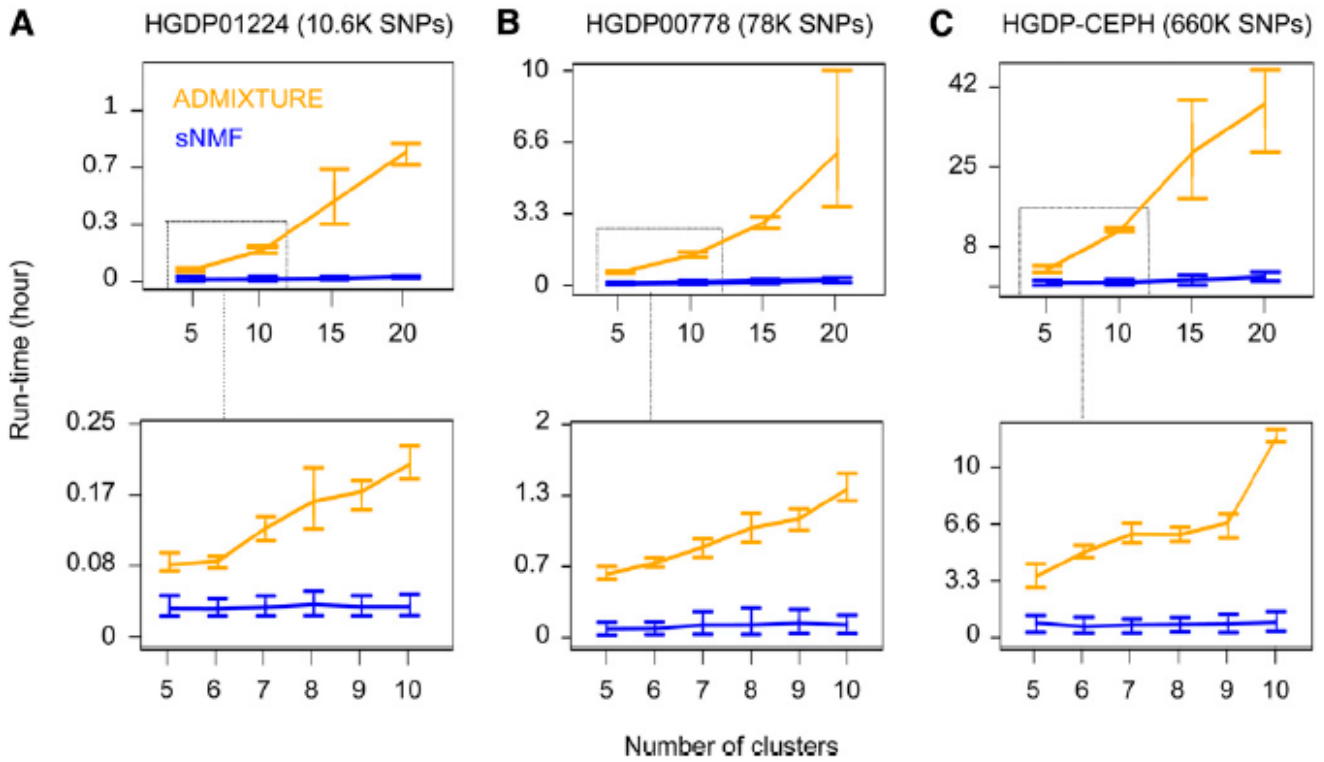


Figure 2 Runtimes for sNMF and ADMIXTURE runs. Averaged time elapsed before the stopping criterion of the sNMF (blue) and ADMIXTURE (orange) programs is met. Time is expressed in unit of hours. (A) Run-time analysis for Harvard HGDP panel 01224 (10,600 SNPs). (B) Run-time analysis for Harvard HGDP panel 00778 (78,000 SNPs). (C) Run-time analysis for the HGDP-CEPH data (660,000 SNPs).

Courtesy: Vitor Sousa; Frichot *et al.* (2014) *Genetics*

sNMF

Admixture models generally suppose that the genetic data originate from the admixture of K ancestral populations, where K is unknown *a priori*. Given K populations, the probability that individual i carries j derived alleles at locus ℓ can be written as

Fraction of individual genome that originates from pop ***k***

$$p_{i\ell}(j) = \sum_{k=1}^{K} q_{ik} g_{k\ell}(j)$$

Probability of the genotype of individual ***i***

Genotype frequency at pop ***k***

where q_{ik} is the fraction of individual i's genome that originates from the ancestral population k, and $g_{k\ell}(j)$ represents the homozygote ($j = 0, 2$) or the heterozygote ($j = 1$) frequency at locus ℓ in population k.

Courtesy: Vitor Sousa

sNMF– how to decide on the best *K* value?

Cross-entropy criterion

We employed a cross-validation technique based on imputation of masked genotypes to evaluate the prediction error of ancestry estimation algorithms (Wold 1978; Eastment and Krzanowski 1982). The procedure partitioned the genotypic matrix entries into a training set and a test set. To build the test set, 5% of all genotypes were randomly selected and tagged as missing values. The occurrence probabilities for

$$H\left(p^{\text{sample}}, p^{\text{pred}}\right) = -\sum_{j=0}^{2} p^{\text{sample}}(j) \log p_{it}^{\text{pred}}(j) \ , \quad j = 0, 1, 2.$$

Compare the actual **sample probabilities** with its **prediction:** the minimum cross entropy corresponds to the best *K* to predict the data

Courtesy: Vitor Sousa; Frichot *et al*. (2014)

sNMF versus ADMIXTURE

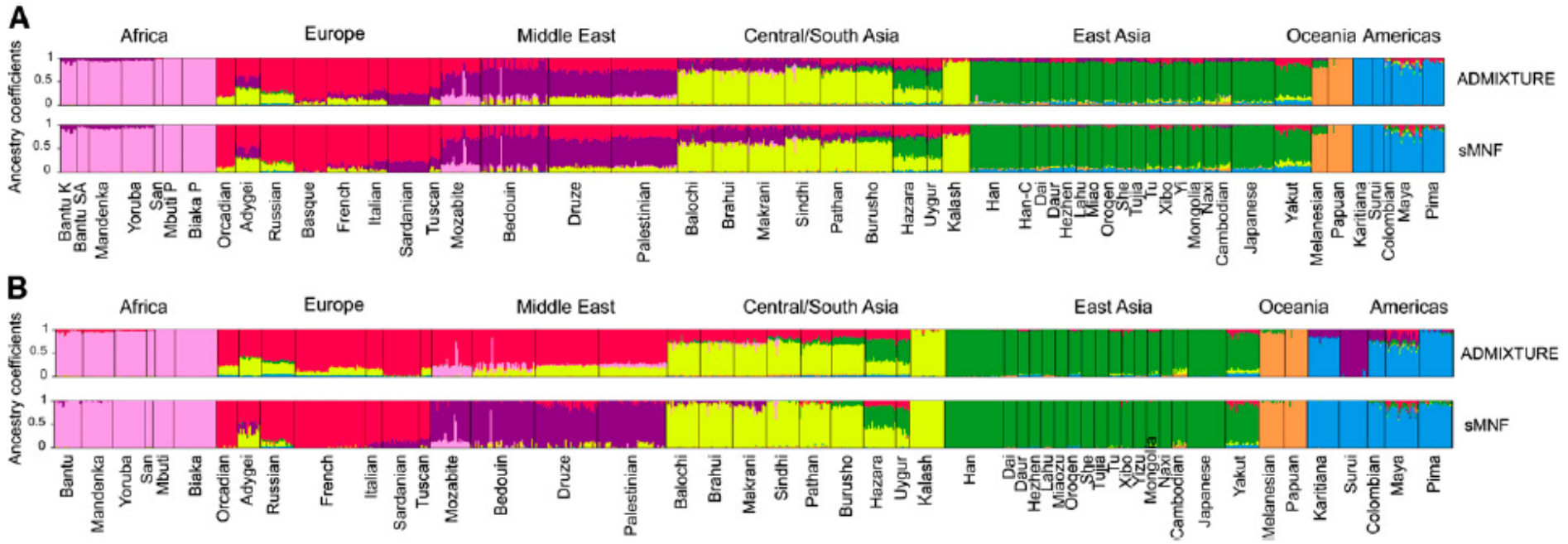


Figure 4 Graphical representation of ancestry estimates obtained for HGDP data sets (K = 7). (A) HGDP00778 panel (78,000 SNPs). Shown are estimated ancestry coefficients using ADMIXTURE (top, cross-entropy = 0.747) and sNMF (bottom, cross-entropy = 0.762 and α = 100). (B) HGDP-CEPH data set (660,000 SNPs). Shown are estimated ancestry coefficients using ADMIXTURE (top, cross-entropy = 0.691) and sNMF (bottom, cross-entropy = 0.704 and α = 100).

Courtesy: Vitor Sousa; Frichot *et al.* (2014)

Barriers to gene flow

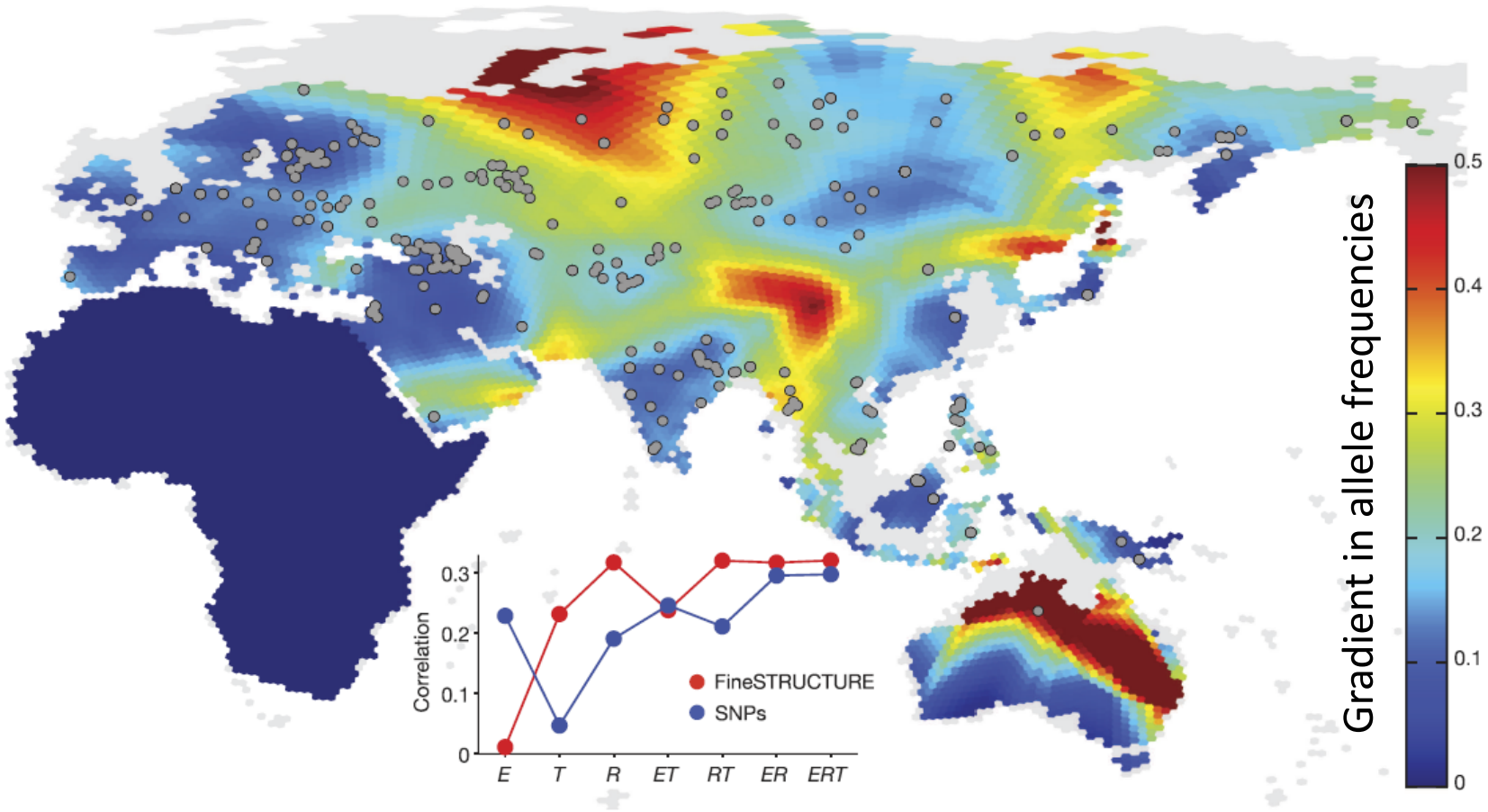
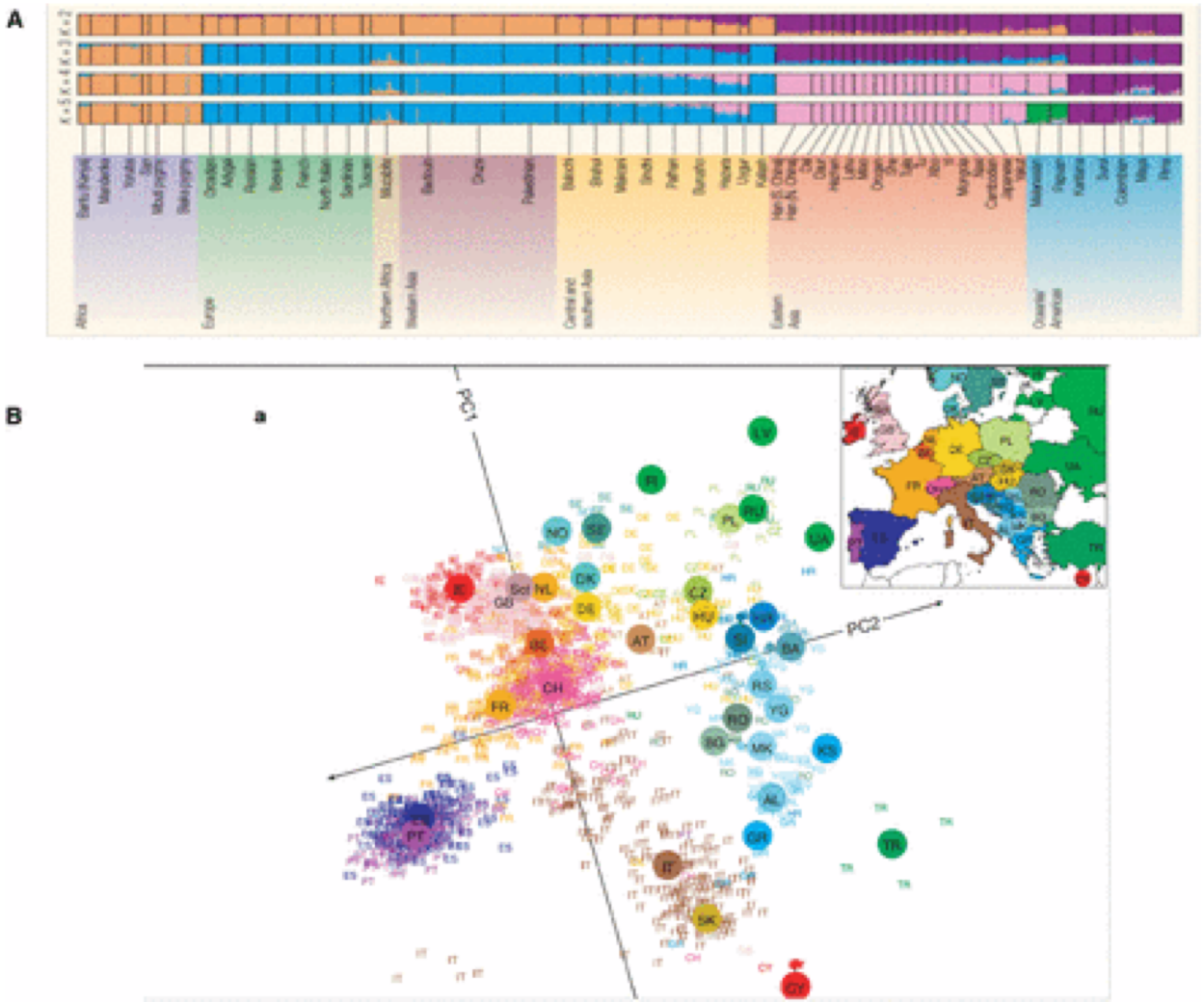


Figure 1 | Genetic barriers across space. Spatial visualization of genetic barriers inferred from genome-wide genetic distances, quantified as the magnitude of the gradient of spatially interpolated allele frequencies (value denoted by colour bar; grey areas have been land during the last glacial maximum but are currently underwater). Here we used a spatial kernel smoothing method based on the matrix of pairwise average heterozygosity and a MATLAB script that plots the hexagons of the grid with a colour coding to represent gradients. Inset, partial correlation between magnitude of genetic gradients and combinations of different geographic factors, elevation (*E*), temperature (*T*) and precipitation (*R*), for genetic gradients from fineSTRUCTURE (red) and allele frequencies (blue). This analysis (Supplementary Information 2.2.2 for details) shows that genetic differences within this region display some correlation with physical barriers such as mountain ranges, deserts, forests, and open water (such as the Wallace line).

