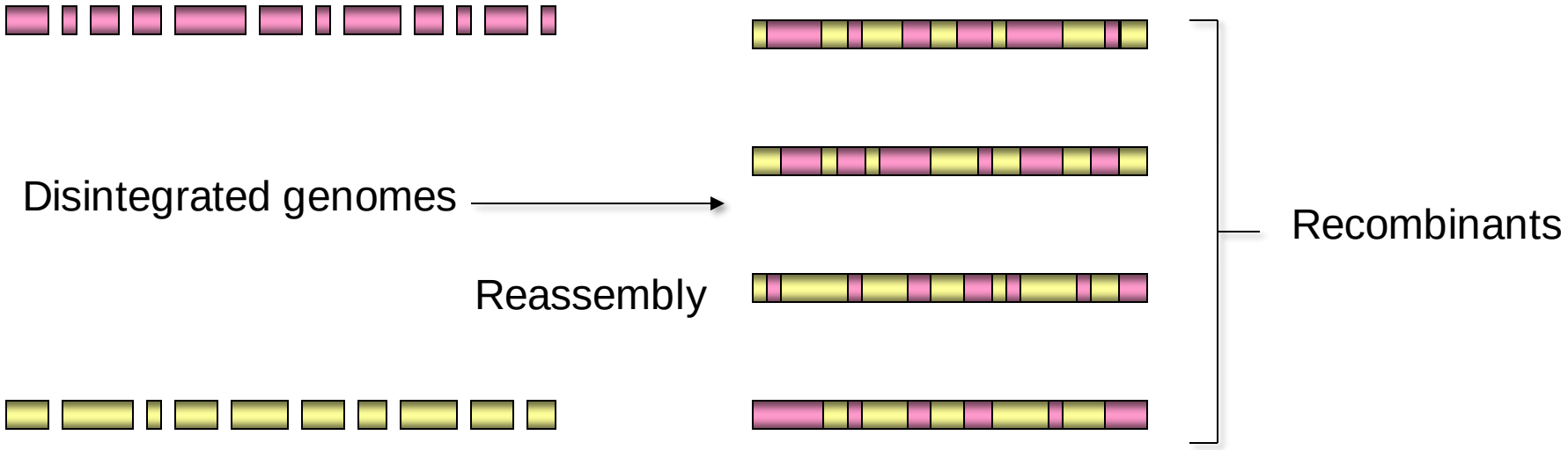
What is recombination and how can it be factored into evolutionary analyses?

How does recombination occur?



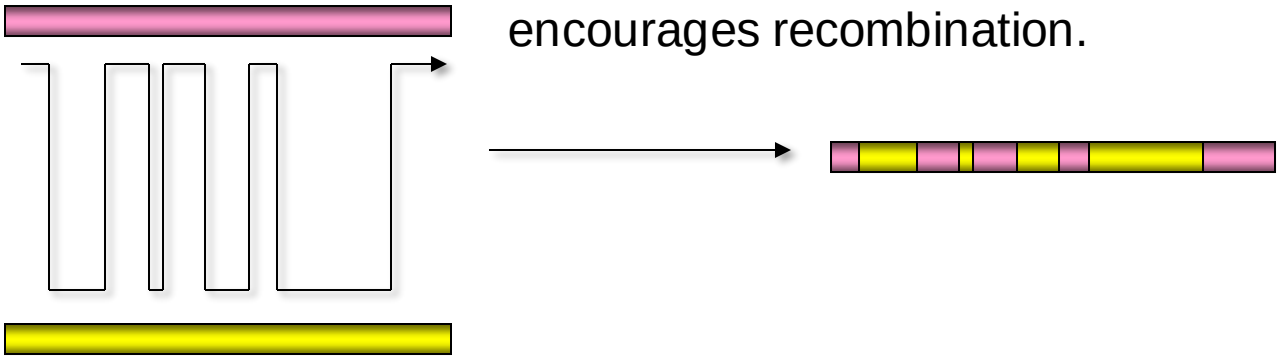
Mechanisms:

Double stranded break and repair - eg Cellular organisms DNA Viruses

Disintegration and repair - eg some bacteria

How does recombination occur?

For HIV packaging of two separate genomes into every virus particle and frequent infection of individual cells with more than one virus particle encourages recombination.



Template switching during reverse transcription

Mechanisms:

Double stranded break and repair - eg Cellular organisms DNA Viruses

Disintegration and repair - eg some bacteria

Template switching during reverse transcription - eg retroviruses like HIV

Why is recombination important?

Almost all genomes undergo recombination

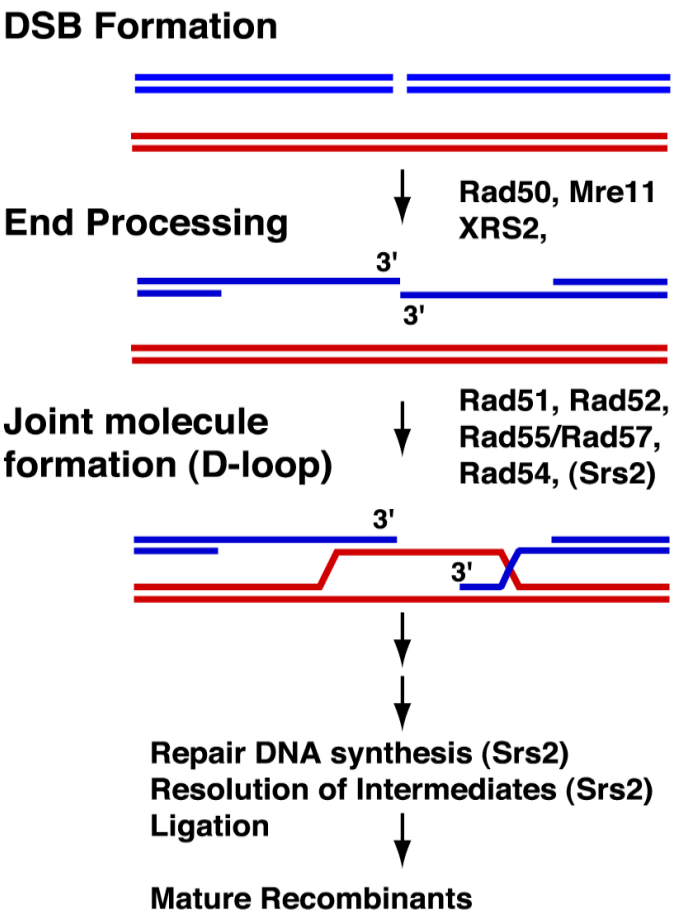
Why do they bother?

Why is recombination important?

Almost all genomes undergo recombination
Why do they bother?

Double stranded break repair?

Large DNA molecules often break and replication forks frequently stall. Recombination is a good way of repairing these.



