Inferring population history

Feb 13th 2019

Rasmus Heller

Session overview

9.15-10.00: Theory behind population history inference.

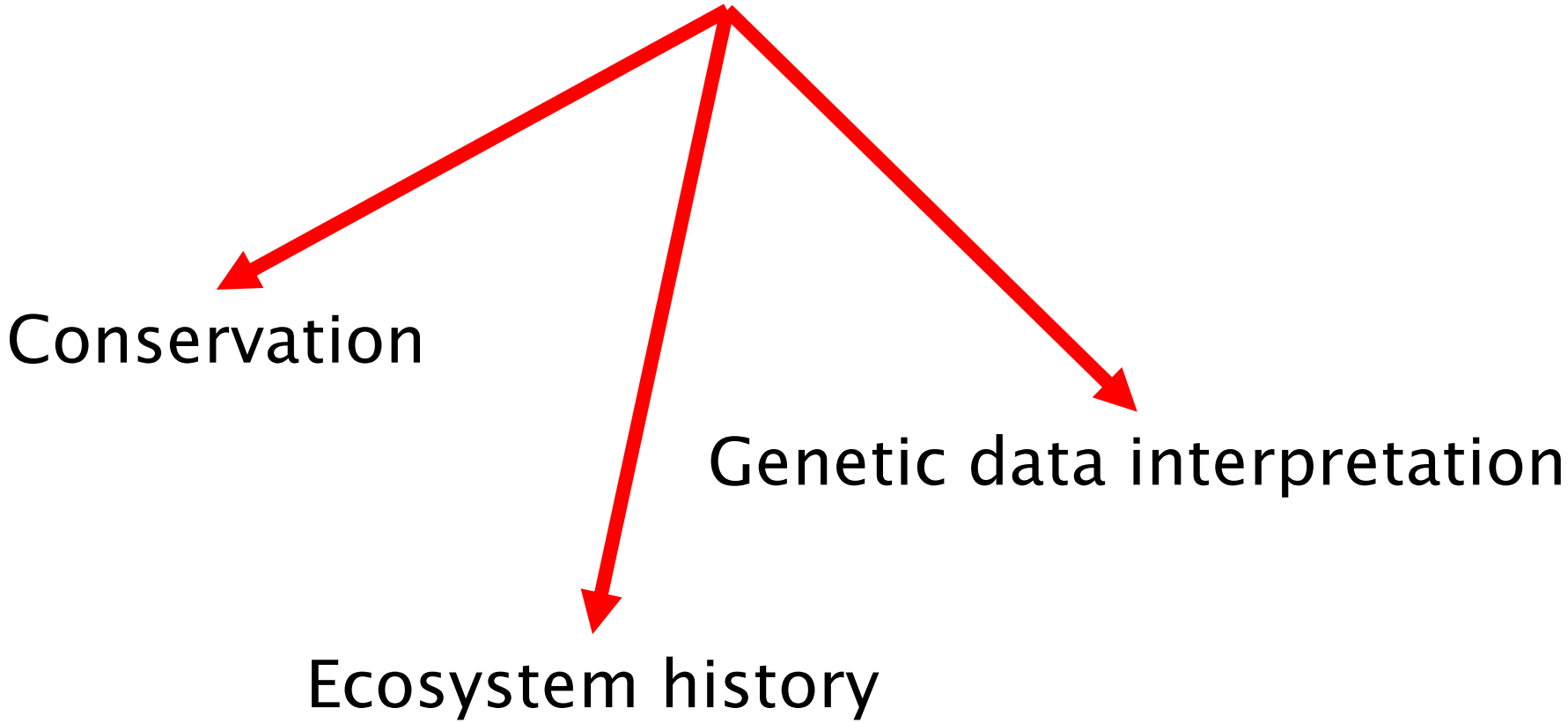
---BREAK---

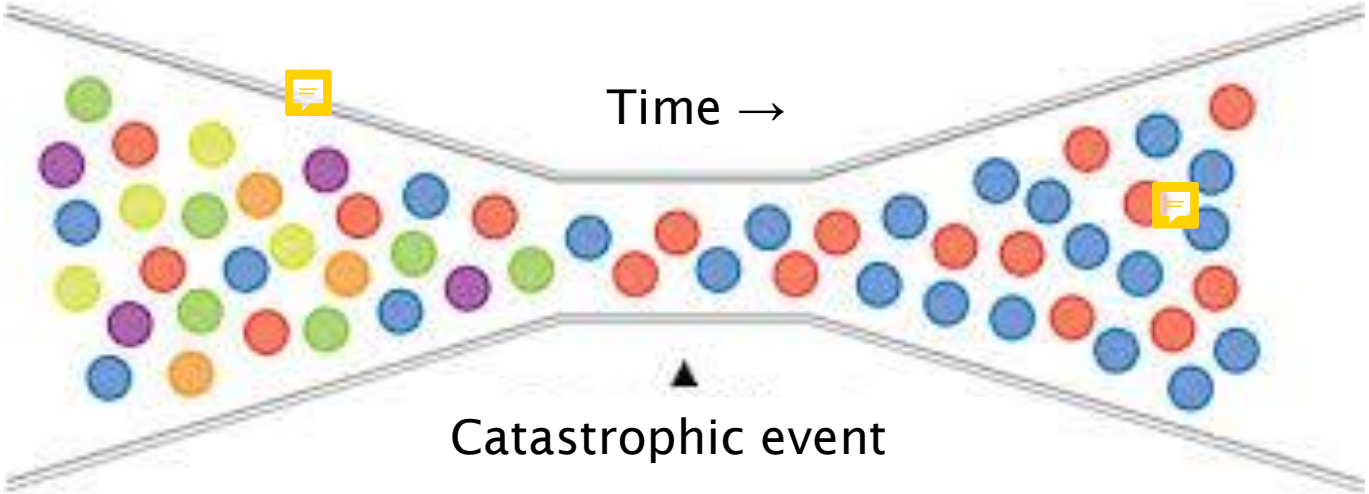
10.15-12.00: Simulations and population history lab.