Courtesy: Vitor Sousa; Pagani *et al.* (2016)

Characterization of population structure with individual-based methods – Example 1

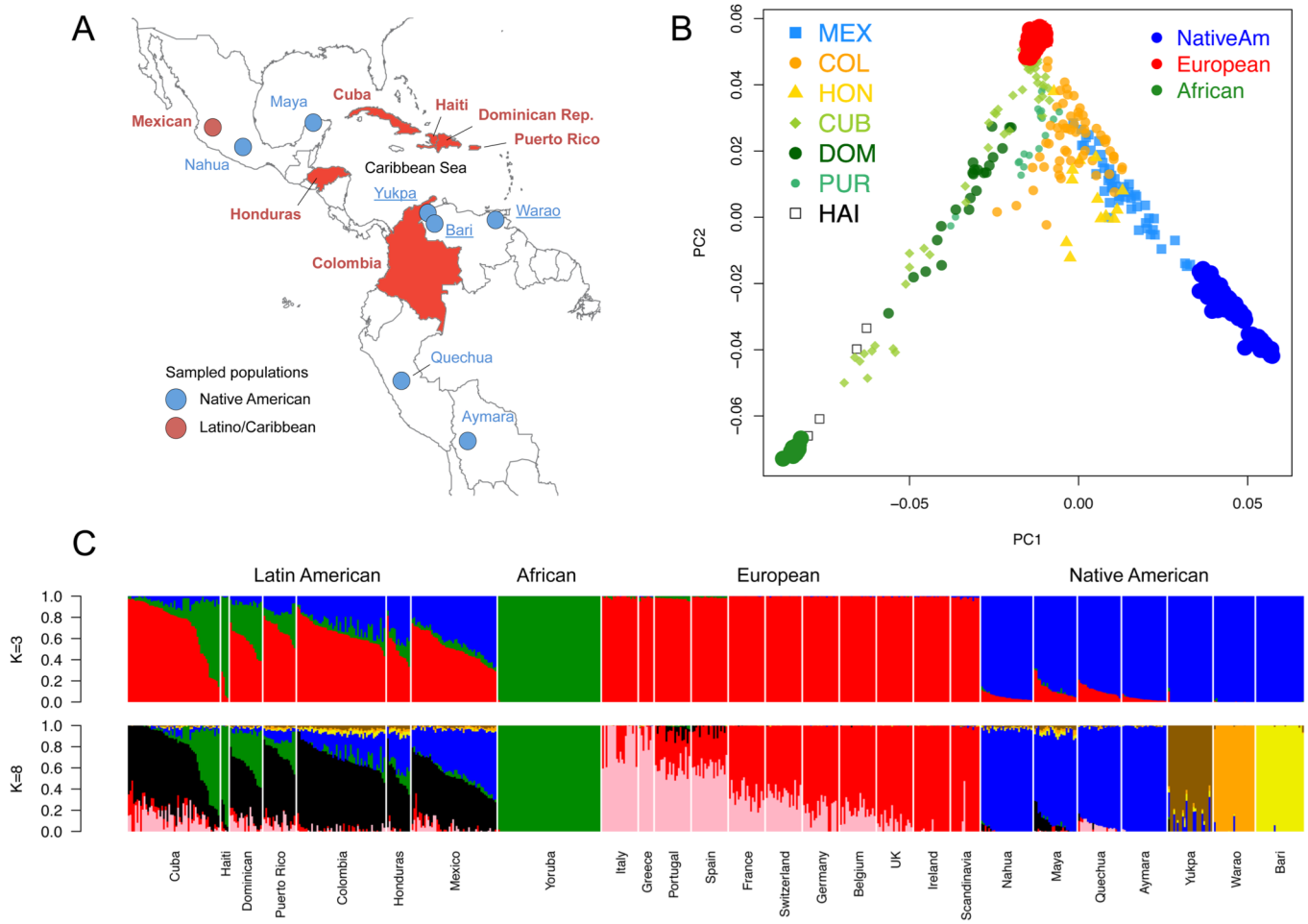
- Population structure in Europe is consistent with isolation by distance



Courtesy: Vitor Sousa; Novembre *et al*. (2008) *Nature*

Characterization of population structure with individual-based methods – Example 2

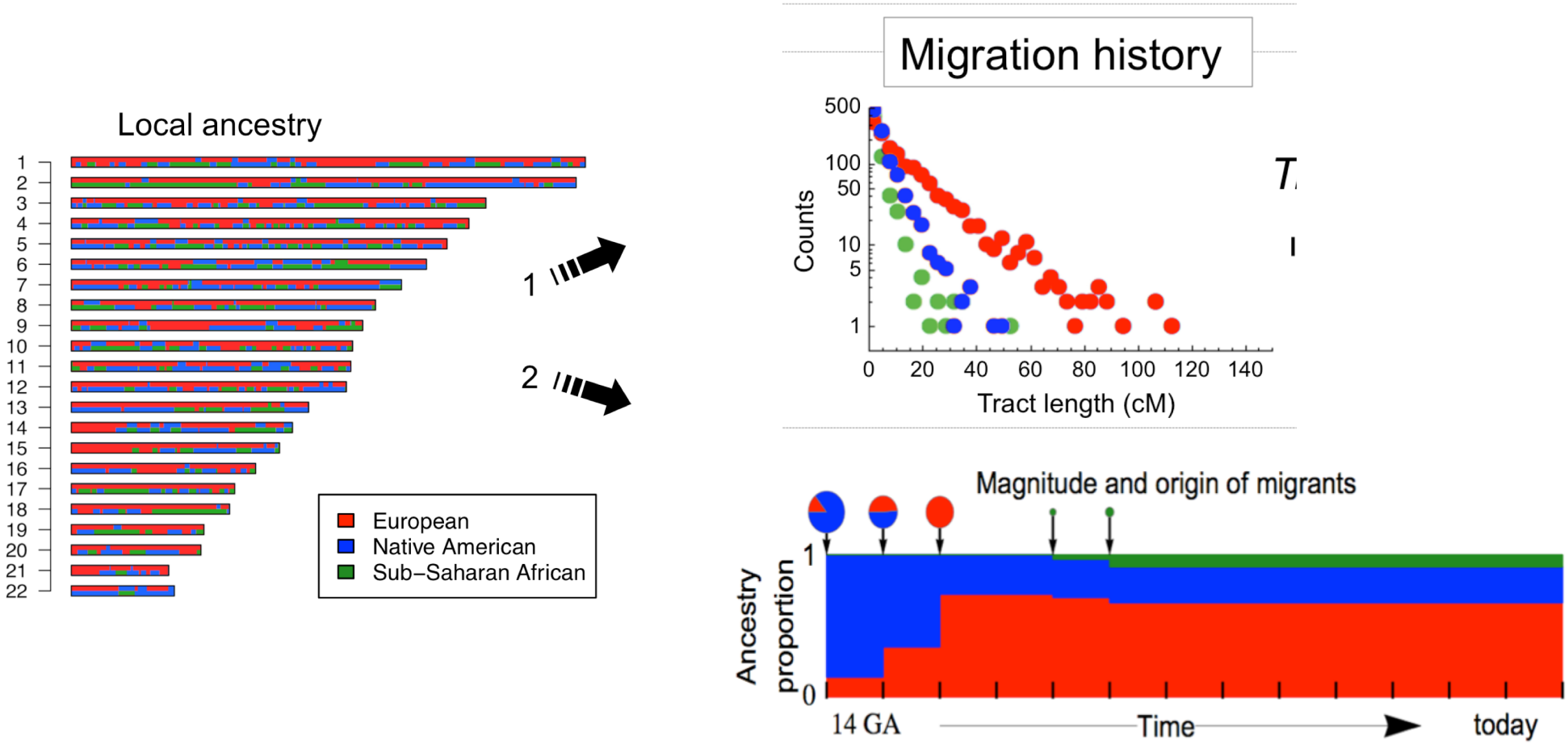
- Population structure of the Caribbean:
  - extensive gene flow across the Caribbean in pre-Columbian times



Moreno-Estrada *et al.* (2013) *PLoS Genetics*

Characterization of population structure with individual-based methods – Example 2

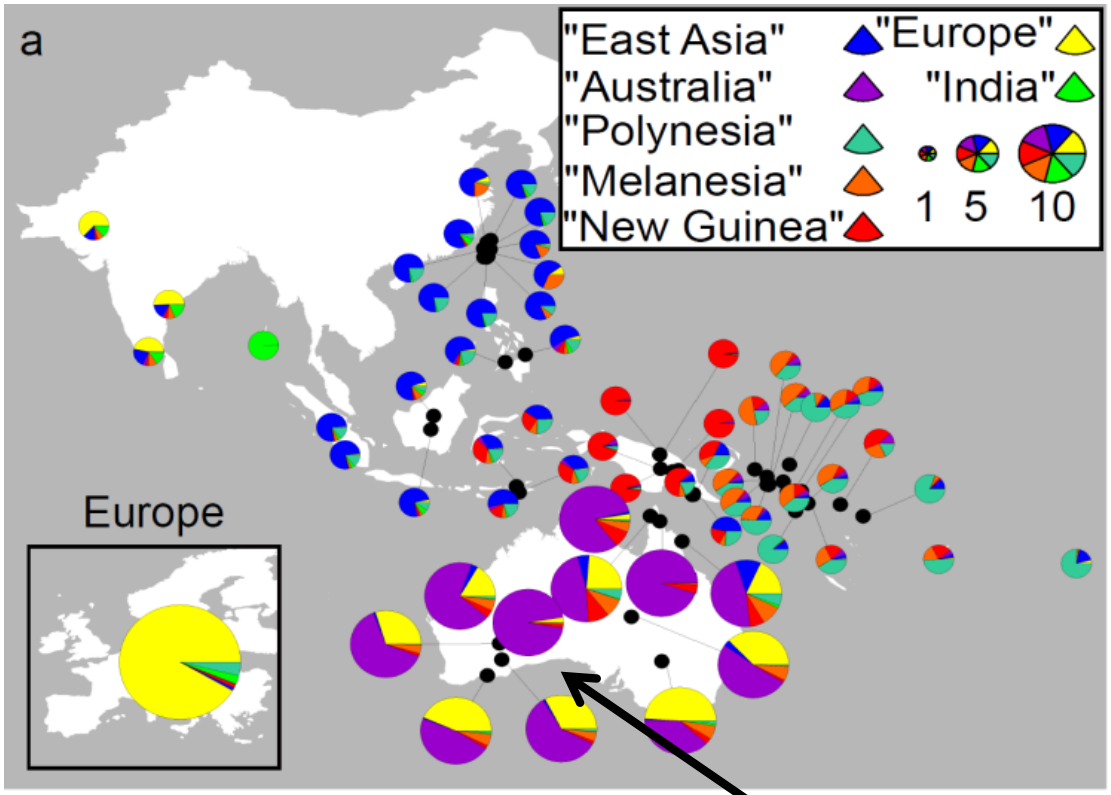
- Population structure of the Caribbean:
  - 2 pulses of african migration with different ancestry tracts sizes



Moreno-Estrada *et al*. (2013) *PLoS Genetics*

Characterization of population structure with individual-based methods – Example 3

- 83 high-coverage Aboriginal Australians genomes
- Average depth of coverage: 65x

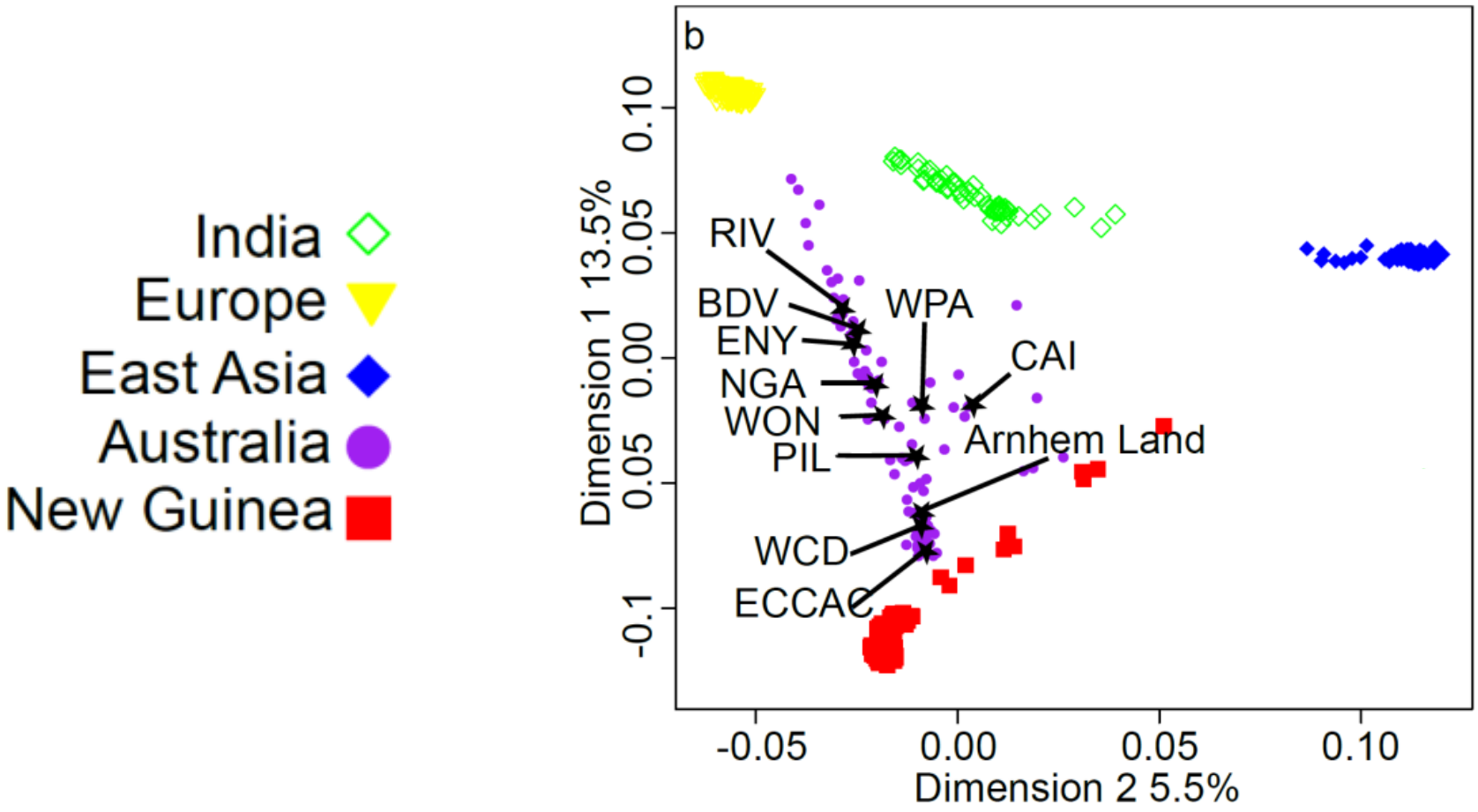


Western Central Desert (WCD)

Courtesy: Vitor Sousa; Malaspinas *et al.* (2016) *Nature*

Characterization of population structure with individual-based methods – Example 3

- Aboriginal Australians closer to New Guinea Papuans



Courtesy: Vitor Sousa; Malaspinas *et al.* (2016) *Nature*

(3) Detecting selection using sequencing data

Thomas Flatt

Department of Biology
Ecology & Evolution
University of Fribourg

Outline of lecture 3

In this lecture we will discuss how molecular data can be used to detect selection. As we shall see, this can be done in different ways, e.g. using

- comparisons of sequence divergence vs. polymorphism
- patterns of genetic variability & the statistics of variant frequency distributions within species, e.g. to infer selective sweeps
- genome scans of selection more generally, e.g. using F_{ST} outlier approaches, etc.