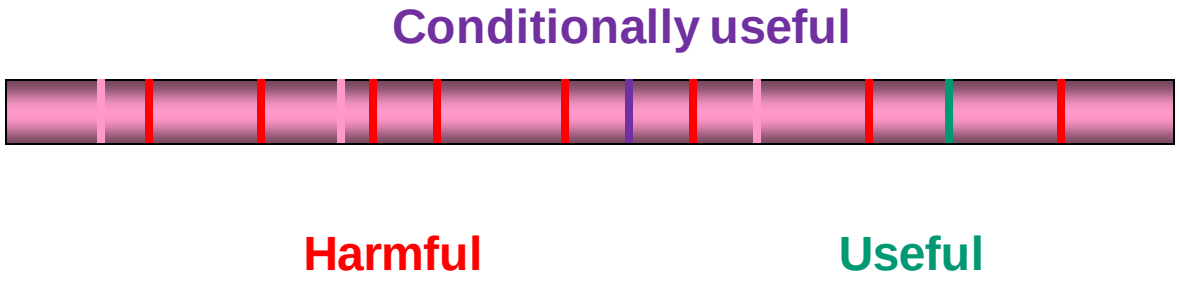
Why is recombination important?

Almost all genomes undergo recombination
Why do they bother?

Double stranded break repair?

Repair of harmful mutations?

Most mutations are harmful. It is much harder to repair harmful mutations by reversion than it is to repair them by recombination.



Why is recombination important?

Almost all genomes undergo recombination
Why do they bother?

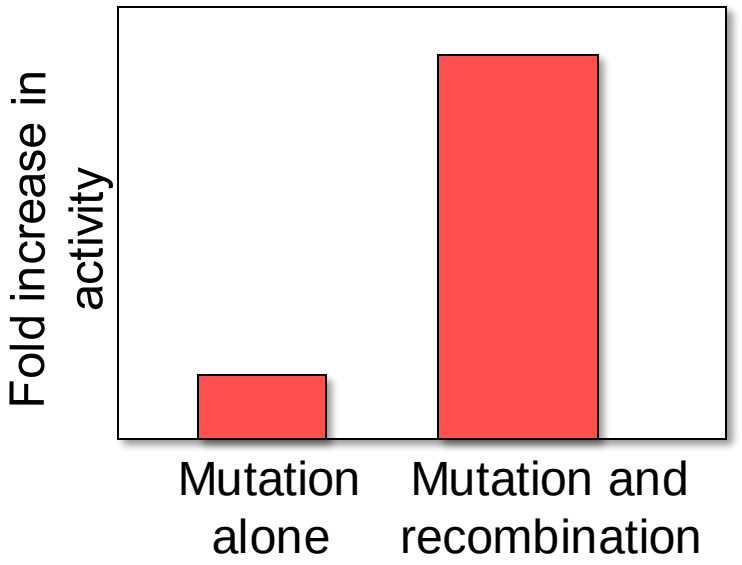
Double stranded break repair?

Repair of harmful mutations?

Better exploration of sequence space?

Given **enough parental sequence diversity and enough template switching** during a single round of replication recombination can provide access to many more locations in sequence space than are accessible by mutation

Eg DNA shuffling experiments



Why is recombination important?

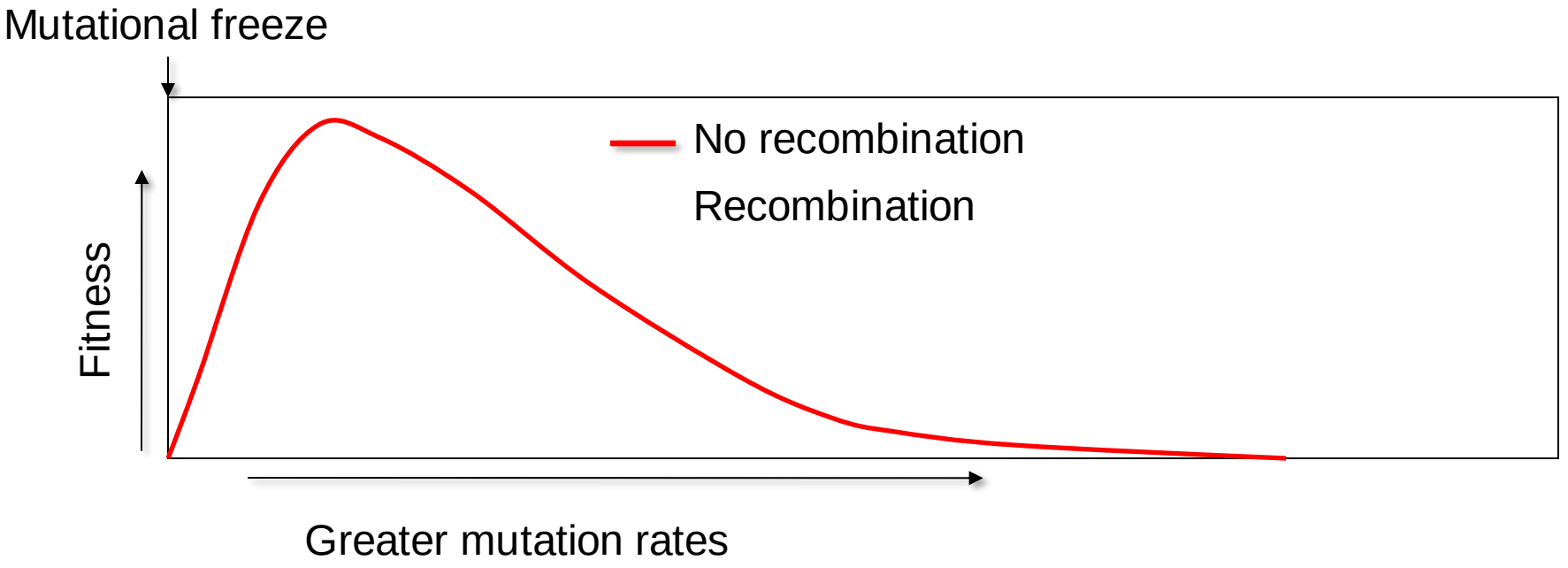
Almost all genomes undergo recombination
Why do they bother?

Double stranded break repair?	Yes
Repair of harmful mutations?	Probably
Better exploration of sequence space?	Probably

Besides its role in double strand break repair the proposed evolutionary benefits of recombination are questionable

