Bioinformatics for Next-Generation Sequencing
ICIPE, Nairobi, November 2014

Evolutionary Genomics 2: Population Genomics

Simon Martin

Department of Zoology
University of Cambridge

Outline

- Questions in population genomics
- Statistics, inference and examples
  - Genetic diversity and population size
  - Population subdivision and speciation
  - Selection
- Study design
  - Sampling and sequencing

Questions in population genomics

- Diversity and population structure
  - How much genetic variation exists in a population?
  - How and why is a species subdivided into populations?
- Demographic history and gene flow
  - How has a population size and distribution changed over time?
  - How much migration and gene flow occurs between populations / species?

Questions in population genomics

- Speciation
  - Which genes/events cause two species to become distinct and remain distinct?
  - What are the relative roles of adaptive and neutral forces in speciation?
  - What is a species?
- Adaptation
  - Where in the genome is natural selection acting?
  - What is the source of beneficial genetic variation?

Diversity and allele frequencies

- Heterozygosity: Proportion of the genome that is heterozygous

```
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGATCTCGTGCCTAAGTAGATCATGATACT
TAGGTCGTCGAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATTATGATAAT
```

  - Inbreeding
  - Selection
  - Population size
  - Humans ~0.001, Fruit Flies ~0.01

Measuring genetic diversity in a population

- *S*: number of variable sites
- Π (nucleotide diversity): Average number of differences between any pair of sequences

```
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGATCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATCATGATAAT
TAGGTCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATTATGATAAT
```

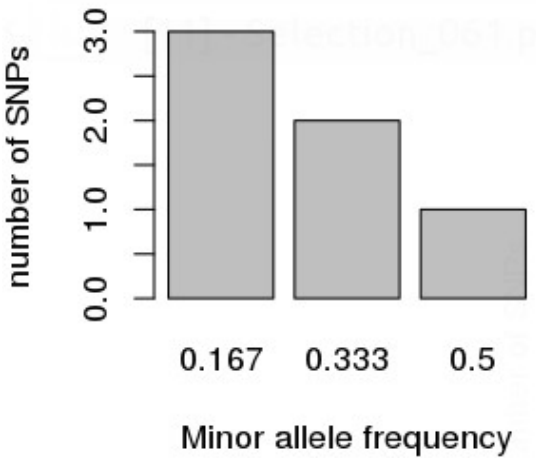
  - Selection
  - Population size
  - If no inbreeding Π ~ heterozygosity

Measuring genetic diversity in a population

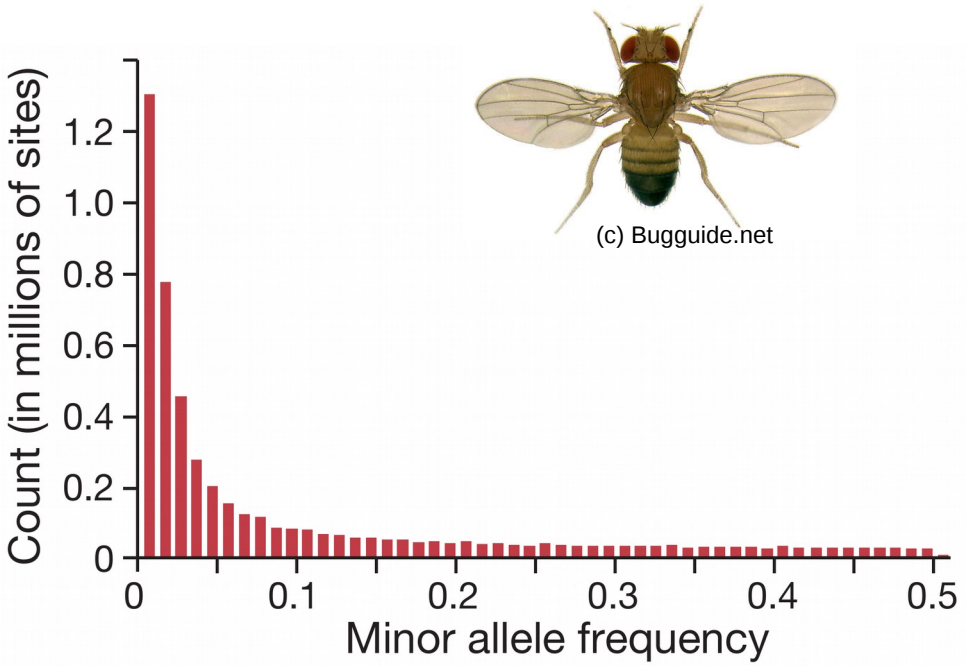
- Site frequency spectrum (SFS) number of occurances of variants at each frequency

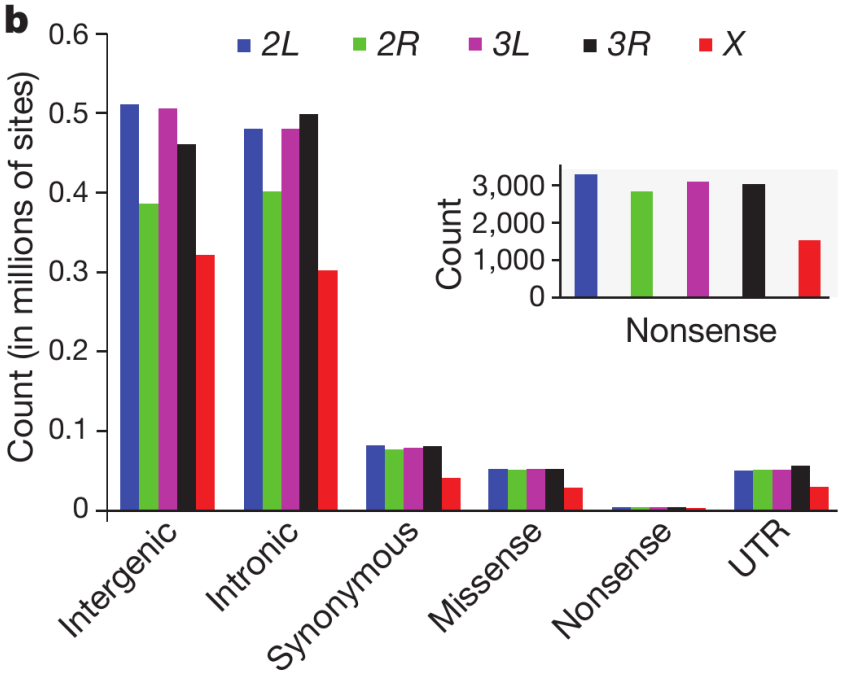
```
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGATCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATCATGATAAT
TAGGTCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTAGATTATGATAAT
```

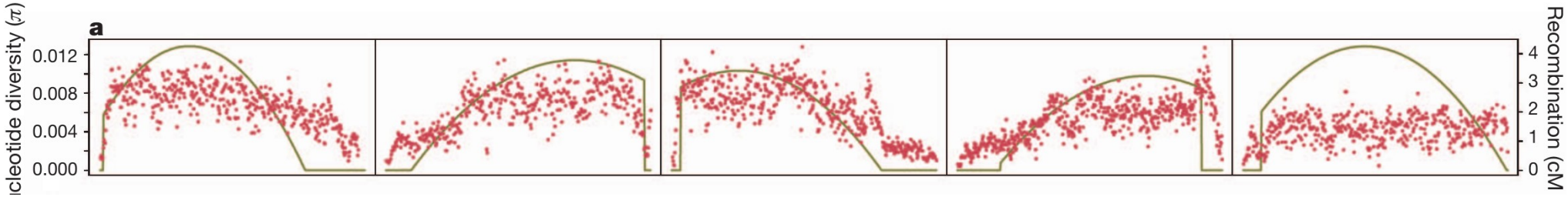
  - Selection
  - Changes in population size
  - Hybridisation



