03: Sequence change and evolution

CSCI4181/6802 Bioinformatics Algorithms
Finlay Maguire (finlay.maguire@dal.ca)

Overview

1. Mutations are changes that are passed from generation to generation

2. Many different types of mutation are possible, each can have different impacts on encoded proteins

3. Mutations can have consequences for the organism, leading to differential fitness, selection, and speciation

DNA is passed from parent to offspring through *replication*
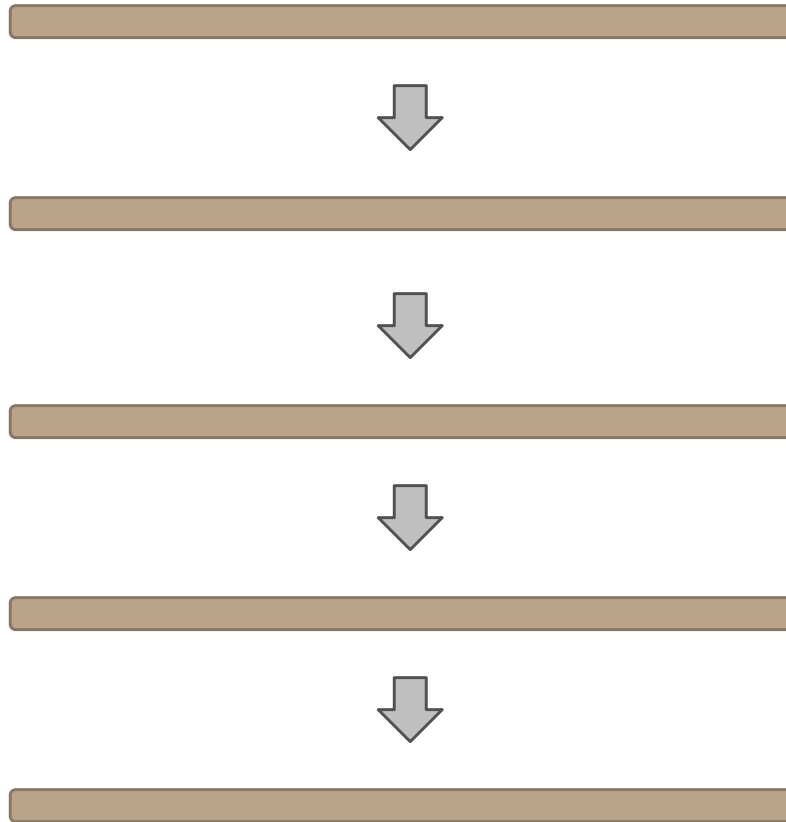
Replication is not always 100% faithful –
mutations can occur when mistakes are made
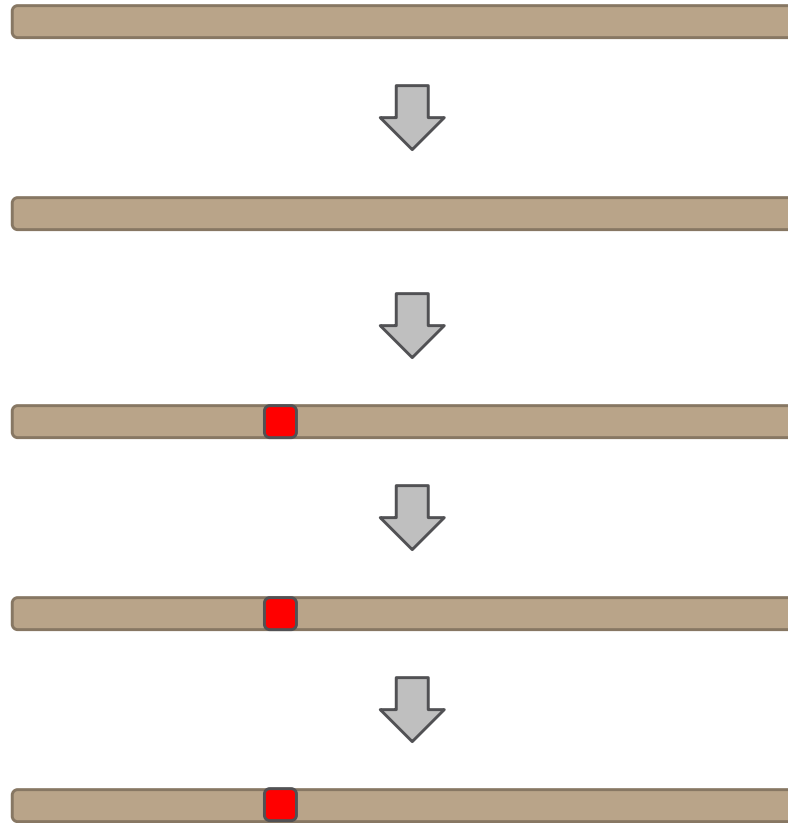
Point mutations in genes…

- May be SILENT (remember the genetic code!)

  GCC (Ala) → GCA
  (Ala)

- Or may change a single amino acid (*missense mutation*)

  GCC (Ala) → CCC
  (Pro)

- *Or* may introduce a stop codon (*nonsense mutation*)

  CAA (Gln) → TAA
  (Stop)

Other types of change

- Insertions

  AAA ACC GAA TCA CCG GAT → AAA ACC G**C**A ATC ACC GGA T..

- Deletions

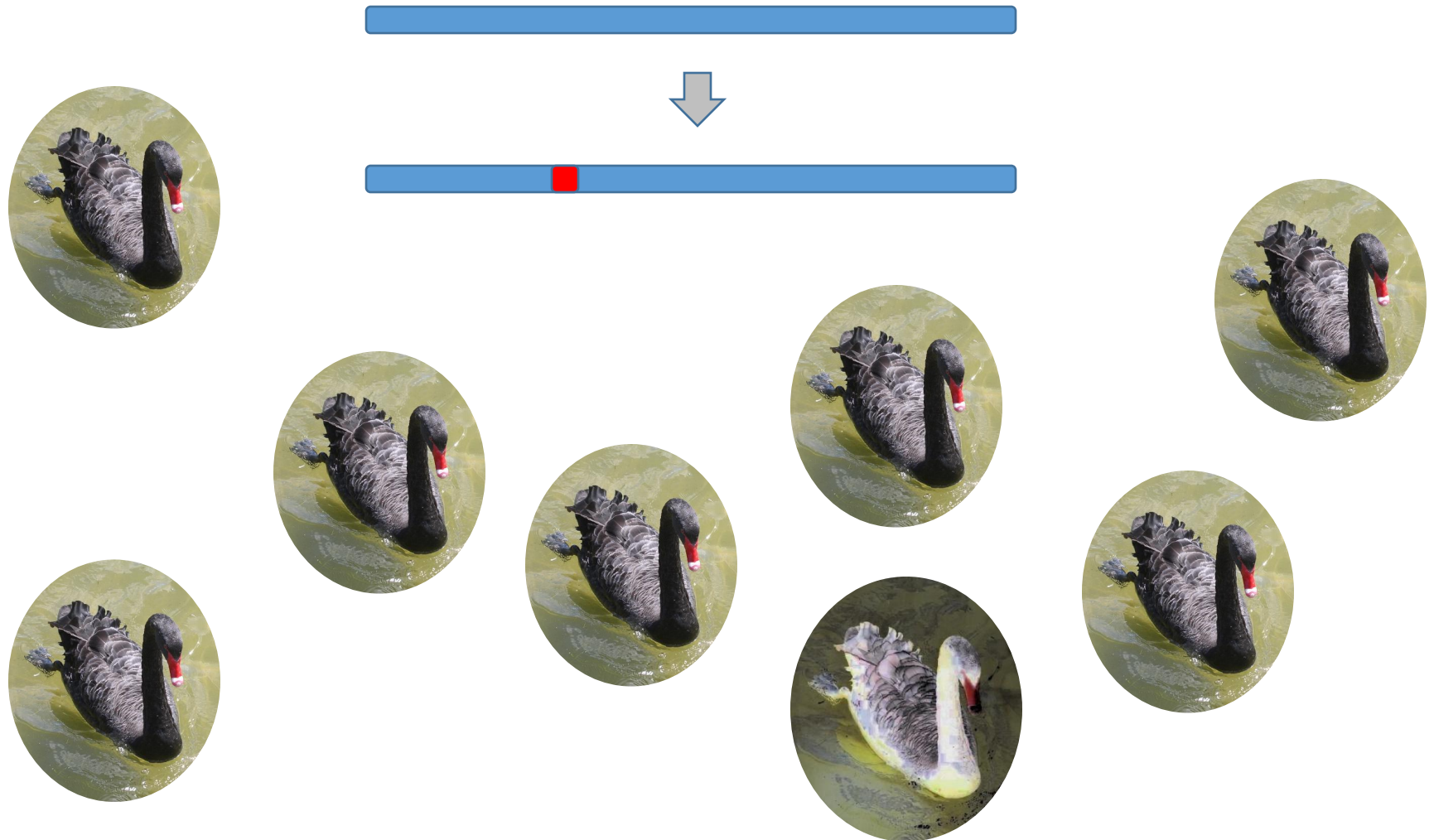 AAA ACC GAA TC**A** CCG GAT → AAA ACC GAA TCC CGG AT.

- Inversions

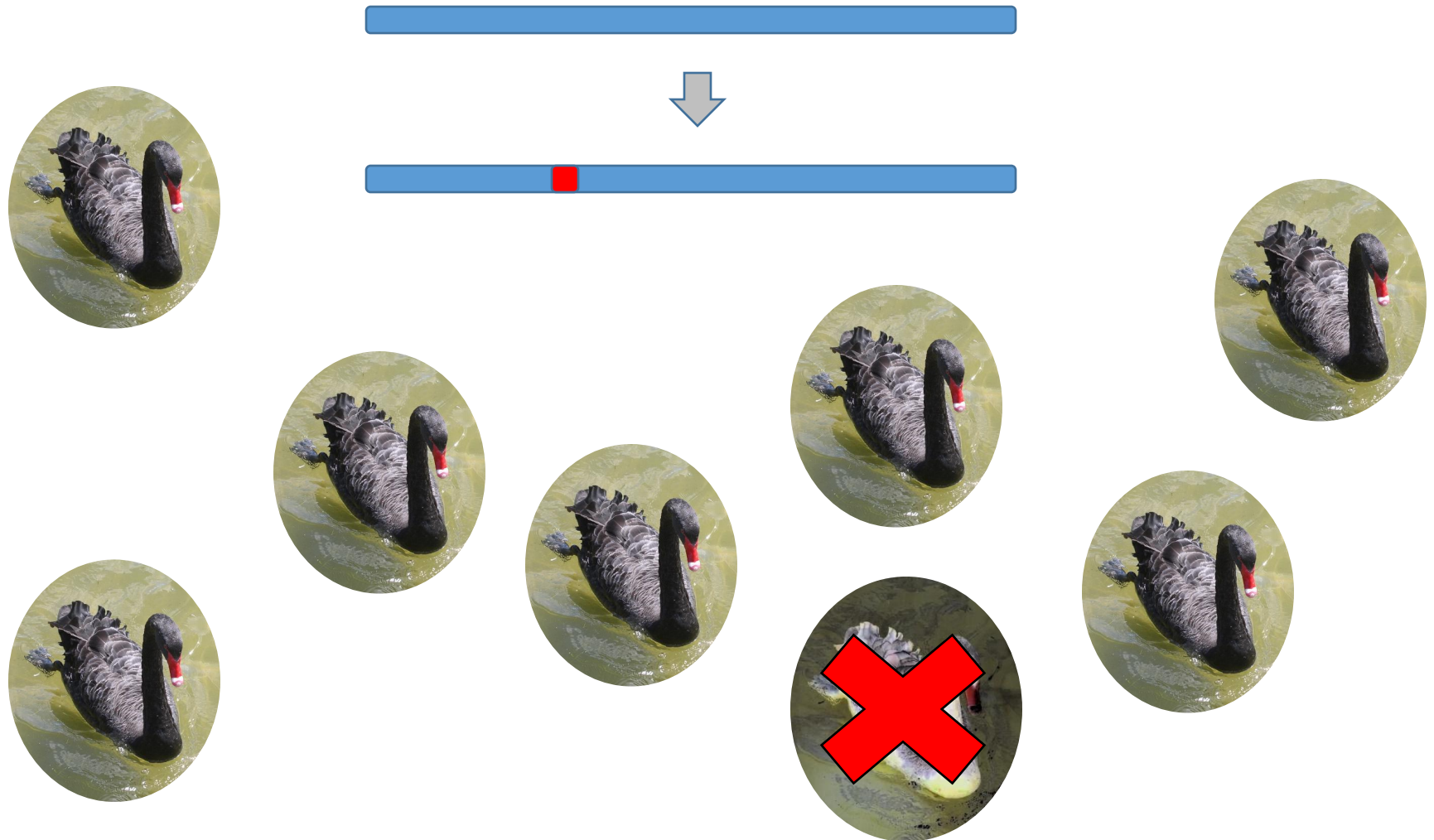 AAA ACC **GAA TCA** CCG GAT → AAA ACC **TGA TTC** CCG GAT