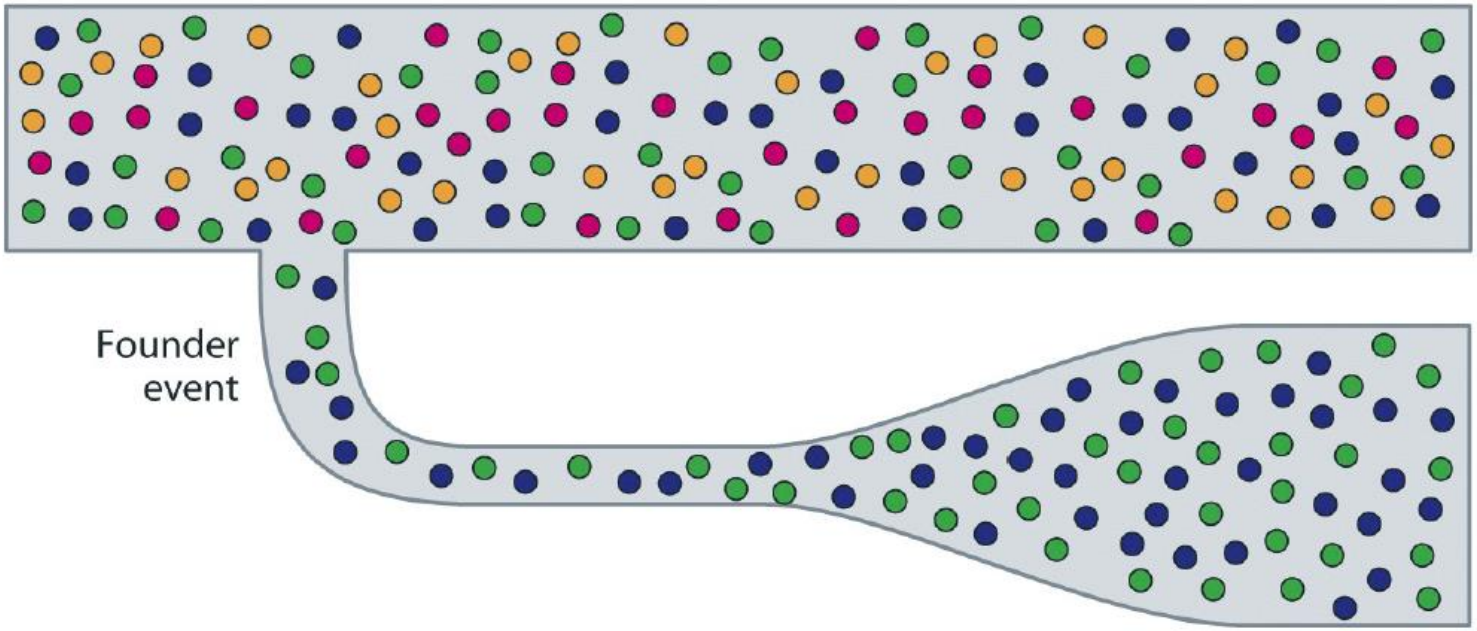
---LUNCH BREAK---

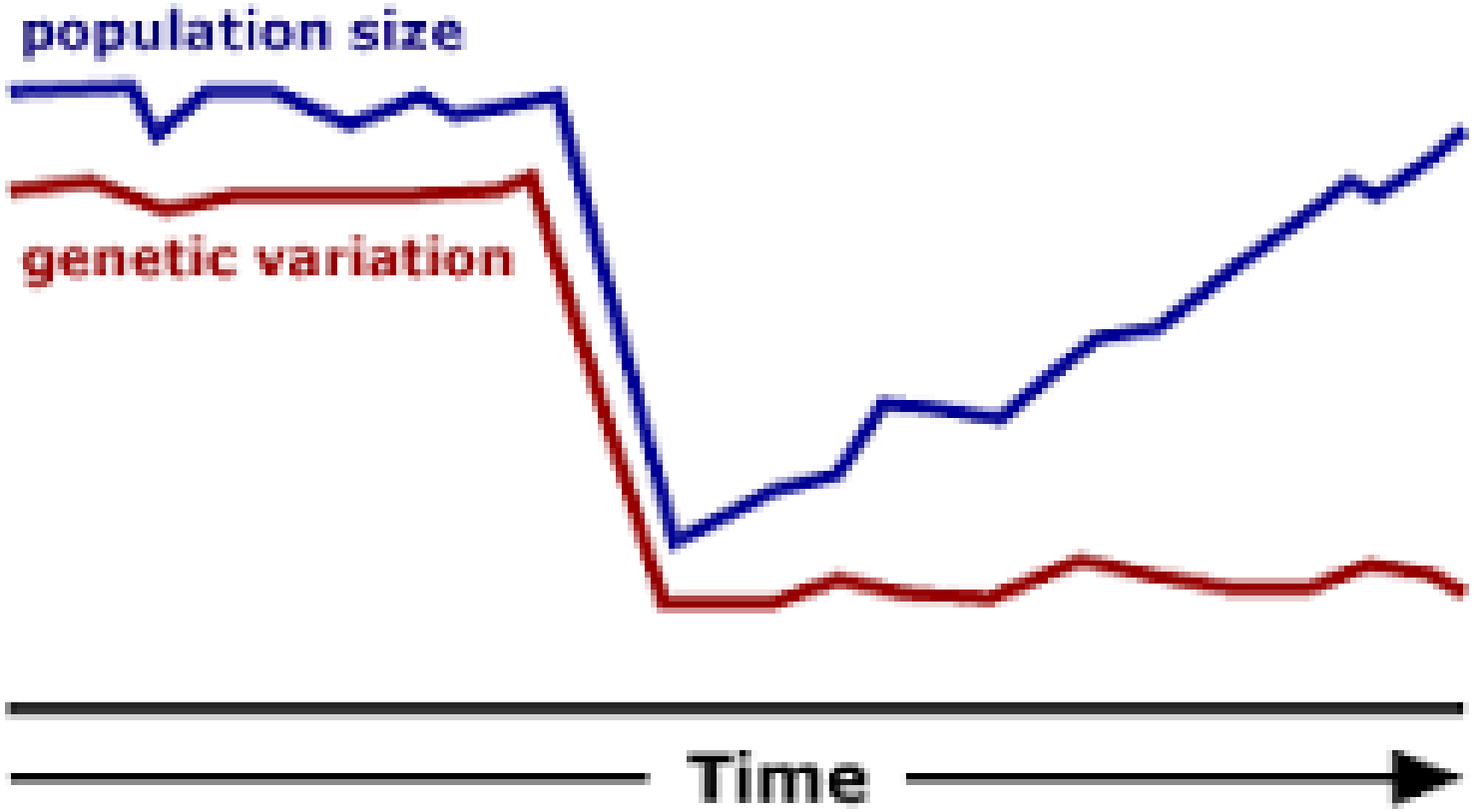
13.15-14.00: Effective population size introduction.

