Detecting signals of natural selection from population genetic data using deep learning

Intented Learning Outcomes

At the end of this session you will be able to

- appreciate the importance of studying natural selection
- understand how selection changes allele frequencies
- describe commonly used methods to detect selection
- illustrate basic principles of ML
- critically read papers using ML to detect selection
- implement simple ML methods to detect selection

Part 1: Motivation and theory
Part 2: “Common” methods
Part 3: ML/DL methods

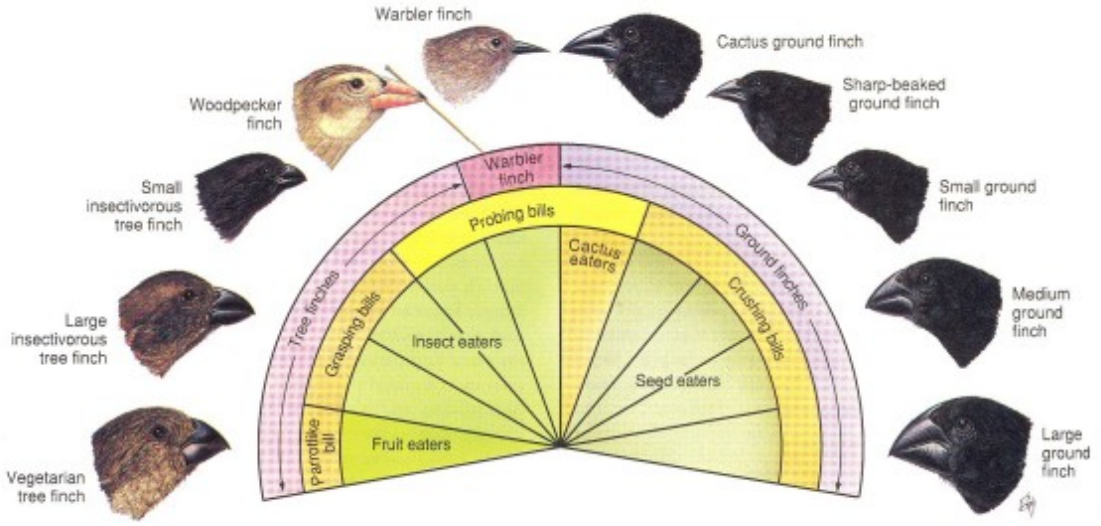
So far so good

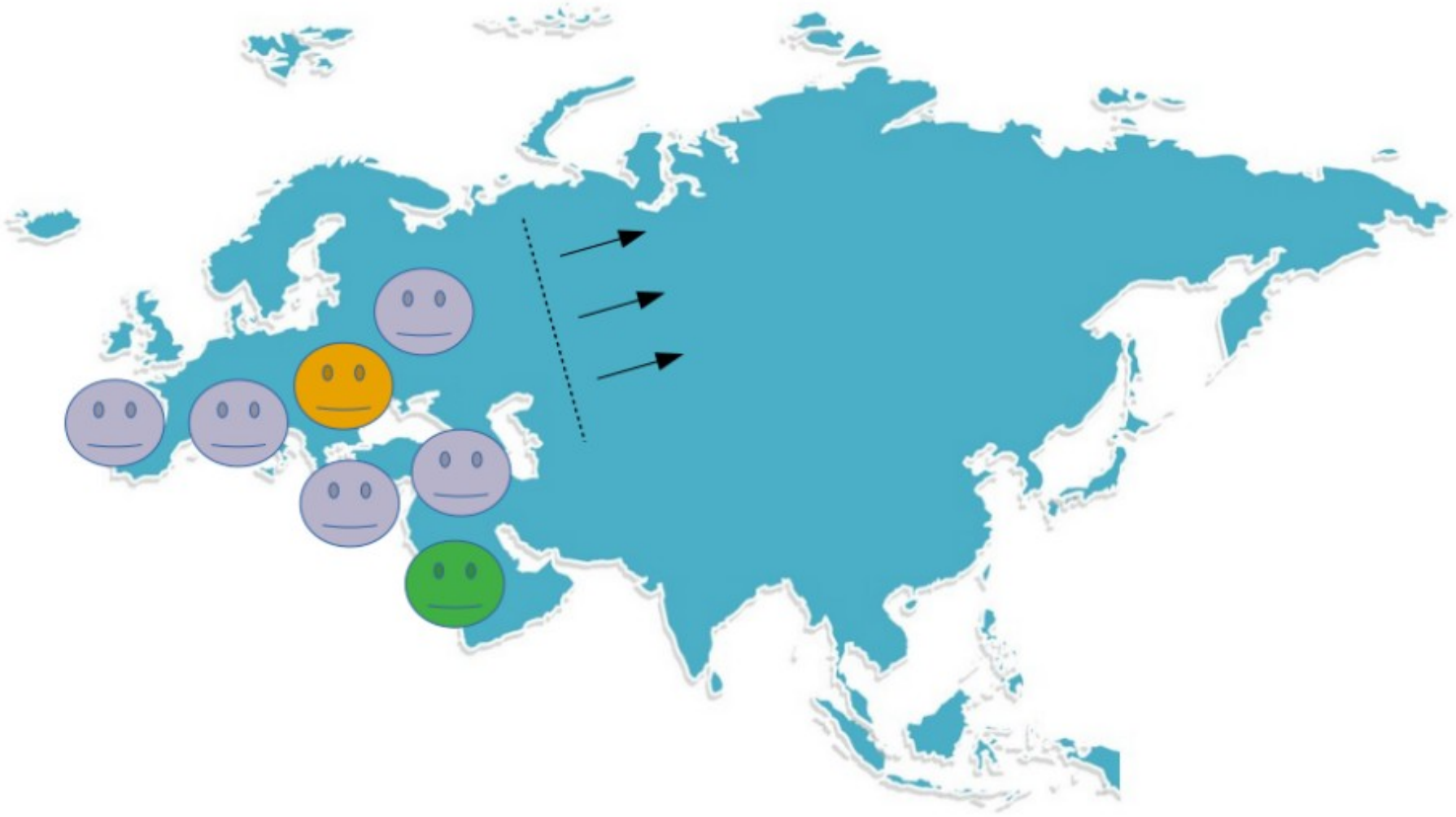
- random mating
- genetic drift
- mutation / recombination