- Duplications

  AAA ACC **GAA TC**A CCG GAT → AAA ACC **GAA TCG AAT** CAC CGG AT.

(this is why we need sequence alignment)

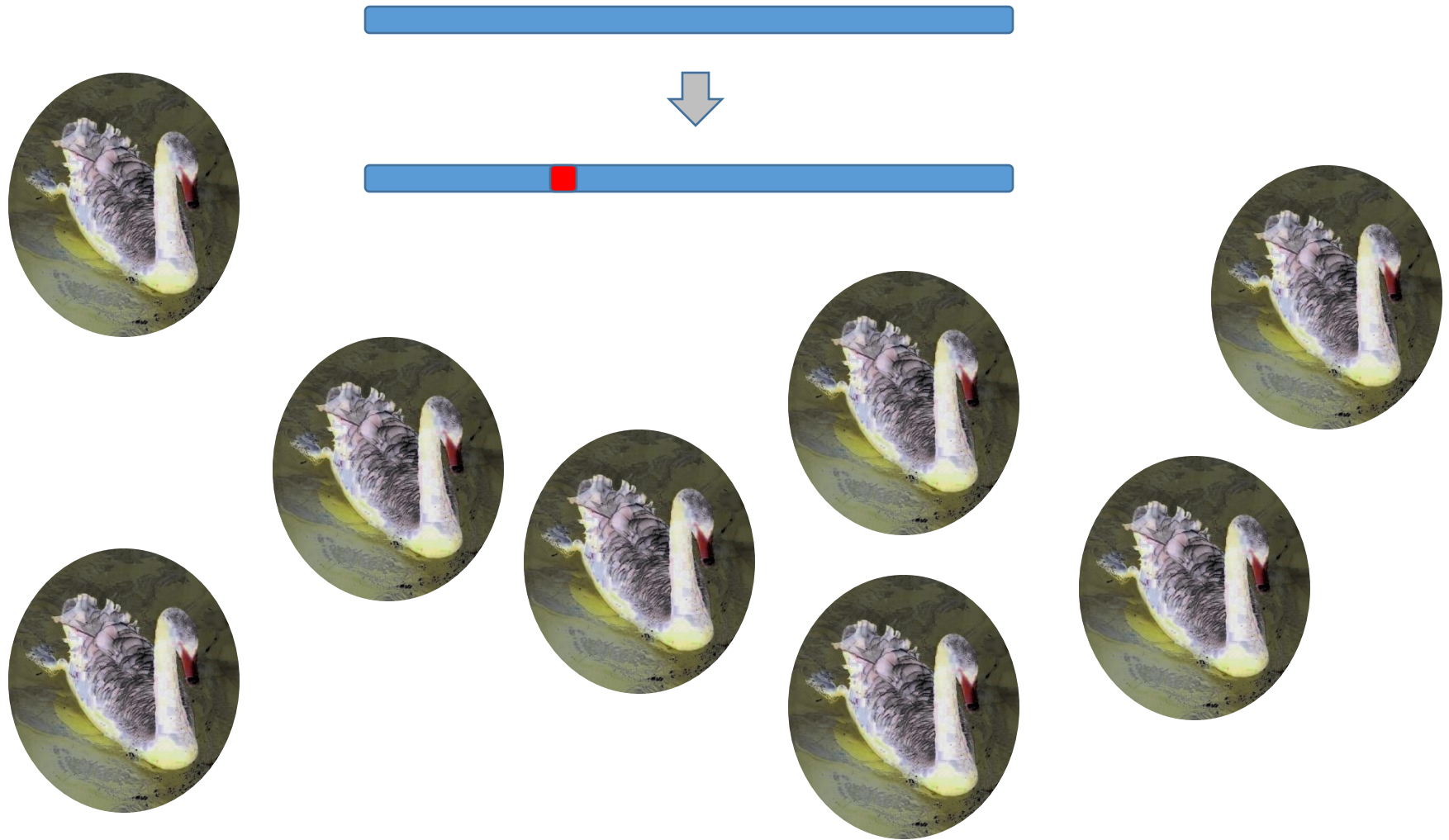# The population context of a mutation
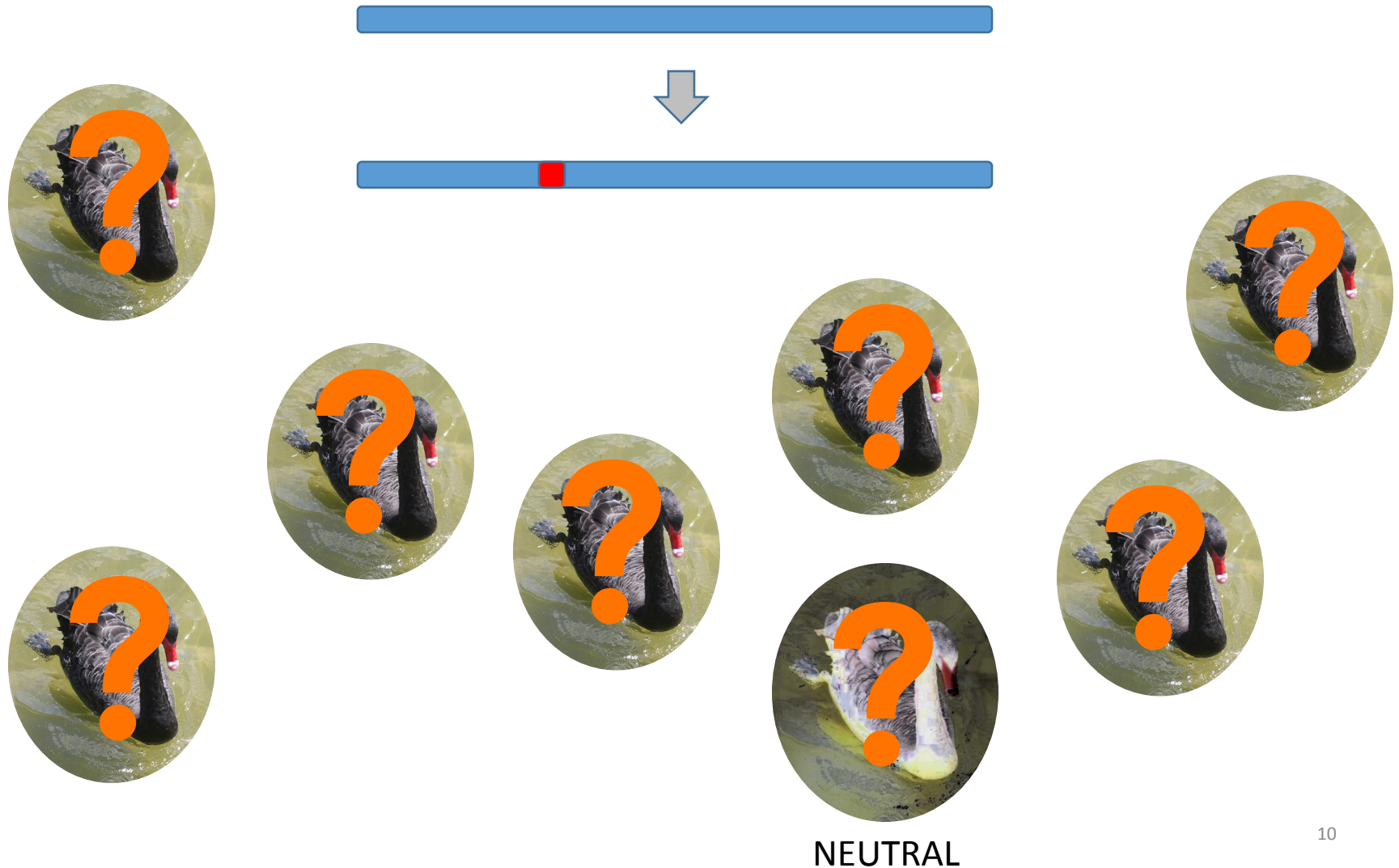
The population context of a mutation
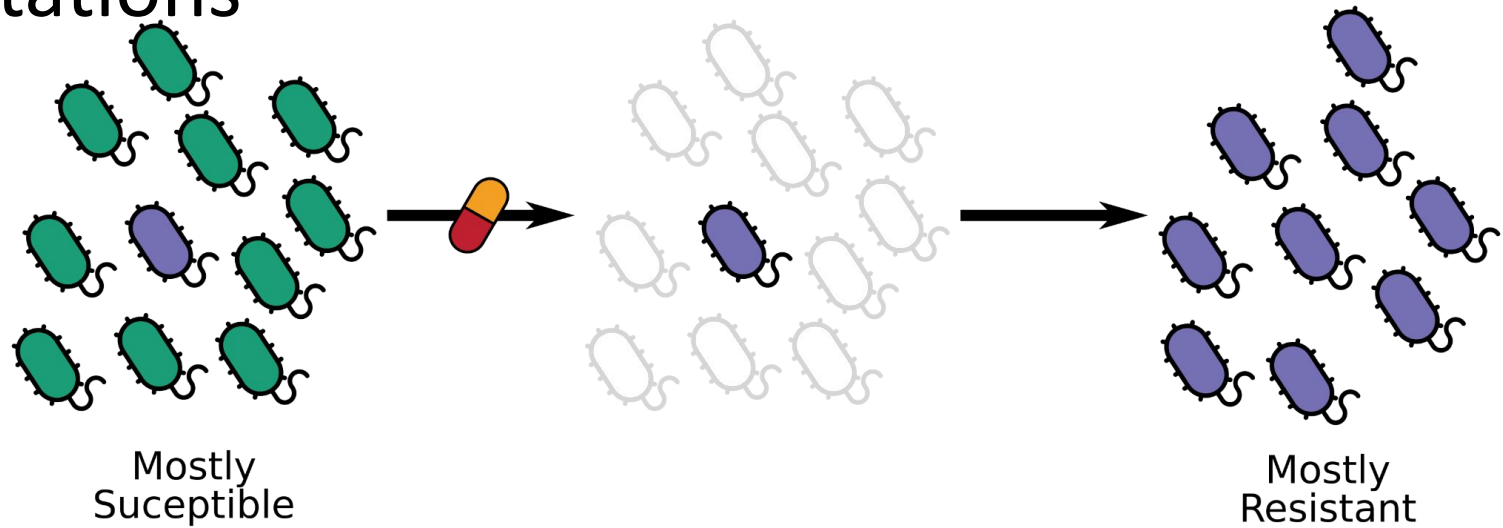