Population history

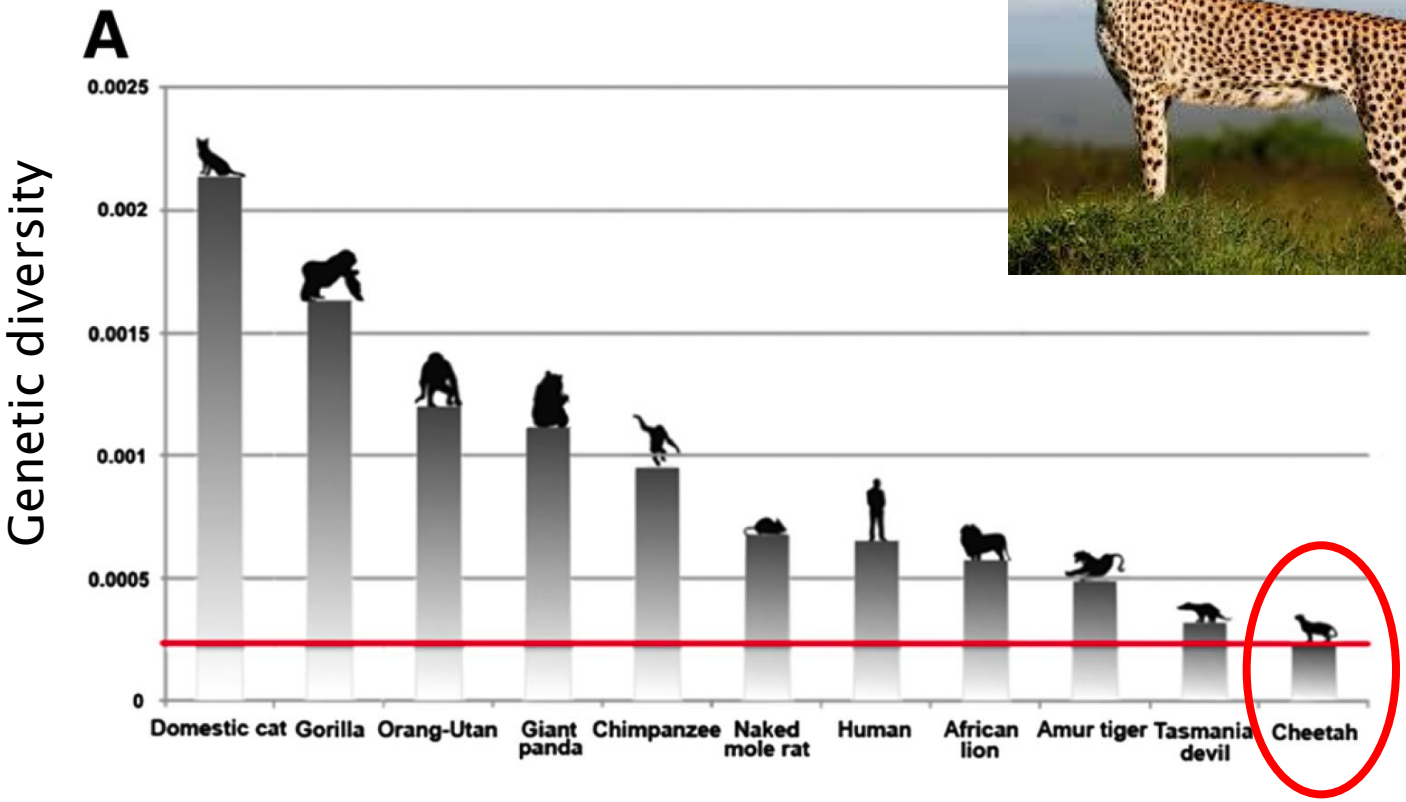



Time →
Catastrophic event


Founder
event

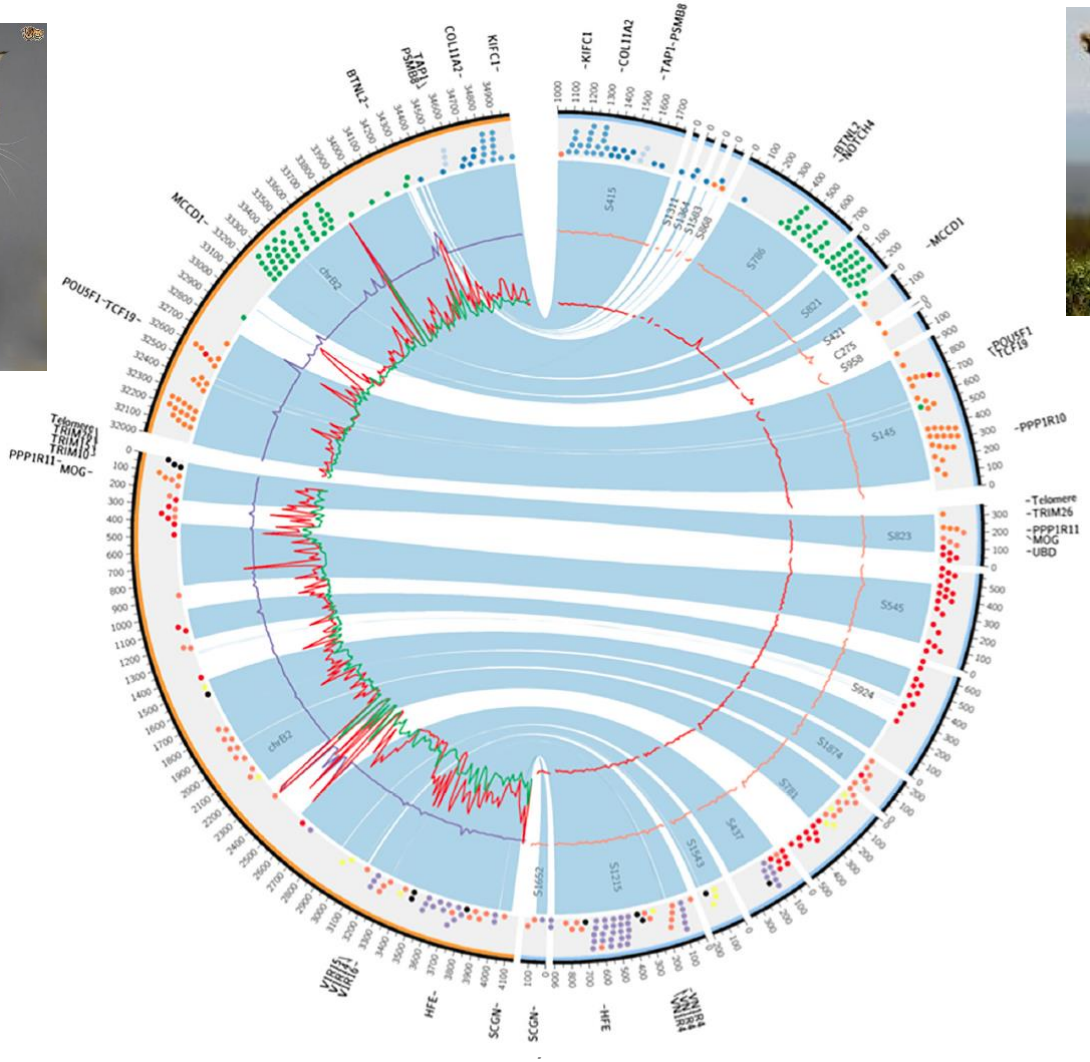

population size
genetic variation
Time

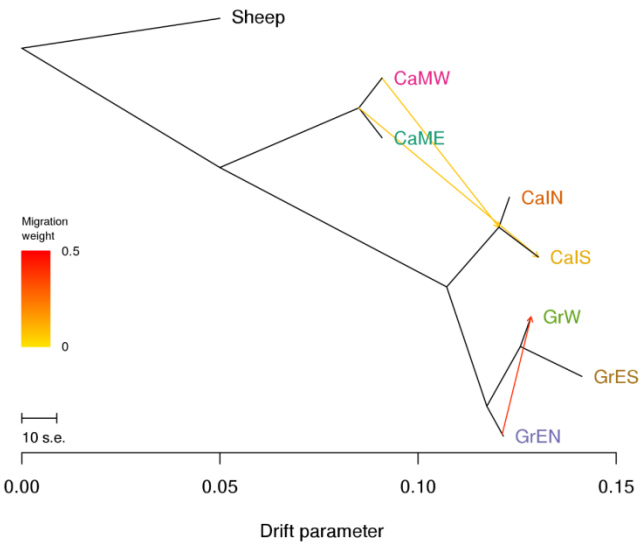
A classic example

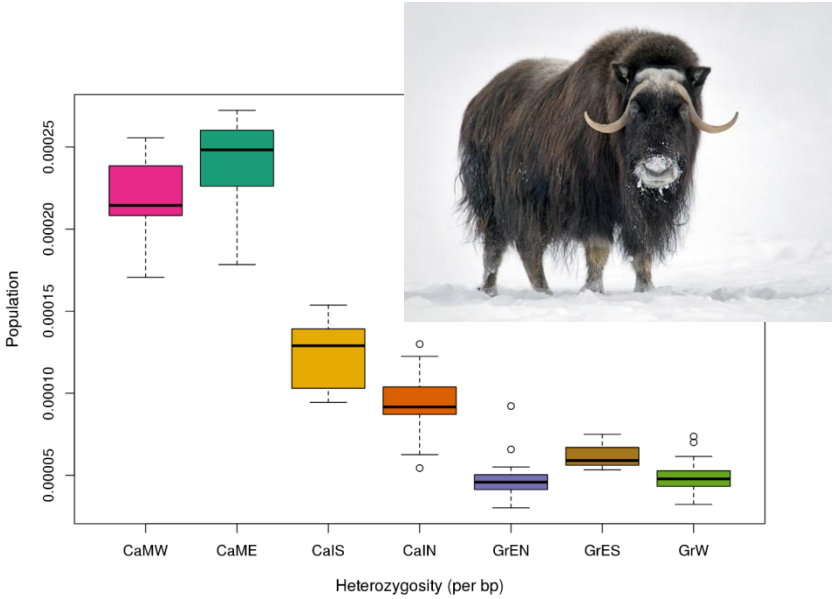


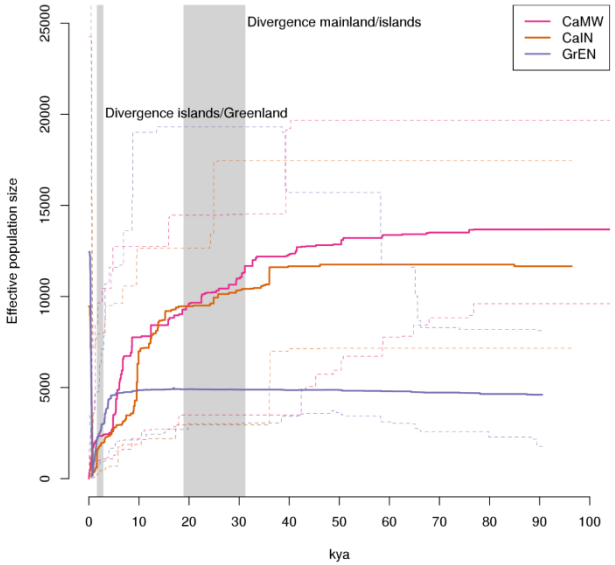
Dobrynin et al. 2015

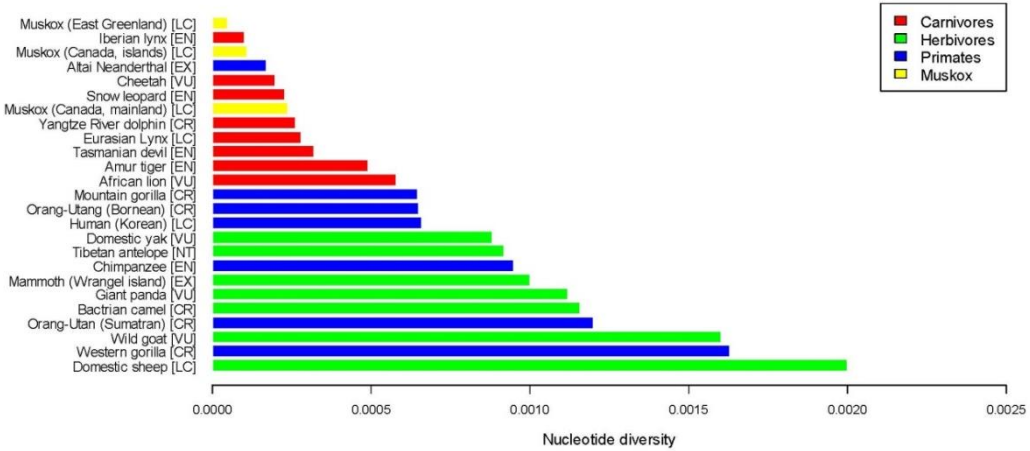
MHC genes diversity



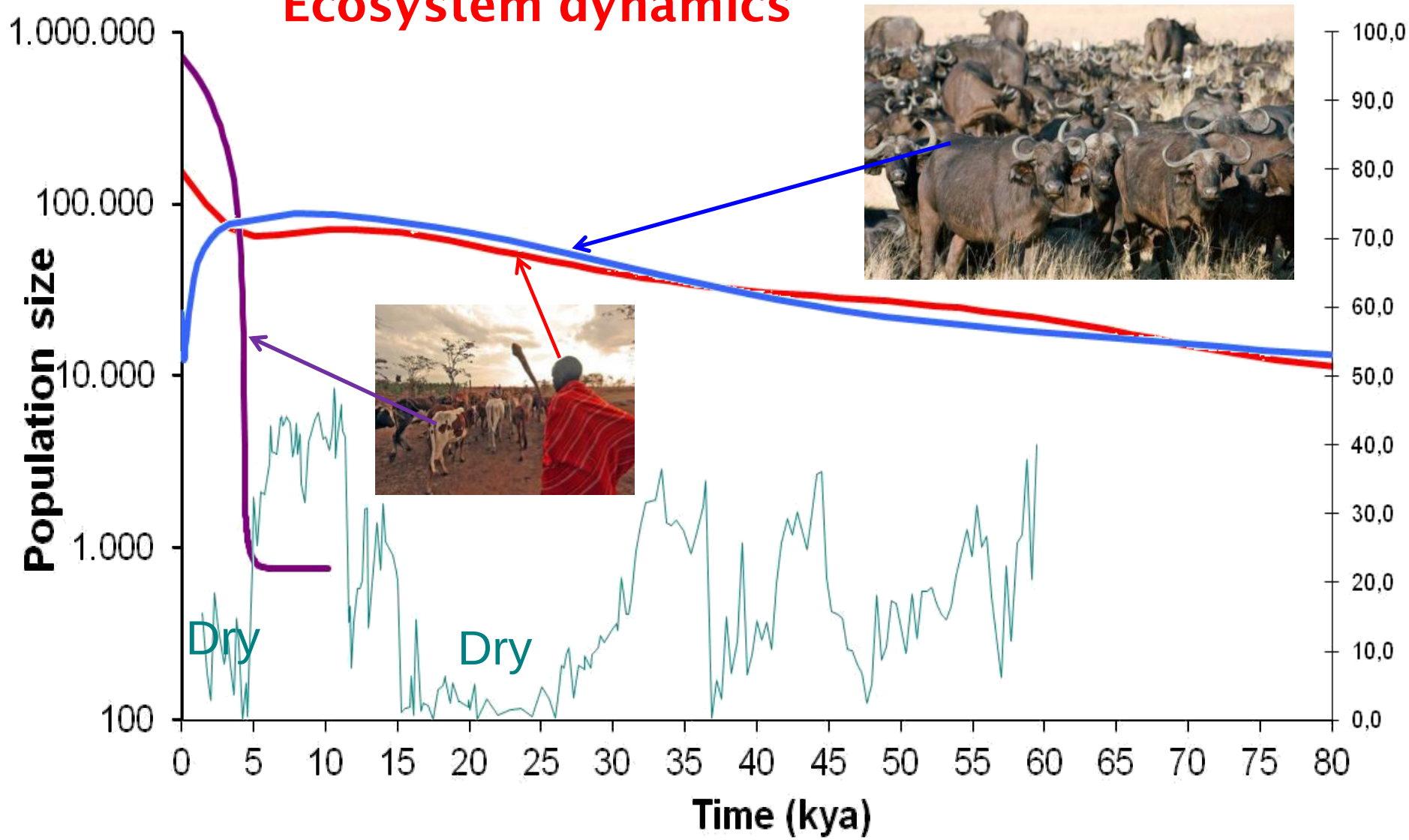

Sheep
CaMW
CaME
CaIN
CaIS
GrW
GrES
GrEN
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
Drift parameter


Population
0.00005
0.00010
0.00015
0.00020
0.00025
CaMW
CaME
CaIS
CaIN
GrEN
GrES
GrW
Heterozygosity (per bp)


Divergence mainland/islands
Divergence islands/Greenland
CaMW
CaIN
GrEN
Effective population size