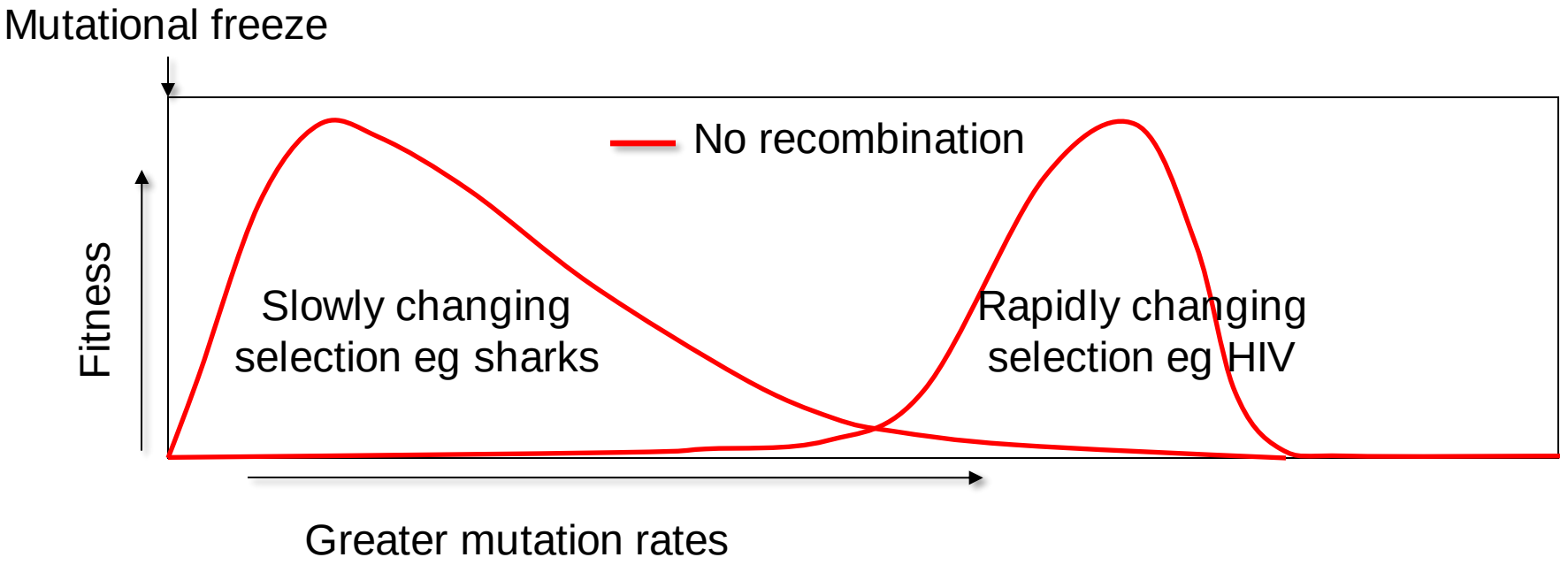
Recombination and mutation

Muller's Ratchet => It is very difficult to repair harmful mutations by mutation



Recombination and mutation

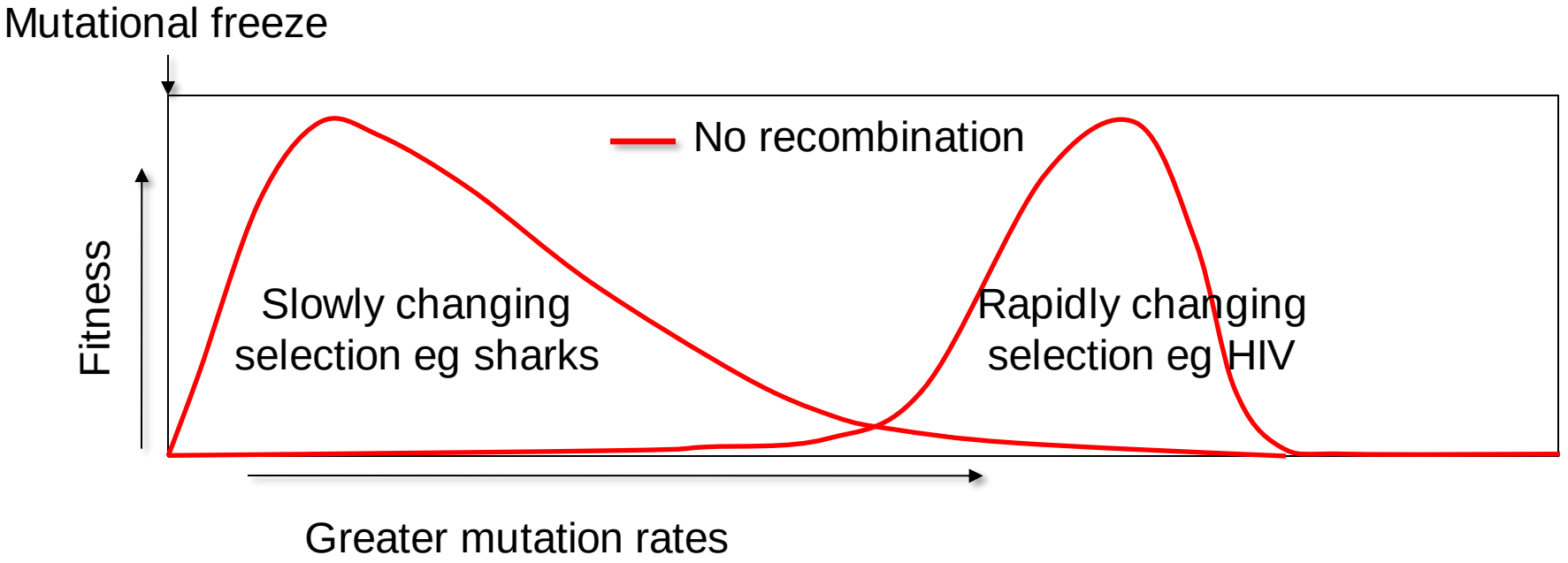
Muller's Ratchet => It is very difficult to repair harmful mutations by mutation



NB. The optimal mutation rate varies from organism to organism

Recombination and mutation

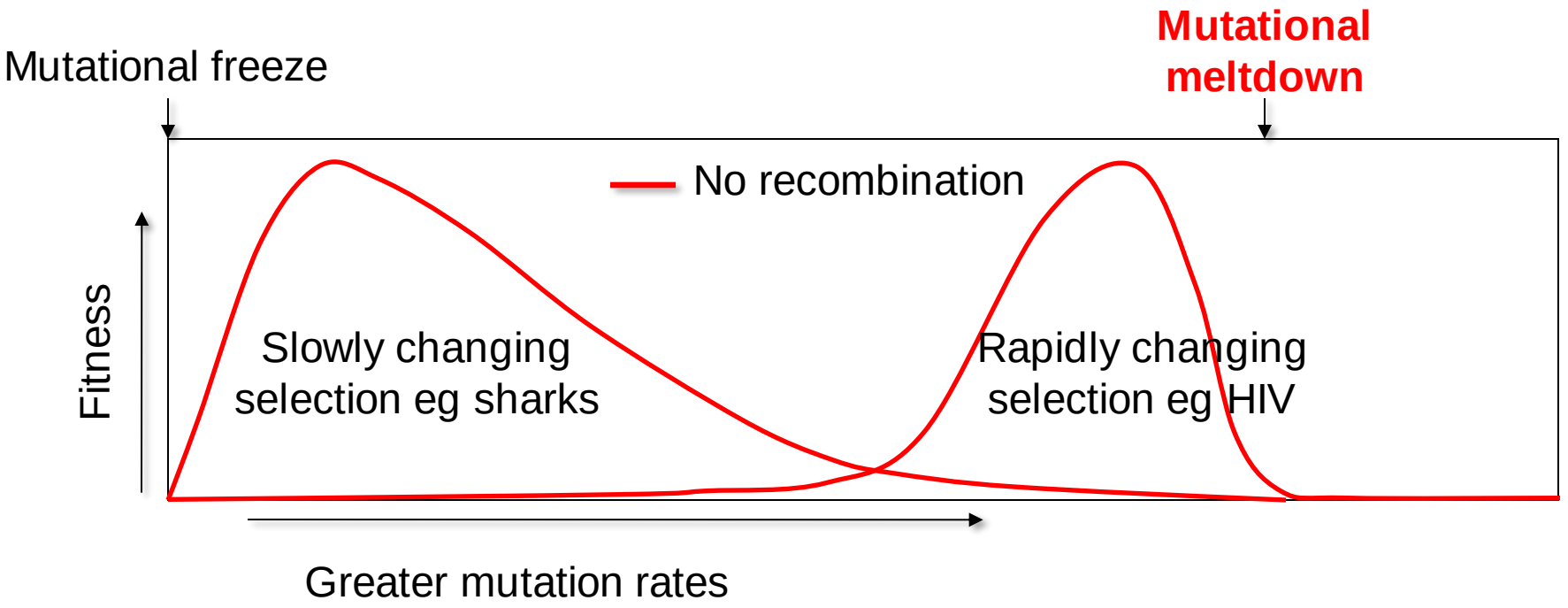
Muller's Ratchet => It is very difficult to repair harmful mutations by mutation



BUT…….. it tends to be 1 mutation per genome every three replication cycles.