DELETERIOUS

The population context of a mutation

ADVANTAGEOUS

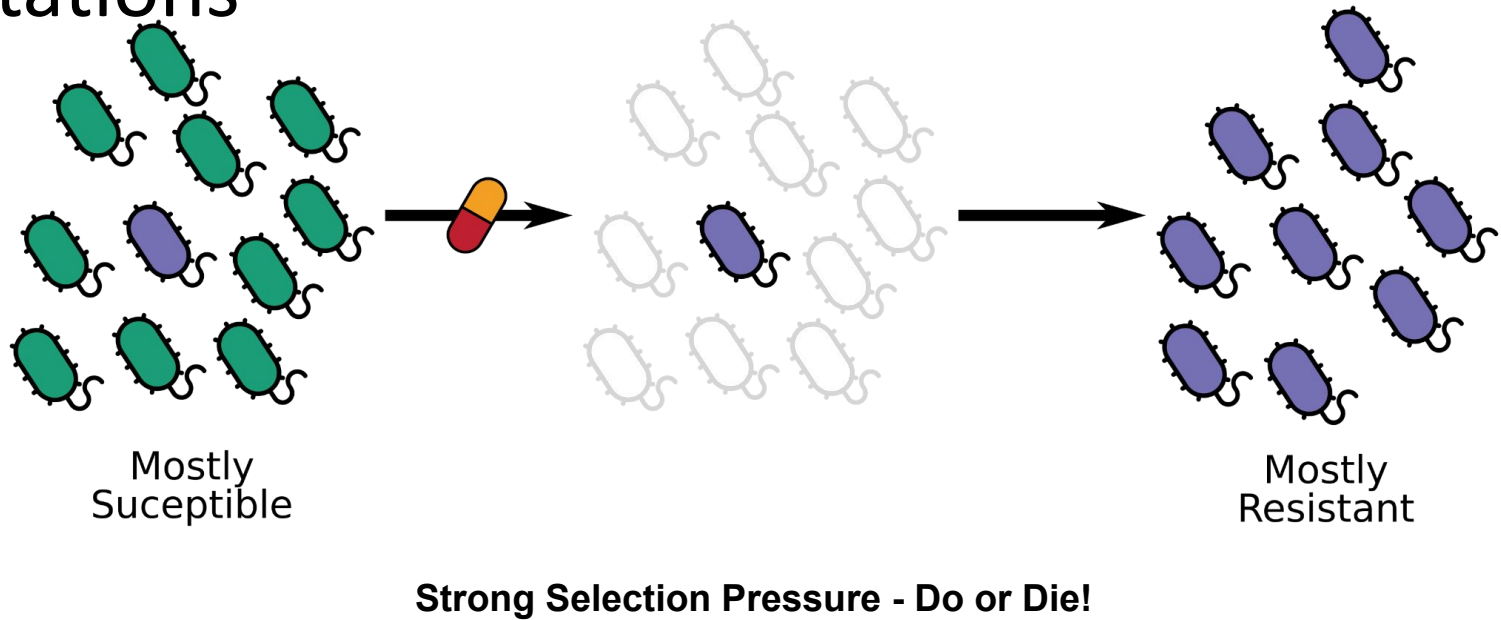# The population context of a mutation

NEUTRAL

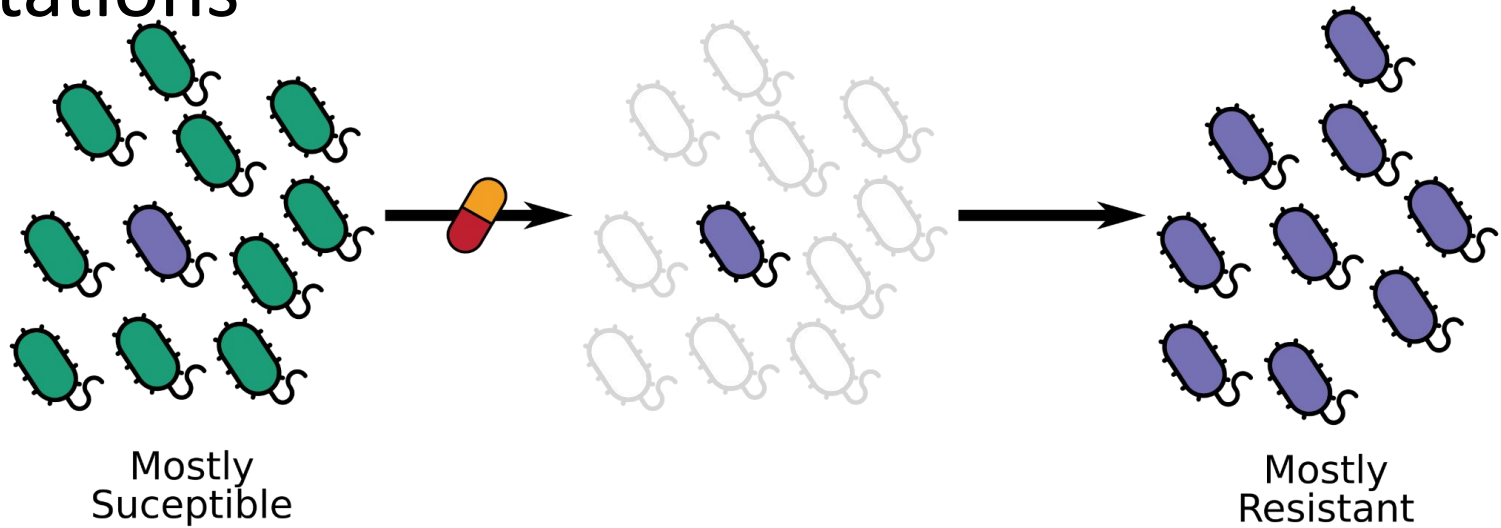Let's work through an example of selection (don't worry if the details are overwhelming!)

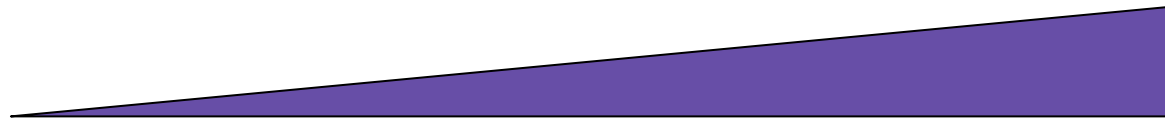# Selection (and other evolutionary forces) act on mutations

Mostly
Suceptible

Mostly
Resistant

Selection (and other evolutionary forces) act on mutations

Mostly
Suceptible

Mostly
Resistant

Strong Selection Pressure - Do or Die!

Selection (and other evolutionary forces) act on mutations
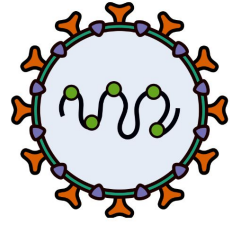

Mostly Suceptible

Mostly Resistant

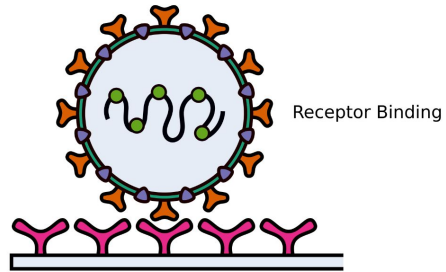**Strong Selection Pressure - Do or Die!**

Proportion of population with resistance mutation

Remdesivir is an important antiviral

Remdesivir is an important antiviral

Receptor Binding

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and
Membrane Fusion

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and
Membrane Fusion

Viral RNA
release

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and
Membrane Fusion

Viral RNA
release

RNA
replication

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and
Membrane Fusion

Viral RNA
release

RNA
replication

N
S
M
E

Transcription

Translation

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and Membrane Fusion

Viral RNA release

RNA replication

Transcription

N
S
M
E

Translation

Virion Release

Assembly and Budding

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and
Membrane Fusion

Viral RNA
release

Virion
Release

Remdesivir

RNA
replication

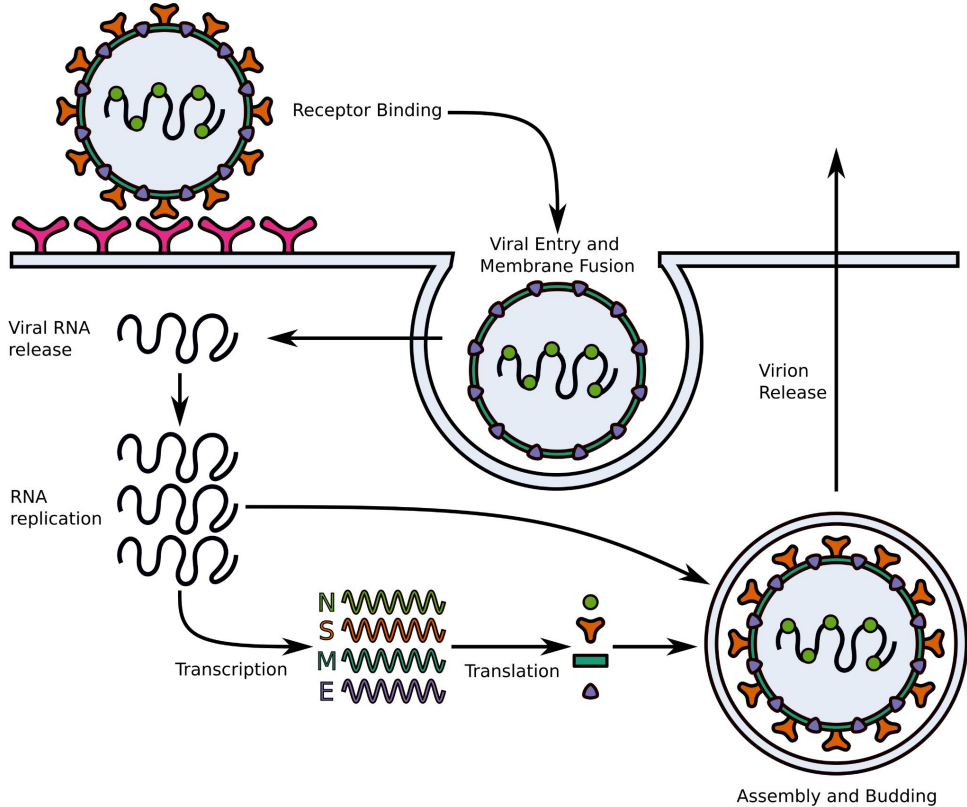Transcription

N
S
M
E

Translation

Assembly and Budding

Remdesivir is an important antiviral

Remdesivir is an important antiviral

Receptor Binding

Viral Entry and
Membrane Fusion

Viral RNA
release

RNA
replication

Remdesivir

Shortened treatment can promote resistance

No resistance

Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)