0
5000
10000
15000
20000
25000
0
10
20
30
40
50
60
70
80
90
100
kya

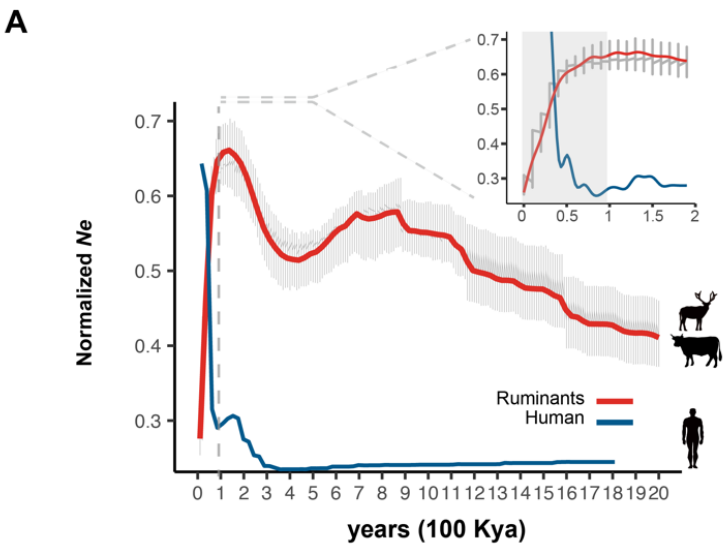

Muskox (East Greenland) [LC]
Iberian lynx [EN]
Muskox (Canada, islands) [LC]
Altai Neanderthal [EX]
Cheetah [VU]
Snow leopard [EN]
Muskox (Canada, mainland) [LC]
Yangtze River dolphin [CR]
Eurasian Lynx [LC]
Tasmanian devil [EN]
Amur tiger [EN]
African lion [VU]
Mountain gorilla [CR]
Orang-Utang (Bornean) [CR]
Human (Korean) [LC]
Domestic yak [VU]
Tibetan antelope [NT]
Chimpanzee [EN]
Mammoth (Wrangel island) [EX]
Giant panda [VU]
Bactrian camel [CR]
Orang-Utan (Sumatran) [CR]
Wild goat [VU]
Western gorilla [CR]
Domestic sheep [LC]
Carnivores
Herbivores
Primates
Muskox
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
Nucleotide diversity