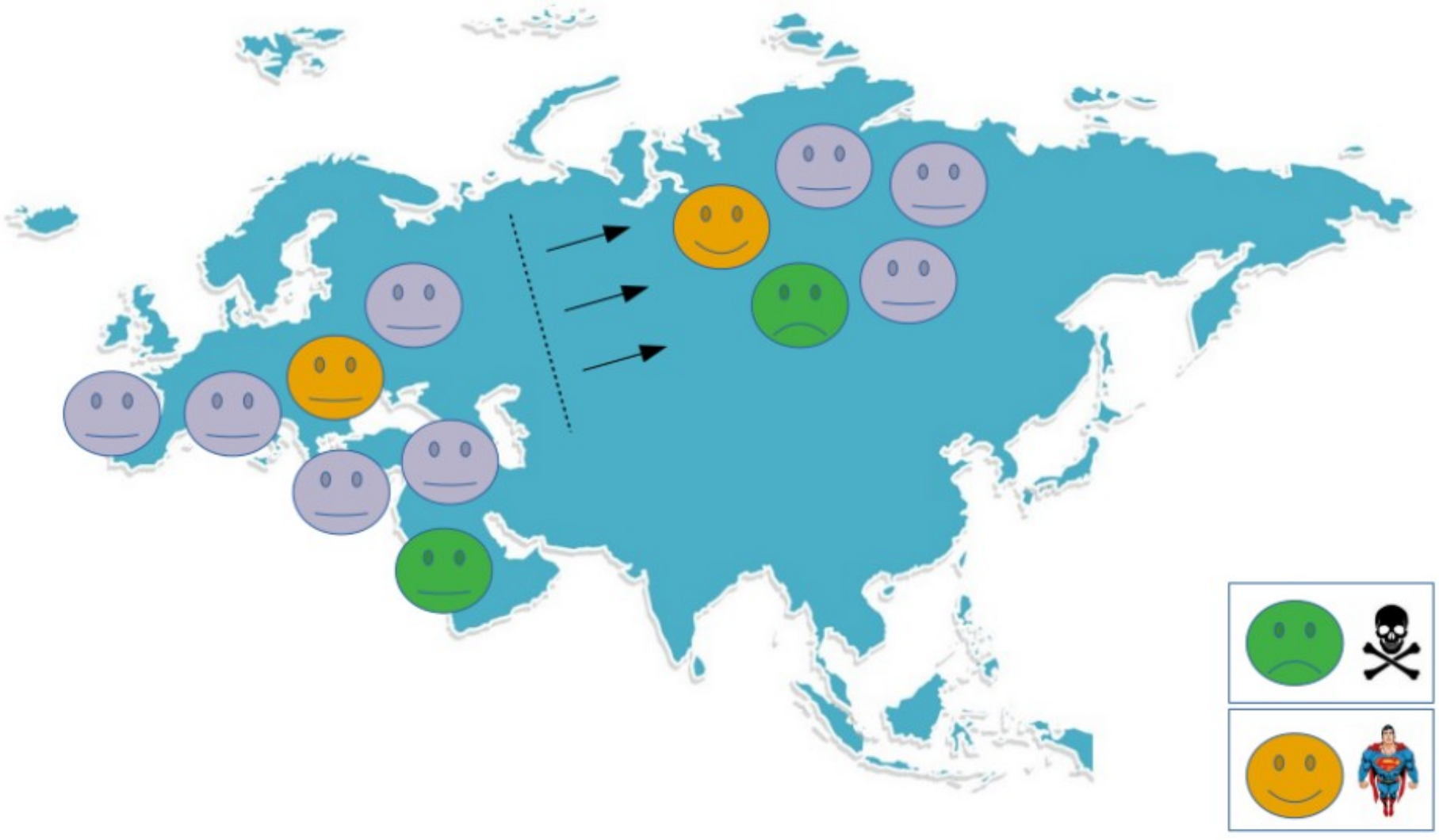
What if different alleles affect survival?