Recombination and mutation

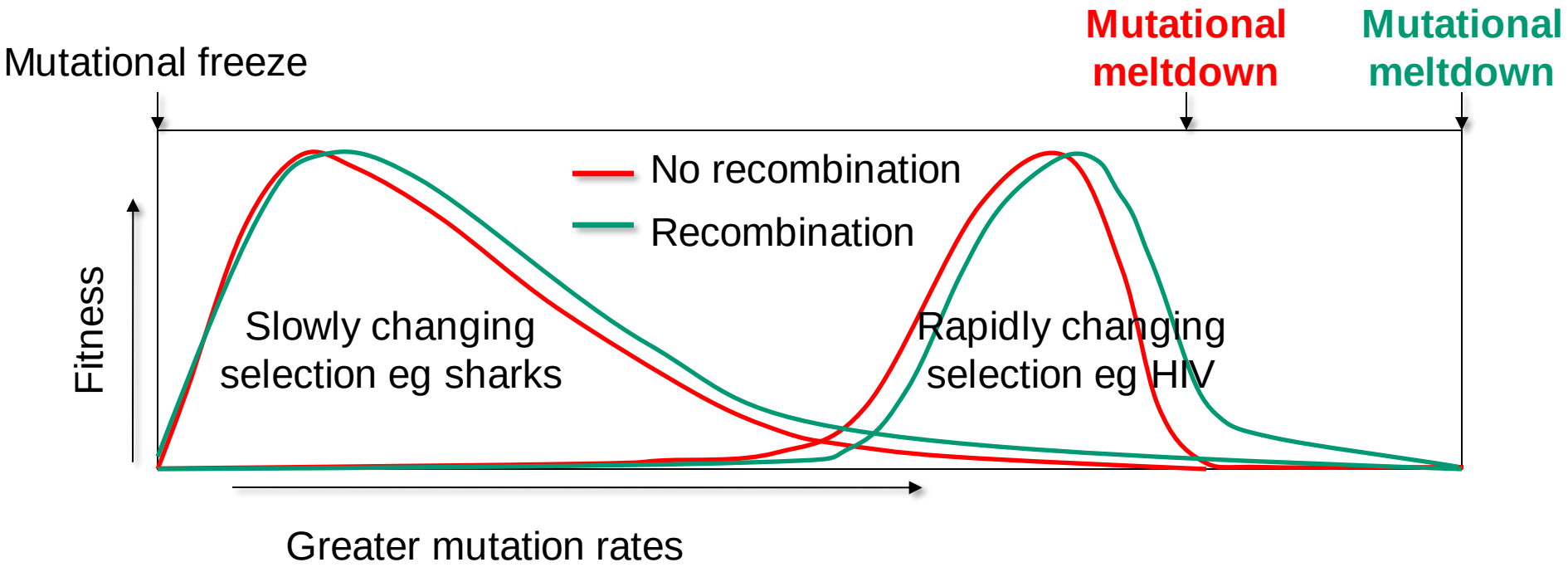
Muller's Ratchet => It is very difficult to repair harmful mutations by mutation



Mutational meltdown would occur if there were ~1 or more mutations per genome per replication cycle.