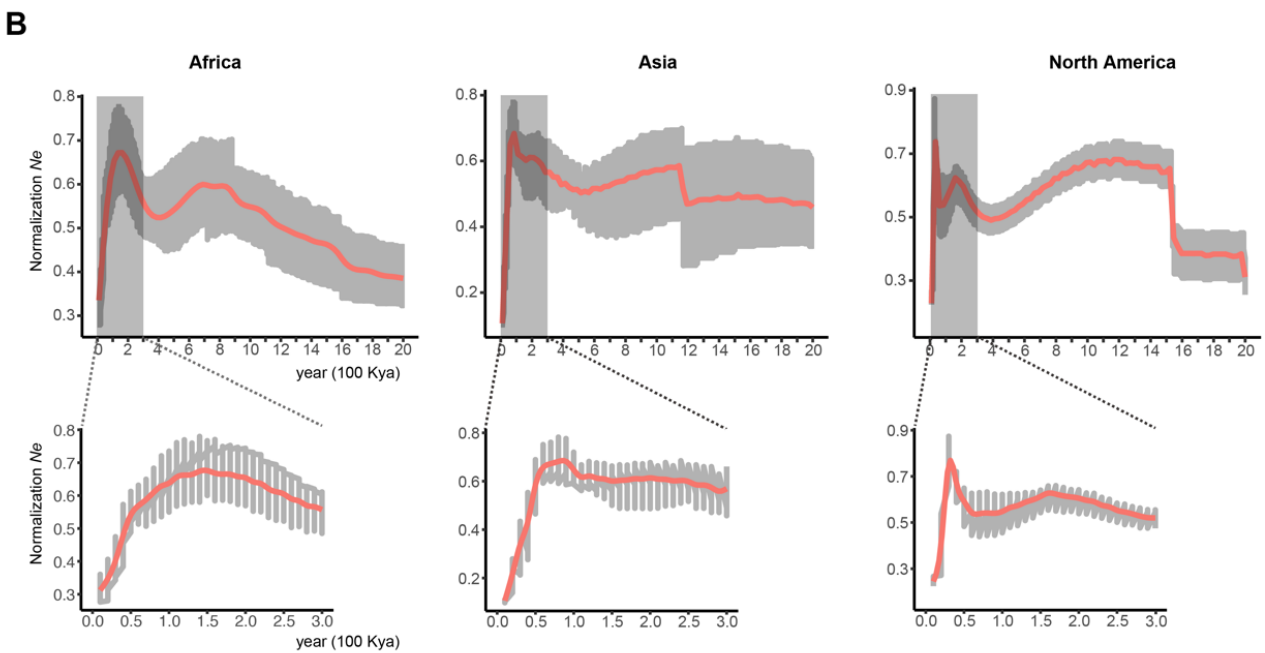
Ecosystem dynamics



Heller, Brüniche-Olsen & Siegismund (2012)

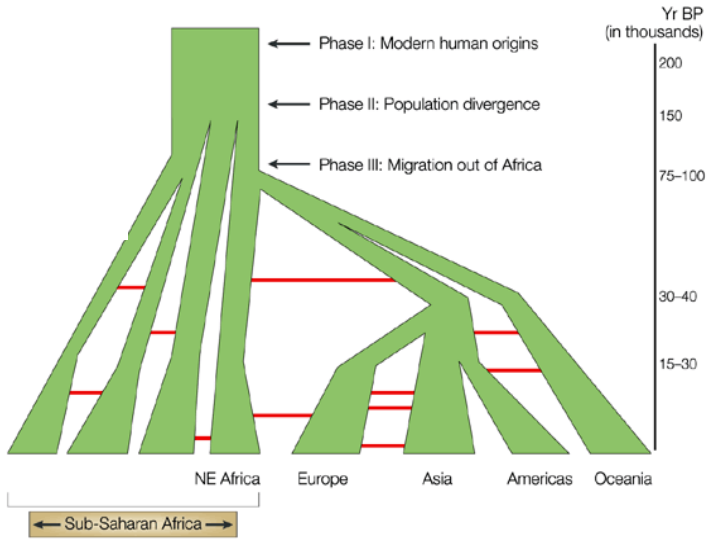
Human influence on wild species





Different approaches

How do we get from data to population history?



Different approaches

How do we get from data to population history?

- **Simple measures:** θ, Tajima's D, site frequency spectrum etc.
- **Tree inference**: estimate history from coalescent trees.
- **Coalescent simulations:** Simulation lab.
- **Maximum likelihood/Bayesian methods**

All of these rely on the coalescent

Simple measures

What do they tell us about population history?

Statistics to estimate *N*

Two simple statistics:

π = average pairwise differences

S = segregating sites

We defined $\theta = 4\boldsymbol{N}\mu$.

Two commonly used estimates of θ:

Tajimas: $\hat{\theta}_T = \pi$

Wattersons: $\hat{\theta}_w = \frac{S}{\sum_{k=1}^{n-1}\frac{1}{k}}$

Segregating sites = 3

Sequence 1	AATGTCAACG
Sequence 2	AATGTCAACG
Sequence 3	ATTGTCAACG
Sequence 4	ATTGTGATCG
	* * *
Site number	1 2 3 4 5 6 7 8 9 10

Segregating sites (S and p_S):

Sites 2, 6, and 8 have variable base pairs among the four sequences (columns marked with *). These are segretating sites. Therefore, for these sequences $S = 3$ segregating sites and $p_S = 3/10 = 0.3$ segregating sites per nucleotide site examined.

Nucleotide diversity (π):

Nucleotide diversity = 1.67

```
1  AATGTCAACG
2  AATGTCAACG
```
$d_{12} = 0$

```
1  AATGTCAACG
3  ATTGTCAACG
```
$d_{13} = 1$

```
2  AATGTCAACG
3  ATTGTCAACG
```
$d_{23} = 1$

```
1  AATGTCAACG
4  ATTGTGATCG
```
$d_{14} = 3$

```
2  AATGTCAACG
4  ATTGTGATCG
```
$d_{24} = 3$

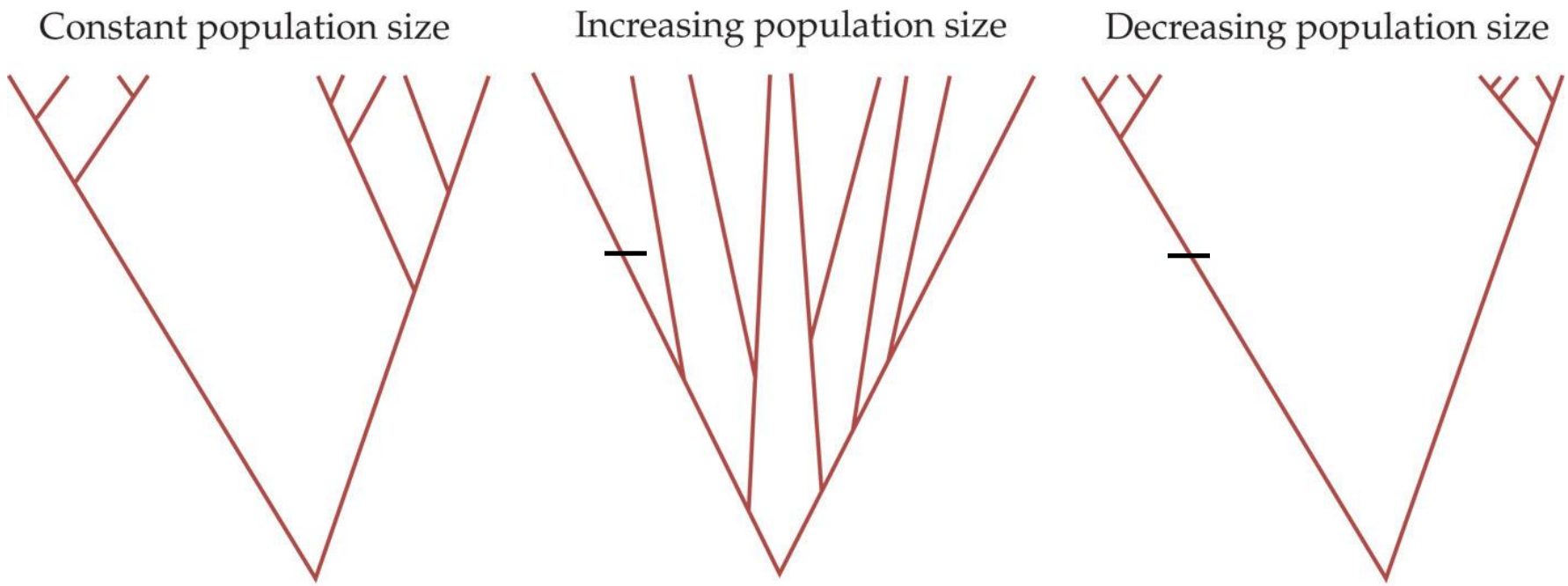
```
3  ATTGTCAACG
4  ATTGTGATCG
```
$d_{34} = 2$

$$\Sigma d_{ij} = 0 + 1 + 3 + 1 + 3 + 2 = 10$$

Number of pairs of sequences compared = $[n(n-1)]/2 = [4(3)]/2 = 6$

$\hat{\pi}$ = 10 differences/6 pairs = 1.67 average pairwise differences

$\hat{\pi}$ = 1.67 avg. differences/10 sites = 0.167 pairwise differences per site