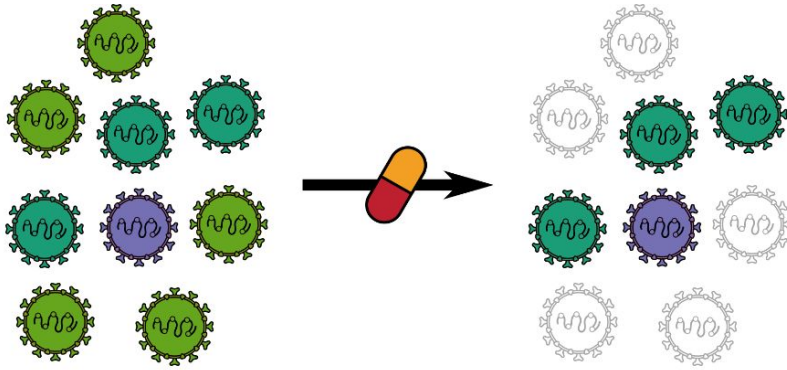# Shortened treatment can promote resistance

No resistance

Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)

Shortened treatment can promote resistance

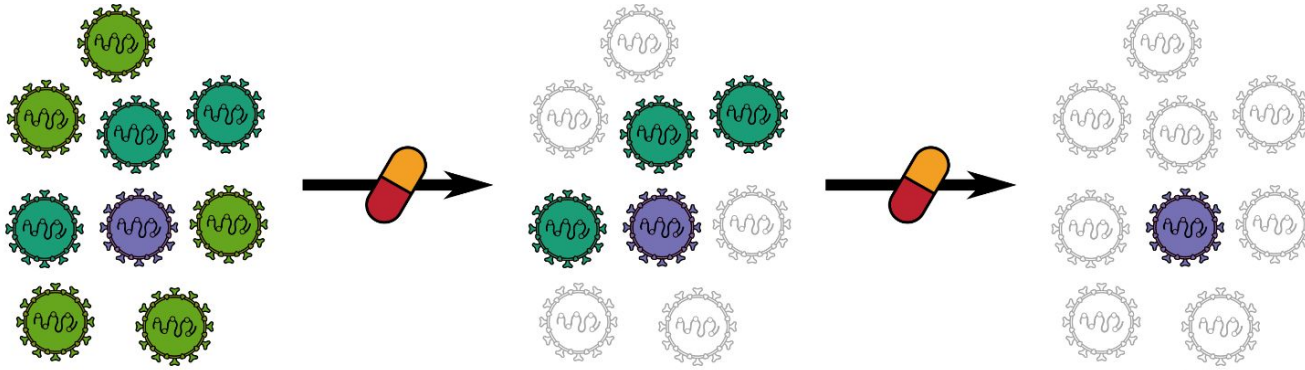No resistance

Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)

Shortened treatment can promote resistance

No resistance
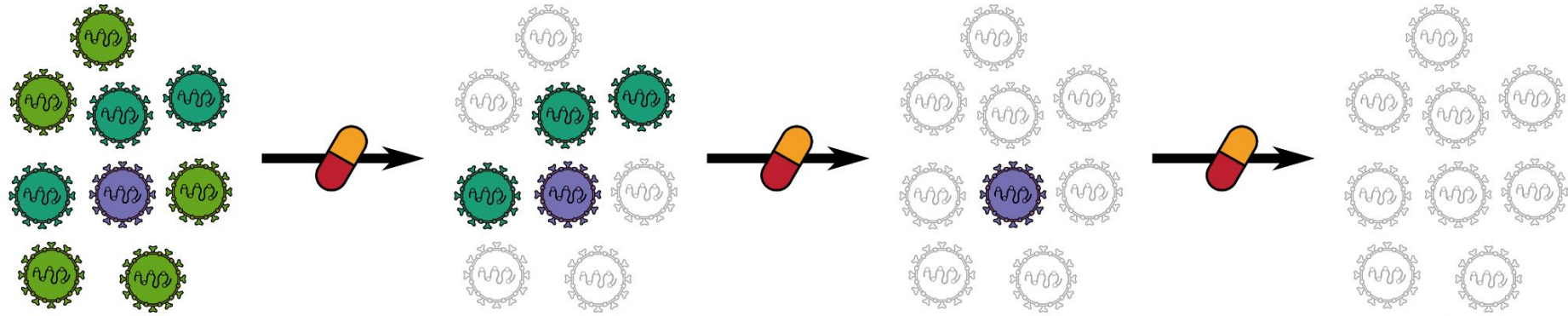
Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)

Shortened treatment can promote resistance

Mostly Susceptible

No resistance

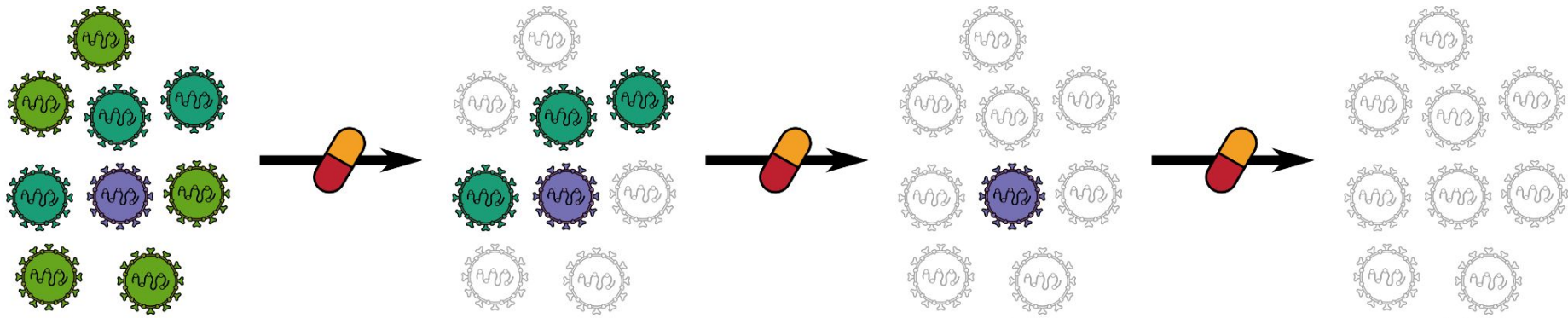Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)

Shortened treatment can promote resistance

Mostly Susceptible

Mostly Resistant

No resistance

Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)
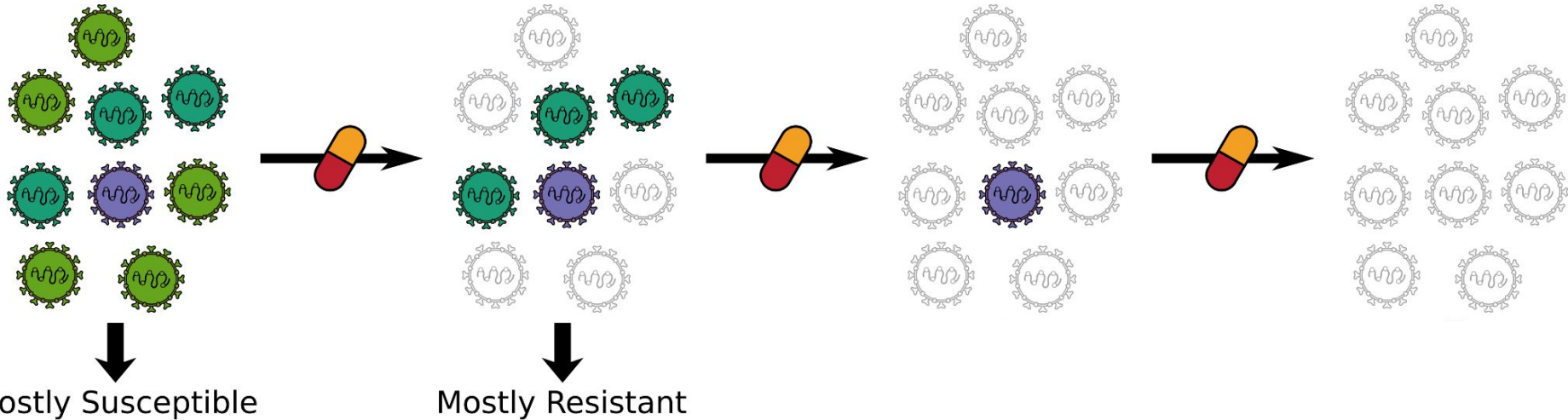
Shortened treatment can promote resistance