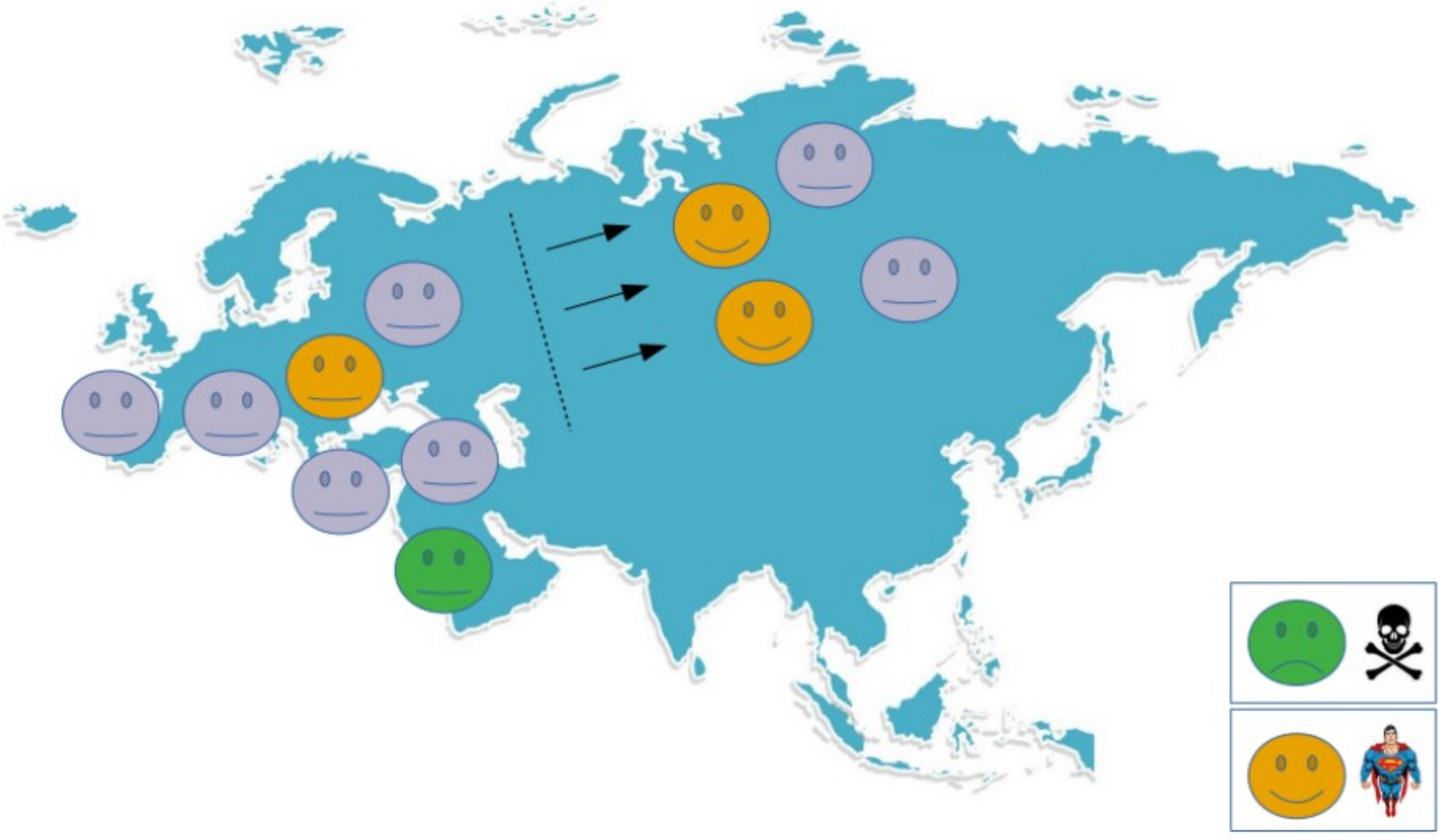
Natural selection

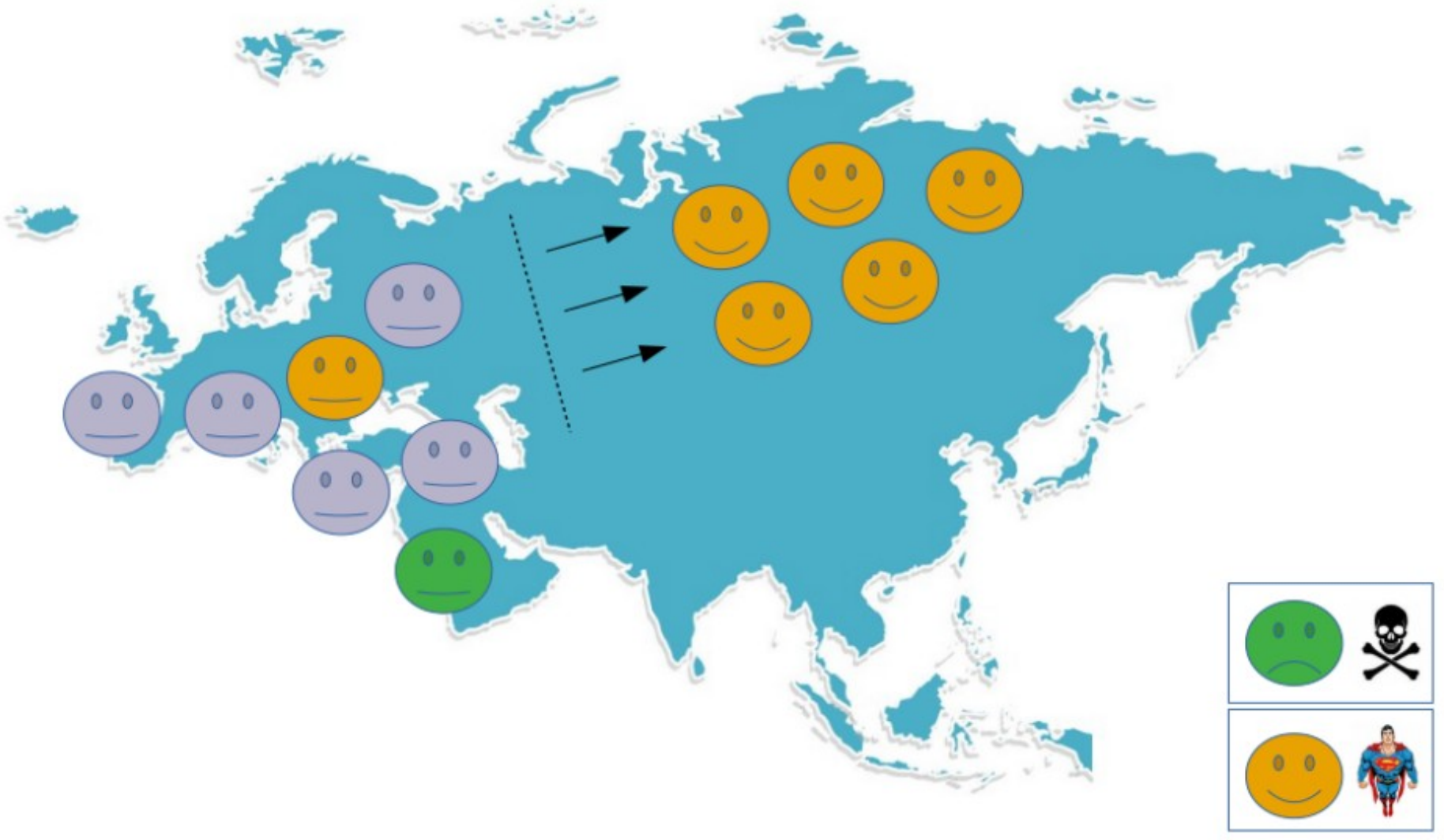
- heritable traits that increase the **fitness** become more common in the population
- mutations evolve accordingly to their **effect on the fitness** of the carrier
- **functionality** is the prerequisite for selection to be effective