Measuring genetic diversity in a population

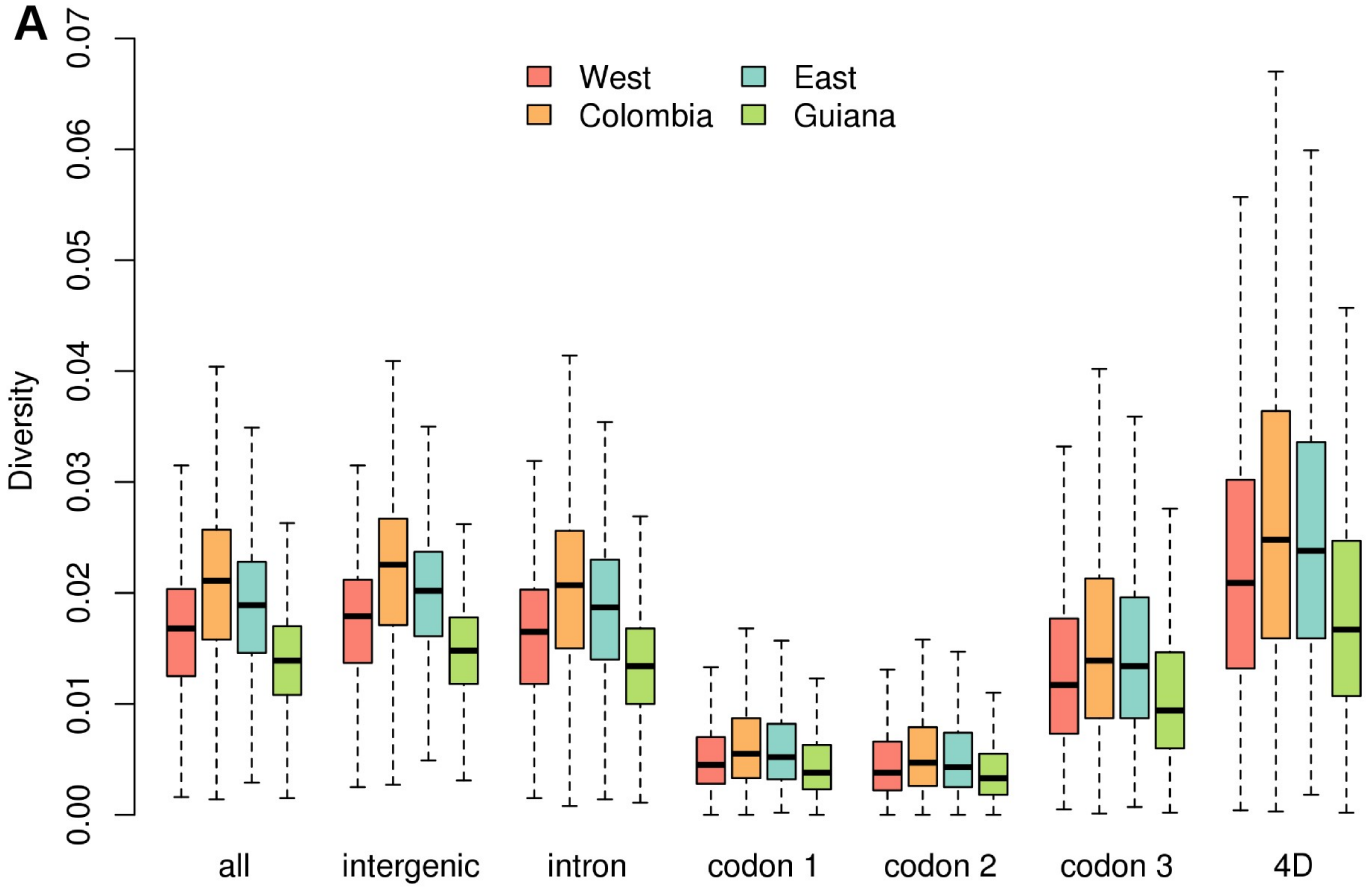


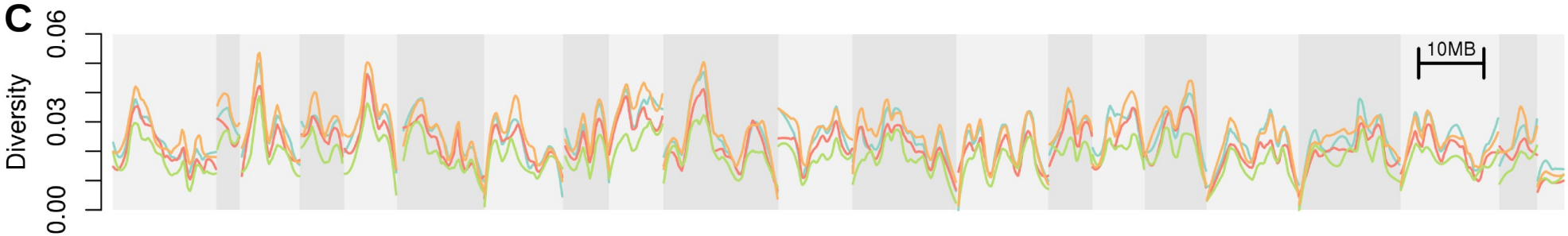




Mackay et al. 2012 Nature

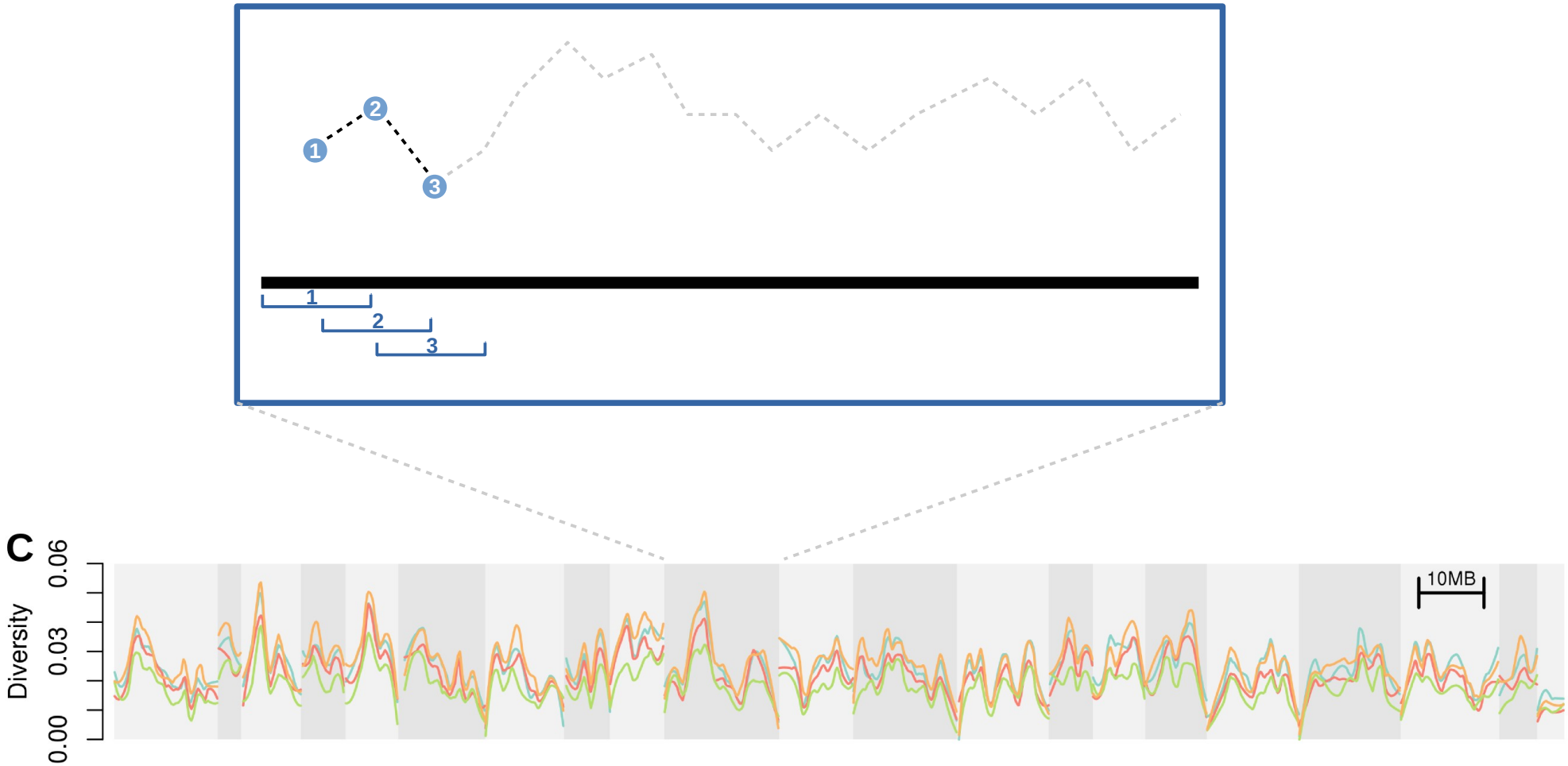
Measuring genetic diversity in a population





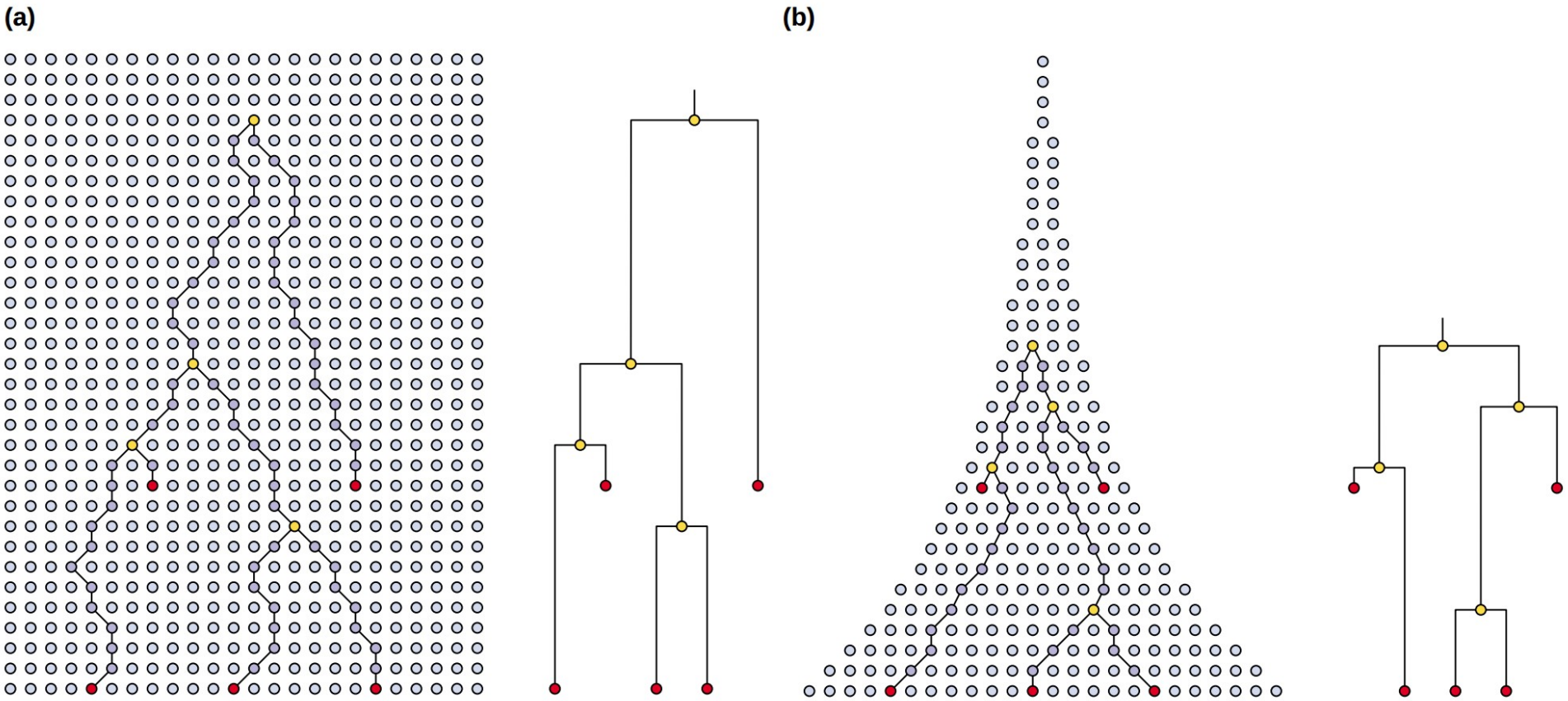
Measuring genetic diversity in a population