Different forms of selection

- Purifying selection, which acts to prevent the spread of deleterious mutations, e.g. those affecting the amino acid sequences of proteins.
- Positive directional selection, which causes an adaptive mutation to spread through a species
- Balancing selection, which maintains alternative variants in the population
- Directional and balancing selection are often collectively referred to as positive selection.

(text book, chapters 2 & 3 & 6)

Basic selection theory: genetic variability as the essential substrate for selection

- See chapter 2 in the Charlesworth & Charlesworth text book.
- Here we just show the basic equation of allele frequency change under selection for a single locus with 2 alleles:
- $\Delta p = pq\,(w_1\text{-}w_2) / w_{\text{average}}$ and $\Delta q = pq\,(w_2\text{-}w_1) / w_{\text{average}}$
- Note the term pq which is familiar from the $2pq$ term in the H-W equilibrium; also note that for binomial sampling, the variance of the binomial distribution is $n\,p\,(1\text{-}p) = n\,(pq)$. Thus, in our case pq can be seen as a measure of genetic variance: genetic variance is the essential substrate for a response to selection.

(see text book, chapters **2** & also 3 & 6)

How to detect molecular signatures of selection?

- One of the major goals of evolutionary genetics is to understand to what extent selection, as opposed to neutral forces of mutation and genetic drift, controls variation and evolution in DNA and protein sequences.

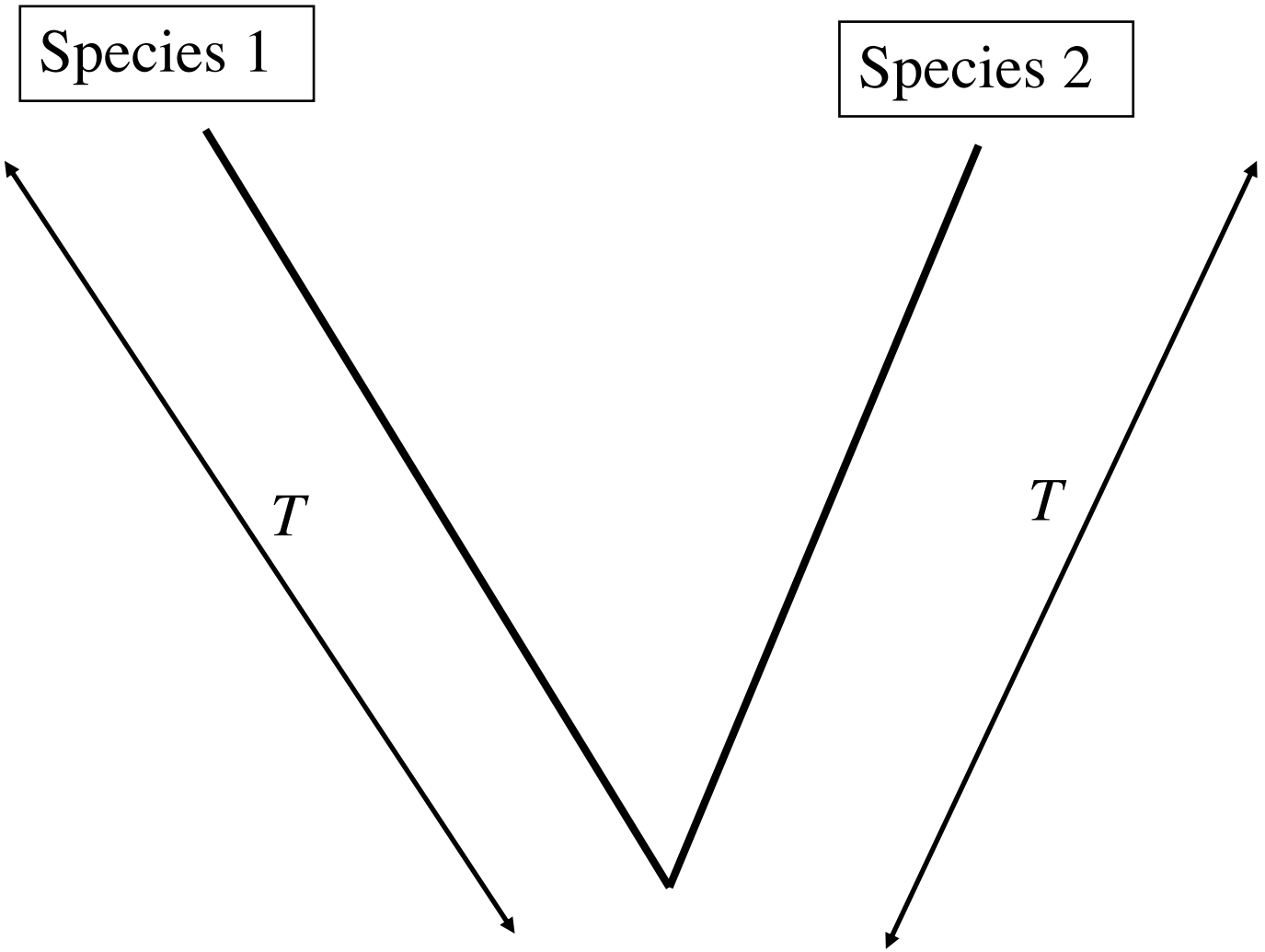
- The methods for doing this often involve combining data on sequence divergence between species with data on genetic polymorphism within species. Sometimes, however, as we shall also see, only polymorphism / variability data from within a species or population are being used.

Use of sequence divergence data

- The simplest situation is when we have two homologous (aligned) DNA sequences from a pair of related species.

- For the purpose of discussion, assume that all evolutionary change occurs by nucleotide substitutions, i.e. the sequence differences are caused entirely by one nucleotide base changing into another by mutation.

- This is usually the case for coding sequences, since insertions or deletions cause disruption of functionality.



The total time separating a pair of sequences from the two species is $2T$

Neutral sequence evolution

- Under neutral evolution, the expected number of changes (mutations) K is expected to be equal to the mutation rate (μ) times the divergence time between the two species, i.e.

$$K = 2\,\mu\,T$$

- The simplest way to understand this is to note that, under neutral evolution, the expected number of mutations that distinguish a pair of sequences is equal to the time separating them ($2T$) times the rate of mutation per unit time (μ). (see text book, pp. 257 and following pages.)

- We compare *K* values for nucleotide sites where mutations can reasonably be assumed to be neutral or nearly neutral with *K* for sites where we wish to test for selection; larger than neutral *K* values indicate directional selection, and smaller than neutral *K* values indicate purifying selection.

- Non-synonymous sites are usually used as the candidates for selection, but there is increasing use of defined types of non-coding sequences.

- This comes from the fact that *K* (and θ) for non-synonymous variants are nearly always much smaller than for synonymous and noncoding sites; this indicates that purifying selection is pervasive.

The K_a/K_s ratio

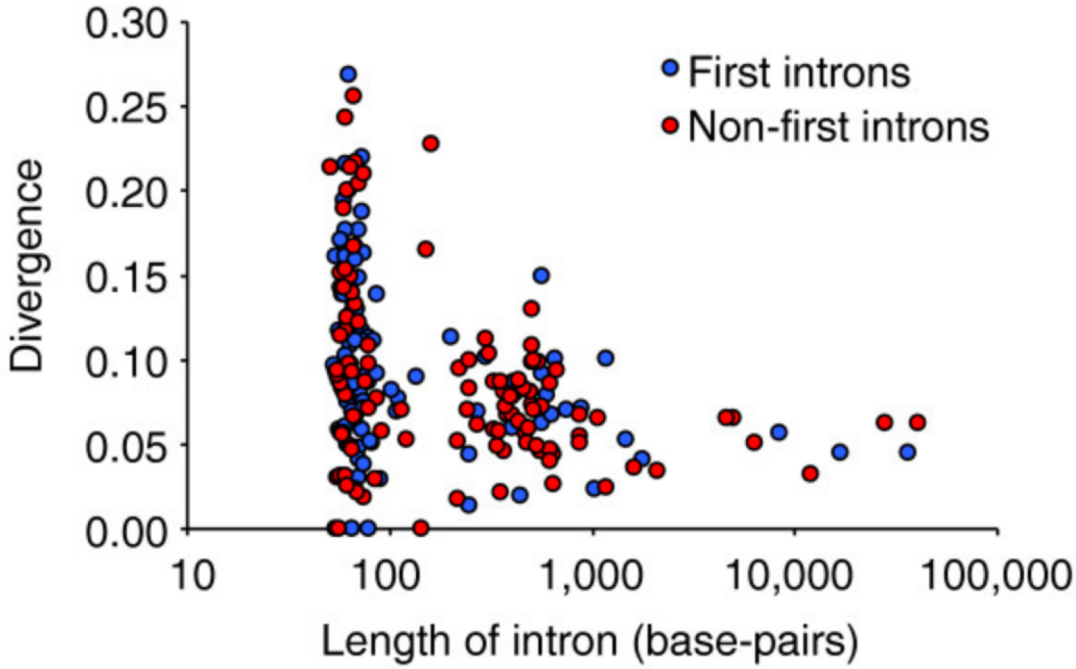
- The K_a/K_s ratio is used to estimate the balance between neutral mutations, purifying selection and beneficial mutations acting on a set of homologous protein-coding genes.
- It is calculated as the ratio of the number of nonsynonymous substitutions per non-synonymous site (K_a), in a given period of time, to the number of synonymous substitutions per synonymous site (K_s), in the same period. The latter are assumed to be neutral, so that the ratio indicates the net balance between deleterious and beneficial mutations.
- A ratio greater than 1 implies positive selection; less than 1 implies purifying selection; and a ratio of exactly 1 indicates neutral (i.e. no) selection.
- The ratio is also known as ω or d_N/d_S.

An example from *Drosophila miranda* and *pseudoobsura*

Means	π_{A1}	θ_{S1}	π_{A2}	θ_{S2}	K_A	K_S
	0.088	0.478	0.206	2.73	2.48	22.2
	(0.044 /	(0.342 /	(0.124 /	(2.31 /	(1.30 /	(19.9 /
	0.141)	0.626)	0.300)	3.14)	3.76)	24.8)

- Statistics on diversity and divergence in *D. miranda* (species 1: 18 loci) and *D. pseudoobscura* (species 2: 14 loci). These species diverged from each other about 2 Mya. All values are percentages.
- The low K_a/K_s ratio suggests strong purifying selection.
- Data from Loewe *et al*. (2006) *Genetics* 172:1079.

Divergence of introns between *Drosophila melanogaster* and *simulans*



The fact that long introns evolve more slowly than average implies that, while the majority of introns in the *Drosophila* genome may experience little or no selective constraint, most intronic DNA in the genome is likely to be evolving under considerable constraint. Short introns seem to be under less constraint.

P. Haddrill *et al*. (2005) *Genome Biol*. 6: R67. 1-8.

Effects of deleterious mutations on fitness

- There are clearly a lot of deleterious mutations entering the population each generation, most of which will eventually be eliminated by selection.

- While the mean level of variability is much lower for non-synonymous than synonymous mutations, this could simply mean that all the deleterious ones are rapidly removed by selection, so that the amino acid variants that we see segregating are in fact selectively neutral.

- It is a topic of current research to try and estimate the distribution of selection coefficients on deleterious amino acid and silent variants in natural populations.

- Estimates for amino acid variants indicate a wide distribution, such that the mean selection coefficient against a heterozygous non-synonymous variant is of the order of 10^{-5} .

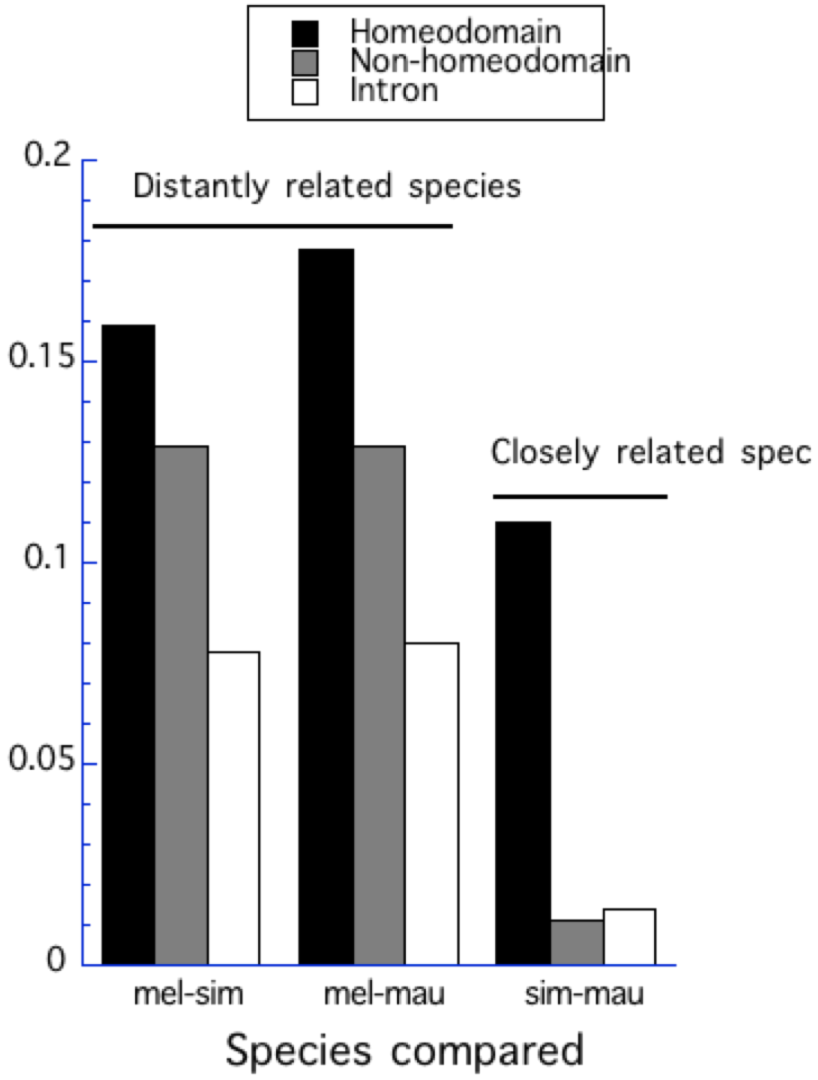
- Values for synonymous or silent variants are much smaller, of the order of 10^{-6}. Again, this indicates that purifying selection against non-synonymous changes is pervasive.

Positive directional selection

Faster divergence in coding than non-coding sequences suggests positive selection

- In the homeobox gene *OdsH* of three *Drosophila* species, divergence in the homeodomain is highly significantly accelerated.

- *"In the past half million years, this homeodomain has experienced more amino acid substitutions than it did in the preceding 700 million years; during this period, it has also evolved faster than other parts of the protein or even the introns. Such rapid sequence divergence is driven by positive selection and may contribute to reproductive isolation."*



C. Ting *et al*. 1998 *Science* 282:1501-1504

The McDonald-Kreitman (and HKA) test

- Compares non-synonymous and synonymous site divergence between species, and non-synonymous and synonymous site diversity within species, in the same gene.
- If variants at both kinds of sites were neutral, the numbers of substitutions at the two kinds of sites between two species should be in the same ratio as the polymorphism within either species, assuming equilibrium between drift and mutation:
- Neutral divergence = $2T\mu$
- Neutral diversity: $\pi = \theta = 4N_e\mu = 2\ (2N_e)\ \mu$ (where $2N_e$ is the mean coalescent time)
- Similar to the HKA test (Hudson, Kreitman, Aguadé, 1987) which compares the number of polymorphisms (SNPs) to the number of divergent sites between 2 species for 2 loci.

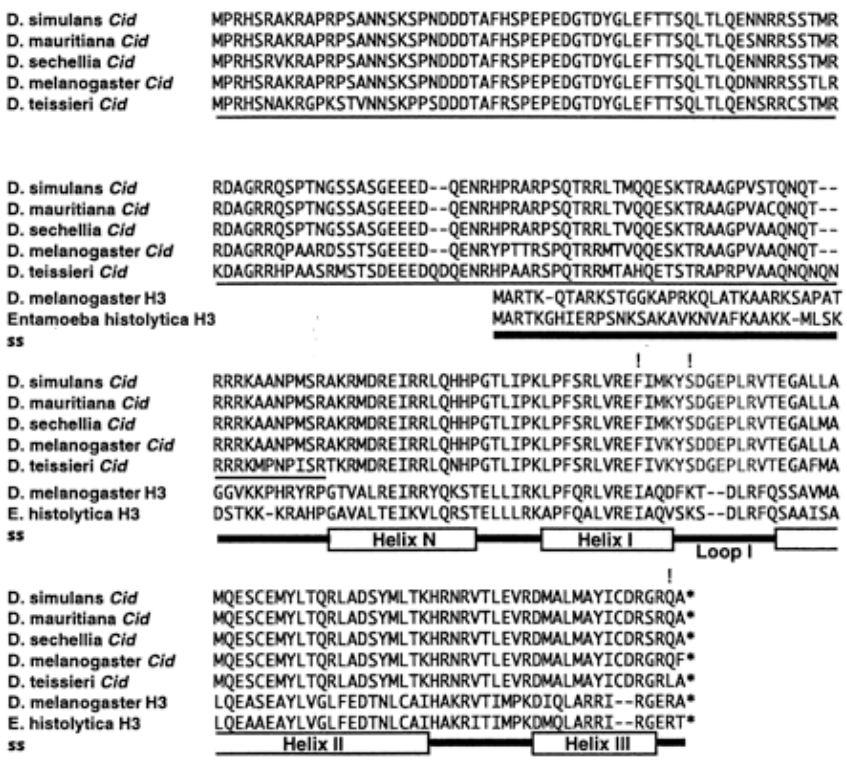
see text book, pp. 280-285.