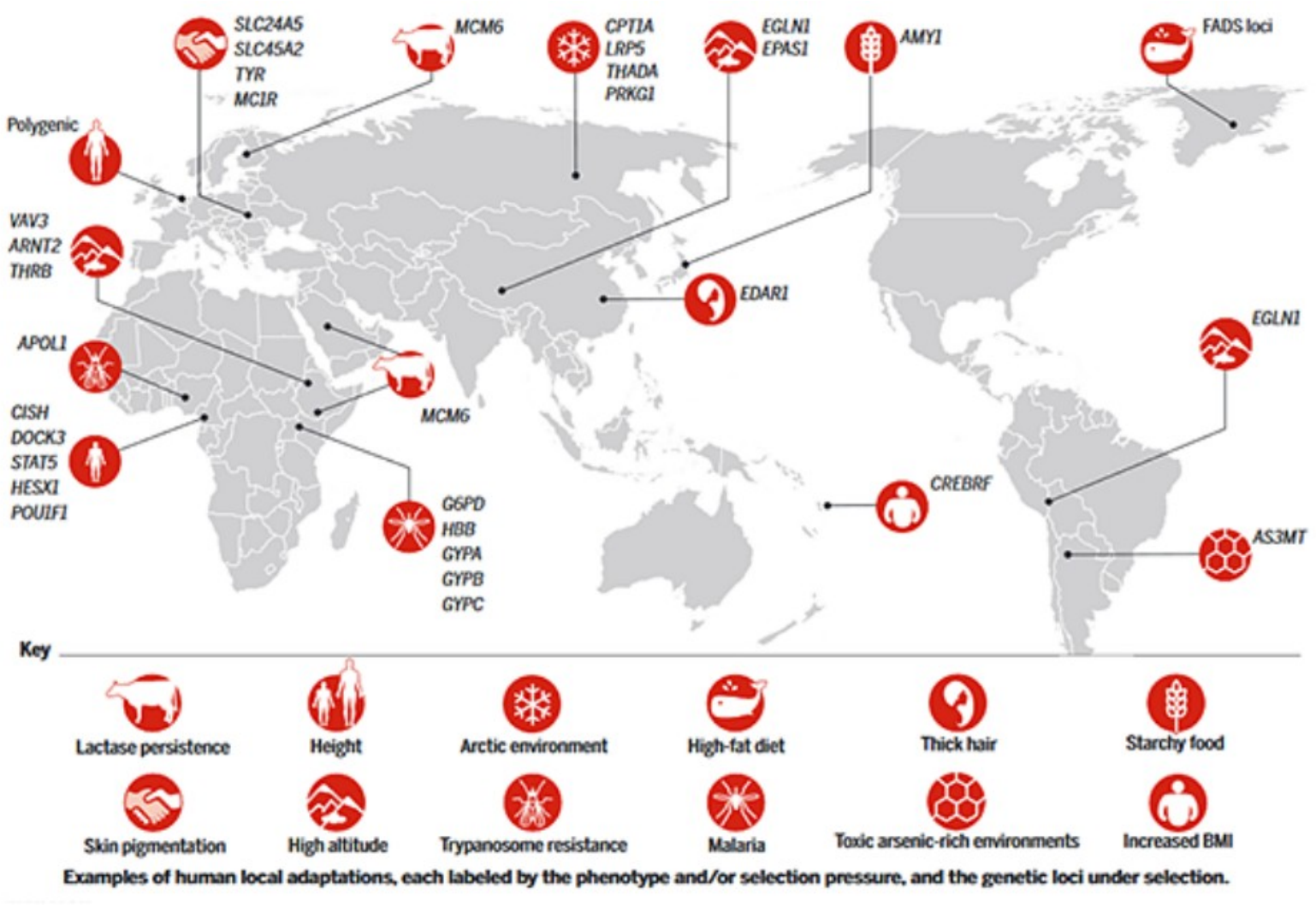








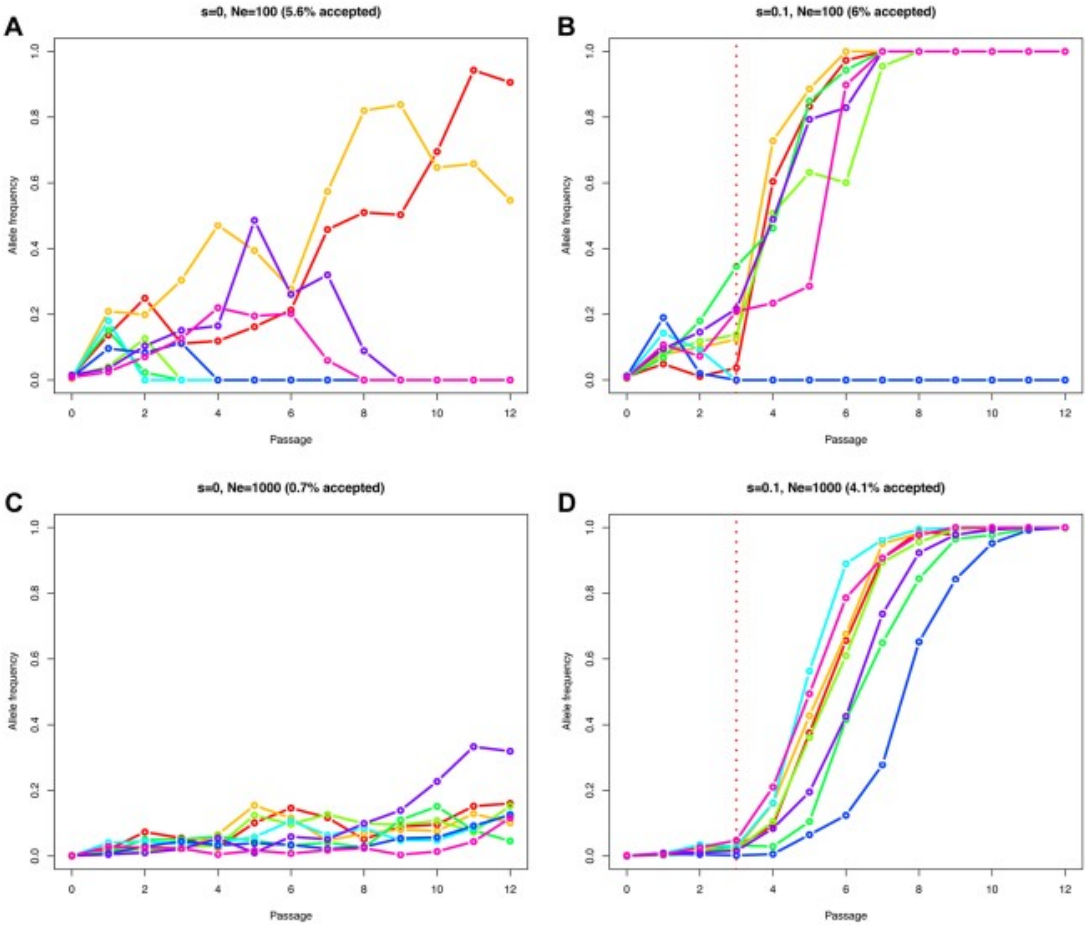




Examples of human local adaptations, each labeled by the phenotype and/or selection pressure, and the genetic loci under selection.

How do we infer genomics signals of selection?

How does selection change allele frequencies?



Foll et al 2014

Let's assume haploid organisms.

At time $t = 0$ we have alleles A and a in the population such that $f_A = N_A/(N_A + N_a)$.

What happens at $t = 1$ if all individuals have the chance contribute to the next generation?

At $t = 1$, A-individuals will have on average w_A descendants. If so,

$$f_A(1) = \frac{w_A f_A(0)}{w_A f_A(0) + w_a f_a(0)}.$$

Example: $f_A(0) = 0.25$, $w_a = 1.5$, $w_A = 1.7$

$f_A(1) \approx$

At $t = 1$, A-individuals will have on average w_A descendants. If so,

$$f_A(1) = \frac{w_A f_A(0)}{w_A f_A(0) + w_a f_a(0)}.$$

Example: $f_A(0) = 0.25$, $w_a = 1.5$, $w_A = 1.7$

$$f_A(1) \approx 0.27$$

What if $w_a = 15$, $w_A = 17$ or $w_a = 0.15$, $w_A = 0.17$?

$$f_A(1) \approx$$

At $t = 1$, A-individuals will have on average w_A descendants. If so,