Sliding windows



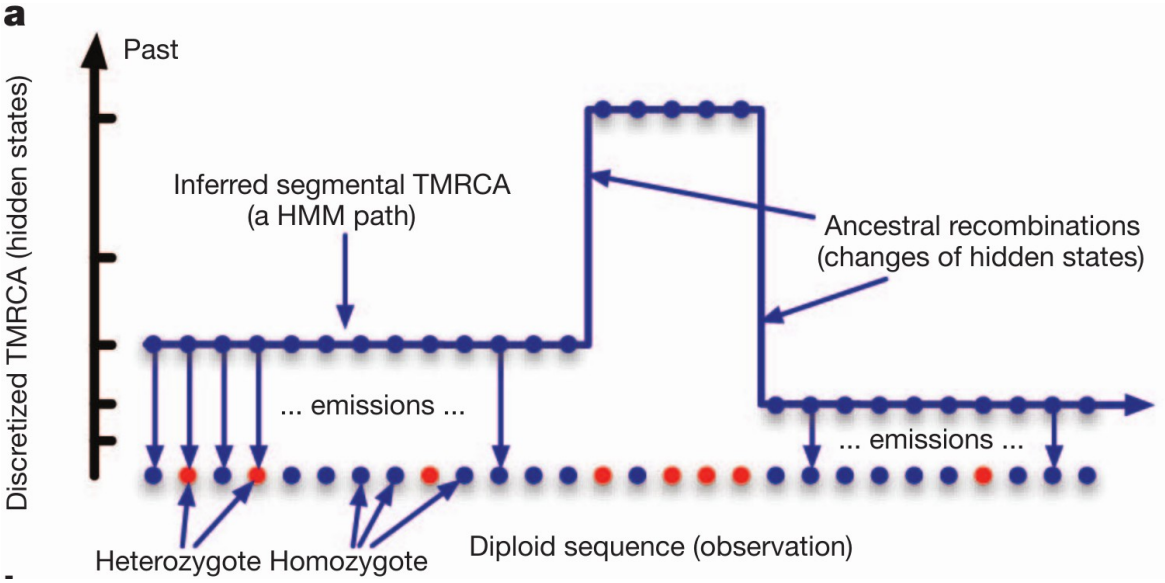
Inferring population size from diversity data

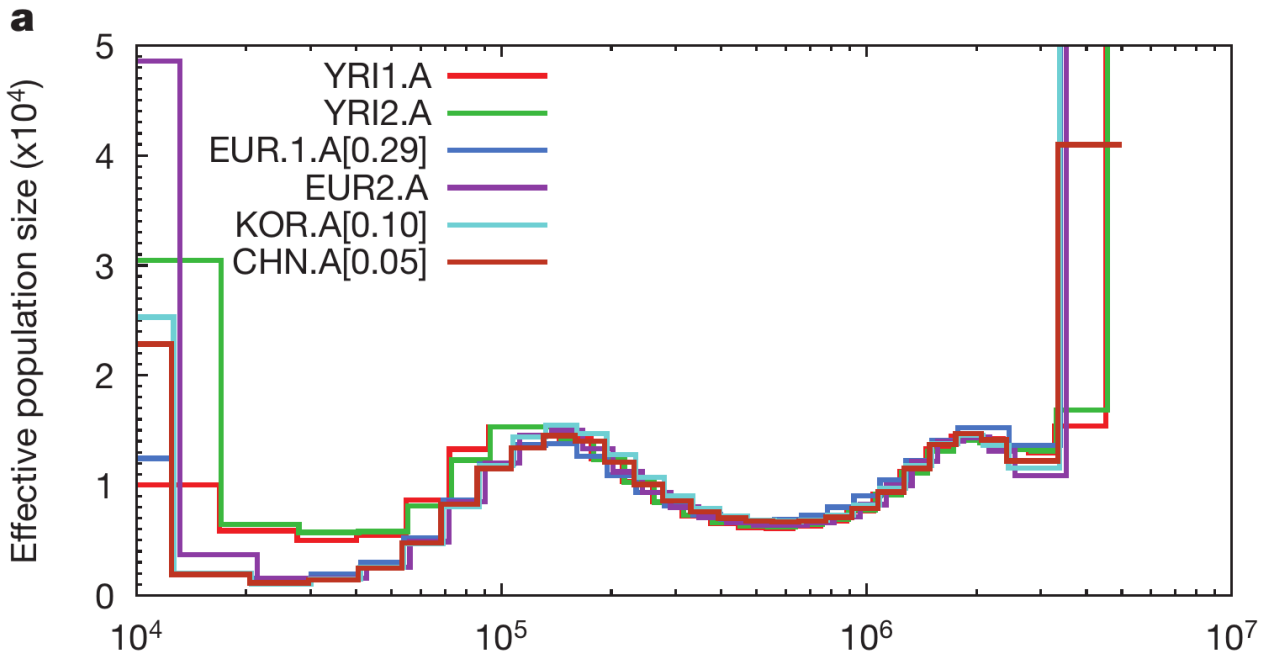
Coalescence: relatedness between individuals relates to population size



TRENDS in Ecology & Evolution

Inferring population size from diversity data





Li & Durbin 2011, Nature

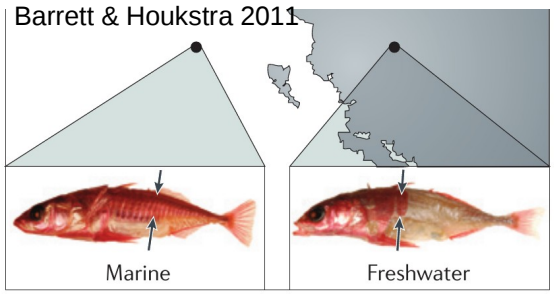
Population subdivision

- Differences in allele frequency between sub-populations
- We can use genetic data to identify boundaries between sub-populations

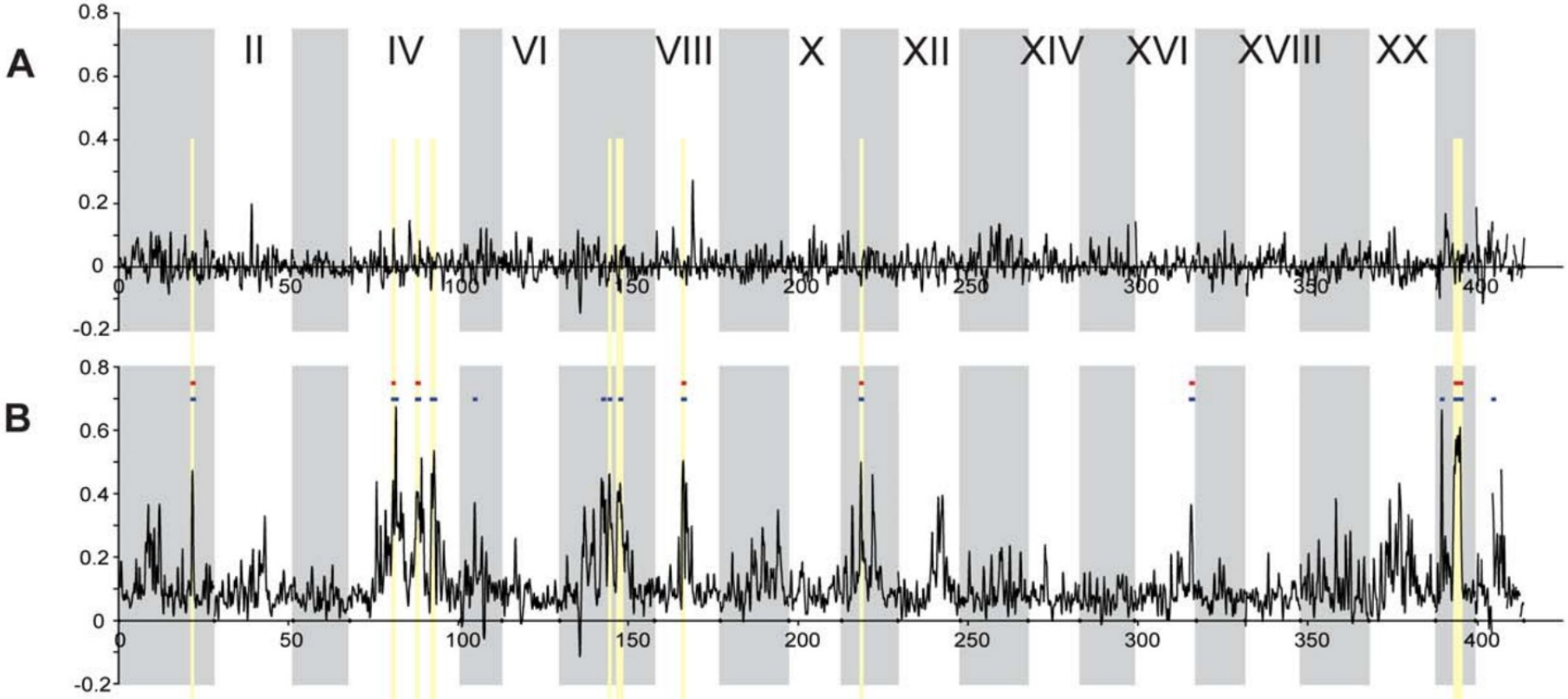
```
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGATCTCGTGCCTAAGTCGATTATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGGTCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTCGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATTATGATAAT

TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCTTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATACT
```