***INTRODUCTION TO POPULATION GENETICS*, Figure 3.10**

”Internal” mutations contribute more to π than ”external” ones.

π is higher under decreasing population size..

Tajima's D

Theta estimated by average pairwise difference

$$(9.1) \qquad \hat{\theta}_W = S \Big/ \left(\sum_{i=1}^{n-1} 1/i \right) \text{ and } \hat{\theta}_T = \frac{\sum_{(i,j):i<j} d_{ij}}{n(n-1)/2}$$

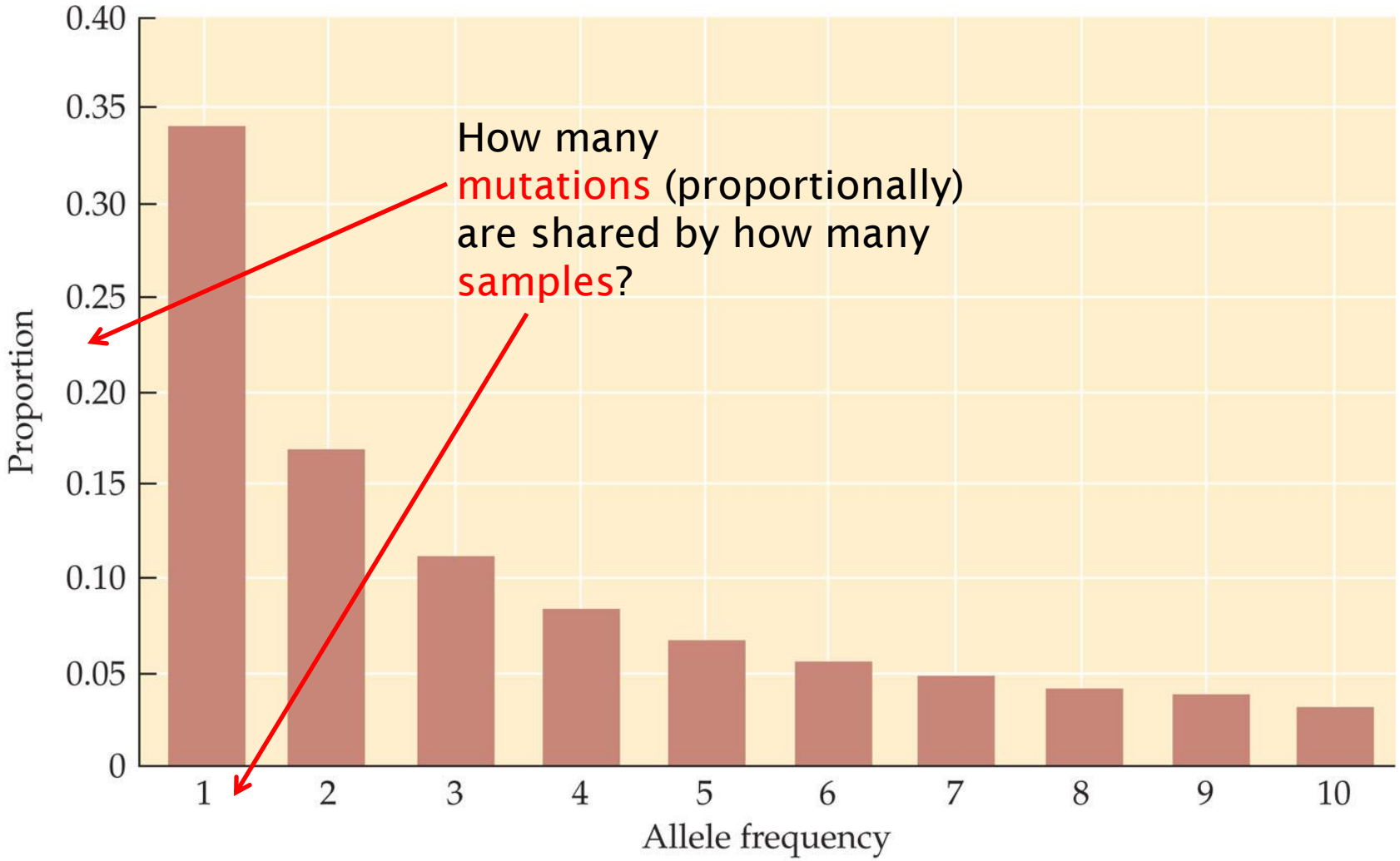
Theta estimated by # segregating sites

$$(9.2) \qquad \text{Tajima's } D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}(\hat{\theta}_T - \hat{\theta}_W)}}$$

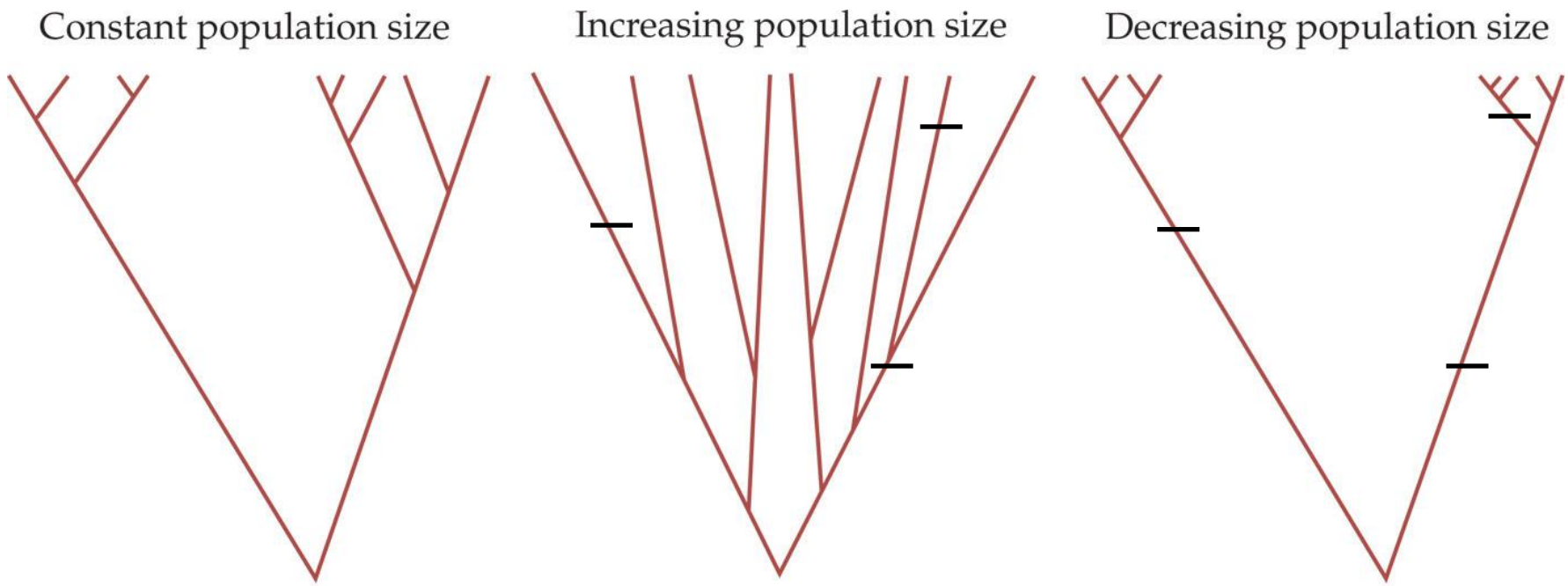
***INTRODUCTION TO POPULATION GENETICS*, Equations 9.1–9.2**

Tajima's D = 0 in a constant population.