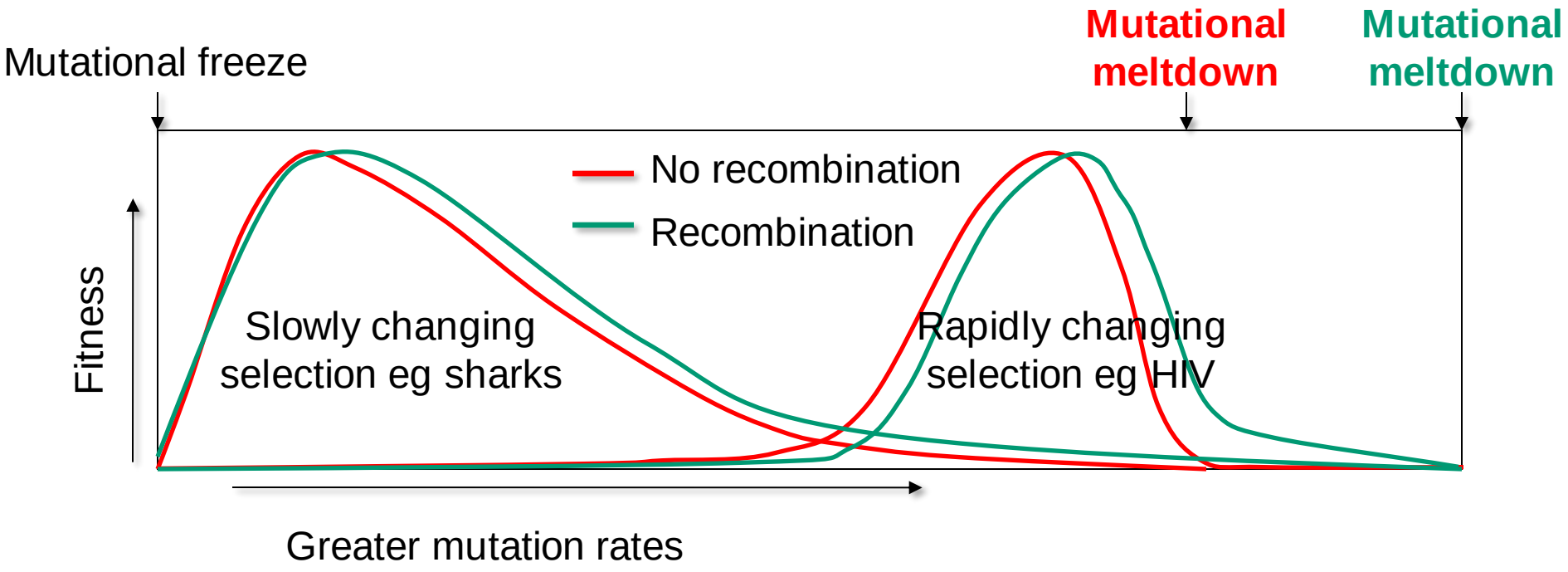
Recombination and mutation

Muller's Ratchet => It is very difficult to repair harmful mutations by mutation



Recombination and mutation

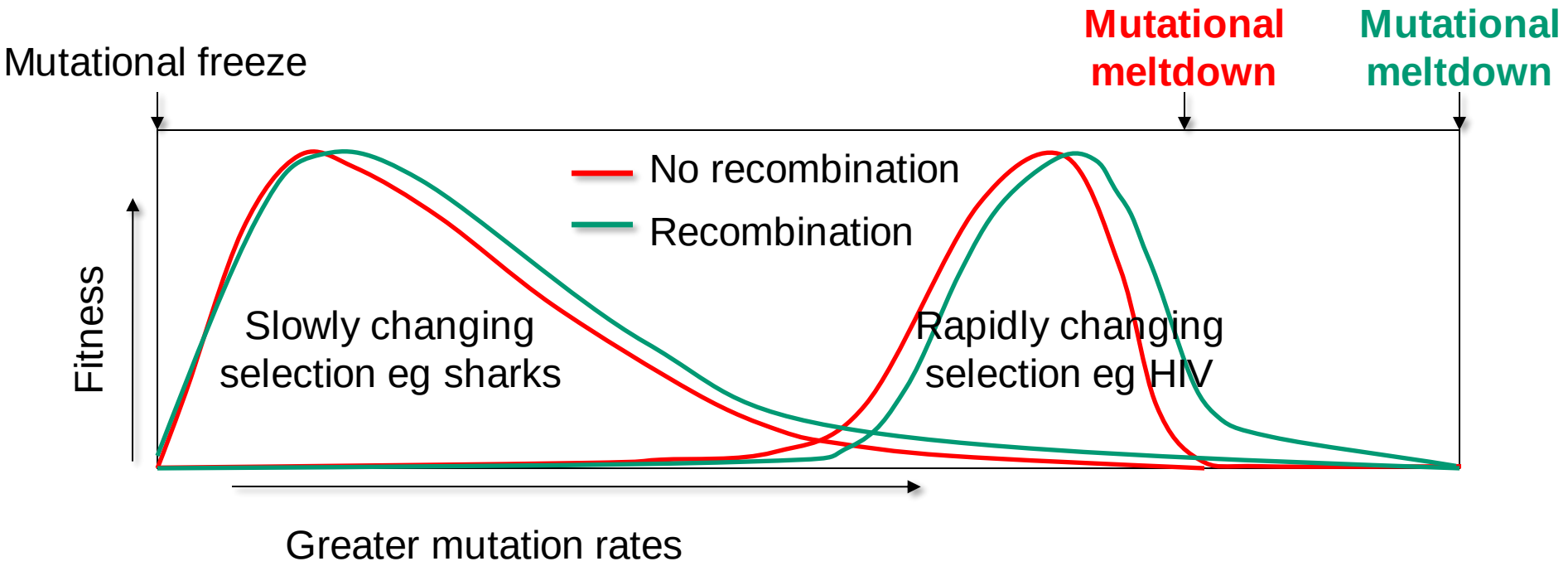
Muller's Ratchet => It is very difficult to repair harmful mutations by mutation



Recombination can "repair" harmful mutations and uncouple beneficial mutations from harmful mutations.

Recombination and mutation

Muller's Ratchet => It is very difficult to repair harmful mutations by mutation

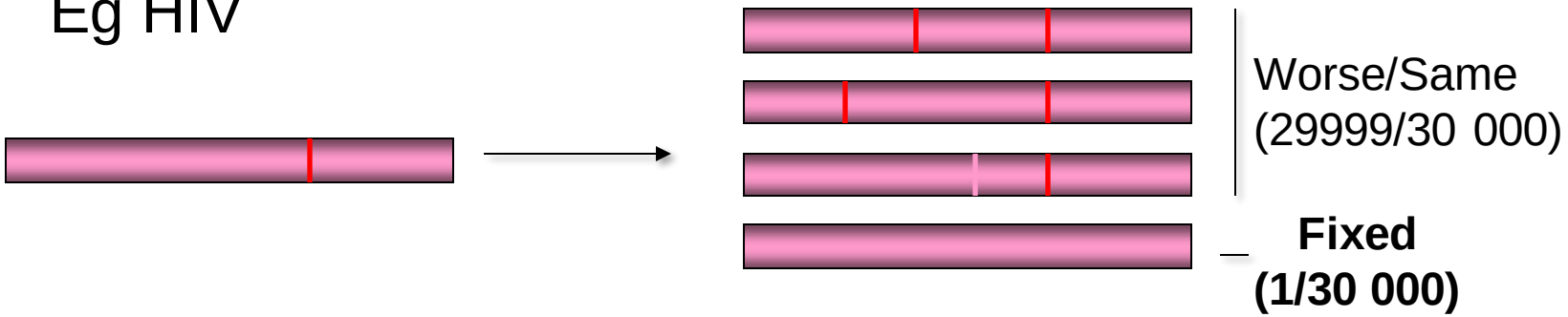


Recombination can "repair" harmful mutations and uncouple beneficial mutations from harmful mutations.

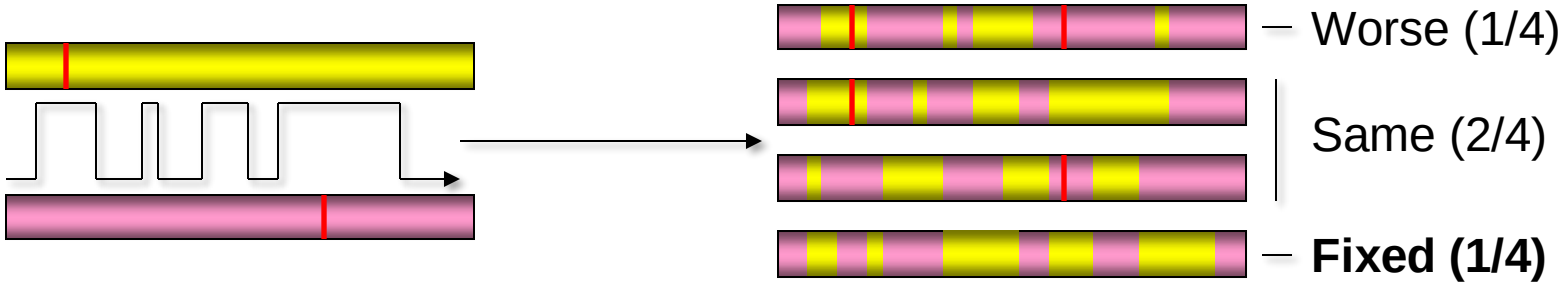
It can, however, also break up beneficial combinations of mutations

“Repairing” harmful mutations

Eg HIV



0.003% success rate with mutation



~25% success rate with recombination

Exploring sequence space

Sequence space = every possible combination of nucleotides in every possible length of DNA