Population subdivision – a genome-wide view

- F_{ST} is a measure of differences in allele frequencies between populations



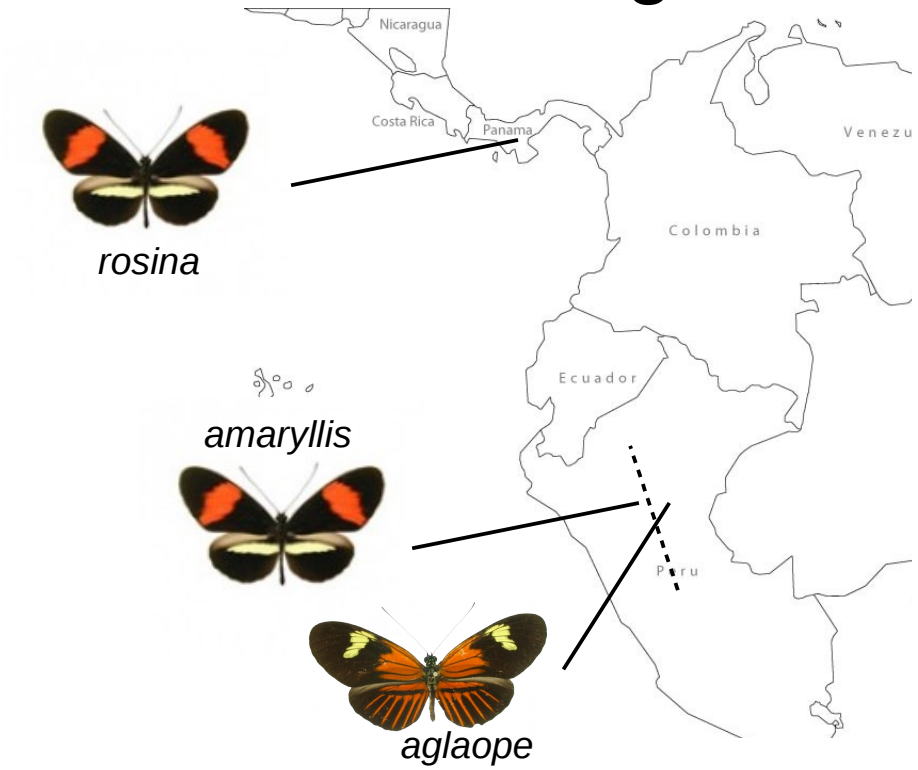
F_{ST} between two marine populations (A) or between marine and freshwater populations (B) (Hohenlohe et al. 2010)

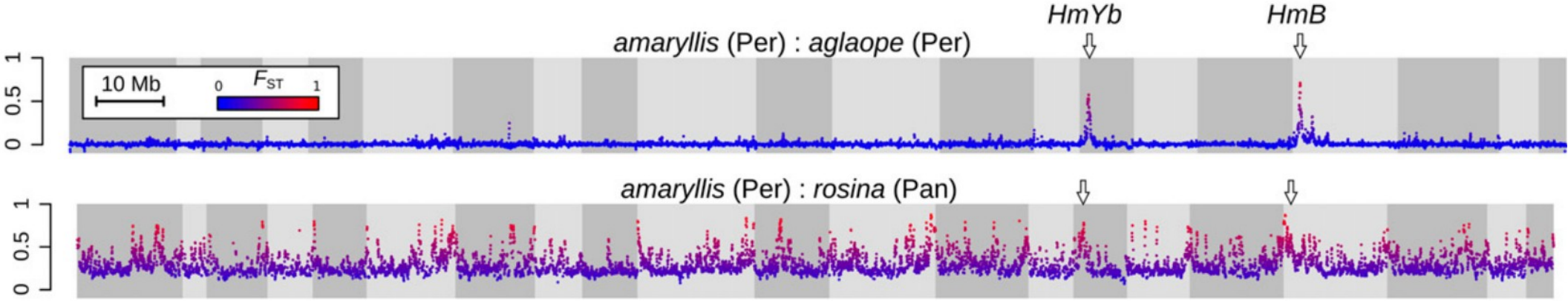


Population subdivision – a genome-wide view

Parapatric wing-pattern races of Heliconius melpomene are only strongly differentiated at two wing patterning loci.

(Martin et al. 2013 Genome Research)





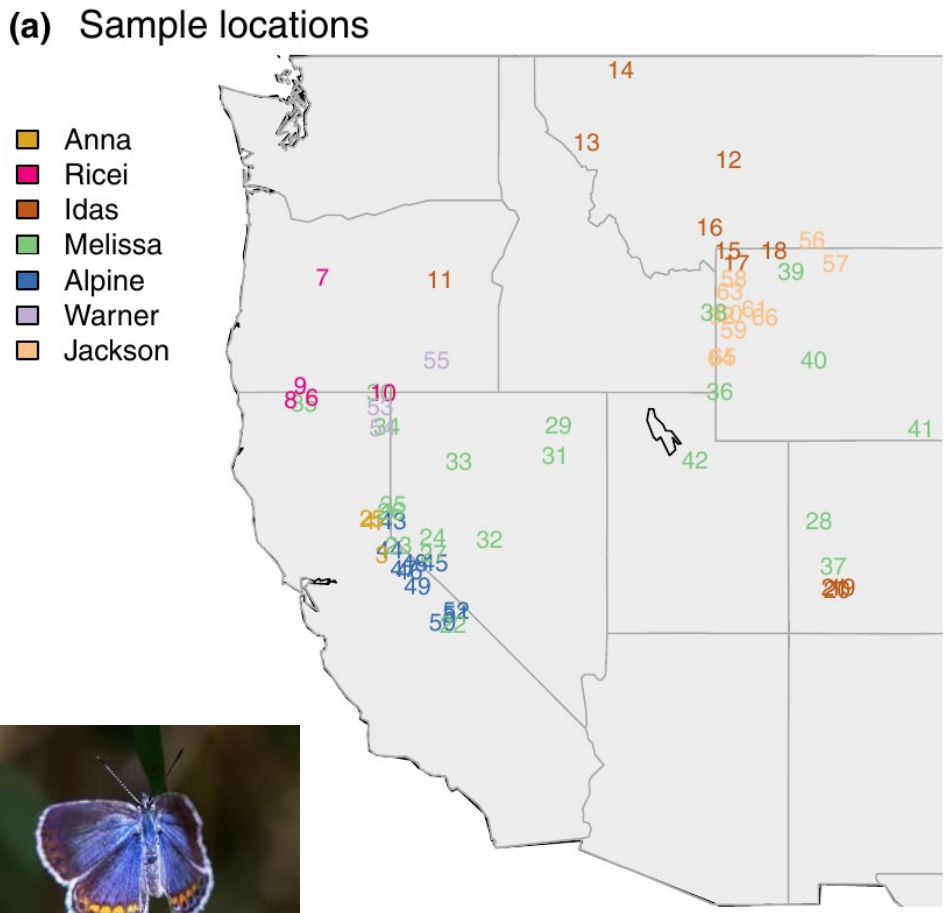
Population subdivision – identifying populations using genomic data

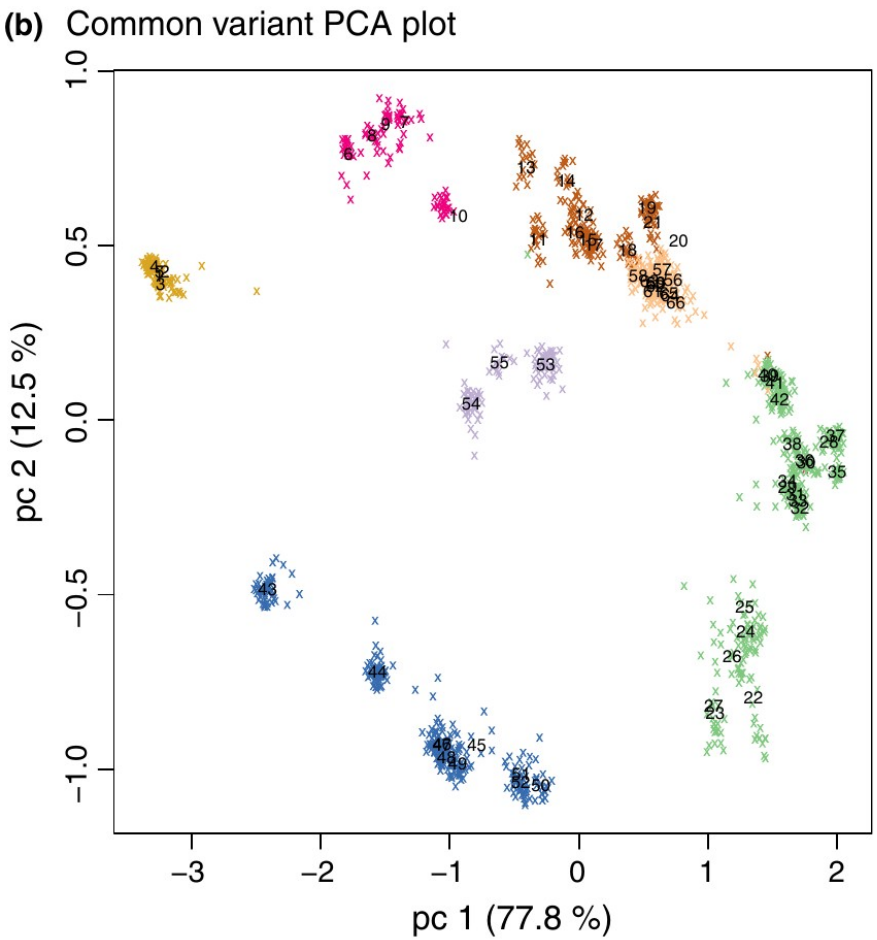
- Given a genetic dataset, can we infer whether there are distinct populations?

```
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGATCTCGTGCCTAAGTCGATTATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGGTCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTCGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCTTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATACT
```

Population subdivision – identifying populations using genomic data

- PCA (Principle Components Analysis) simplifies huge genetic datasets into two (or more) dimensions