$$f_A(1) = \frac{w_A f_A(0)}{w_A f_A(0) + w_a f_a(0)}.$$

Example: $f_A(0) = 0.25$, $w_a = 1.5$, $w_A = 1.7$

$$f_A(1) \approx 0.27$$

What if $w_a = 15$, $w_A = 17$ or $w_a = 0.15$, $w_A = 0.17$?

$$f_A(1) \approx 0.27$$

Only the relative w values are important!
In fact, if we divide by w_A we get

$$f_A(1) = \frac{w_A f_A(0)}{w_A f_A(0) + w_a f_a(0)}$$

$$f_A(1) = \frac{f_A(0)}{f_A(0) + (w_a/w_A) f_a(0)}$$

and we define the **selection coefficient** s as

$$\frac{w_a}{w_A} = 1 - s$$

Only the relative w values are important!
In fact, if we divide by w_A we get

$$f_A(1) = \frac{w_A f_A(0)}{w_A f_A(0) + w_a f_a(0)}$$

$$f_A(1) = \frac{f_A(0)}{f_A(0) + (w_a/w_A) f_a(0)}$$

and we define the **selection coefficient** s as

$$\frac{w_a}{w_A} = 1 - s$$

$$f_A(1) = \frac{f_A(0)}{f_A(0) + (1-s) f_a(0)}$$

If this process persists for a long time t then

$$f_A(1) = \frac{f_A(0)}{f_A(0)+(1-s)^t f_a(0)}$$

which means we can make predictions on the evolution of A and a alleles.

What are the important parameters?

`(example 1 in R)`

Note that $(1-s)^t$ is close to e^{-st}.