Mostly Susceptible

Mostly Resistant

All Resistant

No resistance

Moderate resistance

High resistance

Dead

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)
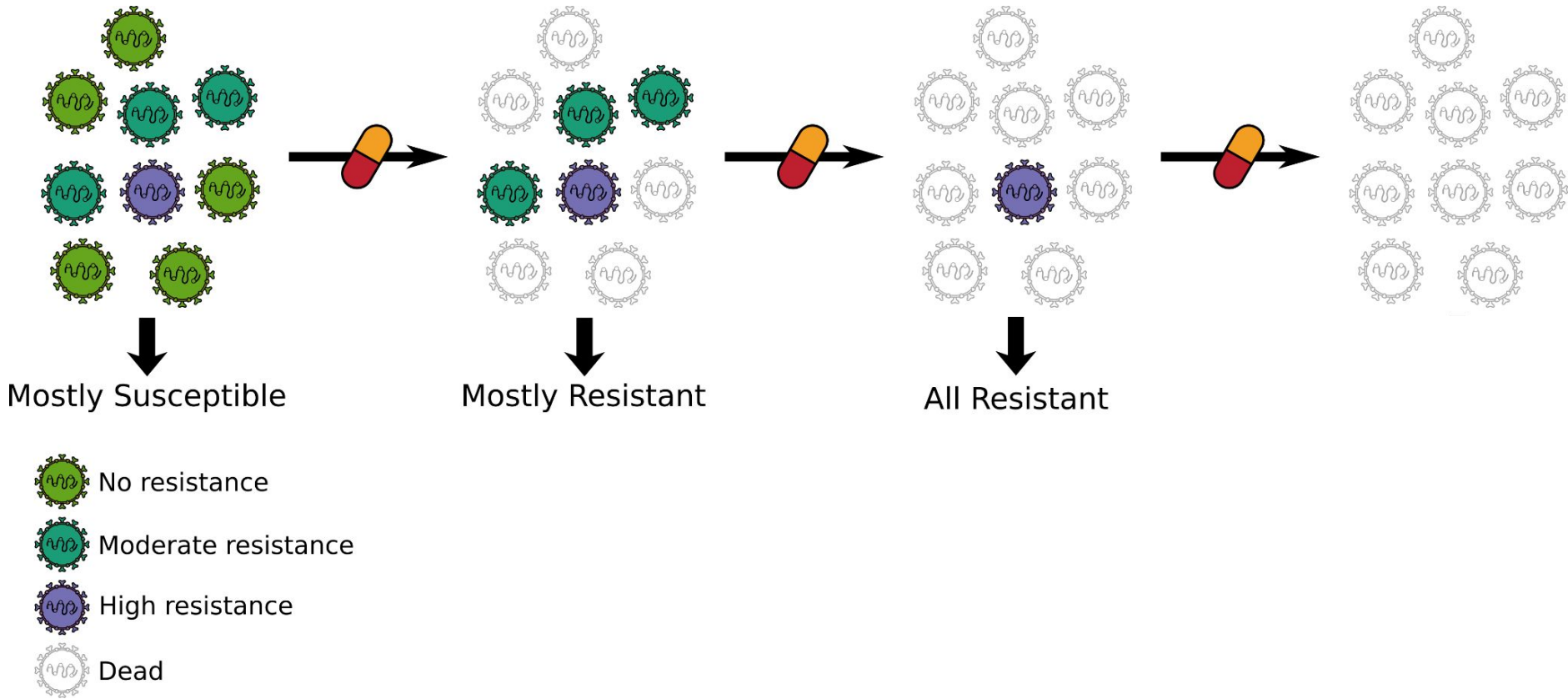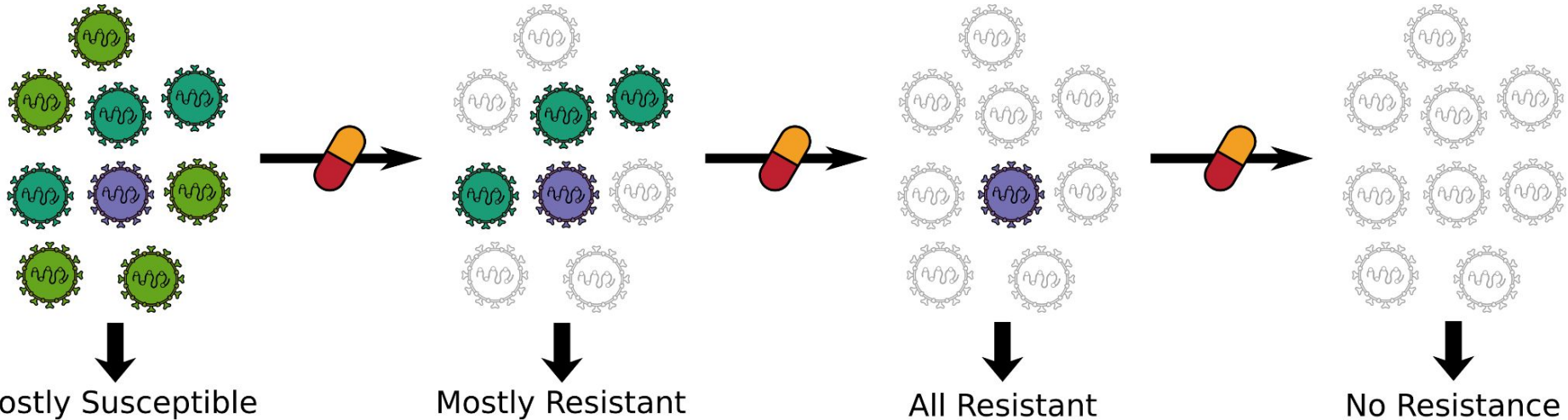
Shortened treatment can promote resistance

Mostly Susceptible → Mostly Resistant → All Resistant → No Resistance (hopefully!)

No resistance
Moderate resistance
High resistance
Dead

10 day course recommendation for remdesivir

Caveat: lots of other dynamics/forces contribute to emergence of resistance (e.g., relative fitness, infection dynamics, use-patterns/contact demographics)

So, how can we investigate this?

Remember mutations can be silent or change the protein:

- May be SILENT (remember the genetic code!)

  GCC (Ala) → GCA
  (Ala)

- Or may change a single amino acid (*missense mutation*)

  GCC (Ala) → CCC
  (Pro)

- *Or* may introduce a stop codon (*nonsense mutation*)

  CAA (Gln) → TAA
  (Stop)

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS ~ 1 : drift/neutral selection

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS > 1 : adaptive/positive selection

dN/dS ~ 1 : drift/neutral selection

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS > 1 : adaptive/positive selection

**dN/dS ~ 1** : drift/neutral selection

dN/dS < 1: purifying/negative selection

dN/dS is one way to detect selection

dN = non-synonymous mutations (normalised)

dS = synonymous mutations (normalised)

dN/dS > 1 : adaptive/positive selection

dN/dS ~ 1 : drift/neutral selection

dN/dS < 1: purifying/negative selection

Challenges:
- Mutation rates vary (across and between genomes)
- Genomes are related (mutations are non-independent)
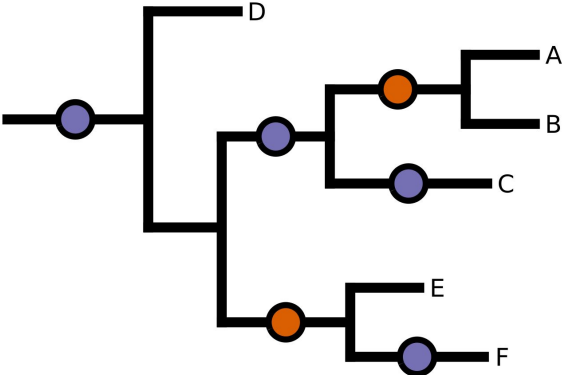
Need trees for statistical analyses
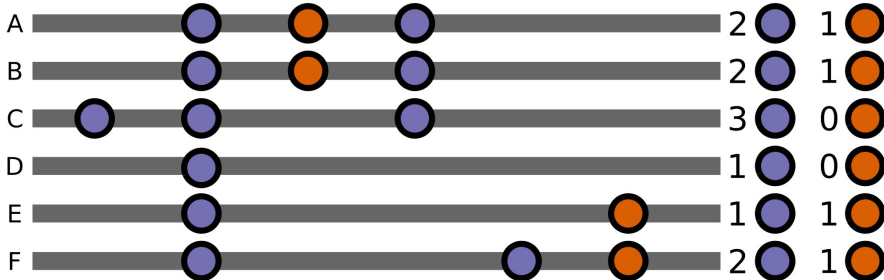
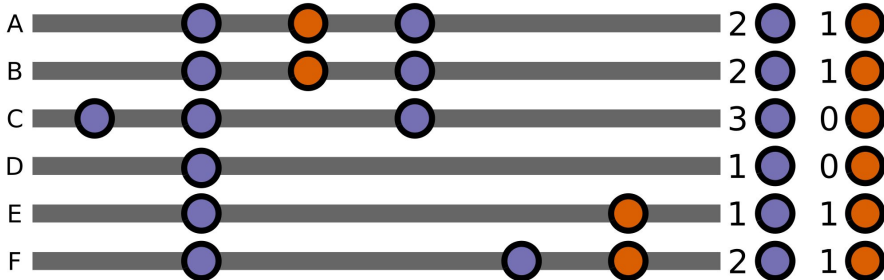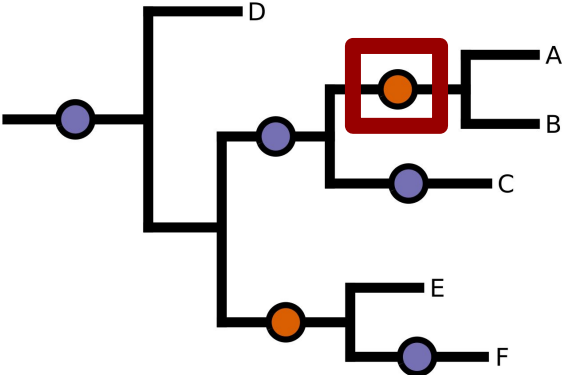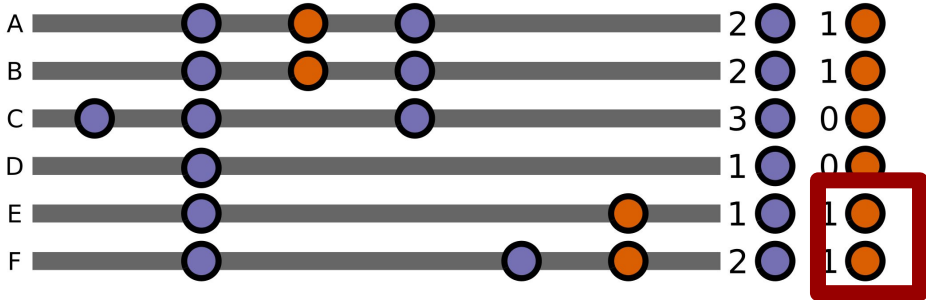# Need trees for statistical analyses

Need trees for statistical analyses
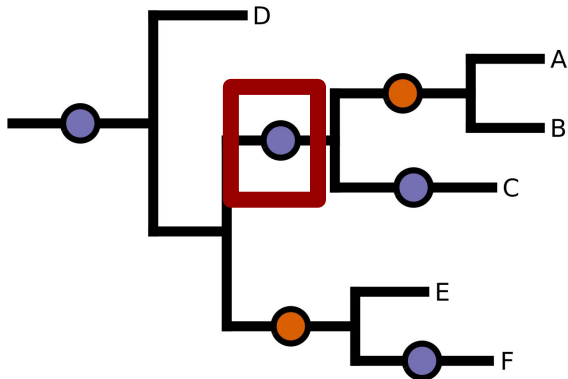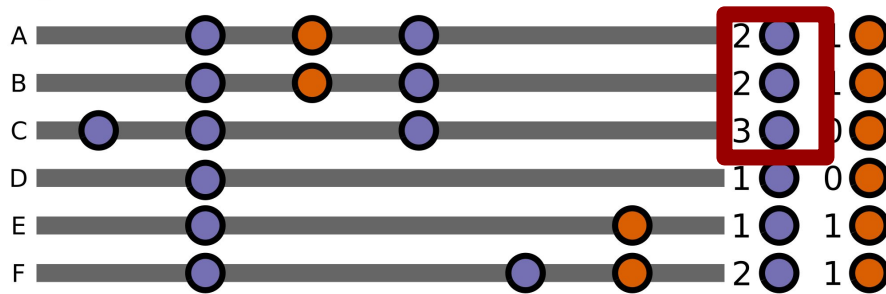
Tree forms basis of statistical analyses

Tree forms basis of statistical analyses
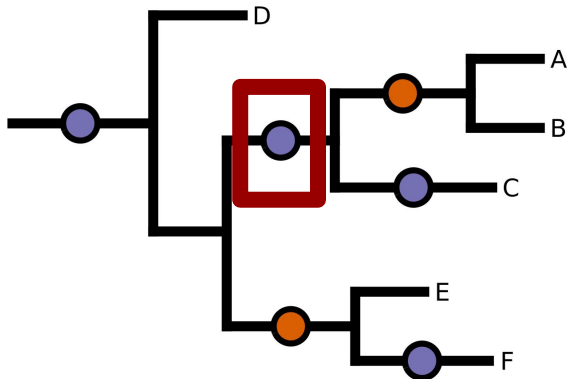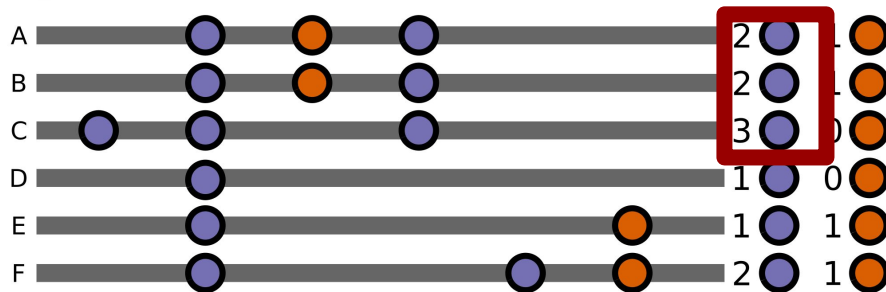
- Phylogeny captures dependency structure of genomic data
- Informs error term for models (e.g., regression)
- Typically joint modelling of phylogeny and epidemiological feature being modelled incorporating sources of uncertainty