The McDonald-Kreitman test

- If the ratio of non-synonymous variants to synonymous variants for differences between species is greater than the ratio for within-species variation, this suggests positive directional selection.

- If the opposite is the case, either purifying selection or balancing selection is acting.

An example: centromeric histone protein evolution



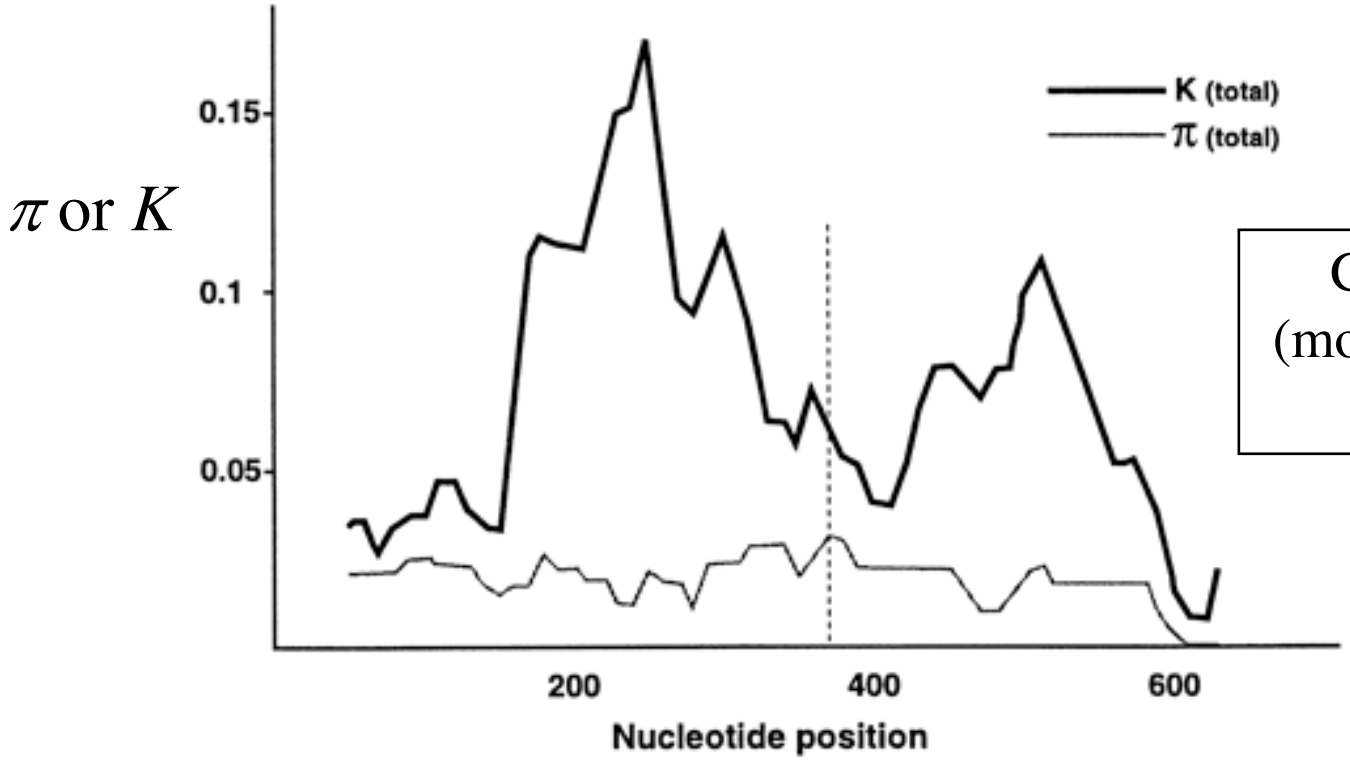
- Alignment of the centromeric Cid proteins of five *melanogaster* subgroup species with histone H3 proteins from *D. melanogaster* (2.3 million years divergence) with *Entamoeba histolytica* (> 1 billion years divergence.
- The most divergent histone H3 sequences have >75% identity to each other, whereas centromeric H3-like proteins are much more diverged (35–50% identical to histone H3).

Malik & Henikoff (2001) *Genetics* 157:1293.

Sliding window analysis of *Cid*

50-nucleotide (nt) window, in steps of 10 nt, using all sites

N-terminal tail region
(mostly non-synonymous)



C-terminal core
(mostly synonymous
substitutions)

intraspecific polymorphism within *D. simulans* (π)

interspecific divergence (K)

Malik & Henikoff (2001) *Genetics* 157:1293.

Evidence for adaptive evolution in *D. melanogaster* and *D. simulans Cid*

- Polymorphism was studied in *D. melanogaster* (15 strains) and *D. simulans* (8 strains), and divergence between them

- McDonald-Kreitman test for the *D. melanogaster* lineage (box): $P < 0.006$

	Fixed differences	**Polymorphic sites**
Non-synonym.	8	0
Synonymous	4	9

Malik & Henikoff (2001) *Genetics* 157:1293.

A large fraction of amino acid differences are under positive selection

- Using data on many different genes, methods have been developed to use the McDonald-Kreitman approach to estimate what fraction of amino acid differences between *D. melanogaster* and *D. simulans* are caused by directional selection.

- This fraction is of the order of 25%, a surprisingly high value.

See: Bierne & Eyre-Walker (2004) *Mol. Biol. Evol.* 21: 1350.

Indirect evidence for selection: selective sweeps

- After an advantageous mutation has spread ("swept") through a population, the level of polymorphism will be reduced across the region (*i.e.*, at closely linked neutral sites).

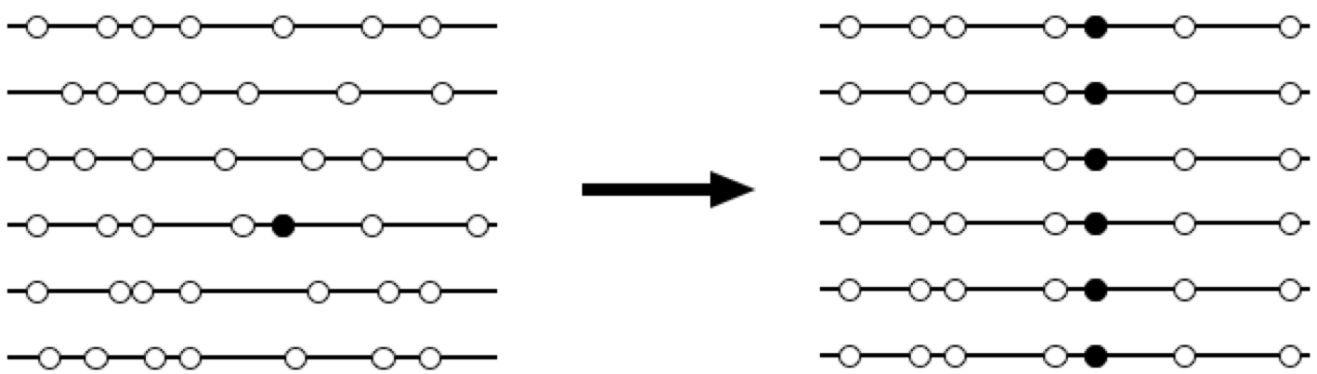
- This is because a unique selectively favourable mutation may arise at a site in a DNA sequence that is completely linked to a polymorphic neutral variant segregating in a population.

- This effect is was called "hitchhiking" by Maynard Smith and Haigh (Maynard Smith & Haigh (1974) *Genet. Res.* 12:12). "Selective sweeps" (Berry *et al.* 1991 in *Genetics*) are a case of what is called linked selection.

- See text book, pp. 407-416

Signature of a selective sweep

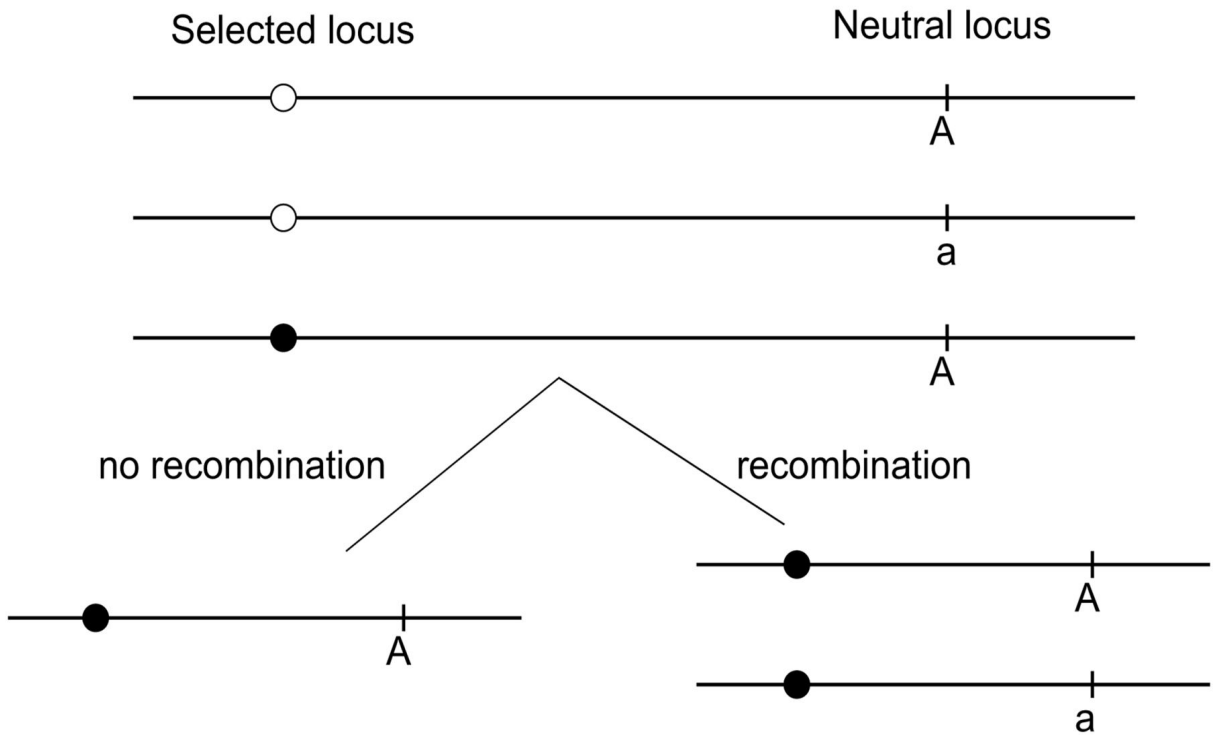
- A selective sweep fixes variants linked to the selected site; it is a form of hitch-hiking:



- As the black (advantageous) variant increases in frequency in a population, it causes low diversity at closely linked sites in a sequence (white circles). This means that π or θ_w will be reduced.

- Such an effect in terms of reduced variability can detectable if the time since selective substitution is sufficiently small (around $0.25N_e$ generations).

The basic hitchhiking model



Basic hitchhiking model. The upper part of the figure shows the three haplotypes present in a population when a beneficial mutation (filled circle) occurs at the selected locus. The wildtype allele at the selected locus is indicated by an open circle. At the neutral locus two alleles A and a are present. The haplotypes after the fixation of the beneficial allele are depicted in the lower part of the figure. If no recombination occurs during the fixation process one haplotype is present (left side). With recombination the neutral locus stays polymorphic and two haplotypes remain (right side).

Stephan (2019) *Genetics* 211:5-13

Molecular signature of a selective sweep

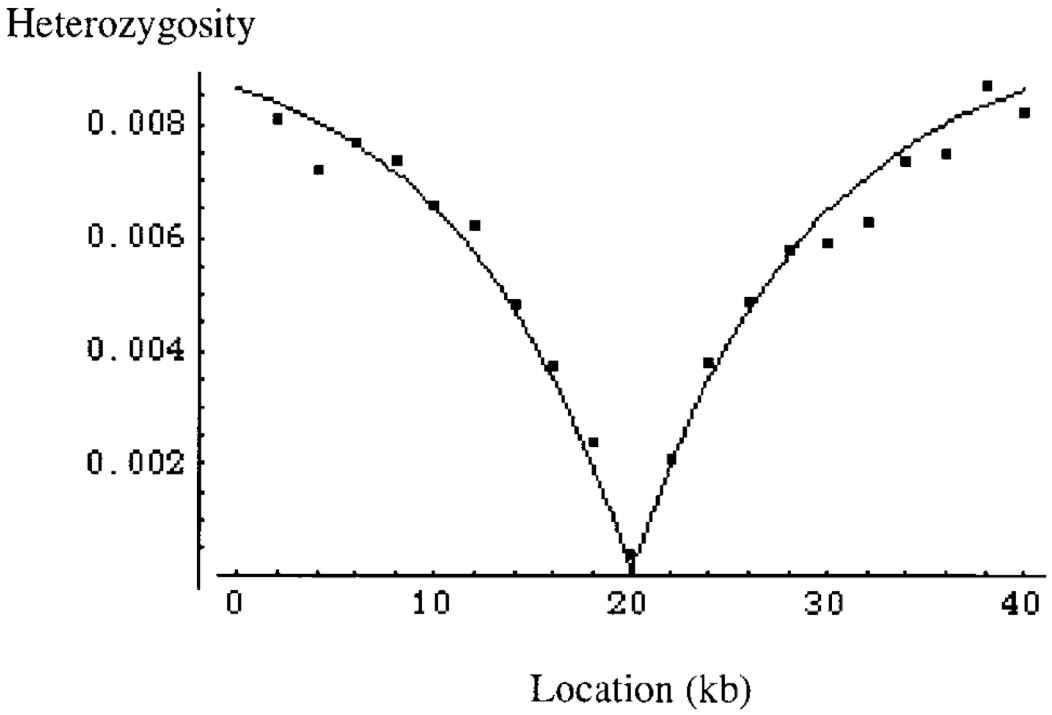
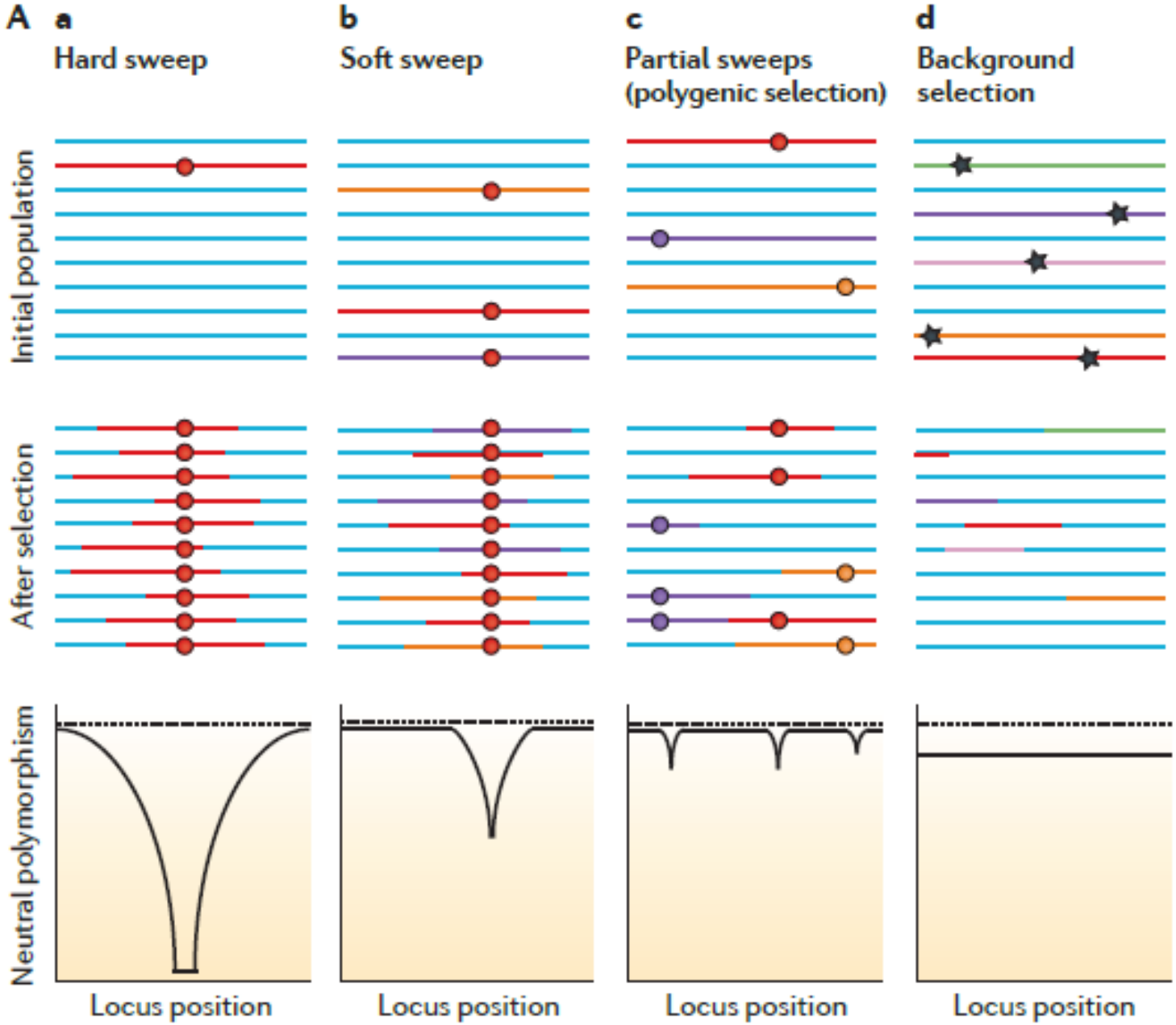