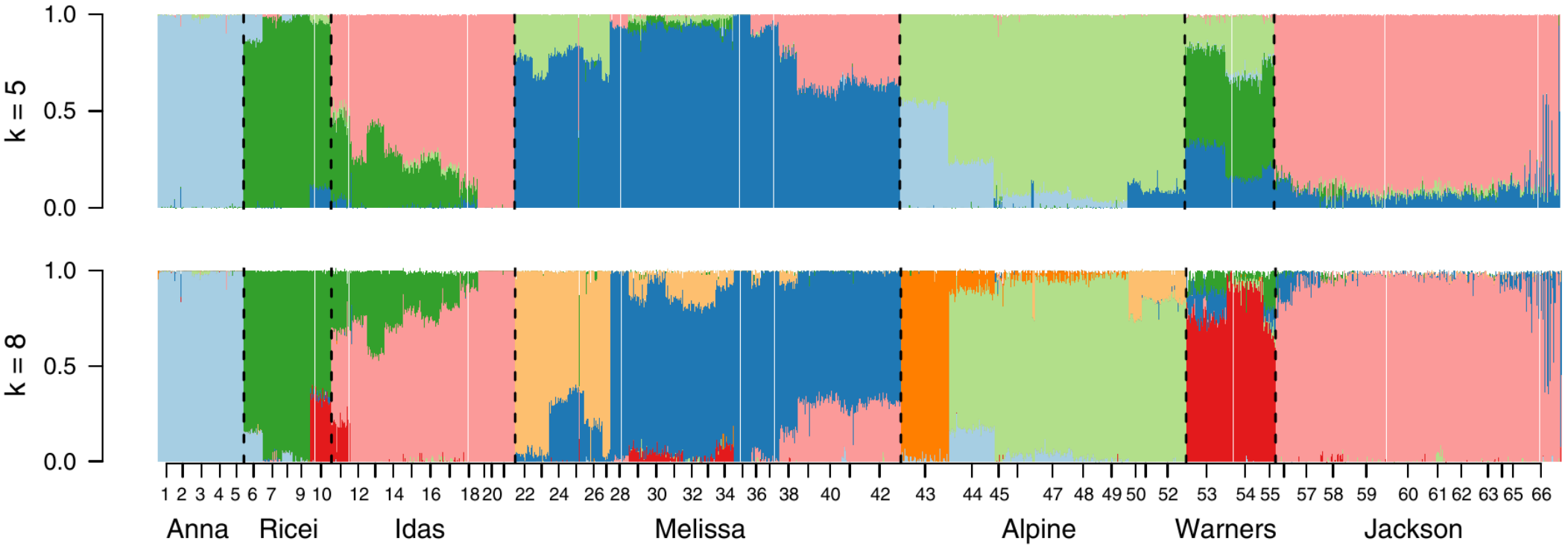




Gompert et al. 2014 Molecular Ecology

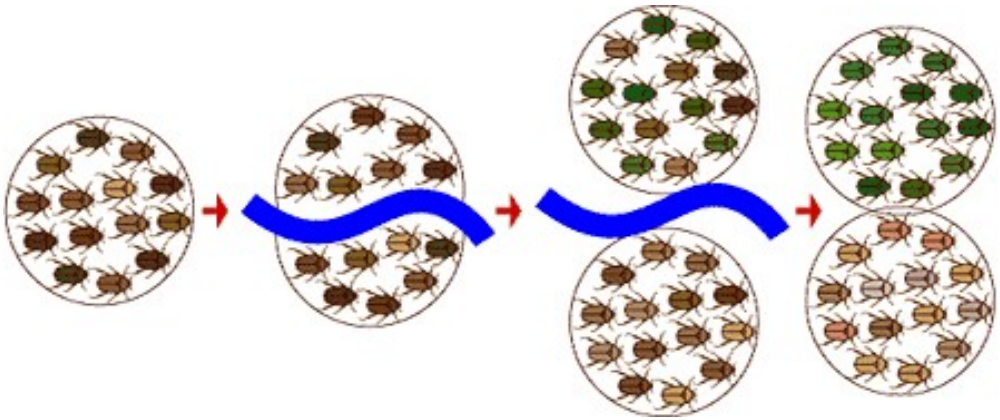
Population subdivision – identifying populations using genomic data

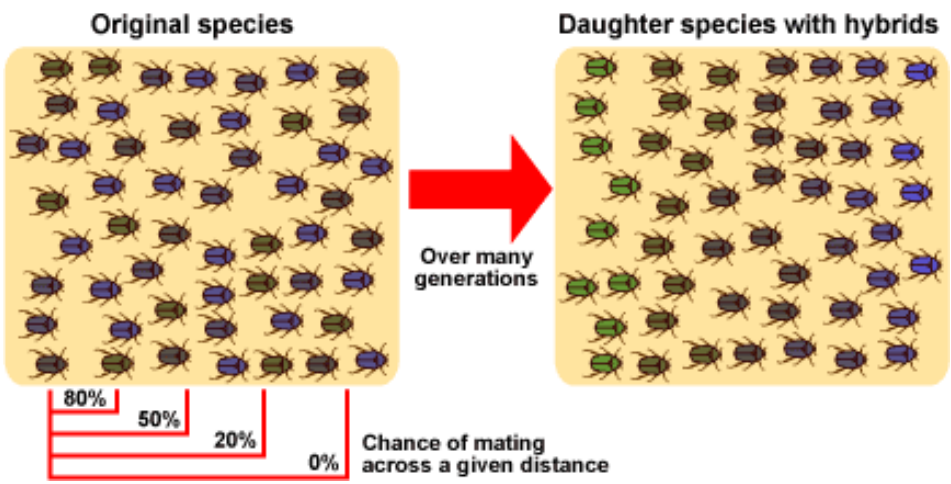
- STRUCTURE (Pritchard et al. 2000) estimates the likelihood that each sample falls into each of *k* clusters