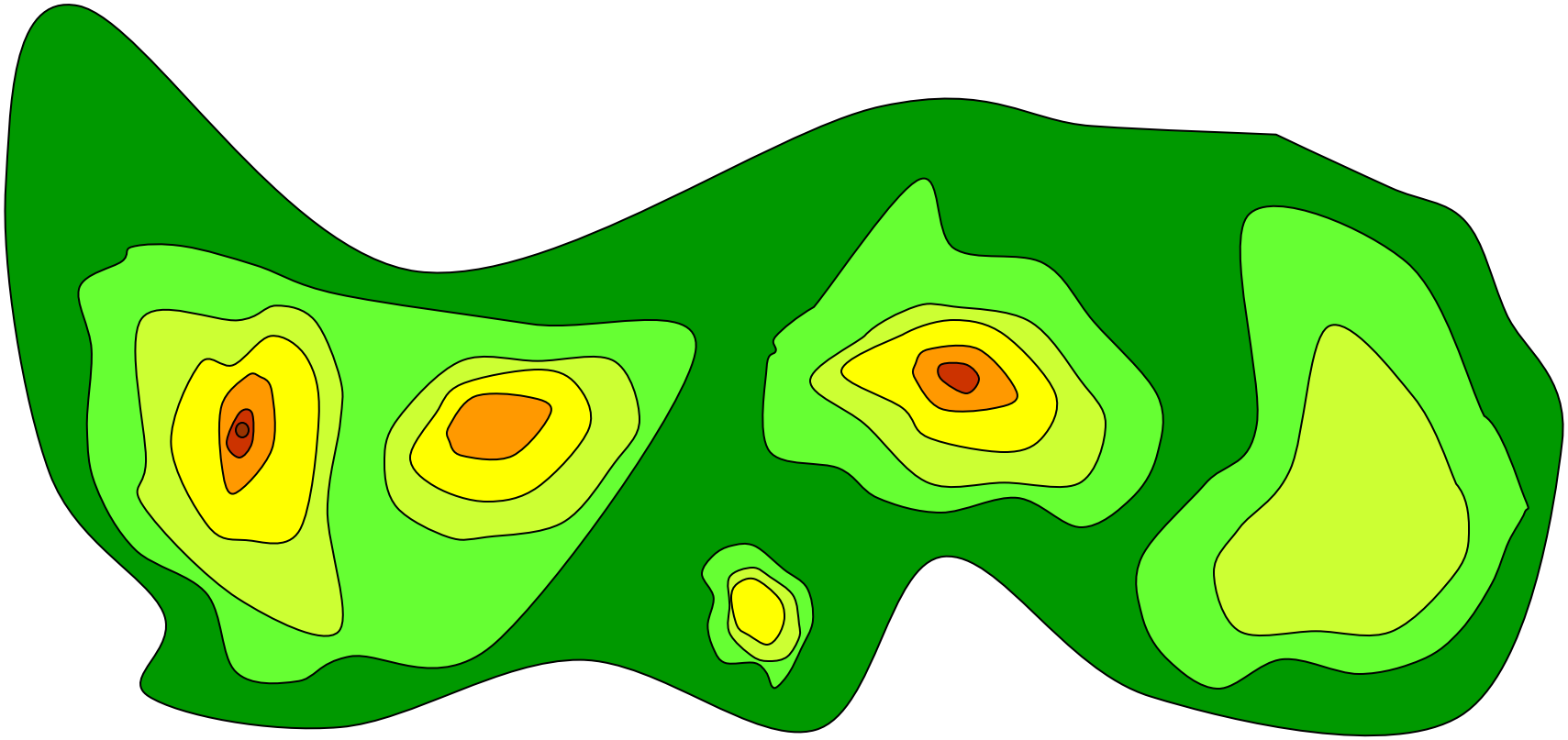
There are 4^{10000} possible combinations of nucleotides in a 10Kb genome.
There are "only" ~4^{170} elementary particles in the universe.
If the universe were one big atomic nucleus it would contain only ~4^{200} elementary particles.

Sequence space is unimaginably large

The proportion of biologically viable positions is unimaginably small

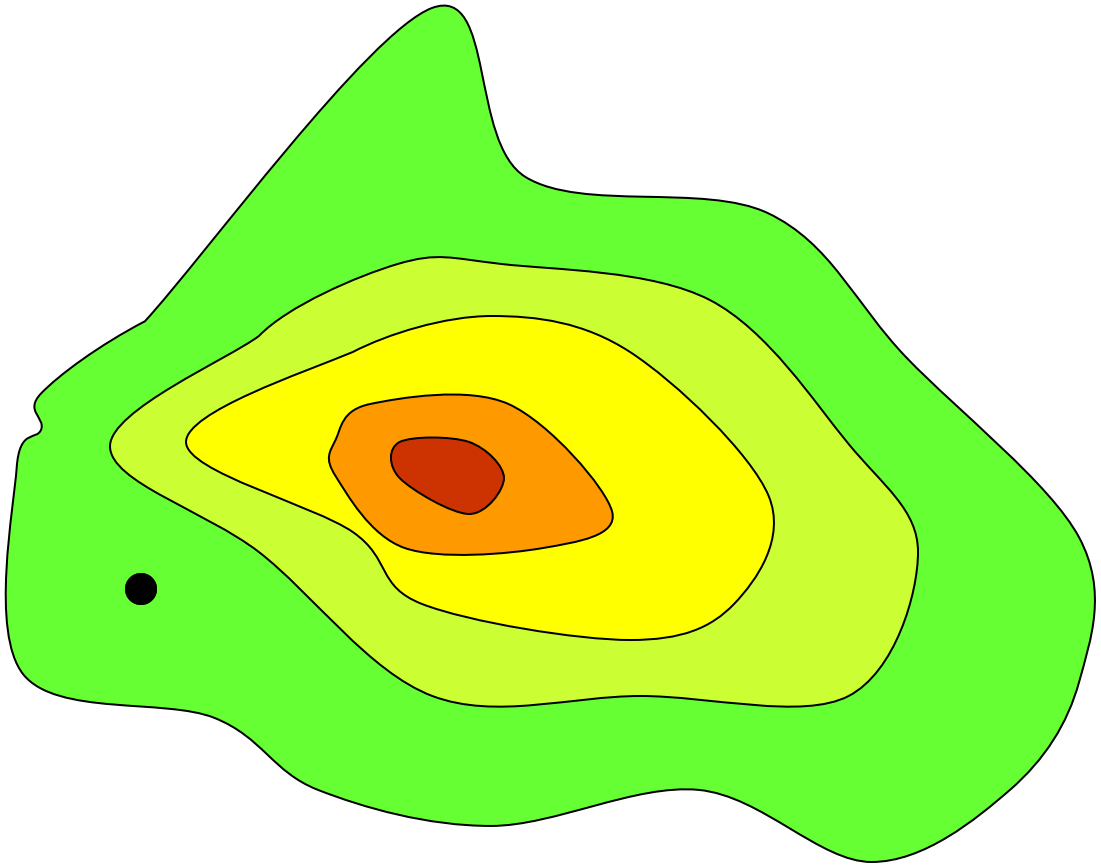
Fitness landscapes

Given a particular biological niche the relative viability of different sequences in a sequence space can be imagined as a fitness landscape