FIGURE 2.—Average nucleotide diversity along a recombining chromosome under the model of genetic hitchhiking, with $n = 5$, $\rho = 10^{-8}$, $N = 2 \times 10^5$, $\theta = 0.01$, and $T_{\text{limit}} = 7.0$. Squares represent average heterozygosity at single nucleotide sites averaged over 50,000 replicates of the simulations. The expected π value (continuous curve) was calculated using Equation 13 of KIM and STEPHAN (2000), with $r = \rho\,|i - 20{,}000|$ as the recombination rate between a nucleotide site i and the site of selection. Directional selection occurs at position 20 kb with $s = 0.001$ and $\tau = 0.005$.

Kim & Stephan (2002) *Genetics*

Types of hitchhiking effects and linked selection



As time goes by and recombination happens between haplotypes with and without selected alleles, we expect to see differences:

- selective sweeps lead to a localized effect, reducing genetic diversity around the selected locus.
- background selection will affect the entire region in a similar way.

Courtesy: Vitor Sousa

Cutter & Payseur (2013) *Nature Reviews Genetics*

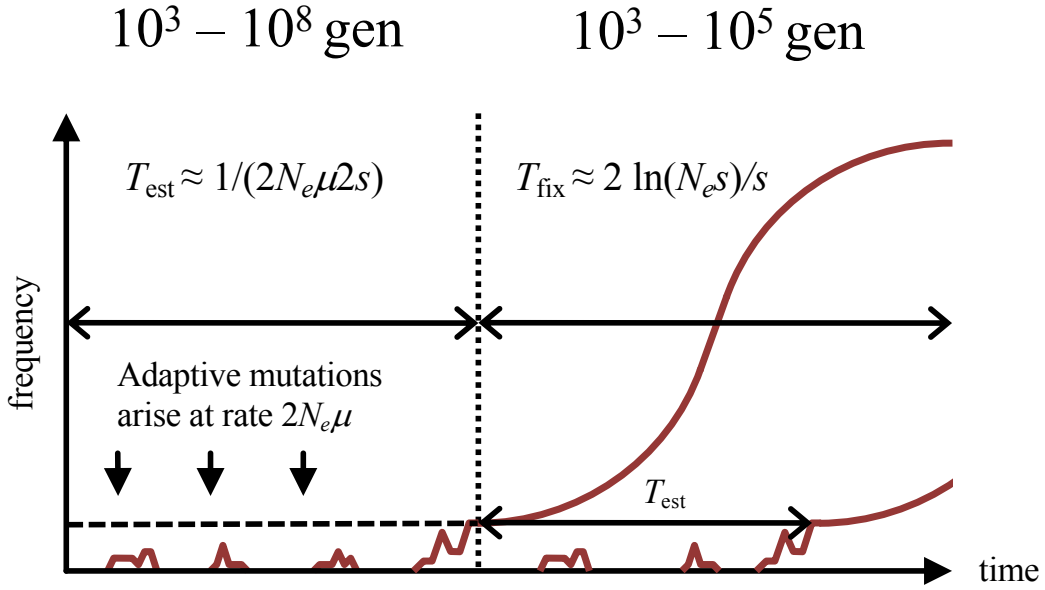
Difference between soft and hard sweeps

- **Hard sweep**: a new advantageous mutation arises and spreads to fixation due to selection. Under this model, neutral variation near to the favored site "hitch-hikes" along with the favored allele. This impacts patterns of variation around the selected site in ways that can be detected using a variety of tests of selection.

- **Soft sweep**: Two different scenarios that contrast with the hard sweep model.
  - Due to a change in selection, an allele that is already segregating in the population (i.e., standing variation) becomes selectively favored, and sweeps up in frequency. It is usually assumed that the allele is neutral or mildly deleterious prior to the change in selection.
  - Multiple independent mutations at a single locus are all favored and all increase in frequency simultaneously. If the favored alleles are all similarly advantageous, then typically none of the favored mutations would fix during the selective event.
  - These two soft sweep scenarios tend to be more difficult than hard sweeps to detect using standard tests of selection.

Speed of Adaptation by New Mutations

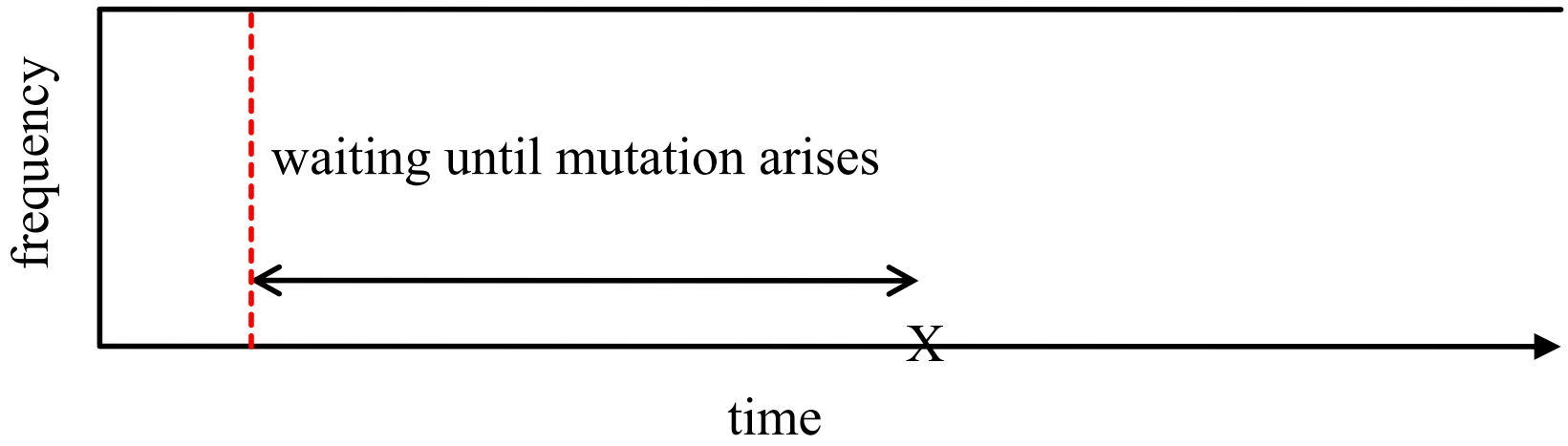
- Mutations arise at a rate of $2N_e\mu \approx \Theta$ = number of beneficial mutations that enter population per generation (haploid case)
- ~ 30% of the time even extremely beneficial mutations will get lost by drift
- mutation needs to 'survive' drift
- only a fraction of ~$2s$ of these will survive drift



- the rate of successful establishment is thus $\approx 2N_e\mu\, 2s \approx 4N_e\mu\, s$ (="population-scaled selection coeffient")
- thus, the time to establishment $\approx 1/(2N_e\mu\, 2s\,) \approx 1/s$
- time to fixation $\approx \ln(N_e s)/s$

Messer & Petrov (2013)

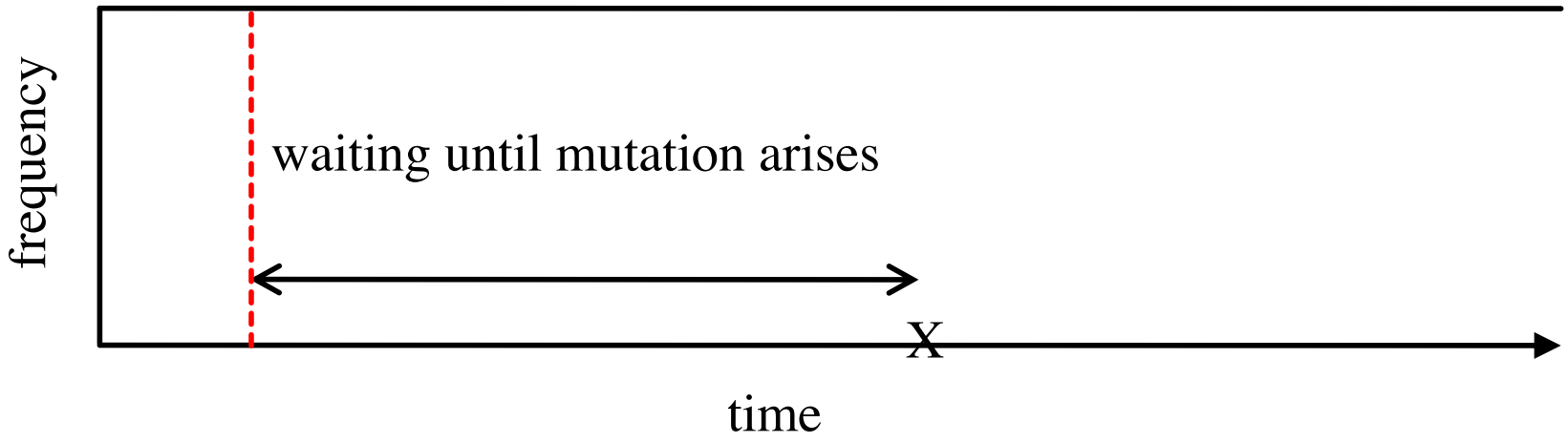
Selective Sweeps are HARD if $4N_e\mu s << 1$

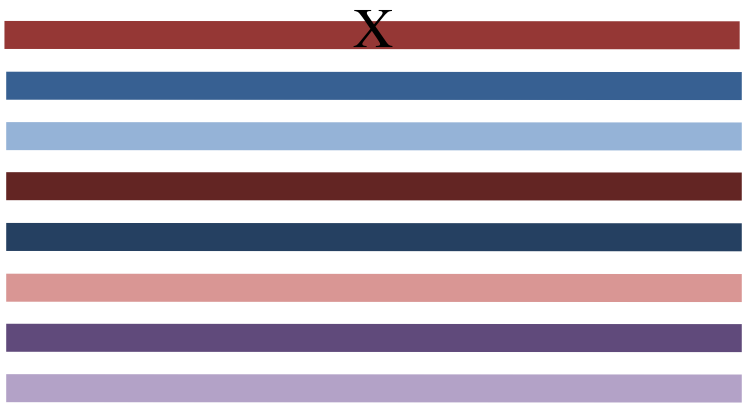


- long waiting time for adaptive mutation to occur
- mutation rates may be low
- effective population size may be low
- adaptation might thus be mutation-limited and rather slow

Messer & Petrov (2013)

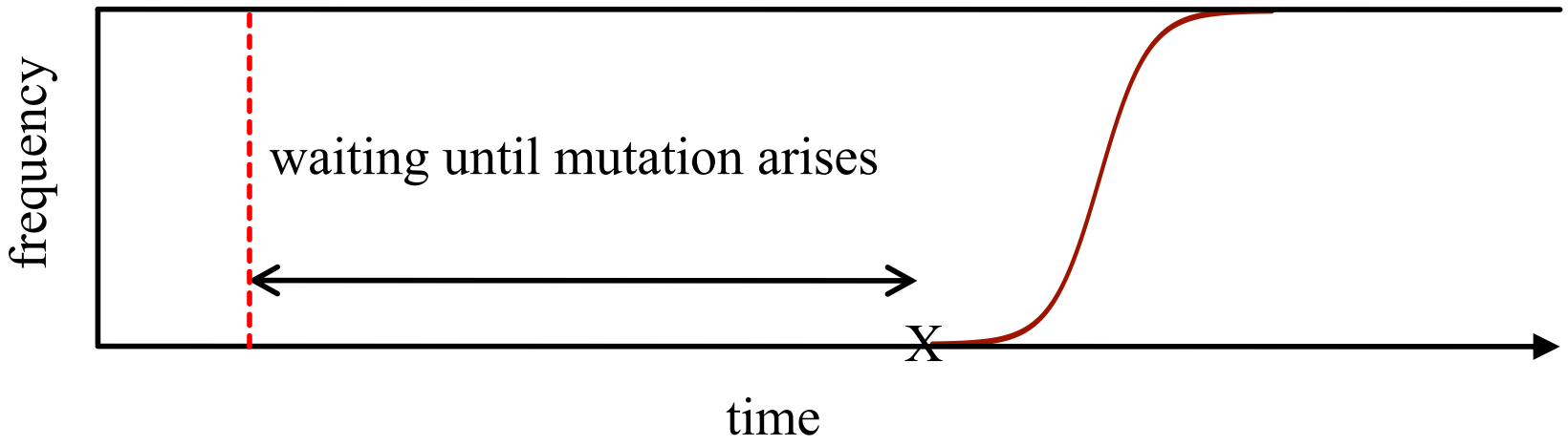
Selective Sweeps are HARD if $4N_e\mu s << 1$

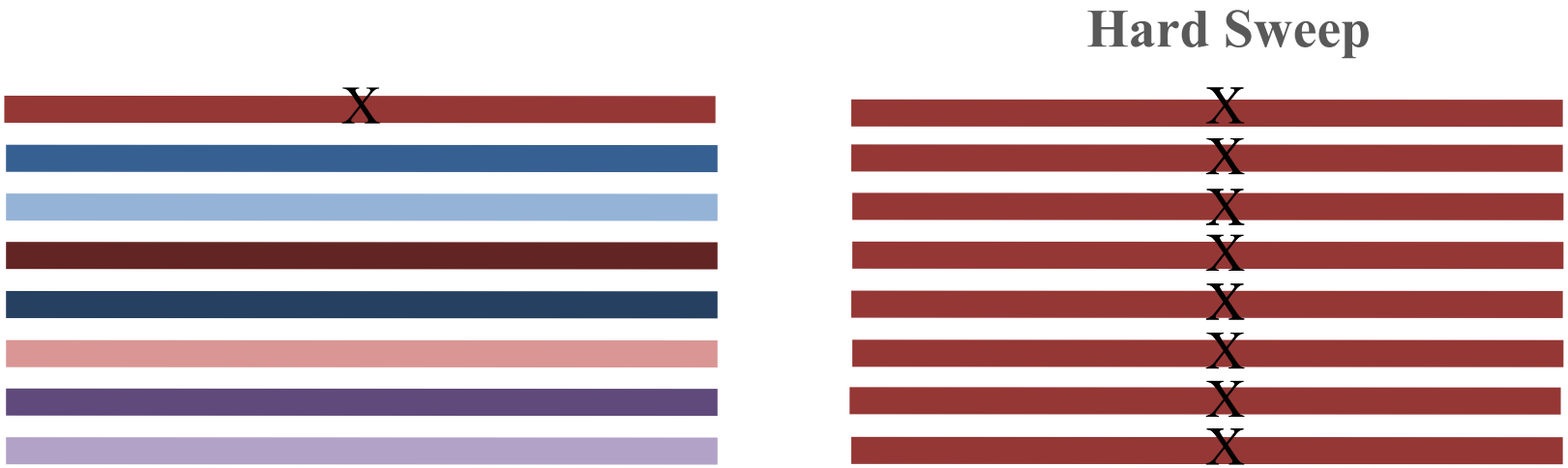




Messer & Petrov (2013)