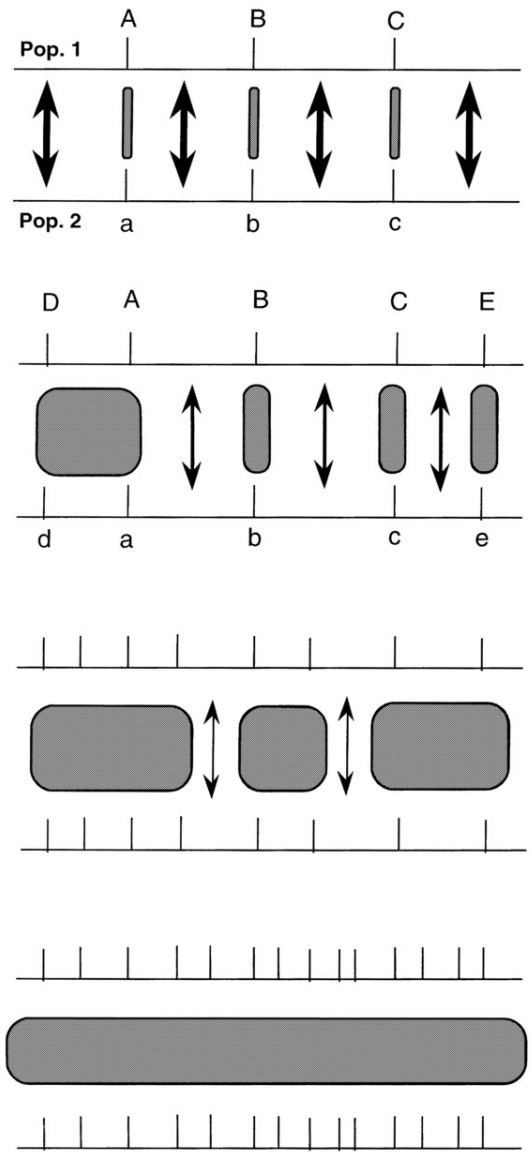


Gompert et al. 2014 Molecular Ecology

Speciation and gene flow



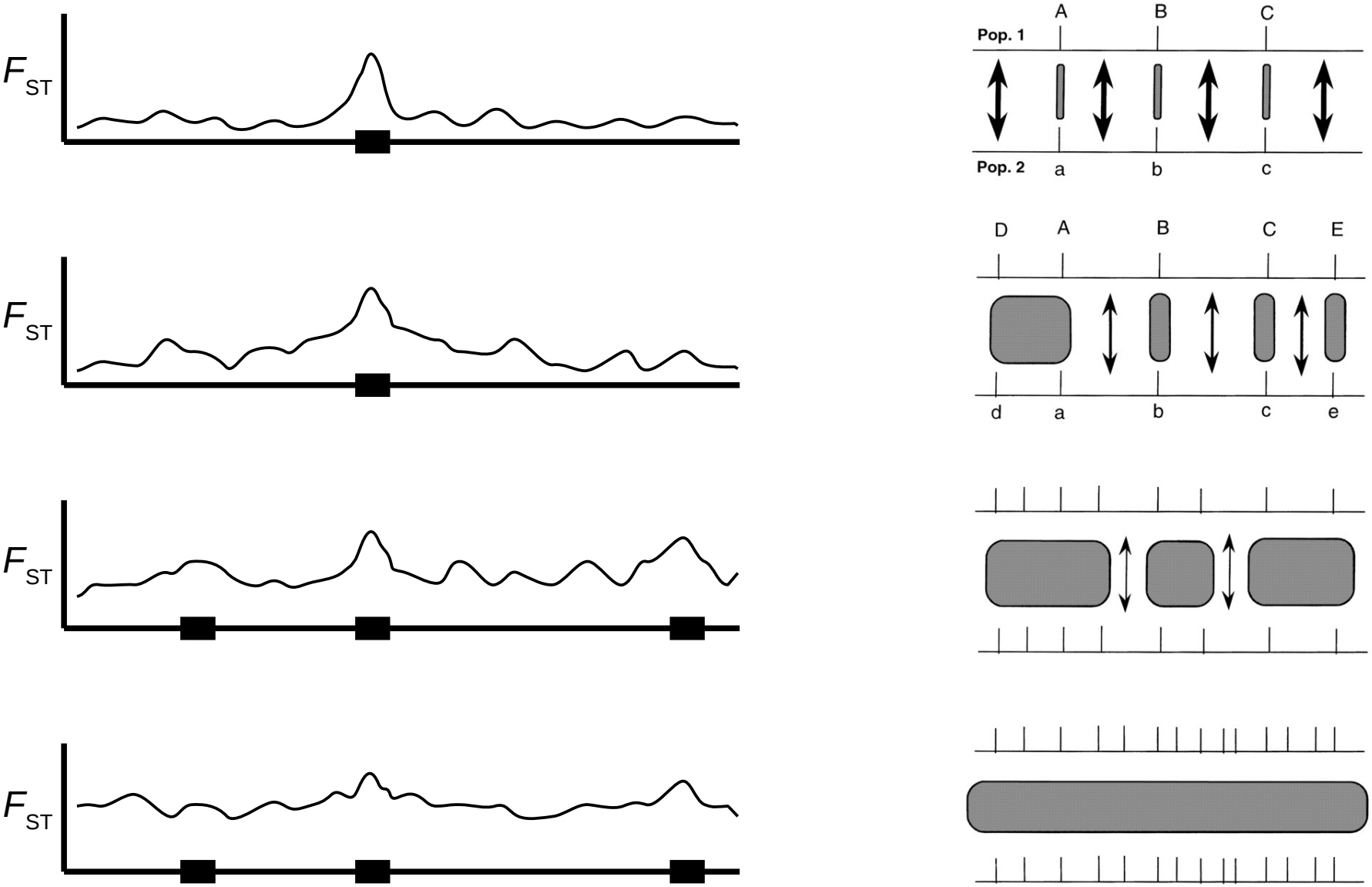


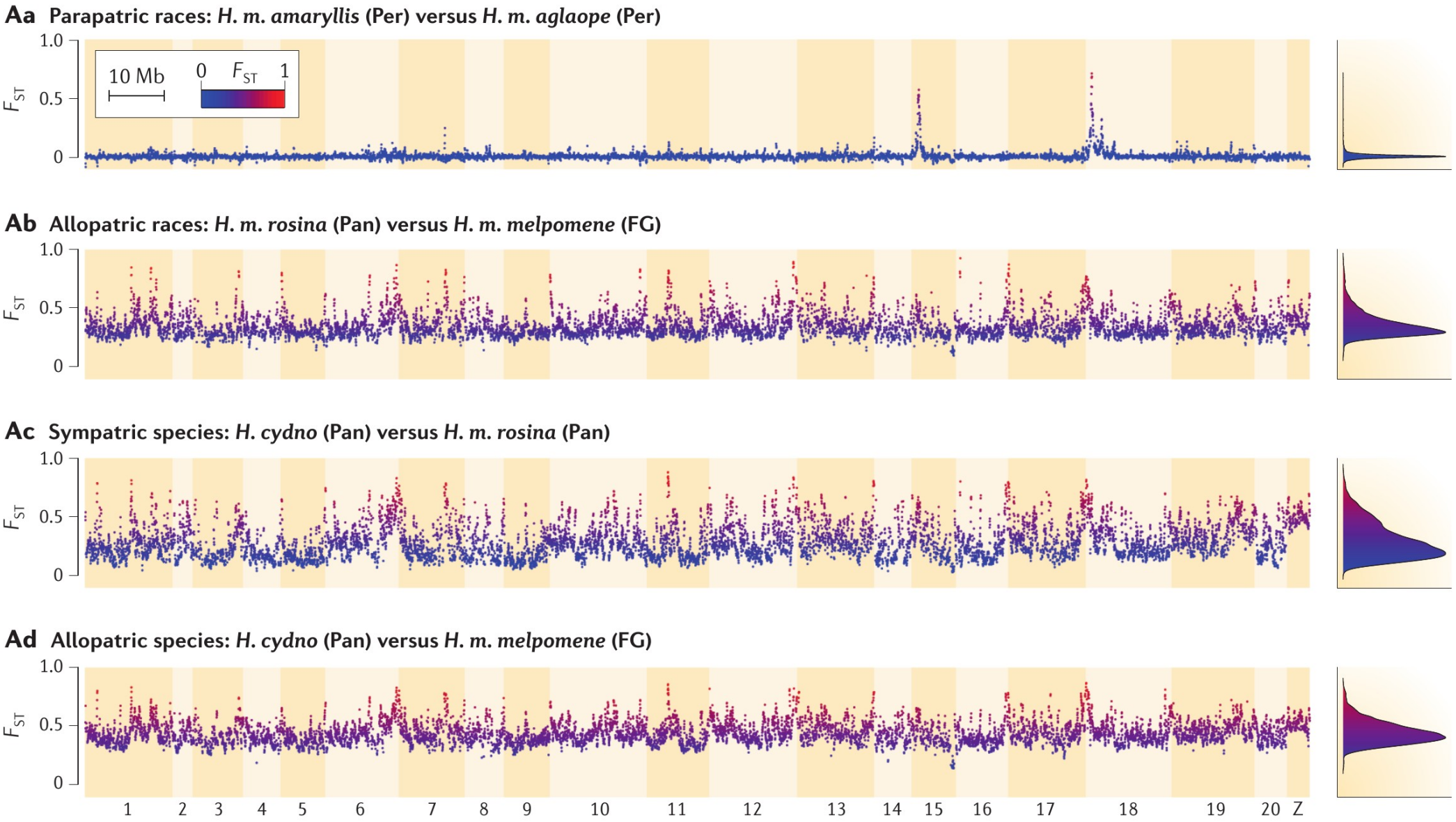


evolution.berkeley.edu

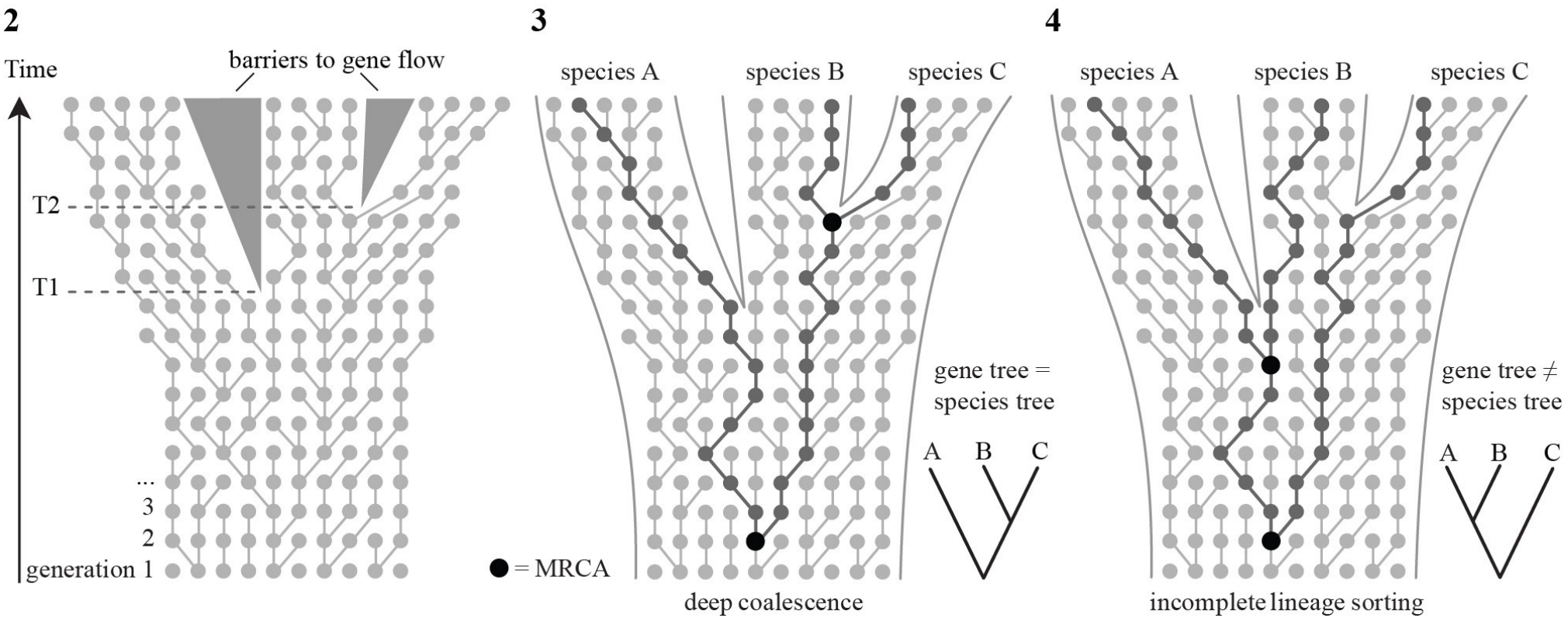
Wu 2001, JEB

Speciation and gene flow




Aa Parapatric races: H. m. amaryllis (Per) versus H. m. aglaope (Per)
Ab Allopatric races: H. m. rosina (Pan) versus H. m. melpomene (FG)
Ac Sympatric species: H. cydno (Pan) versus H. m. rosina (Pan)
Ad Allopatric species: H. cydno (Pan) versus H. m. melpomene (FG)
10 Mb
0 FST 1
FST
1.0
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Z