`(example 2 in R)`

What are the parameters to predict the allele frequency trajectory?

Note that $(1-s)^t$ is close to e^{-st}.

`(example 2 in R)`

$f_A(t)$ depends on the product between s and t. If s is small, then t to generate a certain change in f_A is *inversely* proportional to s.

If $s > 0$ then A-bearing individuals have the advantage, the opposite is true for $s < 0$.

What happens in diploids?

We need to consider the effect of each *genotype* on viability (i.e. the probability to survive from zygote to adult stage).

For one locus with two alleles, each genotype has its viability:

v_{AA} v_{Aa} v_{aa}

If r is the number of zygotes produced on average by each pair of parents, then

$$w_{AA} = rv_{AA};\ w_{Aa} = rv_{Aa};\ w_{aa} = rv_{aa}$$

With random mating, any change in allele frequency is due different genotype viabilities.

$$f'_A = \frac{v_{AA} f_A^2 + v_{Aa} f_A (1 - f_A)}{\hat{v}}$$ with $\hat{v}$ being the average viability.

Note that only relative viabilities matter!

In diploids, selection coefficients are defined for each genotype vs. the largest viability.

$$\frac{v_{Aa}}{v_{AA}} = 1 - s_{Aa};\ \frac{v_{aa}}{v_{AA}} = 1 - s_{aa}$$

if individuals with AA genotype have the highest viability v_{AA}.

Special cases

- additive selection: $w_{Aa} = 1 - s$; $w_{aa} = 1 - 2s$
- dominant advantageous allele: $w_{AA} = w_{Aa}$ and $w_{aa} = 1 - s$
- recessive advantageous allele: $w_{AA} = 1$ and $w_{Aa} = w_{aa} = 1 - s$
- genic selection: each copy of a reduces viability by a factor of $(1 - s)^2$, so that $v_{Aa}/v_{AA} = (1 - s)$ and $v_{aa}/v_{AA} = (1 - s)^2$

`(example 3 in R)`

If s does not change in time, we have three possible scenarios:

- directional selection (additive, dominant or recessive)
- heterozygote advantage
- heterozygote disadvantage

Directional selection

A is the **advantageous allele** if $v_{aa} \leq v_{Aa} \leq v_{AA}$.

f_A will increase every generation and eventually reach 1 (i.e. fixation of A and loss of a).

The rate of change in allele frequency depends on s, selection coefficient.
Even small s can change allele frequency substantially over many generations.

`(example 4 in R)`

Heterozygote advantage

If $v_{Aa} > v_{AA}, v_{aa}$ and we define

$$\frac{v_{aa}}{v_{Aa}} = 1 - s_{aa};\ \frac{v_{AA}}{v_{Aa}} = 1 - s_{AA}$$

f_A will tend to the same value regardless of its initial frequency. In fact, selection won't eliminate either allele (it is a special case of **balancing selection**).

We have learnt how **selection** changes allele frequencies in time. You also know how **genetic drift** changes allele frequencies in time.

What is the combined effect of selection and drift?

Recall that for finite populations and drift lone, the probability u that a new neutral mutation reaches fixations is $1/(2N)$, with N being the **population** size.

`(examples 5 in R)`

For a population of size N, the fixation probability u of a new mutation with selection coefficient s can be defined as:

- strongly deleterious, if $2Ns << -1$ then $u \approx 0$
- nearly neutral, if $-1 < 2Ns < 1$ then $u \approx 1/(2N)$
- strongly advantageous, if $2Ns >> 1$ then $u \approx 2s$

Strongly advantageous mutations are not necessarily fixed. Slightly deleterious alleles have a small but non-zero chance of being fixed.

Whether an allele is strongly selected or nearly neutral depends on both the selection coefficient and the population size.