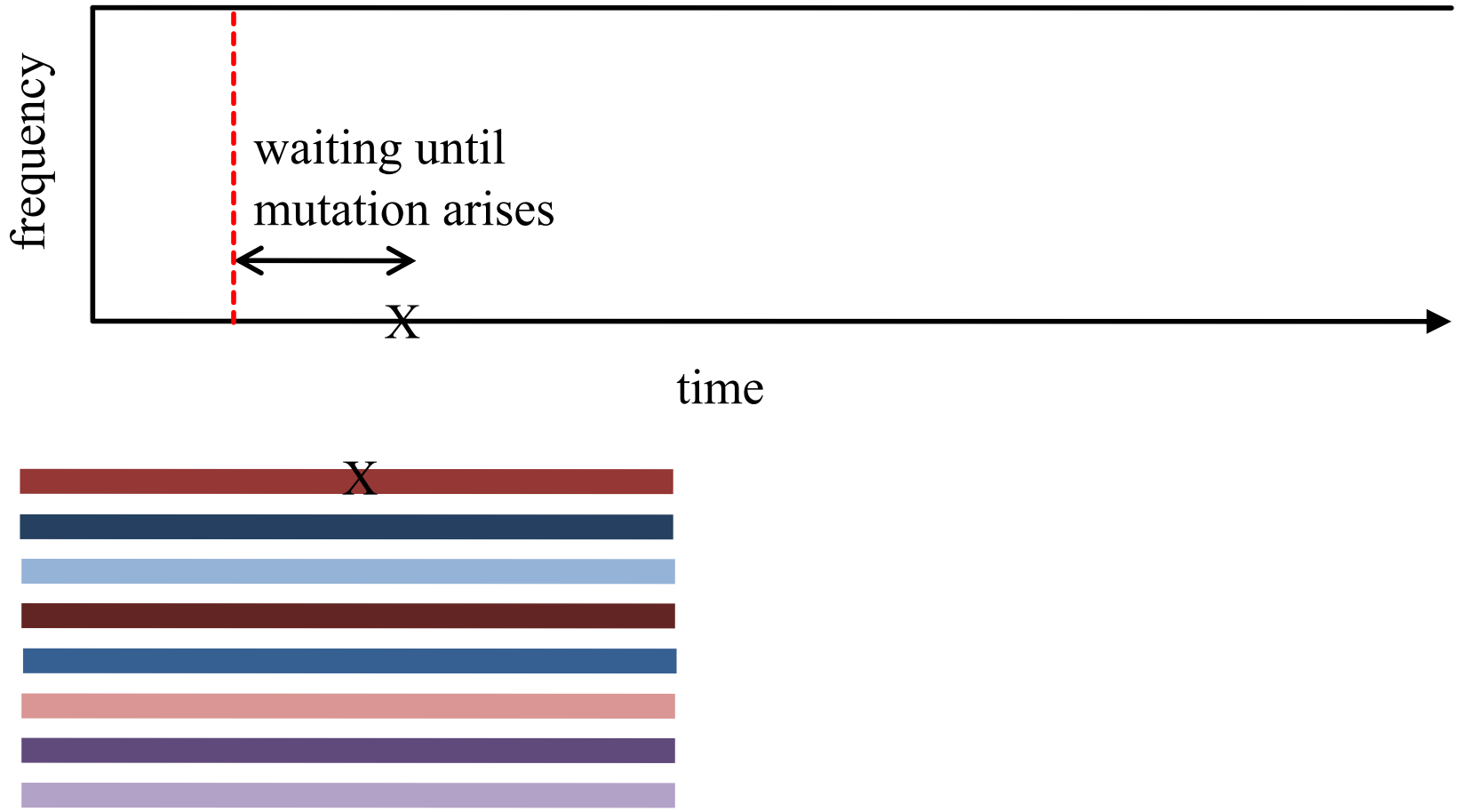
Selective Sweeps are HARD if $4N_e\mu s << 1$





Messer & Petrov (2013)

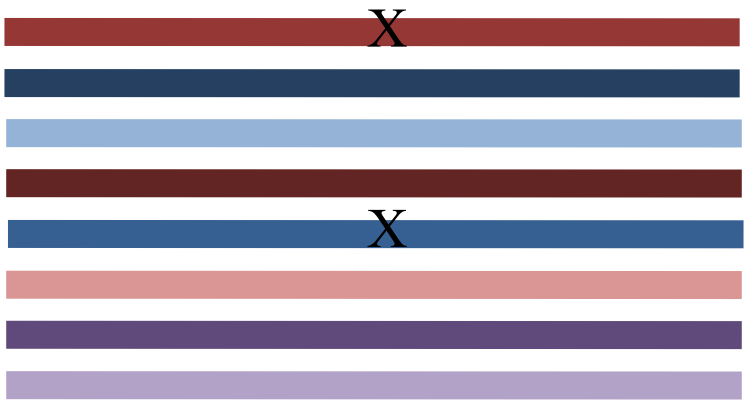
Selective Sweeps are SOFT if $4N_e\mu s \geq 1$



We can have a non-mutation limited regime because either the mutation rate is high, the current population size is large, or standing variation is abundant in the population.

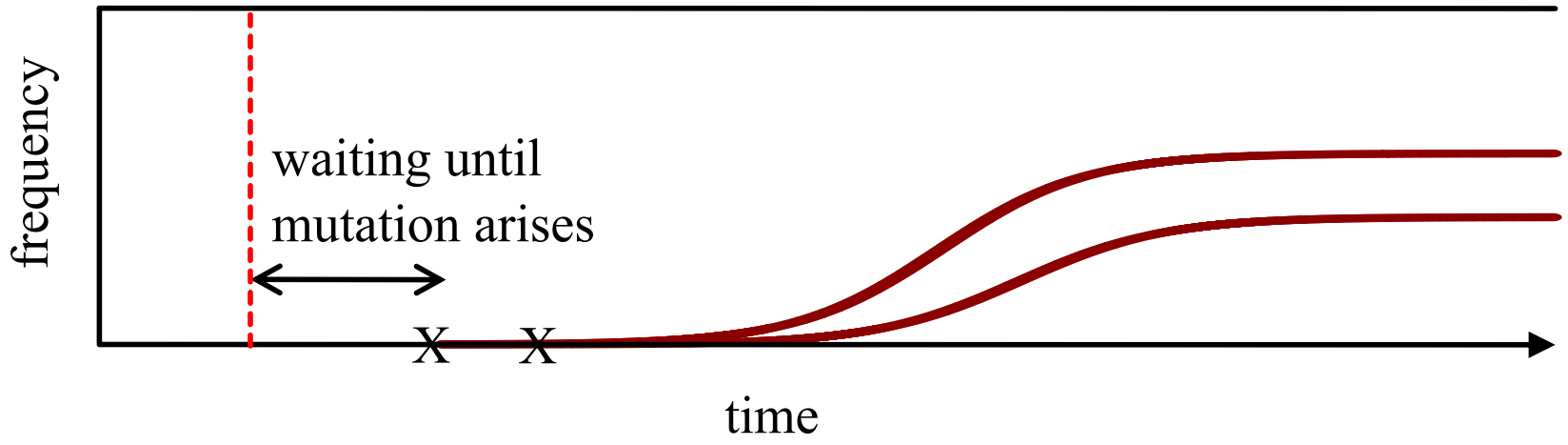
Messer & Petrov (2013)

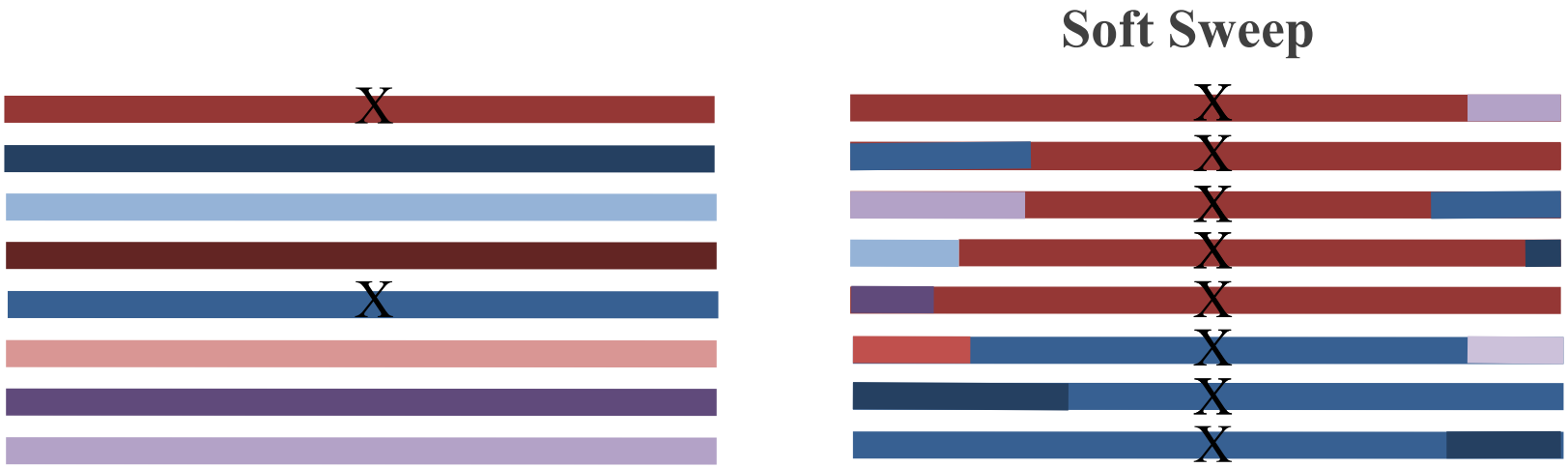
Selective Sweeps are SOFT if $4N_e\mu s \geq 1$



Messer & Petrov (2013)

Selective Sweeps are SOFT if $4N_e\mu s \geq 1$





Messer & Petrov (2013)

When adaptation is not mutation-limited

$$\text{If } 4N_e\mu s \geq 1:$$

- soft sweeps might be common
- multi-step adaptations can arise fast
- often large amounts of standing variation
- mutation rates can sometimes be high; allelic “series” are common
- effective population sizes can be large
- selection can be strong

Messer & Petrov (2013)

Indirect evidence for selection: statistics of variant frequency distributions

- It is also possible to work out the frequencies at which variants are expected to be found in equilibrium populations, under both neutrality and selection.
- Under neutrality, most variants are expected to be quite rare.
- If selection is operating on the sequence, it will affect the frequencies of variants in the sample.
- This forms the basis for some tests for selection, and methods for estimating the intensity of selection.
- See text book, chapter 6, especially pp. 287-290; also see Vitor's slides on allele and site frequency spectra.

Tajima's *D*

- Assuming neutrality and equilibrium, the expected value of both π and θ_w is $4N_e\mu$. Thus, under neutrality, $\pi = \theta_w$. The difference between π and θ_w equals (approximately) a quantity called Tajima's *D*. Under neutrality, $D = 0$.

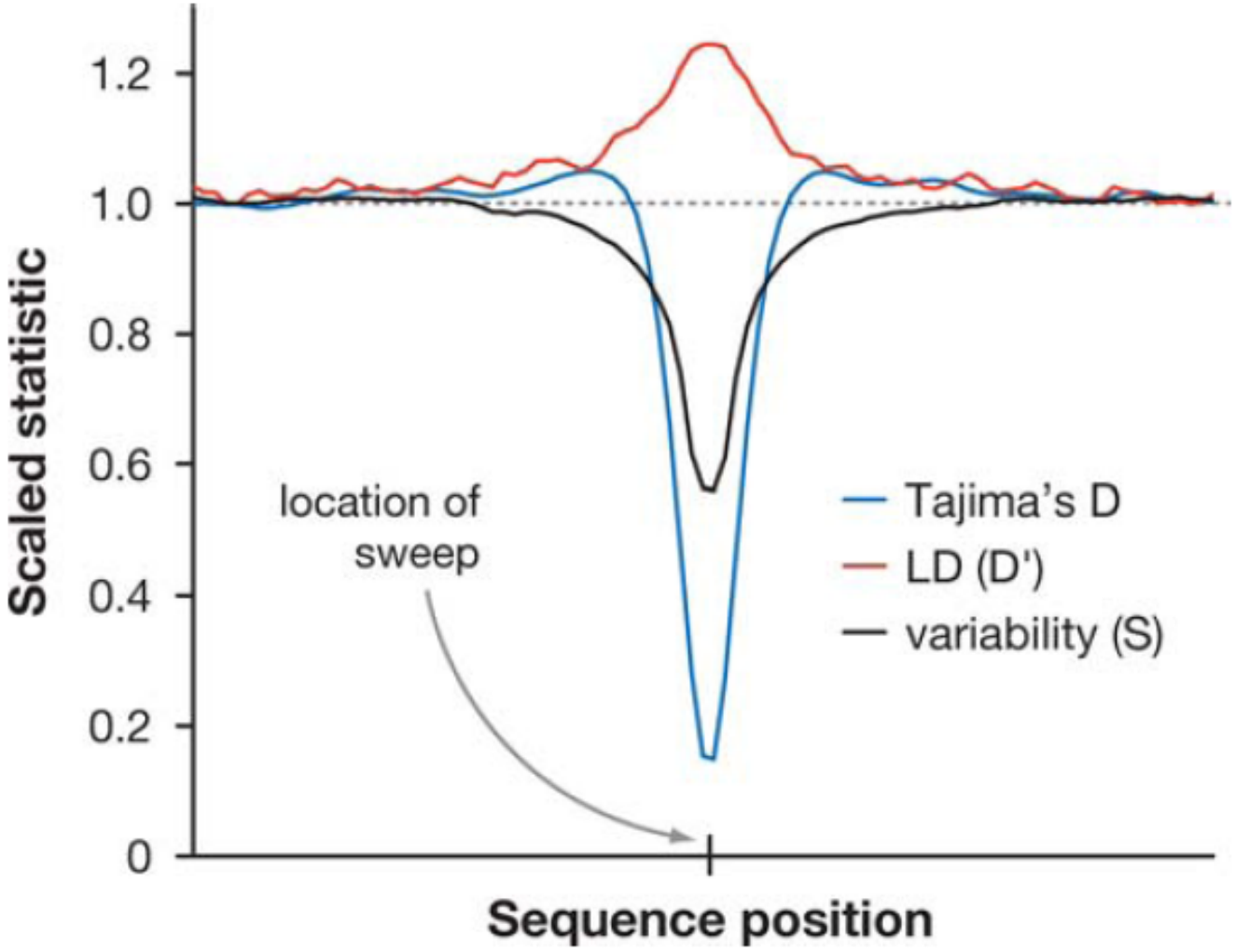
- Under non-neutrality the two quantities can differ, so that $D \neq 0$. While Watterson's method only takes the number of SNPs into account, π also takes into account the frequency of the polymorphisms.

- If $\pi \neq \theta_w$, (so $D \neq 0$), it suggests the possibility of selection:

- $\pi < \theta_w$ (negative Tajima's *D*): an excess rare variants, as compared with what is expected under neutrality, which suggests purifying selection or selective sweeps.

- $\pi > \theta_w$ (positive Tajima's *D*): an excess of high frequency variants might suggest balancing selection.

Strongly reduced Tajima's *D* under a sweep



Problems with interpretation

- But there are several problems:
- First, statistical tests are necessary to decide whether a sample could not have arisen by a chance process of neutral mutation and drift. Only if we can say this, can we conclude that something such as selection has affected the sequences. Neutrality is used as a null hypothesis.
- Second, the population may not have been constant in size, as assumed in the model, and so its demographic history may cause $\pi \neq \theta_{w.}$.
- For example, a negative Tajima's *D* can be caused by population expansion after a recent bottleneck, and a positive value can be caused by a sudden population contraction.

Bottlenecks and selection can both cause negative Tajima's *D*

- The spread of an advantageous mutation affects diversity very much like a bottleneck, but only in the region around the gene.