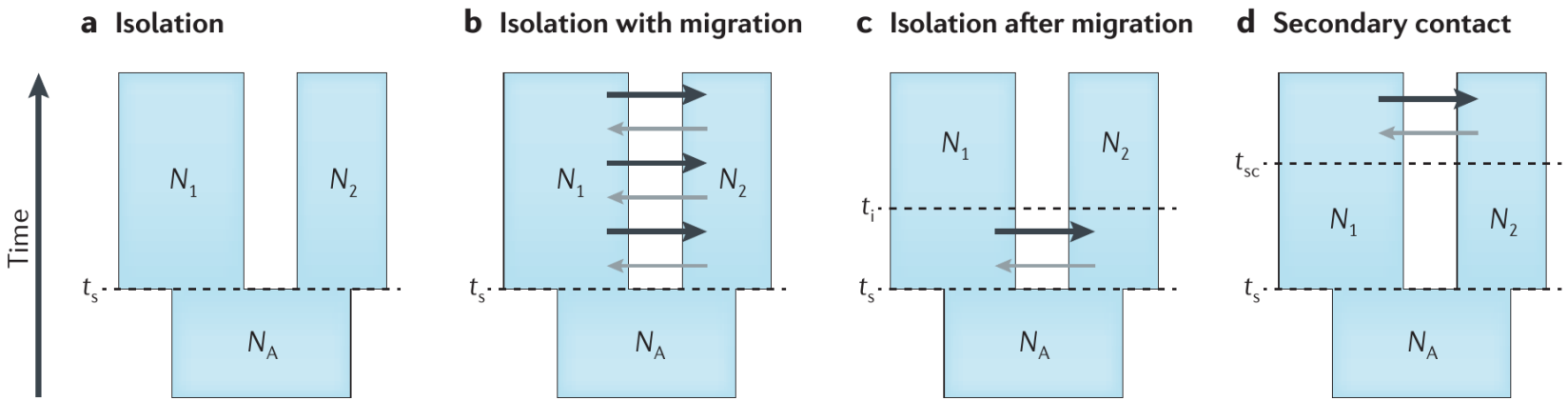
Detecting gene flow



Detecting gene flow

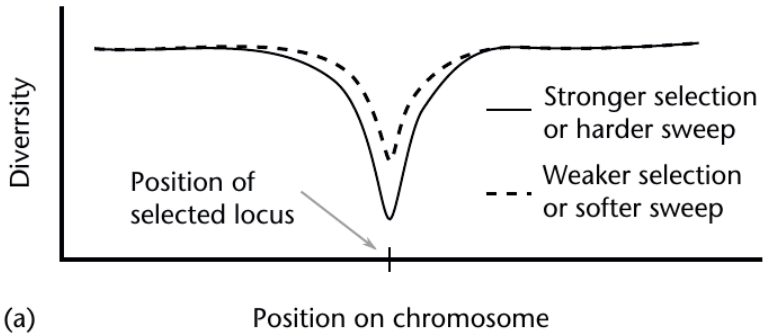
Huge genomic data sets allow sensitive fitting of evolutionary models.

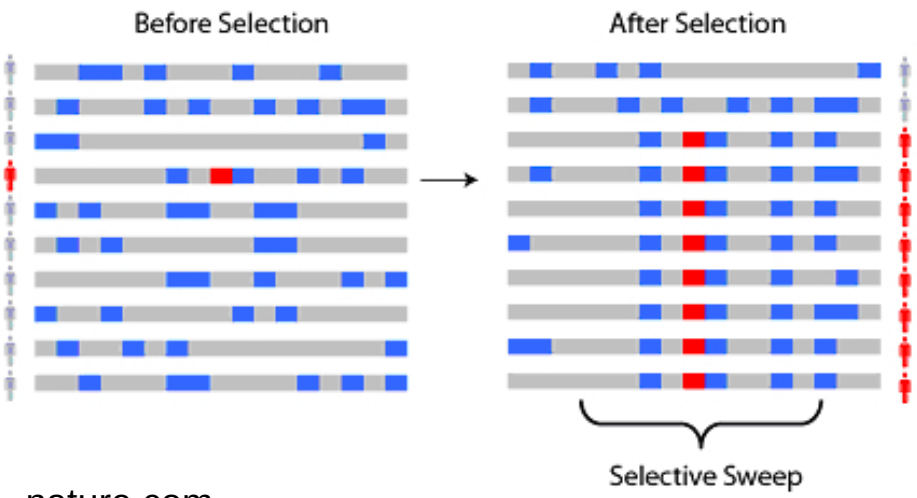
```
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGATCTCGTGCCTAAGTCGATTATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGATCGTCCAGATCGAACTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGGTCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCCTAAGTCGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCCTTTCGCTGAGCTCGTGCTTAAGTCGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCCAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCTTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGATCTCGTGCCTAAGTAGATCATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATAAT
TAGATCGTCGAGATCGATCTAGCCCGTTTCGCTGAGCTCGTGCCTAAGTAGATTATGATACT
```



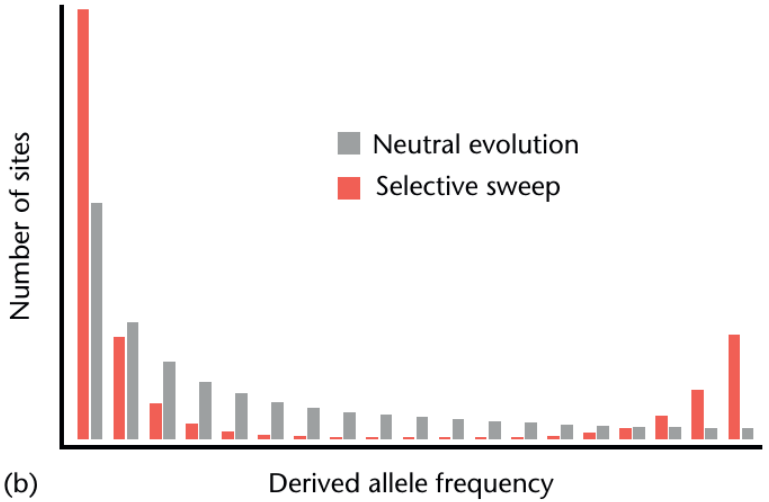
Selective sweeps

- Strong selection rapidly fixes a benefical allele in the population
- Causes reduced diversity at the selected site and a change in the site frequency spectrum



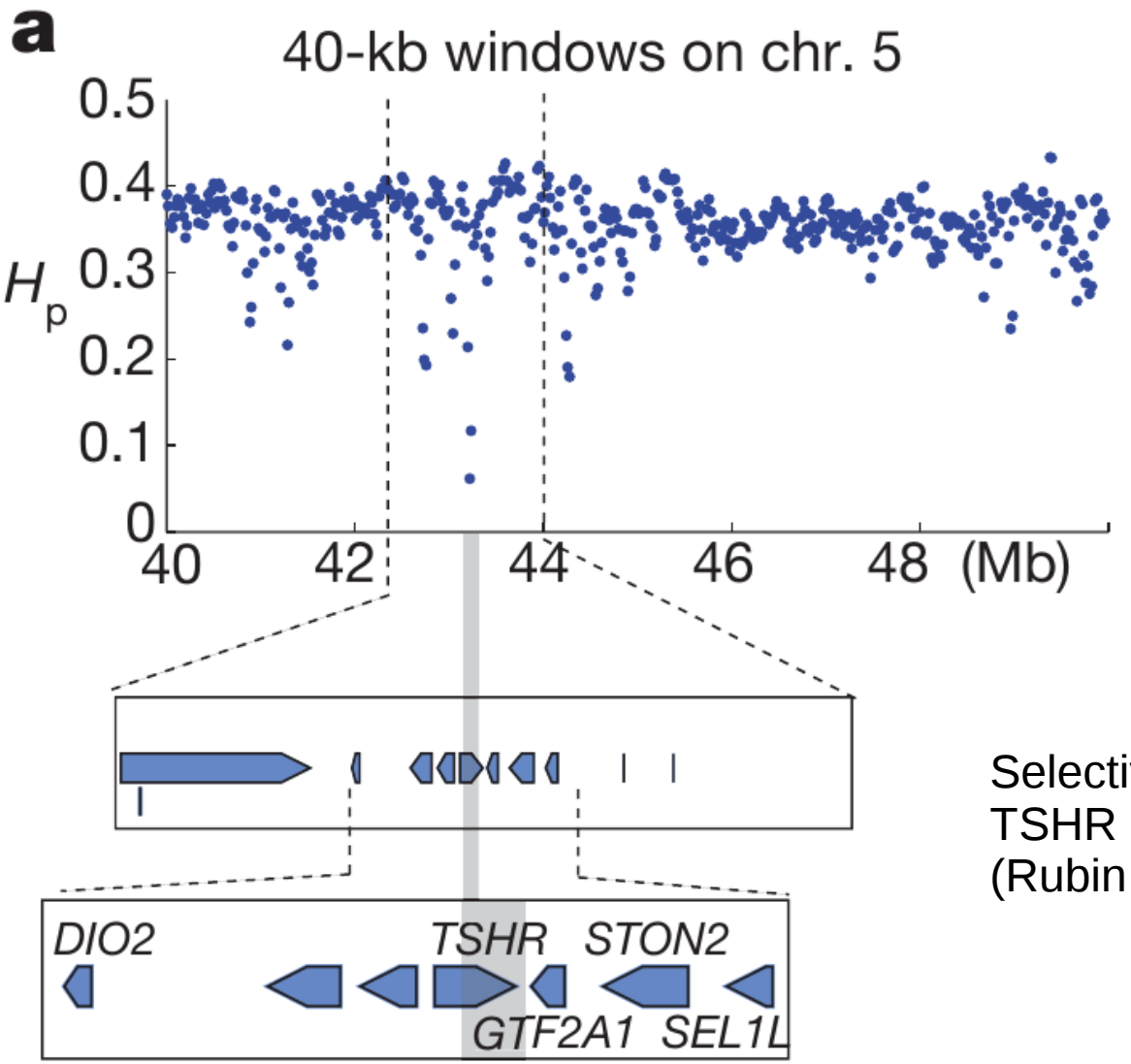


nature.com



Martin et al. 2013 eLS

Selective sweeps



Selective sweep at the TSHR gene in chickens. (Rubin et al. 2012 Nature)