Site frequency spectrum



***INTRODUCTION TO POPULATION GENETICS*, Figure 3.9**

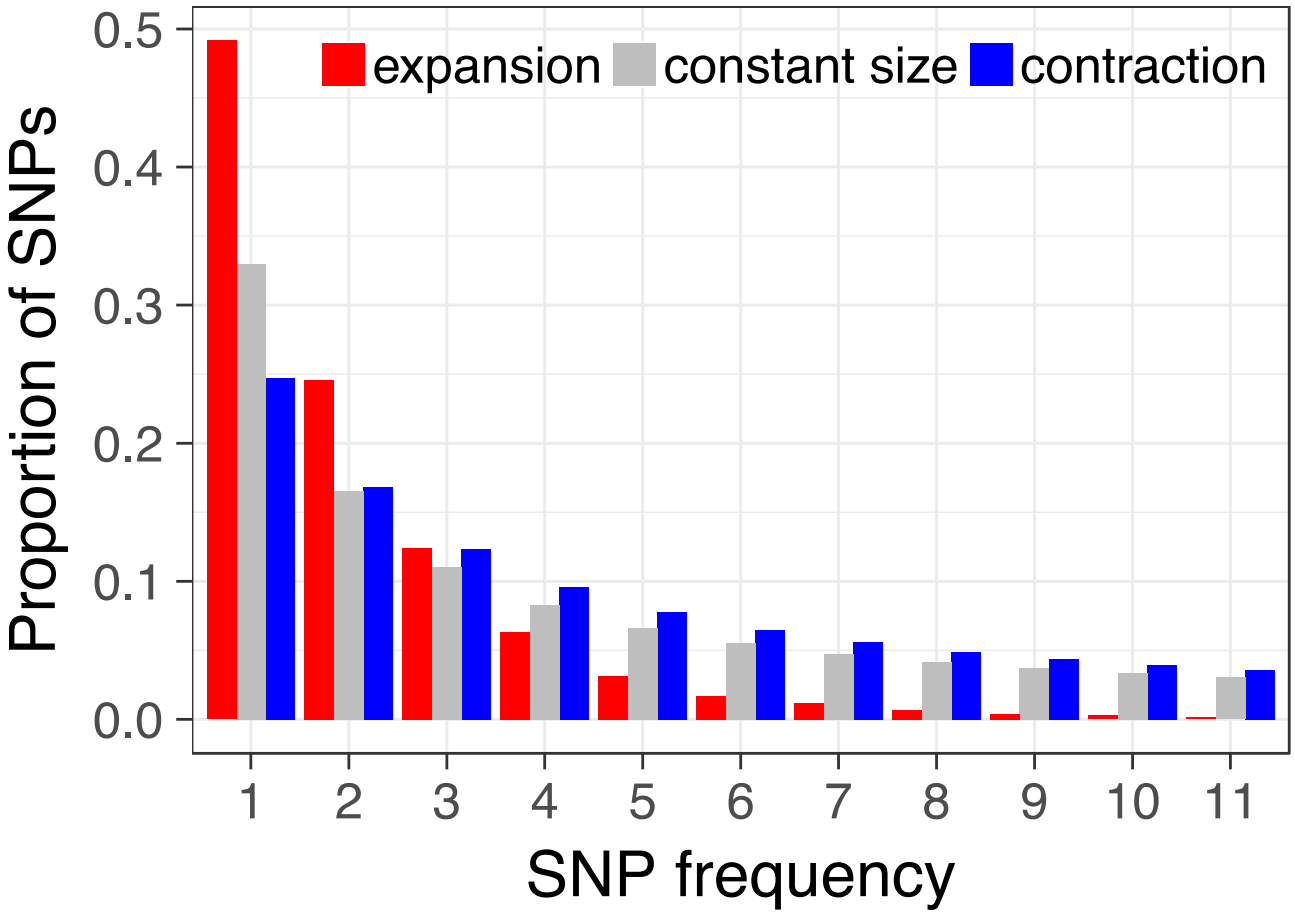


***INTRODUCTION TO POPULATION GENETICS*, Figure 3.10**

Can you draw the site frequency spectrum?

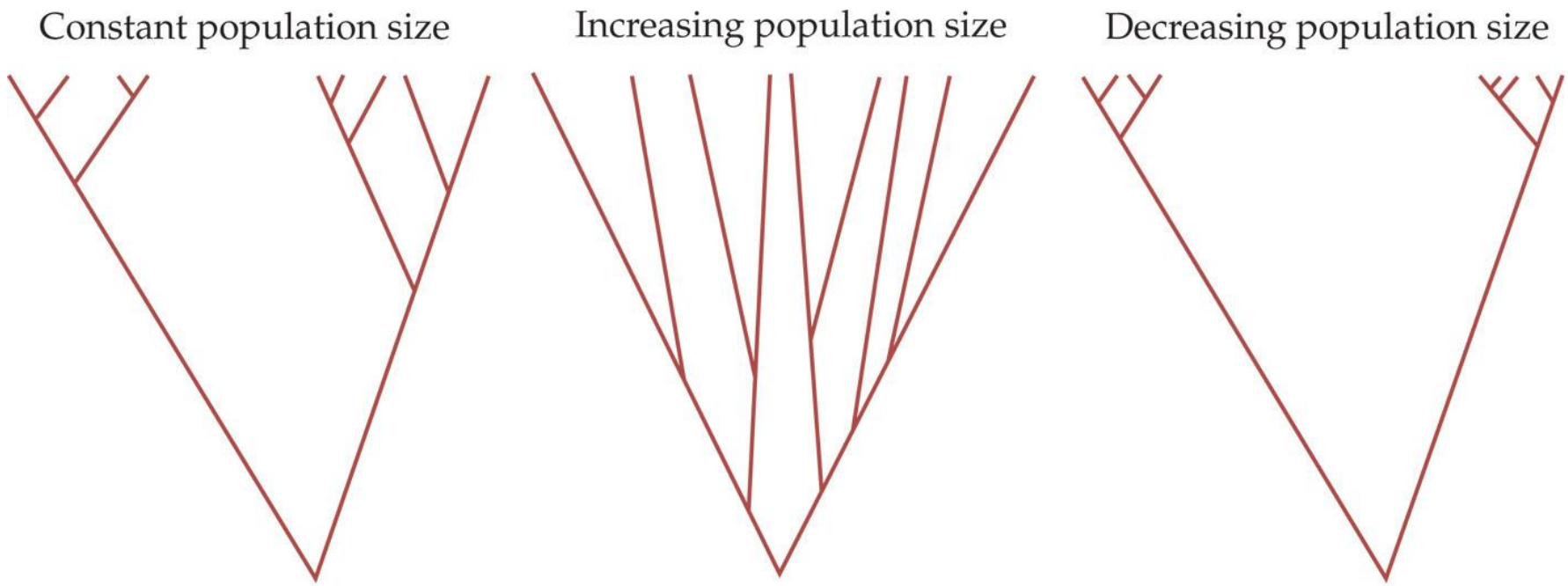
How would SFS'es tend to differ?