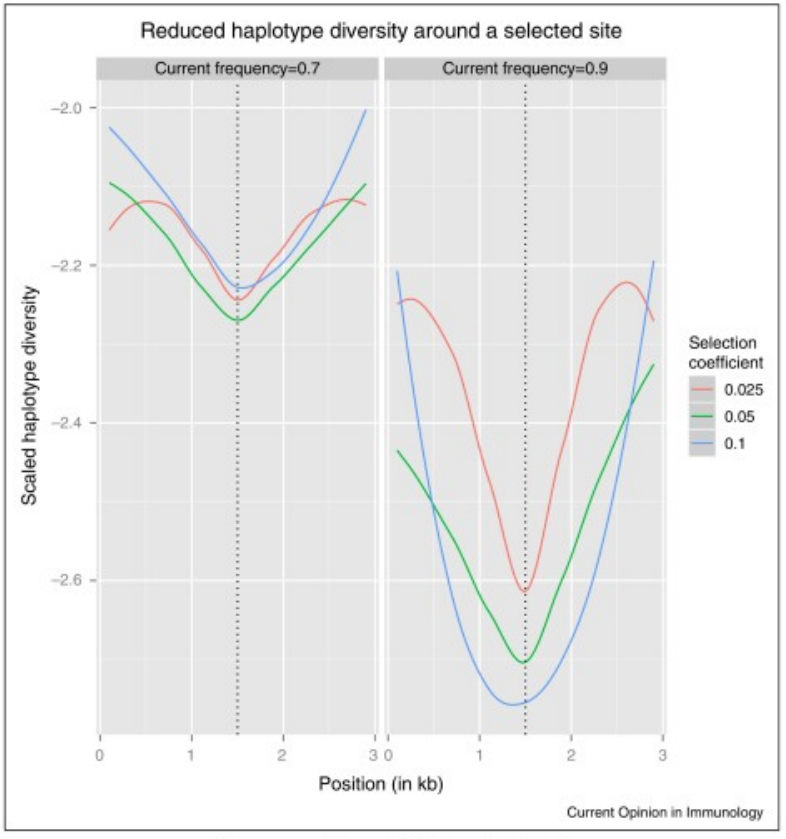
- What is the effect of selection at nearby (“linked”) sites?

  (example in msms)

Genetic hitchhiking

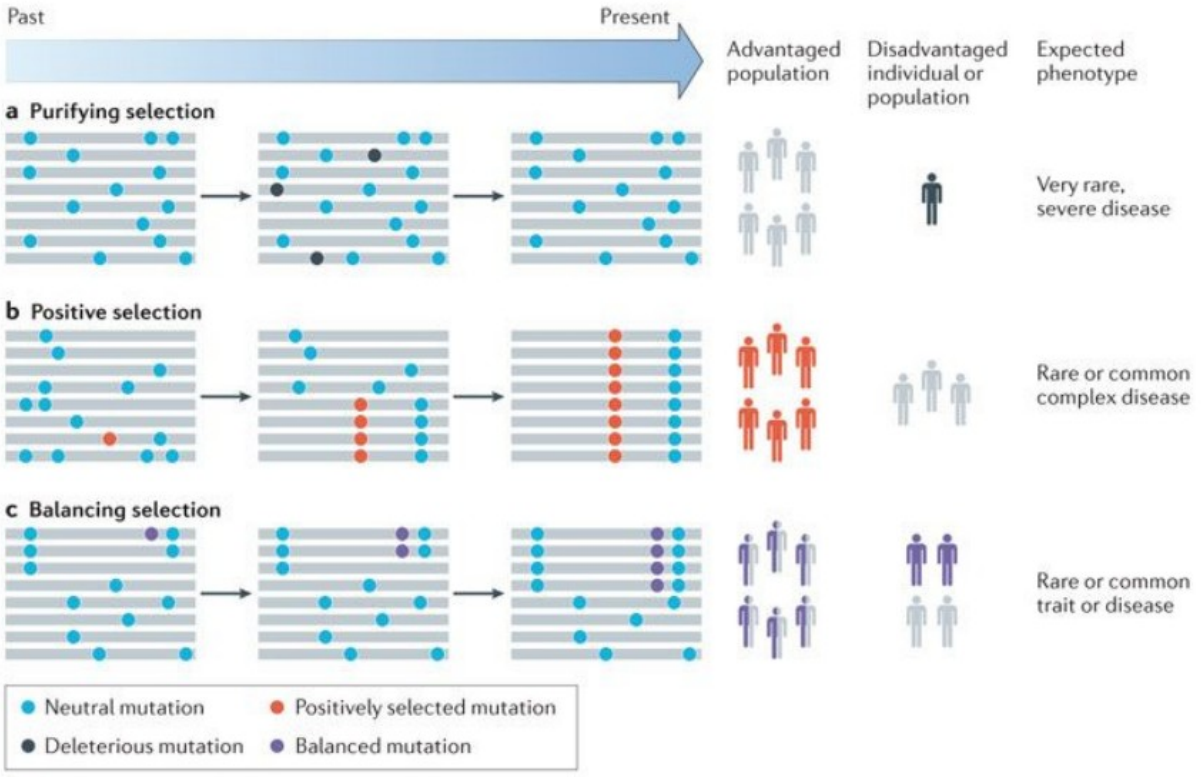
What is the effect at nearby neutral loci?

- selective sweep
- partial sweep
- soft sweep
- assortative overdominance (balancing selection)



Fumagalli and Sironi 2014

Under selective sweeps we expect a local decrease in heterozygosity (diversity) near the selected locus.



Quintana-Murci *et al.* 2013

How do we infer signals of natural selection?

- How do we infer signals of natural selection from genomic data?

  (let’s move to part 2)