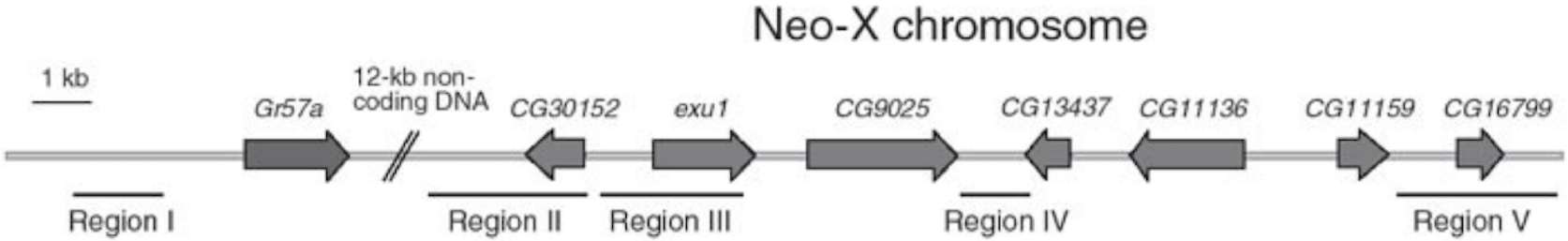
- Extreme bottleneck: only one haplotype present, then new neutral variants occur. This leads to $\pi < \theta_w$ so that Tajima's $D < 0$.

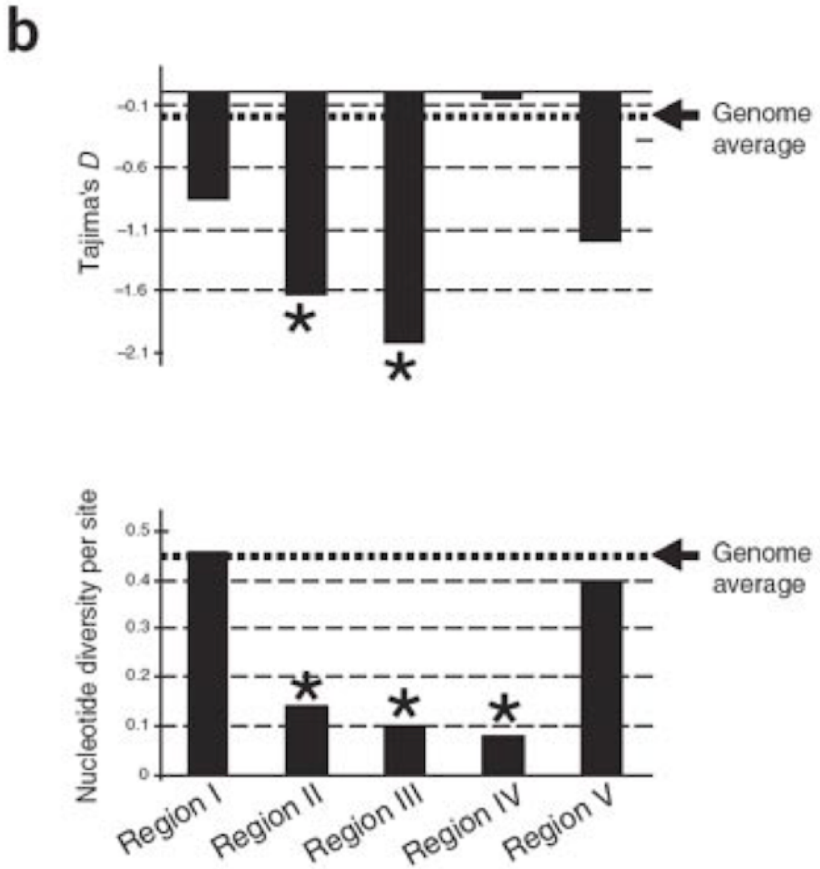
- Fixed advantageous mutation: a single beneficial haplotype selected, then new neutral variants occur $\pi < \theta_w$, so that Tajima's $D < 0$. In contrast to a bottleneck, Tajima's *D* is expected to be reduced only in the local genomic region around the selected locus.

Genome scans for selective sweeps

- There is currently a lot of interest in using scans of variability across the genome, to look for patterns that suggest a recent selective sweep.

- The hope is that this will lead to identification of the mutations that have been favoured by selection.





Evidence for a selective sweep on the neo-X chromosome of *D. miranda*

Bachtrog (2003) *Nat. Genet.* 34: 215.

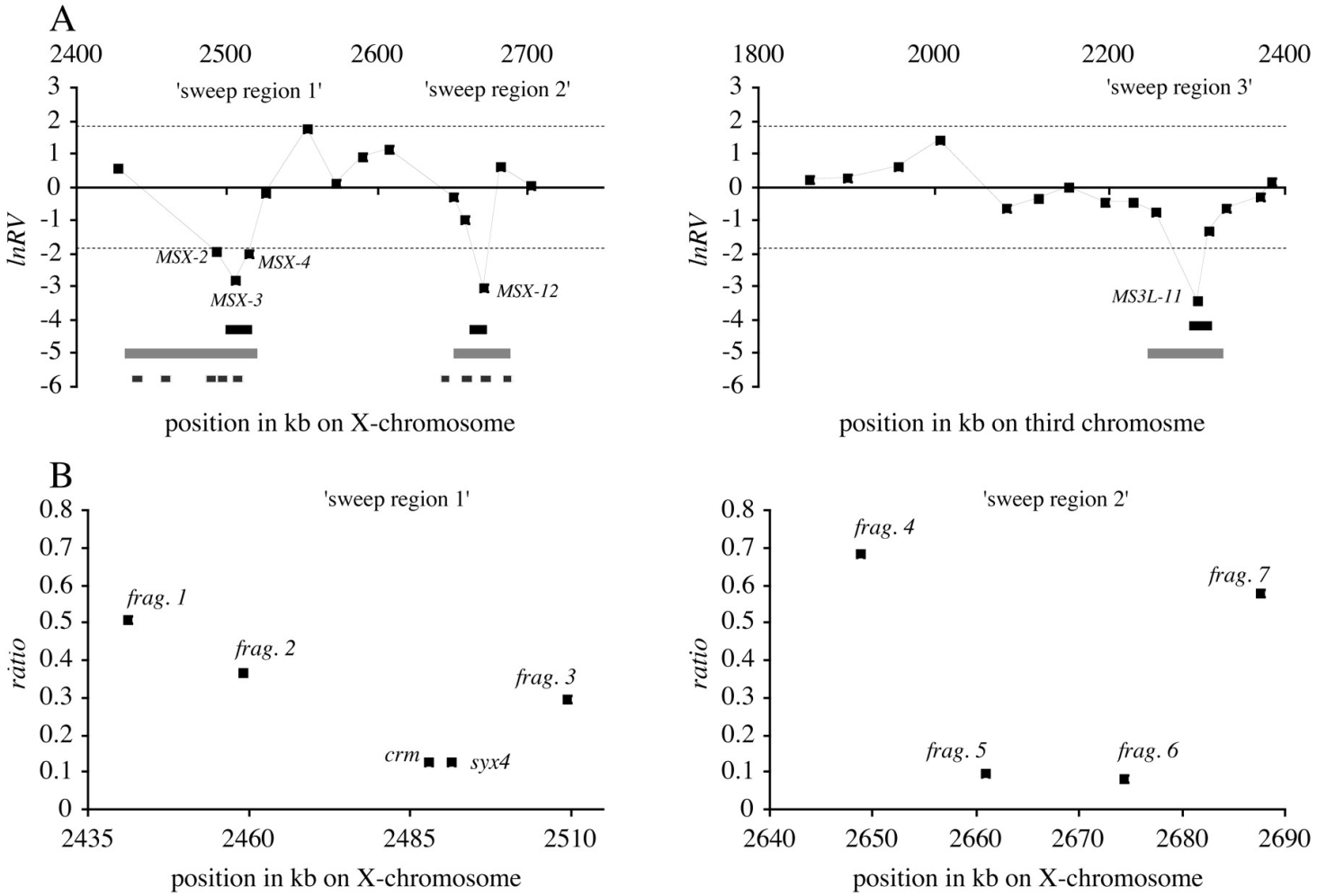
Genome scans for sweeps in *D. melanogaster*

- One subject of study are non-African populations of *D. melanogaster*, which have migrated relatively recently (probably 10,000 – 20,000 years ago) out of Africa.

- They must have adapted to their new environments. It should be possible to see which regions of the genome show evidence of selective sweeps.

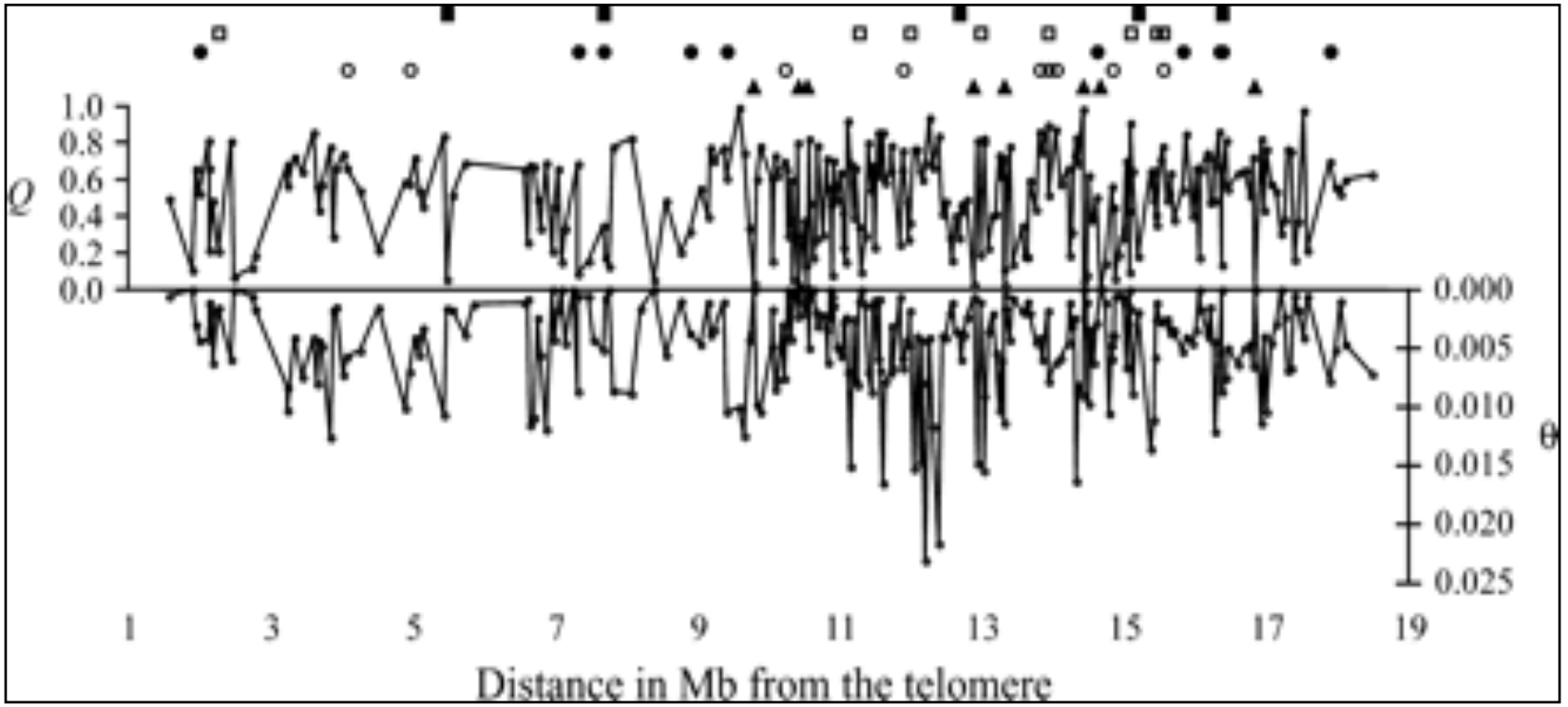
- The problem is that they have also gone through bottlenecks of small population size, which has similar effects to sweeps, but are distributed over the whole genome.

Estimates of microsatellite (A) and sequence diversity (B) in non-African and African populations of *D. melanogaster*



B. Harr *et al*. (2002) Proc. Natl. Acad. Sci. USA 99, 12949-12954 PNAS

Genome scan of 250 approximately 500 bp non-coding sequences across the X chromosome of *Drosophila melanogaster*



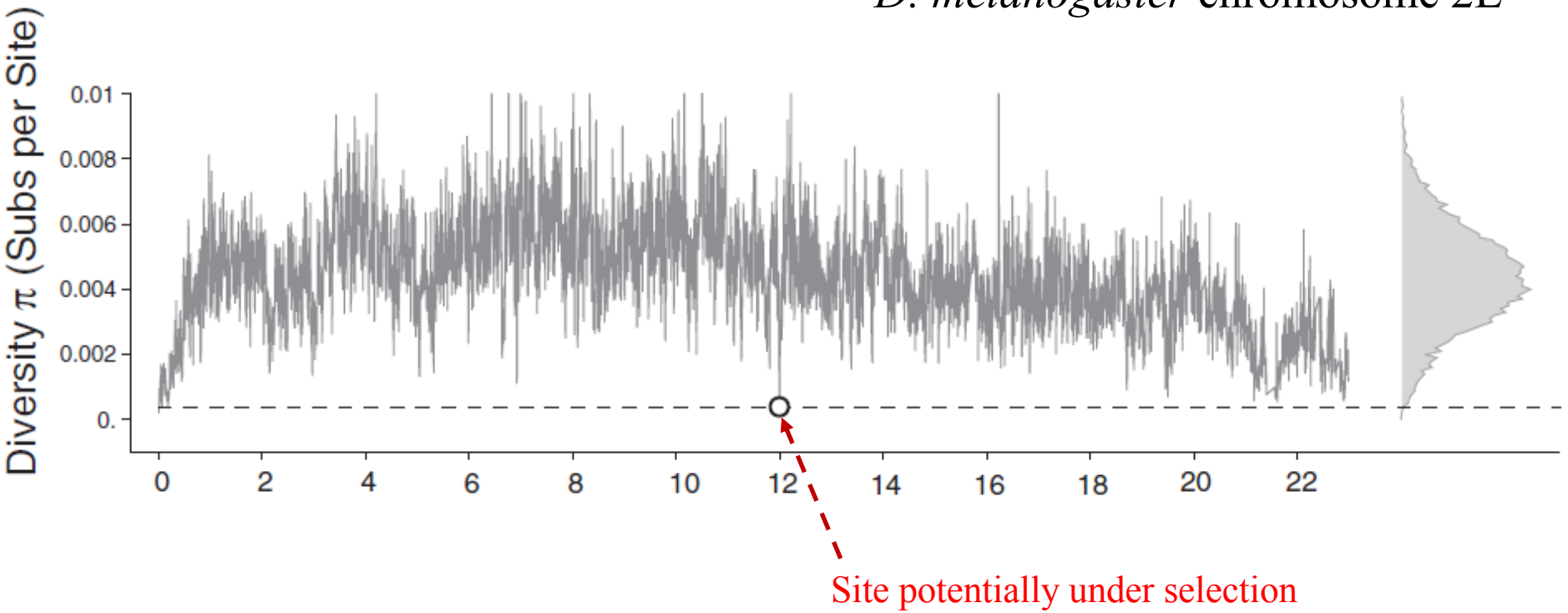
Q is the probability of getting as many as the observed number of polymorphisms in the European sample using a bottleneck model. Empty and filled circles indicate significantly negative or positive Tajima's *D*.