The SFS reveals population history



Population history through tree inference

Coalescent intervals and tree shapes

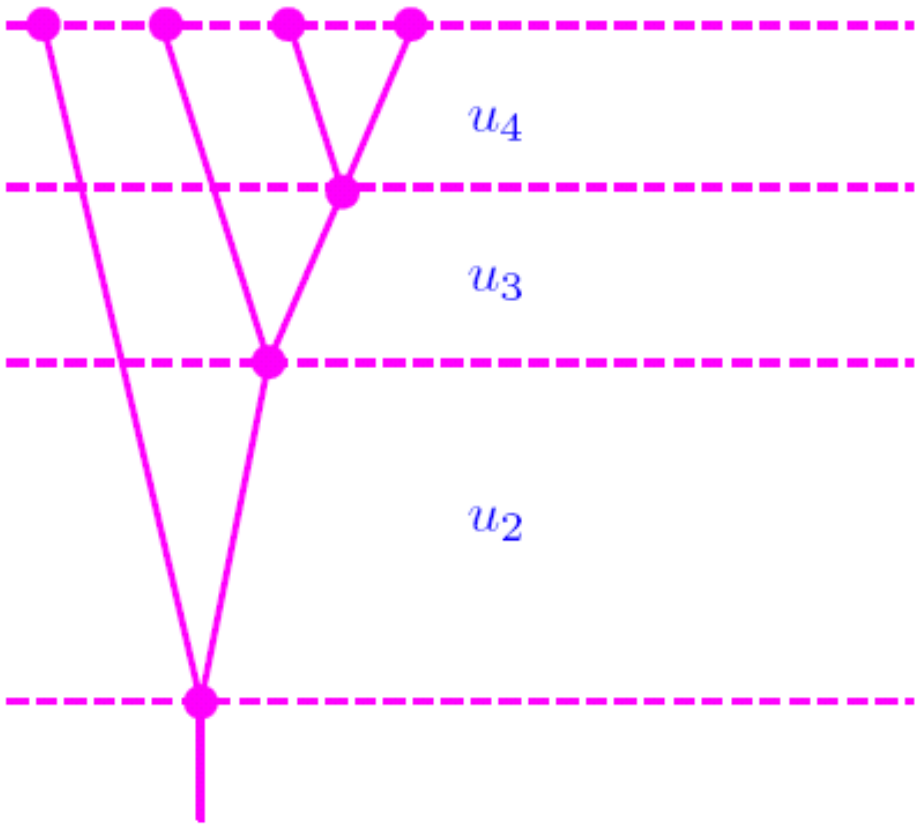


***INTRODUCTION TO POPULATION GENETICS*, Figure 3.10**

Tree ”shapes” are connected to demography

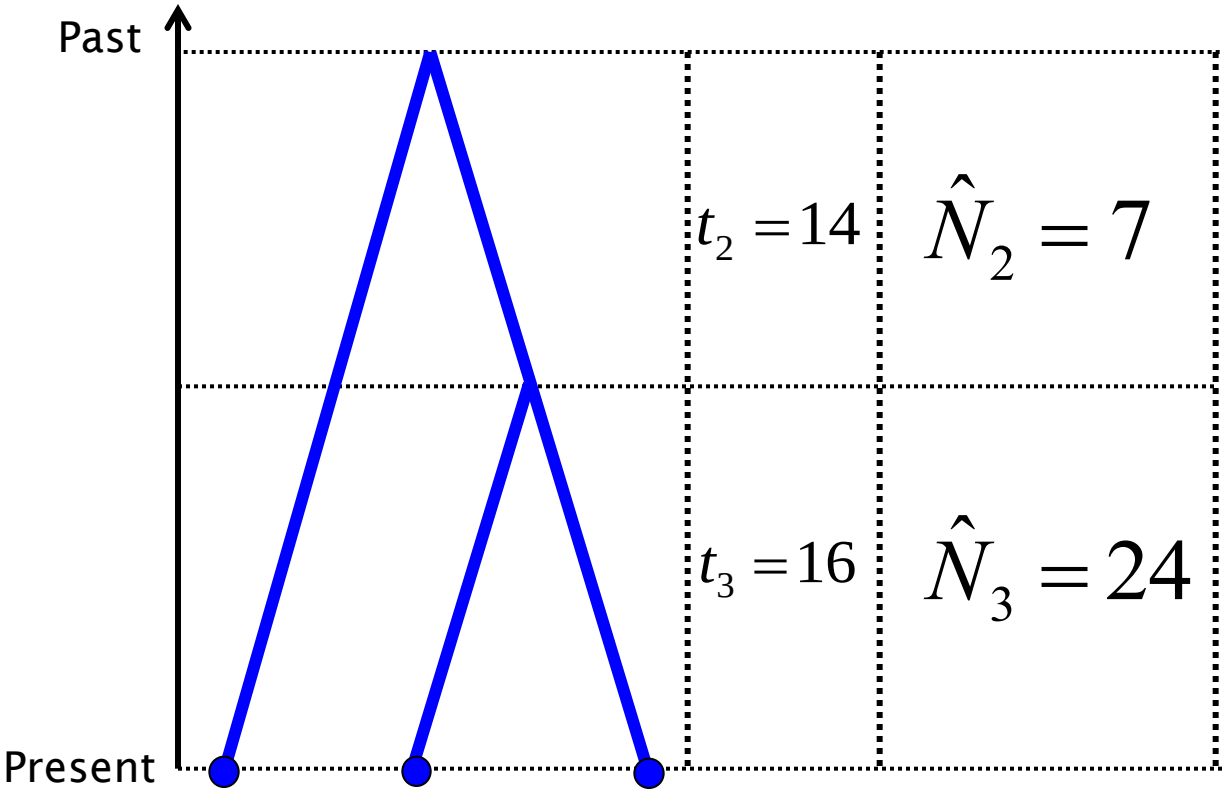
Changing population size: skyline plots

The expected time between coalescence events: $E(u_k) = \frac{4N}{k(k-1)}$



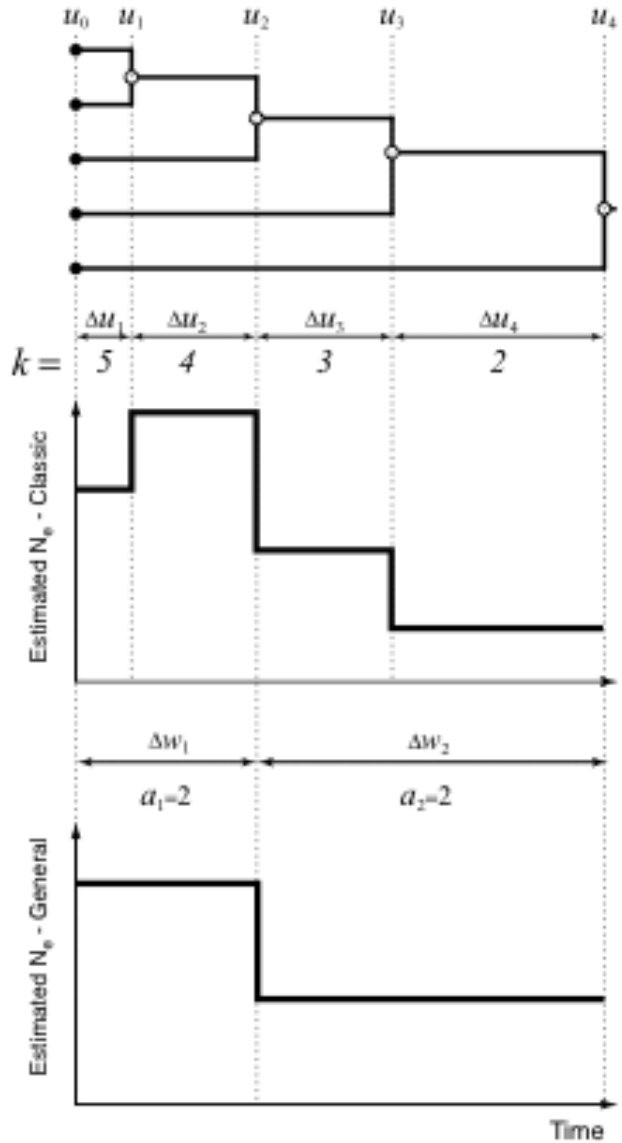
$$N = \frac{k(k-1)}{4} u_k$$

Changing population size: skyline plots



$$\hat{N}_k = \frac{k(k-1)}{4} t_k$$

Changing population size: skyline plots



Conclusions so far

- Many simple data measures contain information about population history.
- More sophisticated information can be extracted from simulations or modeling.
- Knowing population history helps interpretation of genetic data (e.g. selection).