No sign of positive selection in short RDV treatment
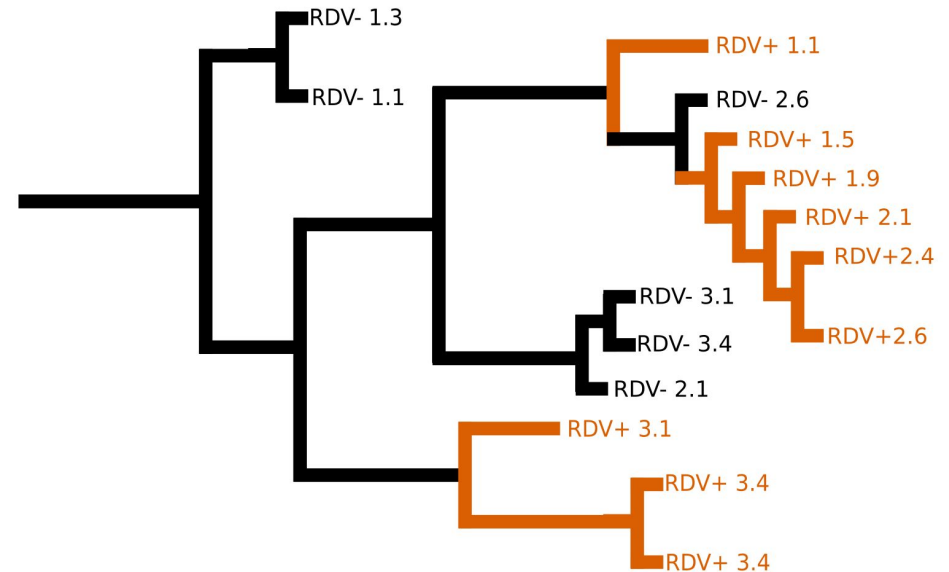
dN/dS:

   >1 positive/diversifying selection

   ~1 neutral evolution

   <1 negative/purifying selection

Codon-alignment gene-based phylogenies

Hyphy:

-   **abSREL**: positive selection in proportion

    of branches

-   **RELAX**: selection changing in set of branches

RDV- 1.3
RDV- 1.1
RDV+ 1.1
RDV- 2.6
RDV+ 1.5
RDV+ 1.9
RDV+ 2.1
RDV+2.4
RDV+2.6
RDV- 3.1
RDV- 3.4
RDV- 2.1
RDV+ 3.1
RDV+ 3.4
RDV+ 3.4

No sign of positive selection in short RDV treatment

dN/dS:

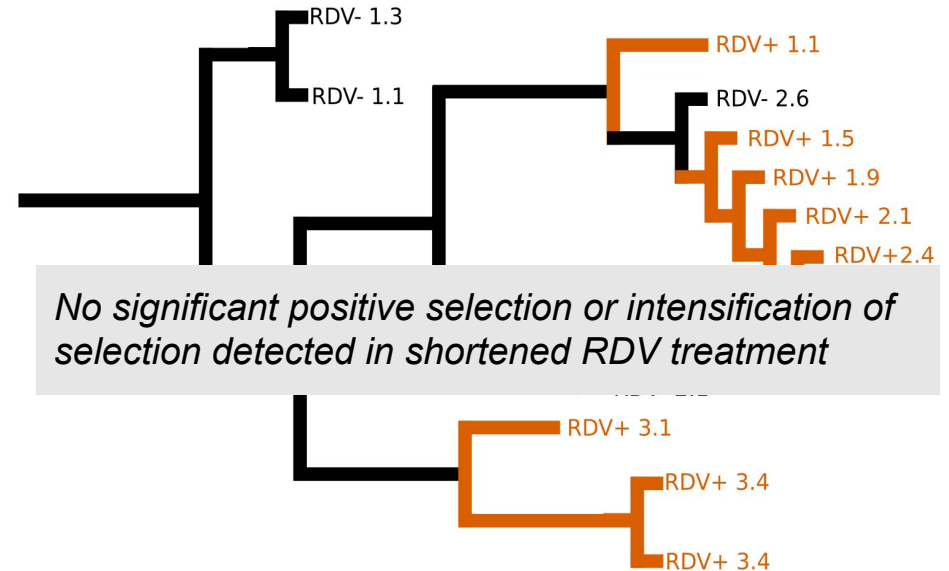   >1 positive/diversifying selection

   ~1 neutral evolution

   <1 negative/purifying selection

**Codon-alignment gene-based phylogenies**

Hyphy:

- **abSREL**: positive selection in proportion

   of branches

- **RELAX**: selection changing in set of branches

No significant positive selection or intensification of selection detected in shortened RDV treatment

RDV- 1.3
RDV+ 1.1
RDV- 1.1
RDV- 2.6
RDV+ 1.5
RDV+ 1.9
RDV+ 2.1
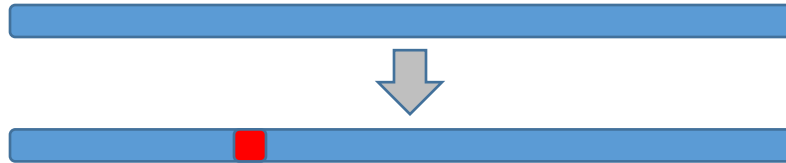RDV+2.4
RDV+ 3.1
RDV+ 3.4
RDV+ 3.4

Let's go back to the swans to the question of how evolution can lead to new species!
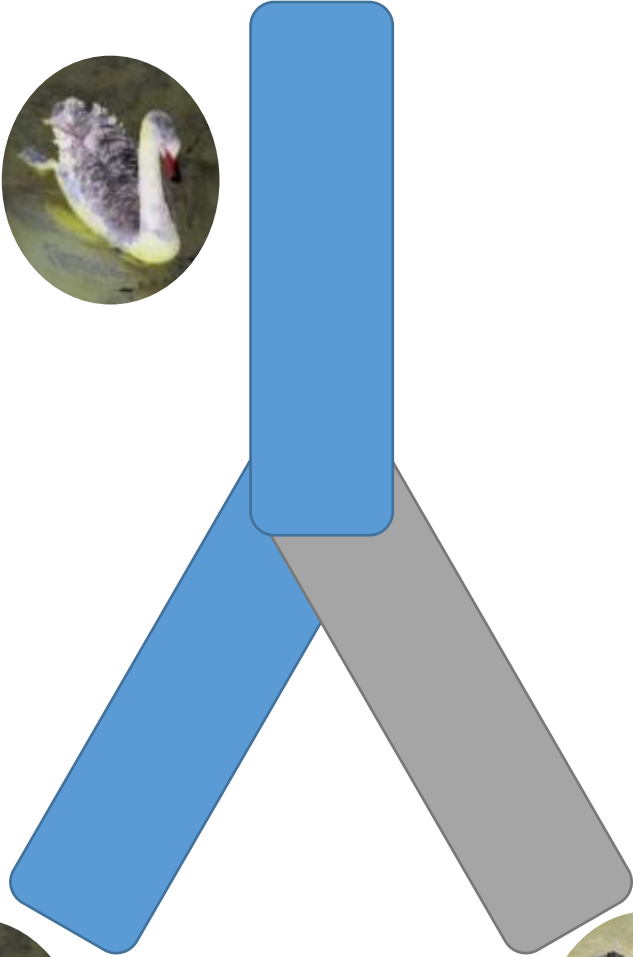
The population context of a mutation
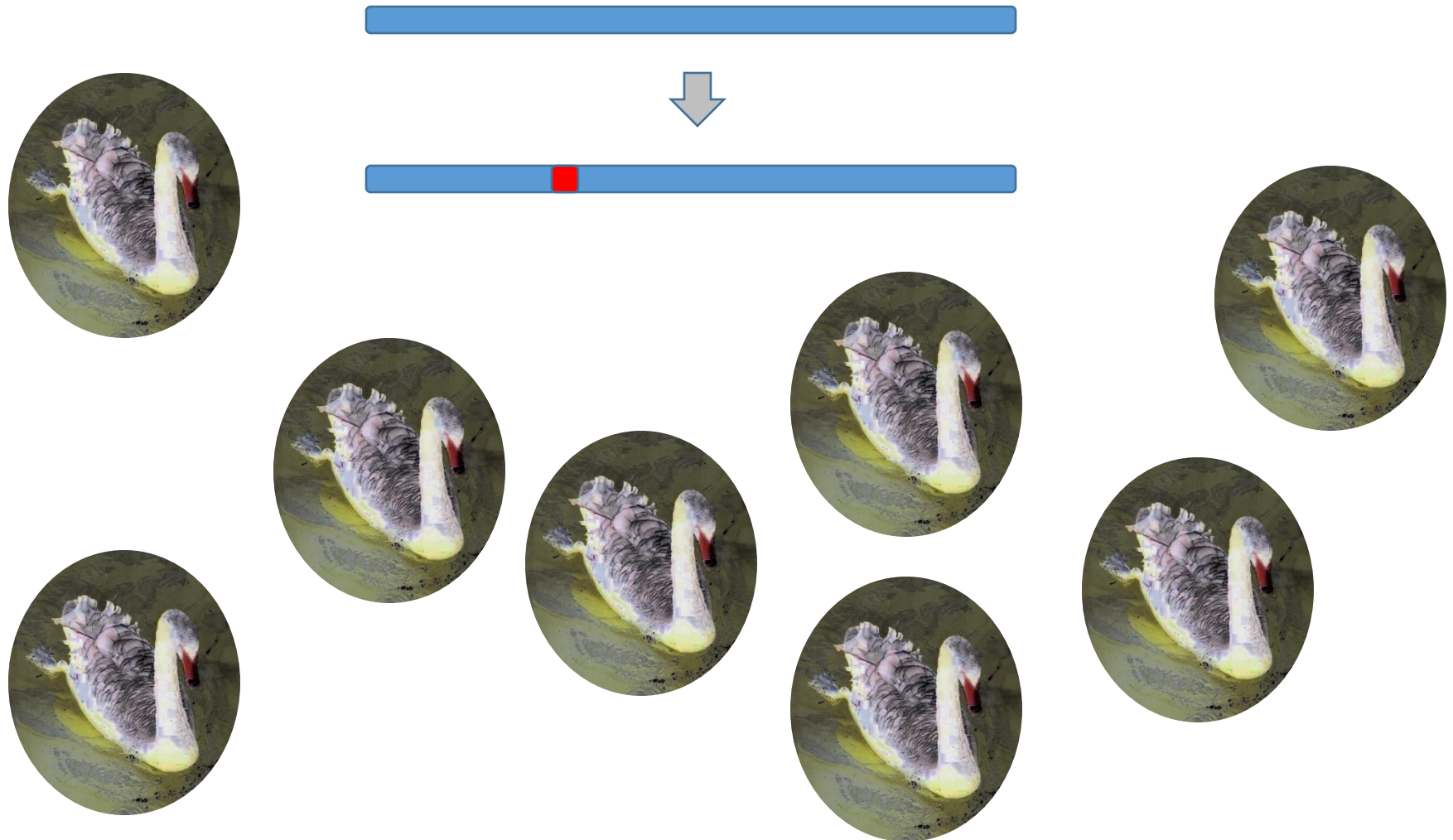
ADVANTAGEOUS