Ometto *et al*. (2005) *Mol*. *Biol*. *Evol*. 22:2119

Adaptive impact of the chimeric gene *Quetzalcoatl* in *Drosophila melanogaster*

θ_π measured in windows of 10 kb

D. melanogaster chromosome 2L

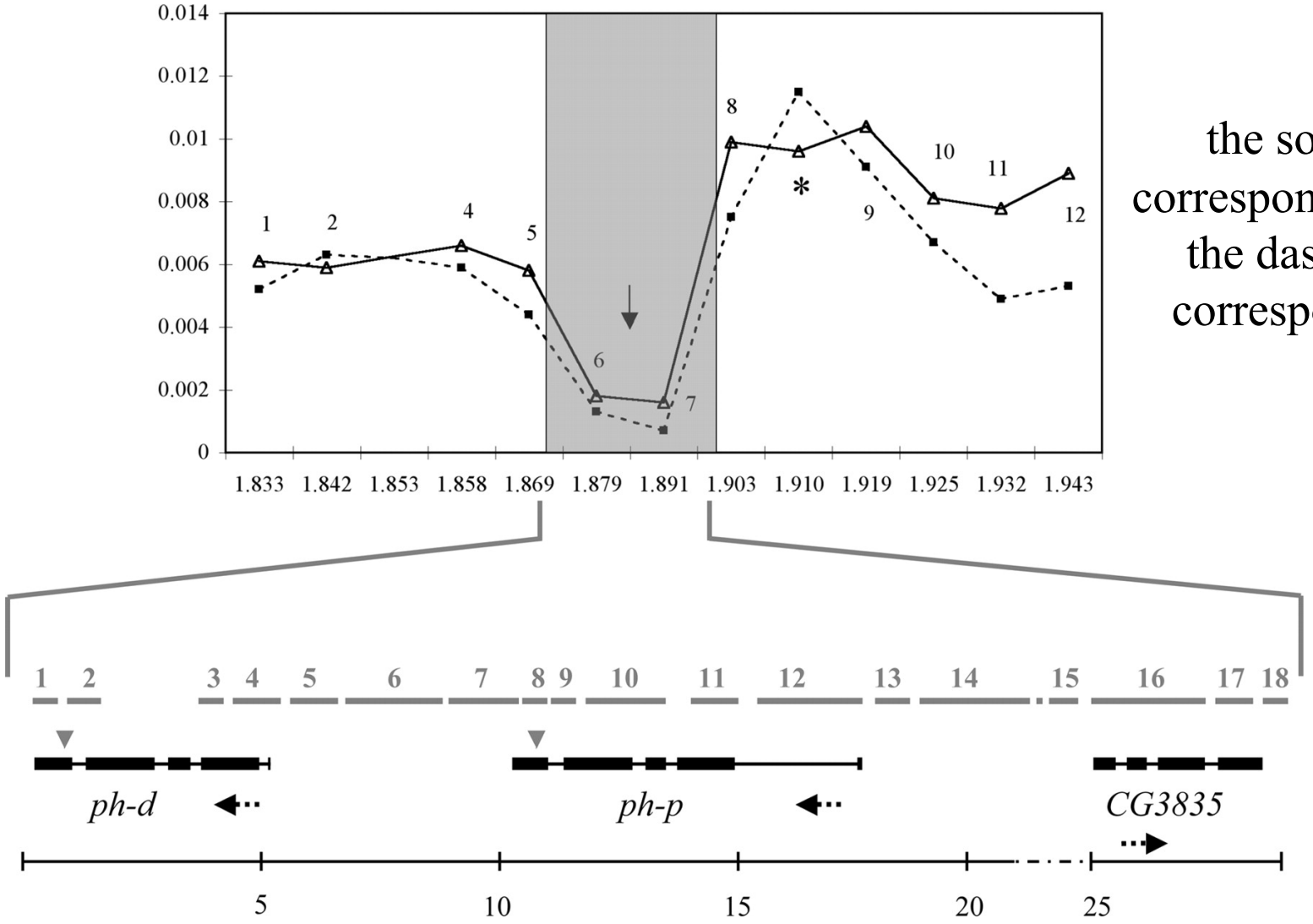


Quetzalcoatl gene formed by the fusion of two genes and which swept to fixation recently

Courtesy: Vitor Sousa

Rogers *et al*. 2010 PNAS

Selective sweep in the polyhomeotic gene region of *Drosophila melanogaster*

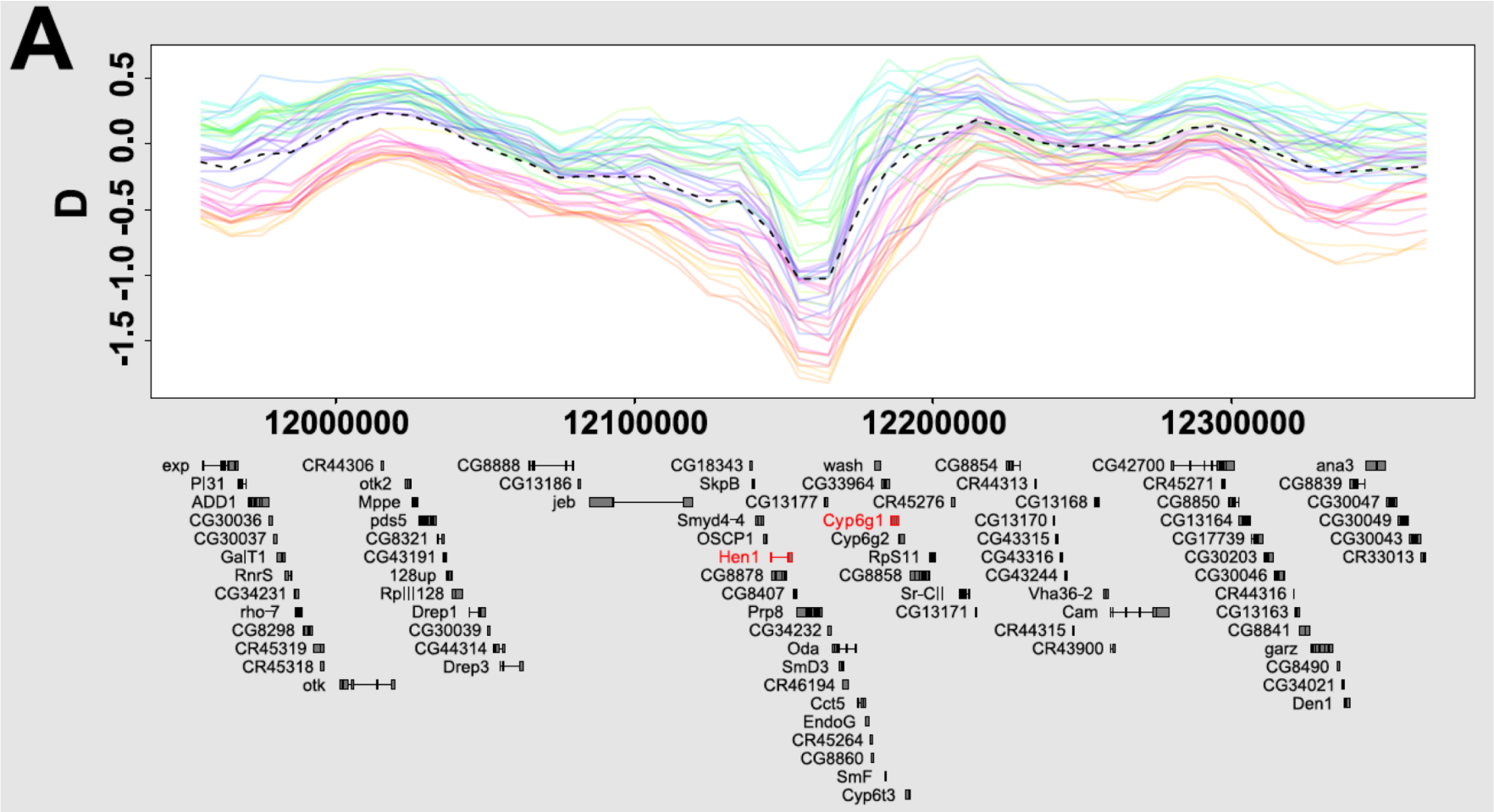


the solid line corresponds to θ and the dashed line corresponds to π

Beisswanger & Stephan PNAS 2008;105:14:5447-5452

PNAS

Selective Sweeps in European *D. melanogaster*



- 48 population samples from 32 locations across Europe
- Well-supported sweeps in the proximity of *Hen1, Cyp6g1*, *wapl / ph-p*, and the chimeric gene *CR18217* plus evidence of previously unknown sweeps.

Kapun *et al*. (2020) *Mol*. *Biol*. *Evol*., in revision

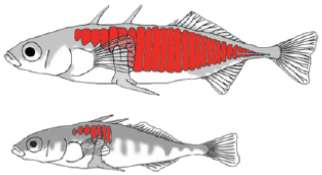
Many methods to detect loci under selection

- **Patterns of diversity within populations**
  - Levels of genetic diversity (*S*, π, HKA)
  - Levels of LD (*D*, *D'*, *r*, EHH, iHS, LDD)
  - Skew in allele frequency spectrum (Tajima's *D*, *H*, *D**, *F*, *Fs*)
- **Patterns of diversity between populations**
  - Levels of genetic diversity (*S*, π, HKA)
  - Levels of LD (*D, D', r,* EHH, iHS, LDD)
  - Skew in allele frequency spectrum (Tajima's *D*, *H, D*, F, Fs*)
  - F_{ST}, d_{xy}, population branch statistic (PBS)

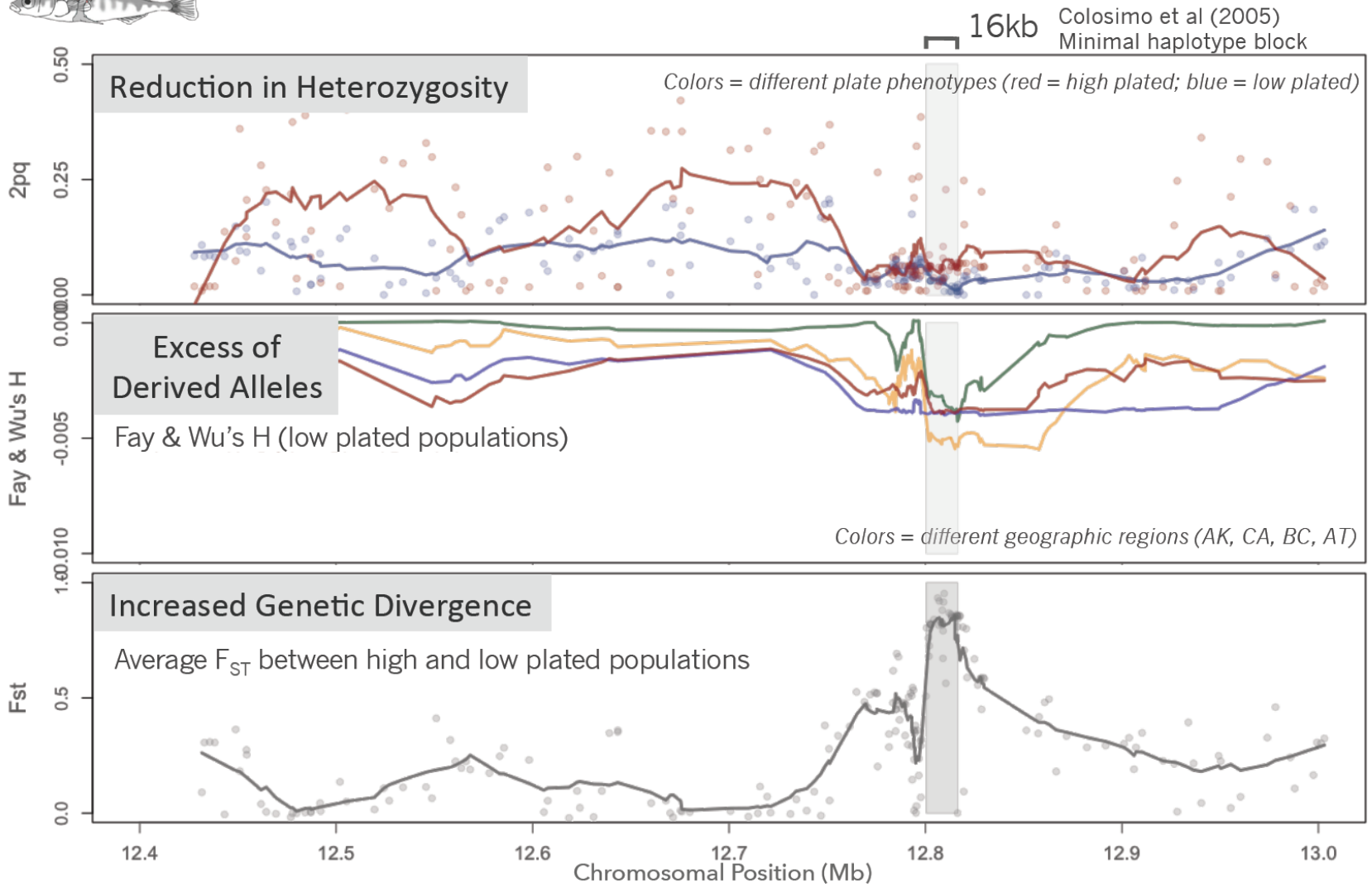
Test	Data	Pattern	Requires multiple loci	Robust to demographic factors?
Tajima's D and related	Population genetic data	Frequency spectrum	No	No
Modeling of selective sweep—spatial pattern	Population genetic data	Frequency spectrum/spatial pattern	No	No
Tests based on LD	Population genetic data	LD and/or haplotype structure	No	No
F_{ST} based and related tests	Population genetic data	Amount of population subdivision	Yes	No[a]
HKA test	Population genetic and comparative data	Number of polymorphisms/substitutions	Yes	No
Macdonald-Kreitman-type tests	Population genetic and comparative data	Number of nonsynonymous and synonymous polymorphisms	No	Yes
d_N/d_S ratio tests	Comparative data or population genetic data without recombination (6)	Nonsynonymous and synonymous substitutions	No	Yes

Courtesy: Vitor Sousa

Using multiple methods simultaneously

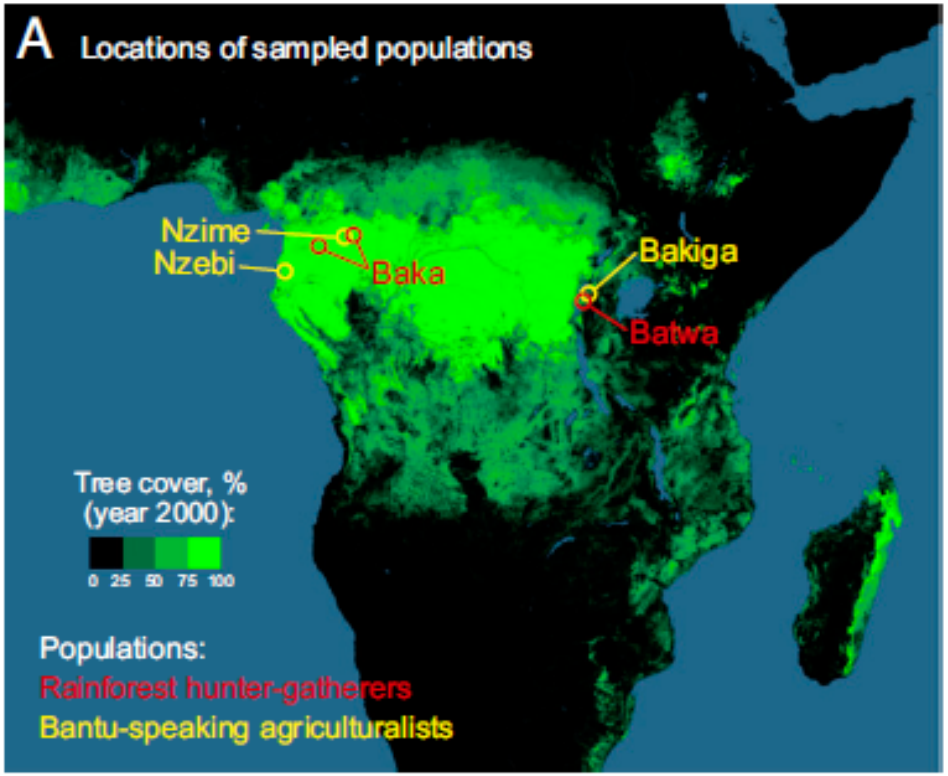


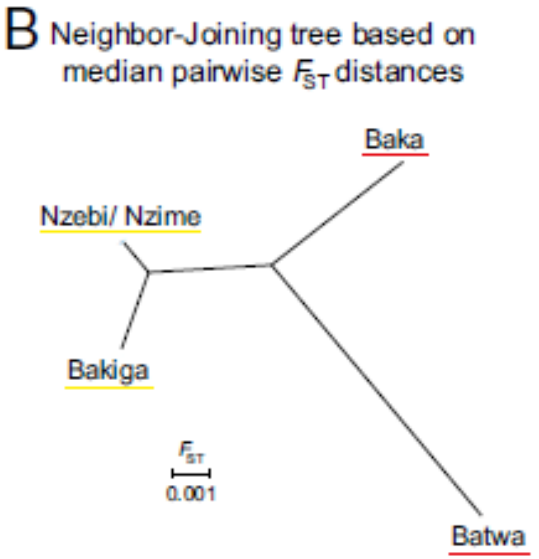
Molecular signatures of selection around *EDA* using high density genotyping arrays



Courtesy: Vitor Sousa & Felicity Jones

9 April: Practical on detecting selection: Data from rainforest Pygmy-Bantu people





- Perry *et al.* (2014) PNAS
- Compare Batwa vs Bakiga

See you on

9 April at 09:15

for the exercises in R