Exploring the landscape

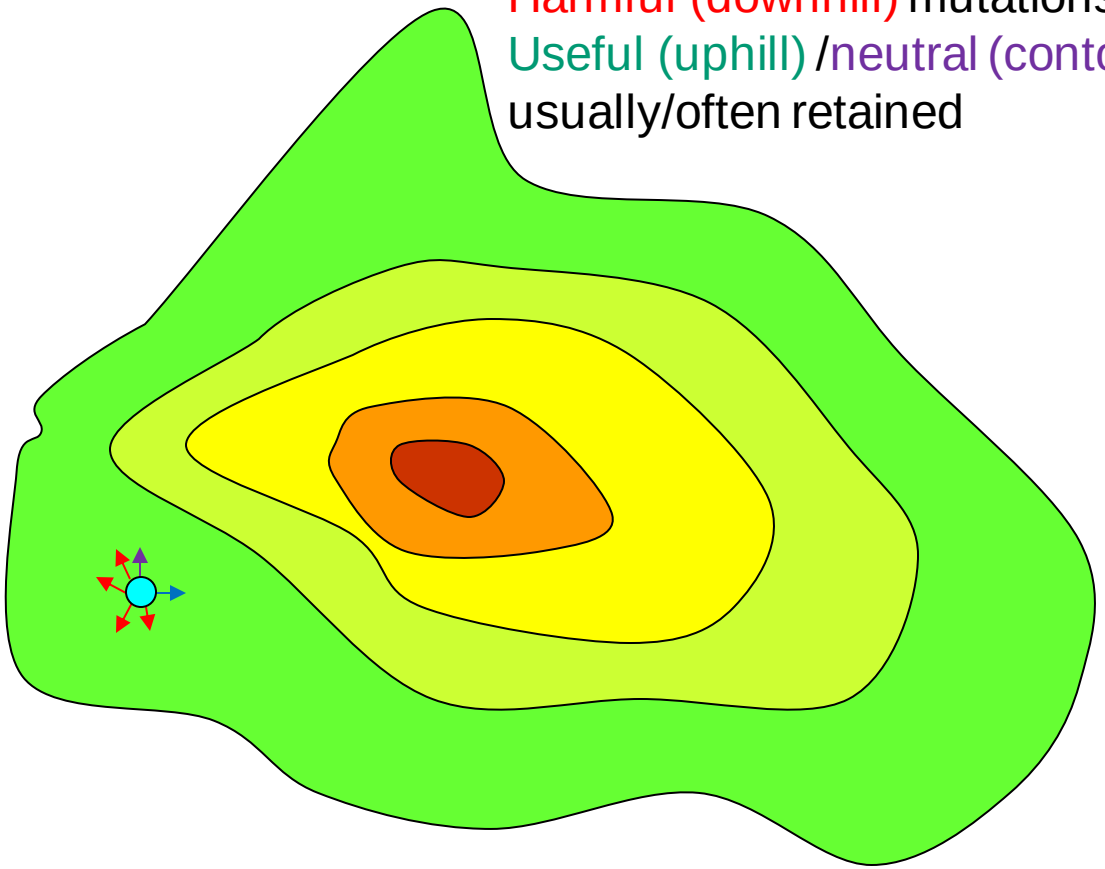
Exploration by mutation occurs in small steps.
Given a 10Kb genome there are 30 000 possible positions in sequence space that are accessible by a single mutation.



Exploring the landscape

Exploration by mutation occurs in small steps.
Given a 10Kb genome there are 30 000 possible positions in sequence space that are accessible by a single mutation.

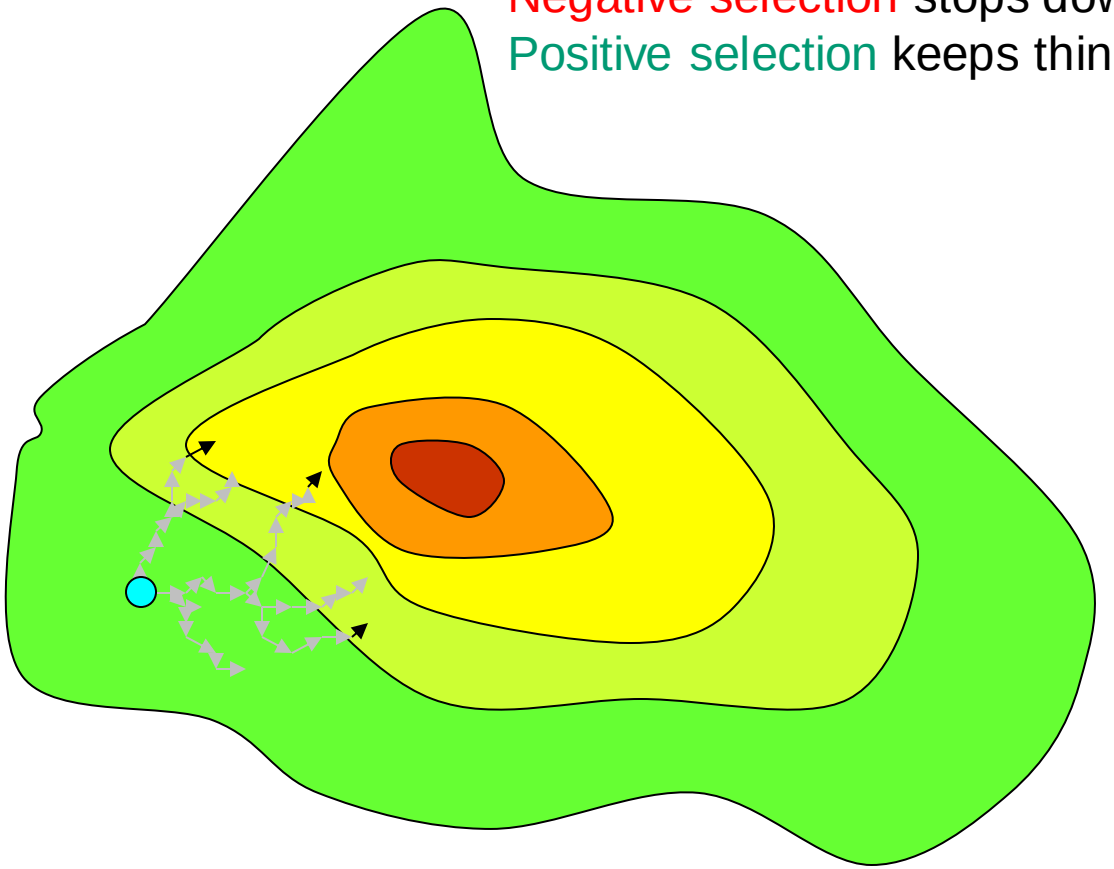
Harmful (downhill) mutations usually purged
Useful (uphill) /neutral (contour) mutations usually/often retained