Advantageous, with boundaries

https://www.geographicus.com/P/AntiqueMap/Australia-johnson-1864
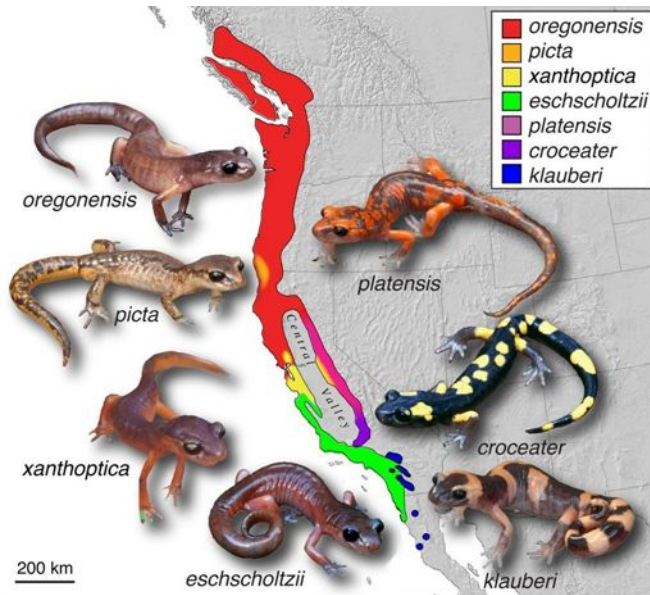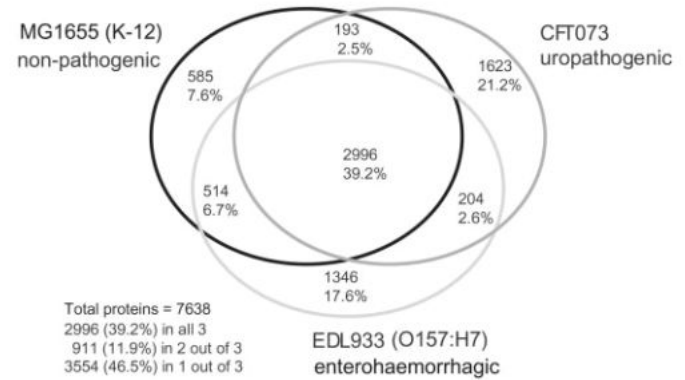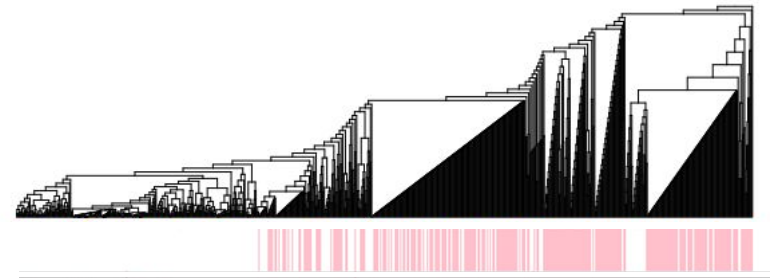
ANCESTOR

DESCENDANTS
(speciation!)

Speciation ain't pretty

Ring species: most neighbours can interbreed,
more-distant populations cannot

Fewer than half of genes in these *E. coli* are
found in all three of them

The erythromycin resistance gene *ErmB* is found
in a scattered subset of *Enterococcus* genomes
(pink = present in genome)

Devitt et al (2011) *BMC Evolutionary Biology*
Welch et al (2002) *PNAS*
Sanderson et al (2022) *MGen*

Relationships among sequences
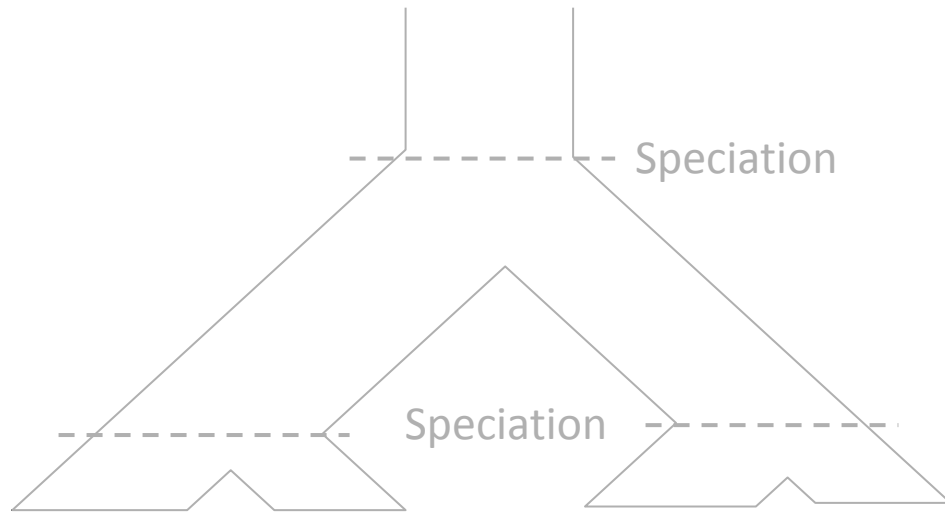
- HOMOLOGY
- ORTHOLOGY
- PARALOGY

Remember: duplications

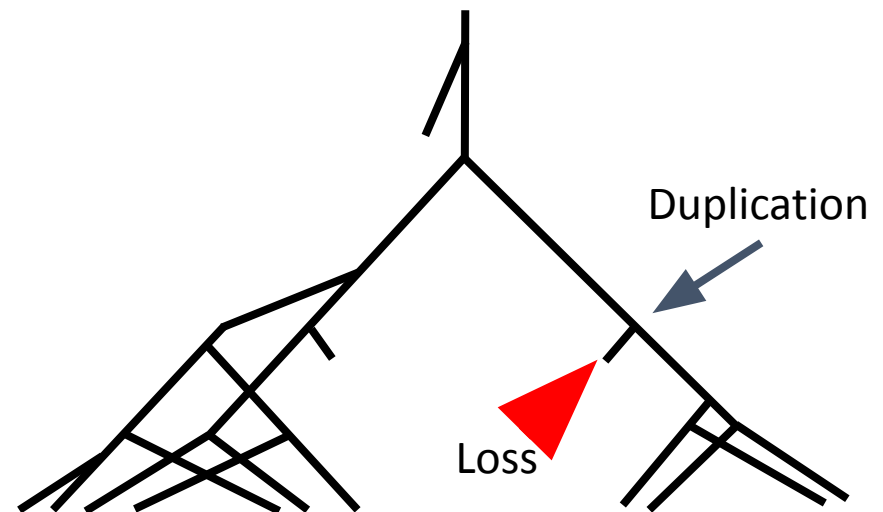AAA ACC **GAA TC**A CCG GAT → AAA ACC **GAA TCG AAT C**A CCG GAT

We can have whole-gene duplications too!

Bifurcations at the gene and species level

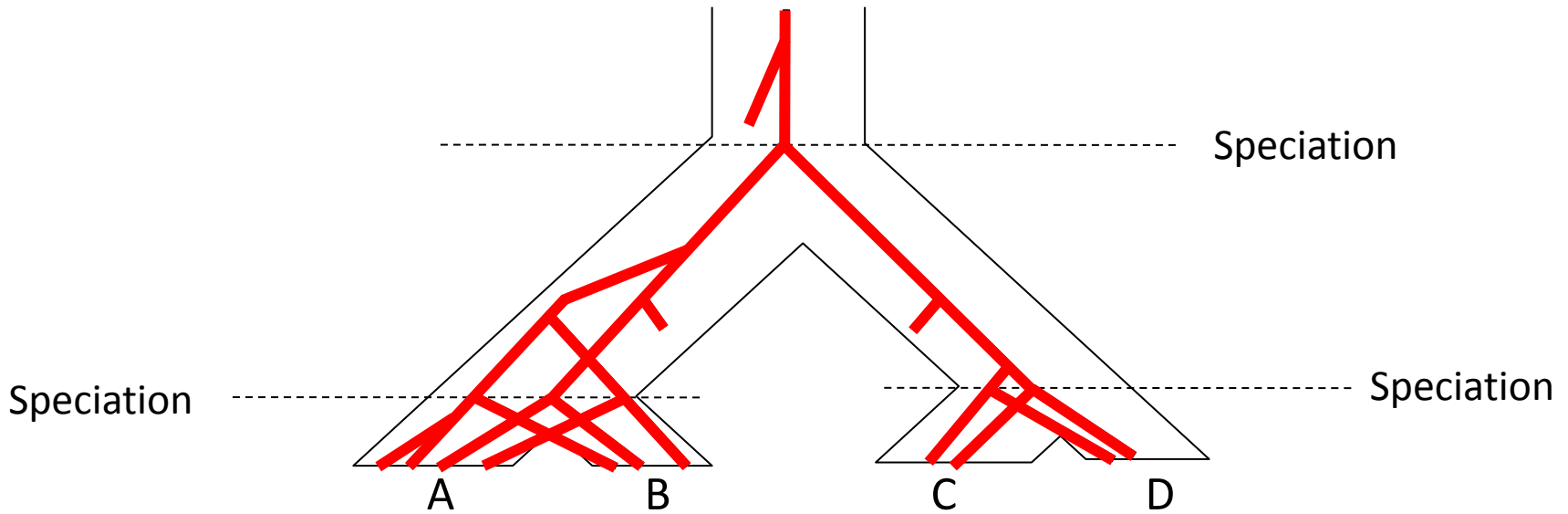When speciation occurs, most genes are inherited by both descendants

Speciation

Speciation

Genes can duplicate and disappear at any time, independently of speciation

Duplication

Loss

Homology

- Any set of sequences (or, more generally, traits) that are descended from a common ancestor