Selective sweeps

(c) Bugguide.net

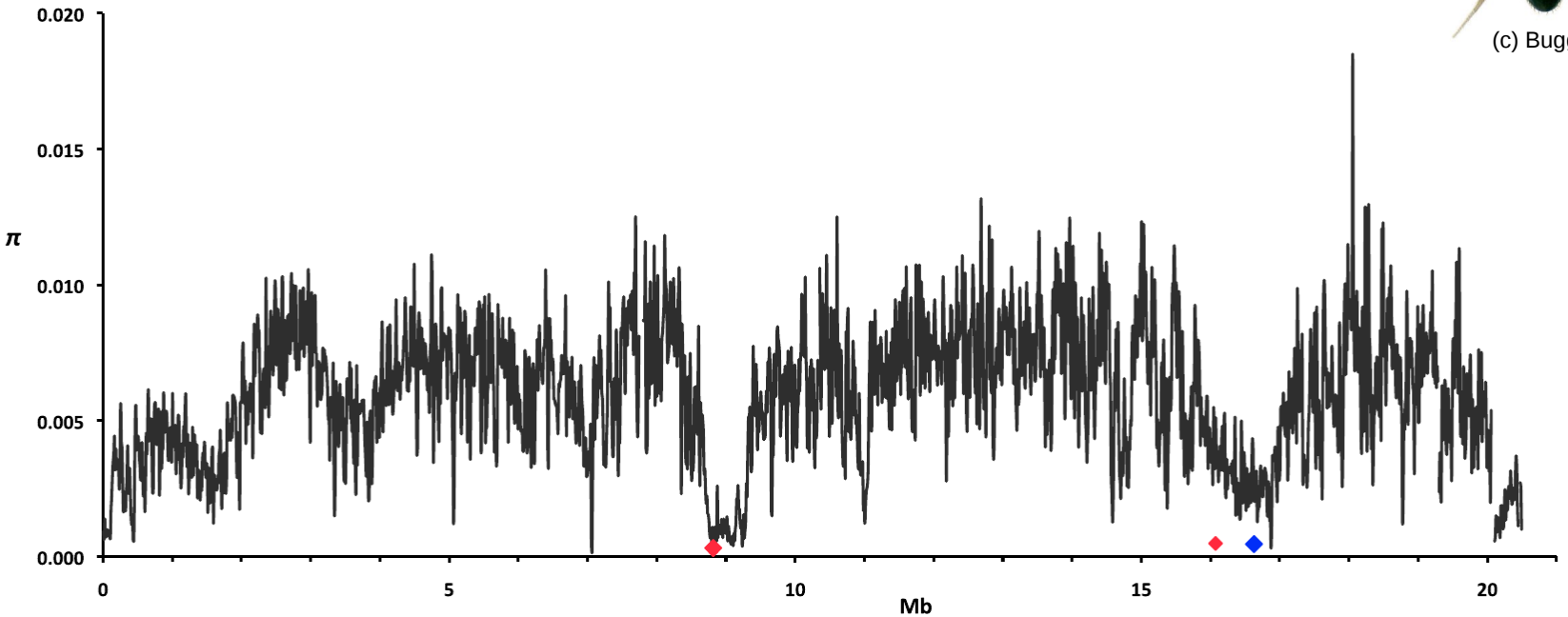


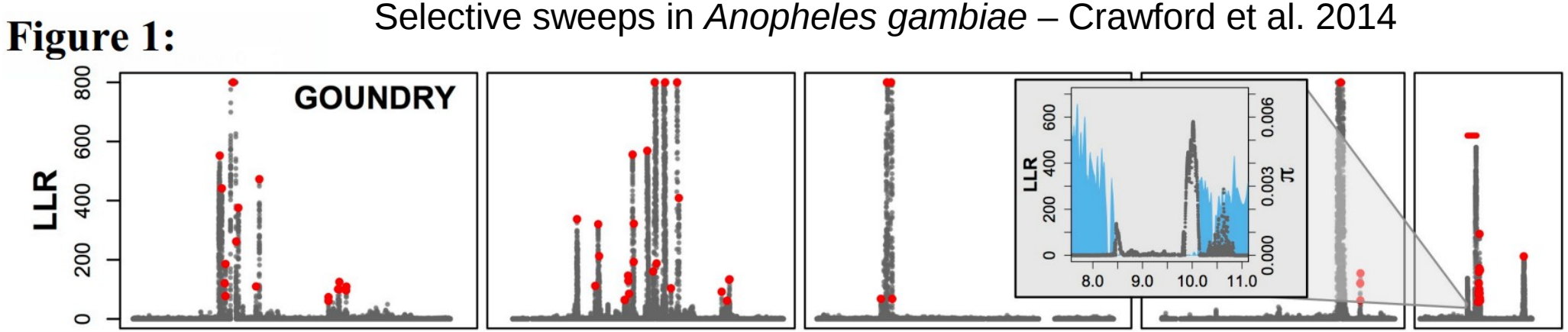
Figure 5. Nucleotide diversity (π) along the *D. mauritiana* X chromosome. The location of genes potentially causing the two selective sweeps are indicated: (large red diamond) *MDox/Dox*; (large blue diamond) *OdsH*; (small red diamond) *E(Dox)*. Nucleotide diversity (π) is plotted in nonoverlapping 10-kb windows.

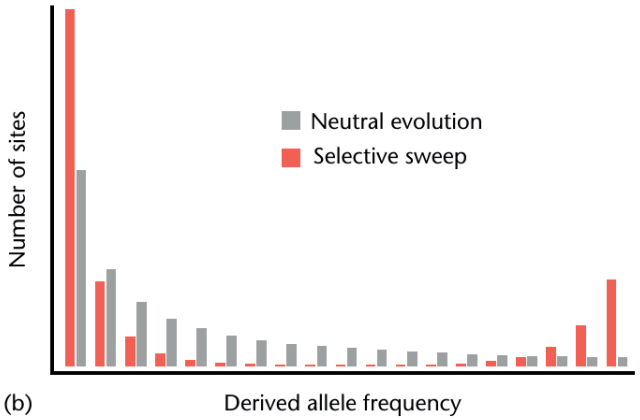
Nolte et al. 2013 Genome Research

Selective sweeps

- **Sweepfinder** (Nielsen et al. 2005) identifies regions showing a strong skew in the site frequency spectrum.

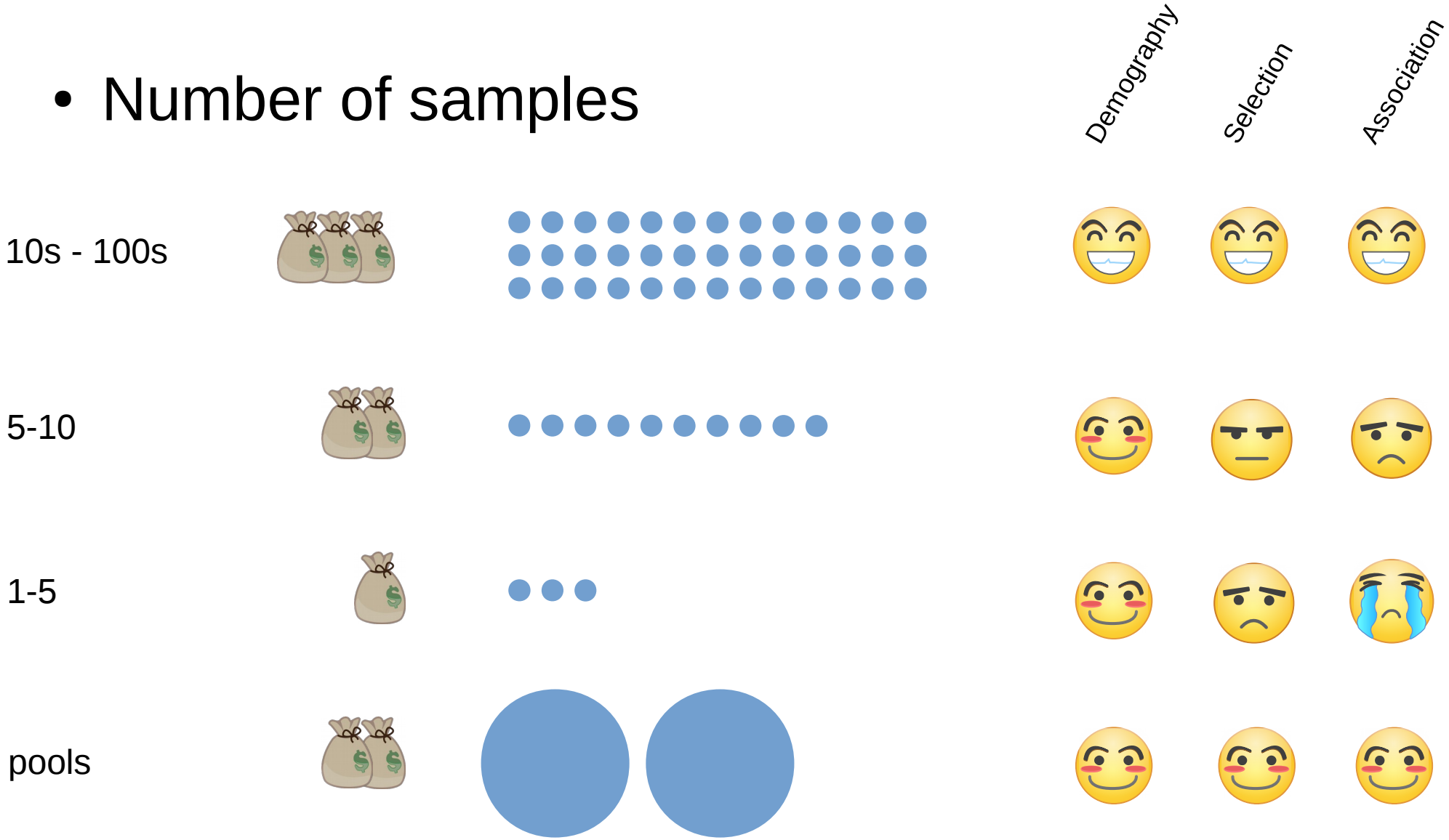
Selective sweeps in *Anopheles gambiae* – Crawford et al. 2014





Designing a population genomic study

- Number of samples



Designing a population genomic study

- Type of sequencing