Speciation
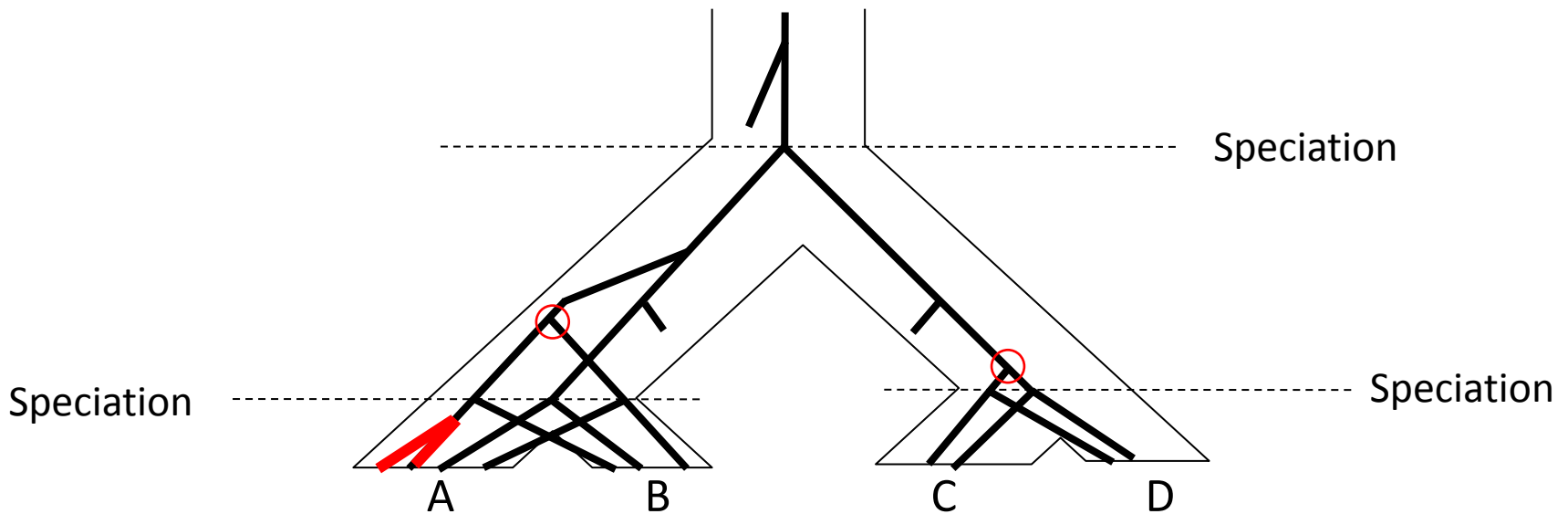
Speciation

Speciation

A   B   C   D

Orthology

- Genes whose last common ancestor occurred at a speciation event
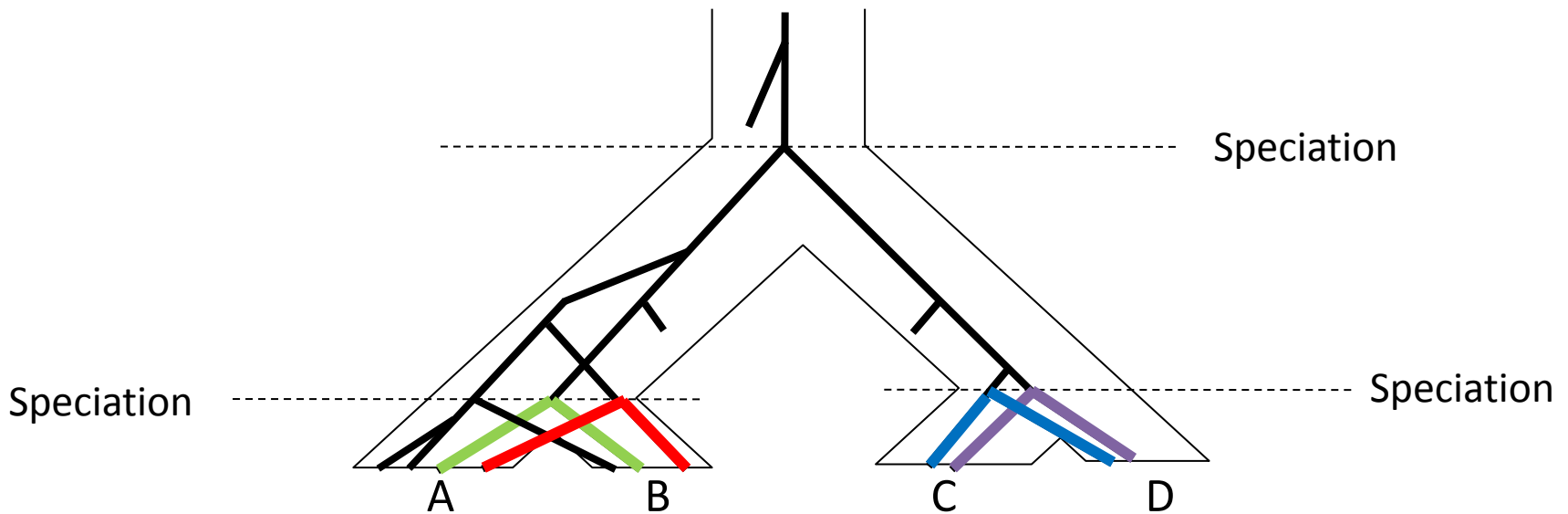
Paralogy

- Two genes whose last common ancestor arose via a gene duplication event

Why do we care about the difference?

- The "ortholog conjecture"

- Orthologs are (kinda) the "same gene" in different species, so they are more likely to have evolved in similar ways and retained the same or similar function

- Paralogs are descendants in a single lineage – they are often lost, but since someone is still doing the original job, they may mutate to specialize over time, or "explore" new functions
  - Some antibiotic-resistance genes are paralogs of much more "innocent" genes
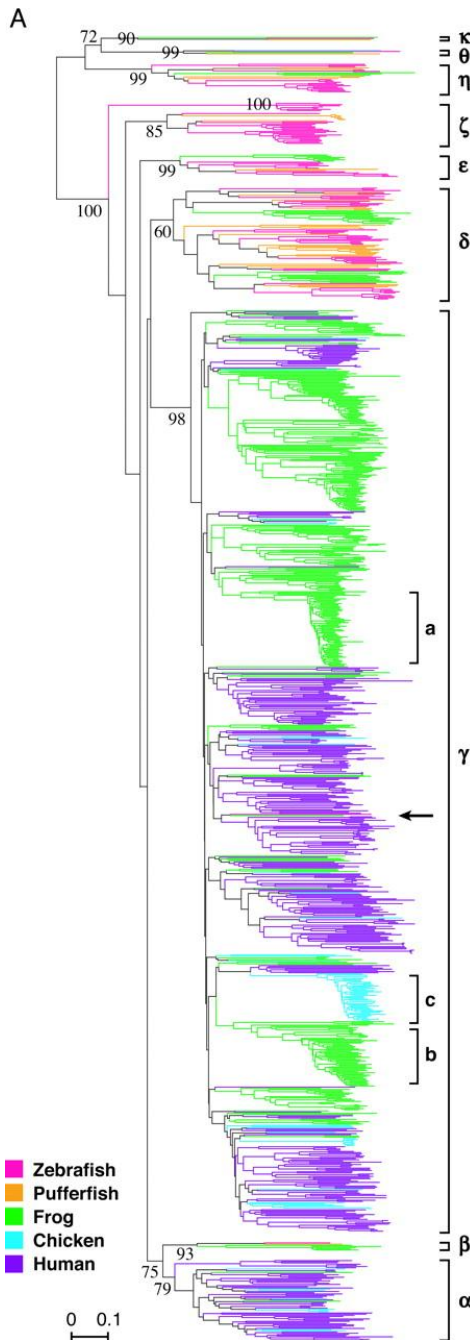
Olfactory receptors

Many different target molecules, but all ORs:
- Sit in the cell membrane
- Recognize a specific set of odor molecules
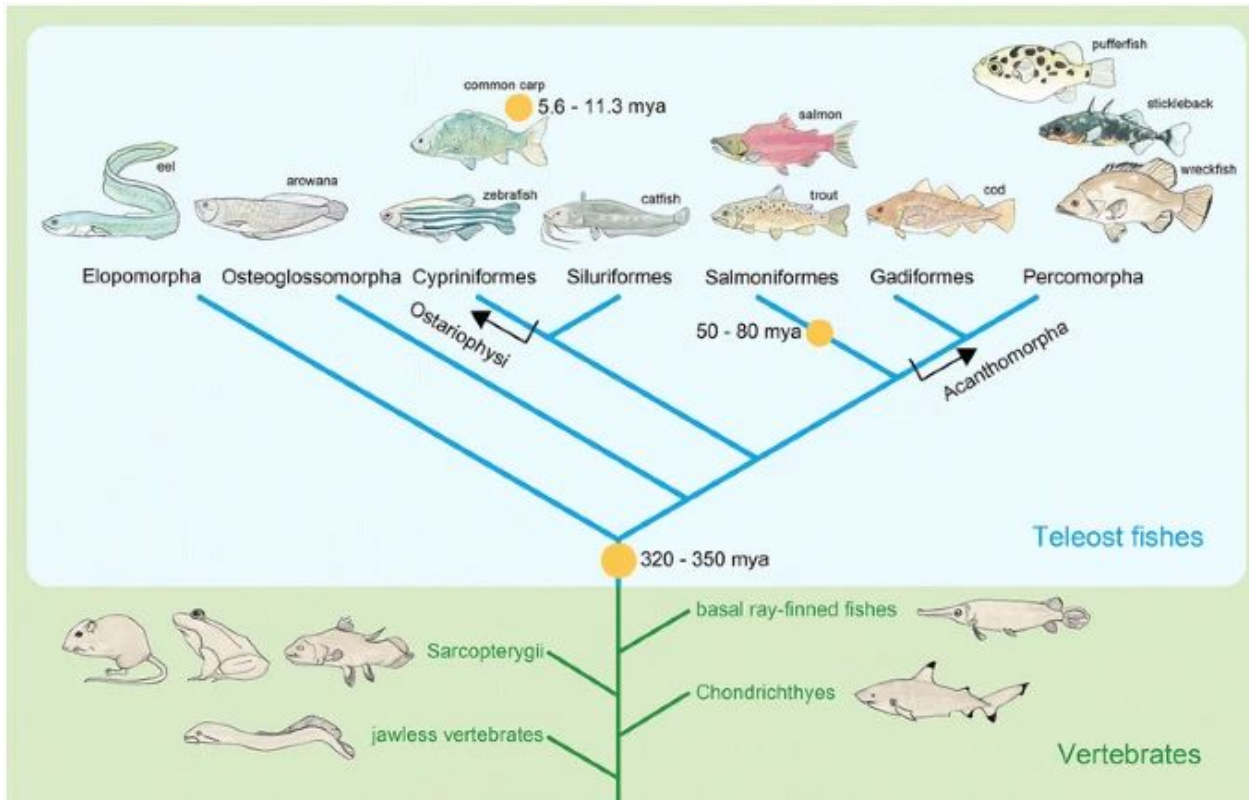- Trigger neural signalling
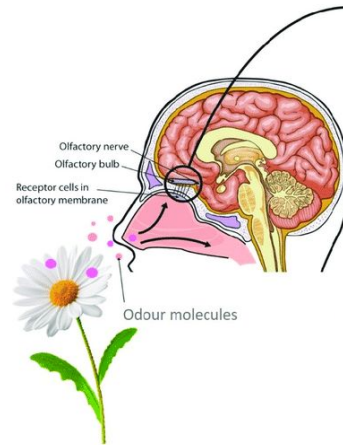
These sequences are all <u>homologous</u>, but trying to disentangle paralogy and orthology is kind of a mess

Niimura and Nei (2005) *PNAS*
Vilela et al (2019) *Foods*

58

Whole-genome duplication – mega paralogy!

Descendants have 2x the genetic material of the ancestor

Many of the newly duplicated genes are lost or mutate and gain new / specialized functions

Glasauer and Neuhauss (2004) *Molecular Genetics and Genomics*

Summary

Mutations change the gene sequence, and often the protein sequence as well

Selection weeds out less-effective versions of the protein

By tracing the history of mutation and selection, we can recover the evolutionary relationships among biological sequences

Summary

All genes that are related by descent from a common ancestor are termed homologous

Orthologs are sometimes thought of as 'the same gene in different species'

Paralogs arise by duplication, and can lead to proteins with new functions