Exploring the landscape

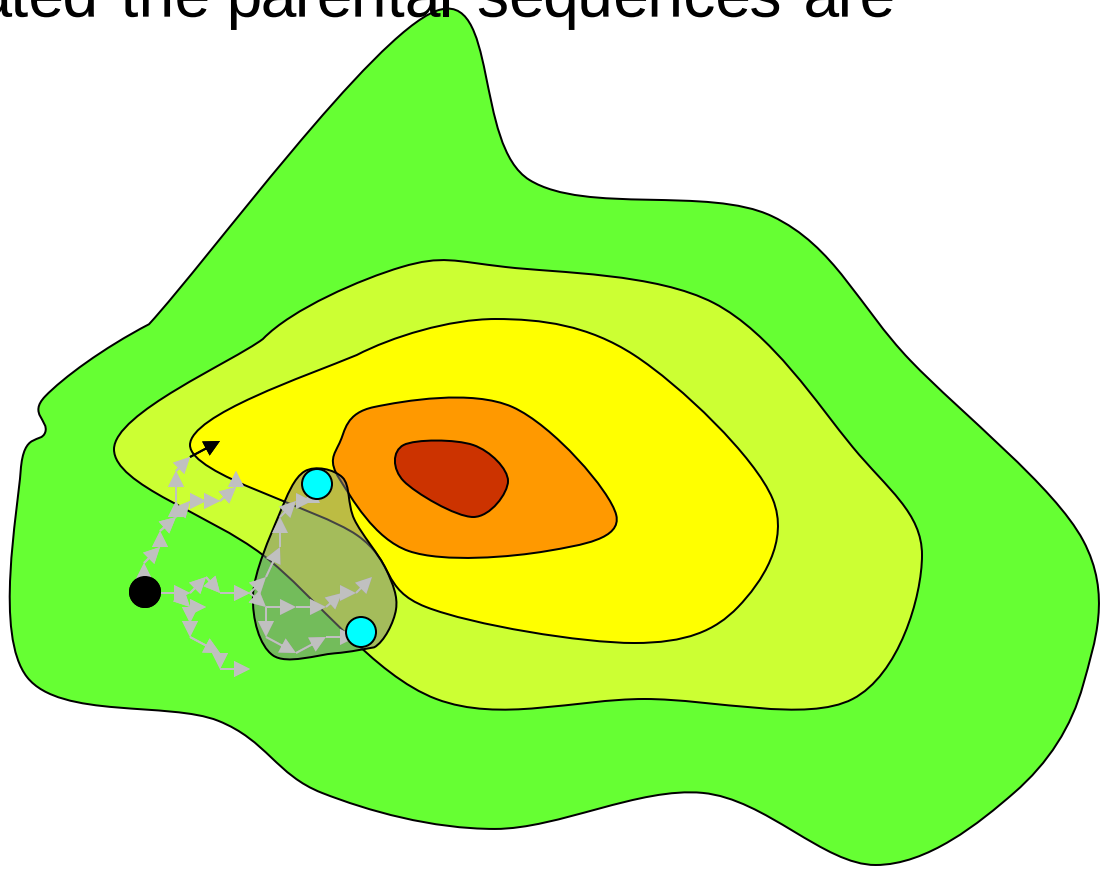
Peaks can slowly be scaled by the accumulation of useful (or adaptive) mutations under natural selection.

Negative selection stops downhill movement
Positive selection keeps things going uphill



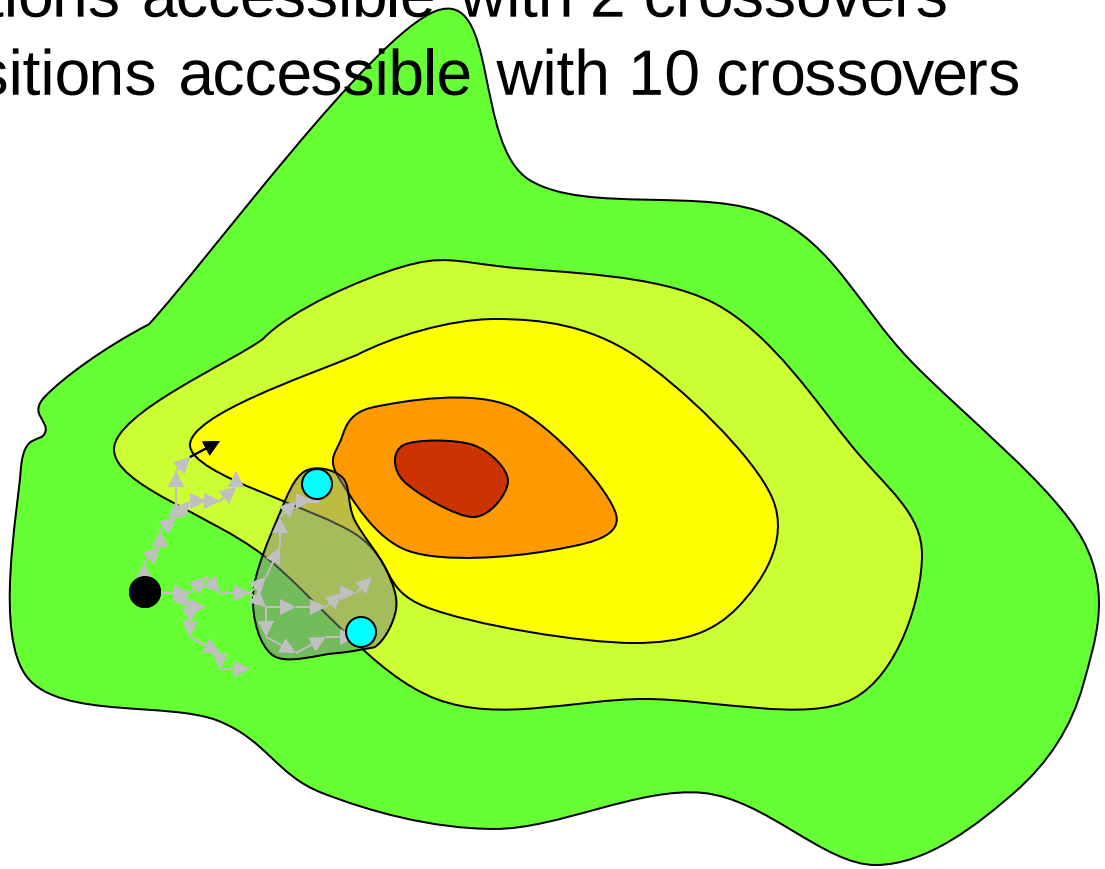
Exploring the landscape

The number of positions in sequence space that are accessible through recombination depends on:
(1) The number of breakpoints per replication cycle
(2) How related the parental sequences are



Exploring the landscape

Exploration by recombination can occur in big jumps
Given 2 10Kb genomes differing at 300 positions there are:
600 positions accessible with 1 crossover
90 000 positions accessible with 2 crossovers
2 X 10^{18} positions accessible with 10 crossovers



Problems with recombination

Sequences interact best with other coevolved sequences - if parental sequences are too diverged **sequence specific interactions may be compromised** and recombinants will have decreased fitness.

High recombination rates can also **break up beneficial combinations of mutations**.

Access to more positions in sequence space is not necessarily more beneficial – **if more positions are accessible by mutation than can be explored then recombination doesn't really offer any benefit**.

Why is recombination important?

It is important in population genetic studies

Recombination moves chunks of sequence between genomes.
Within chunks, alleles with high fitness value will become fixed along with neutral alleles that hitchhiked a ride on the same chunk.
Indirect selection of neutral alleles results in decreased variability of neutral alleles.
Decreased variability (1) decreases estimates of **effective population sizes** (2) could be interpreted by population genetic tests of natural selection as evidence of either recent **selective sweeps** or recent **population expansion**.

Recombination may also preserve genome-wide variability by spreading high fitness alleles to different genomes

Why is recombination important?

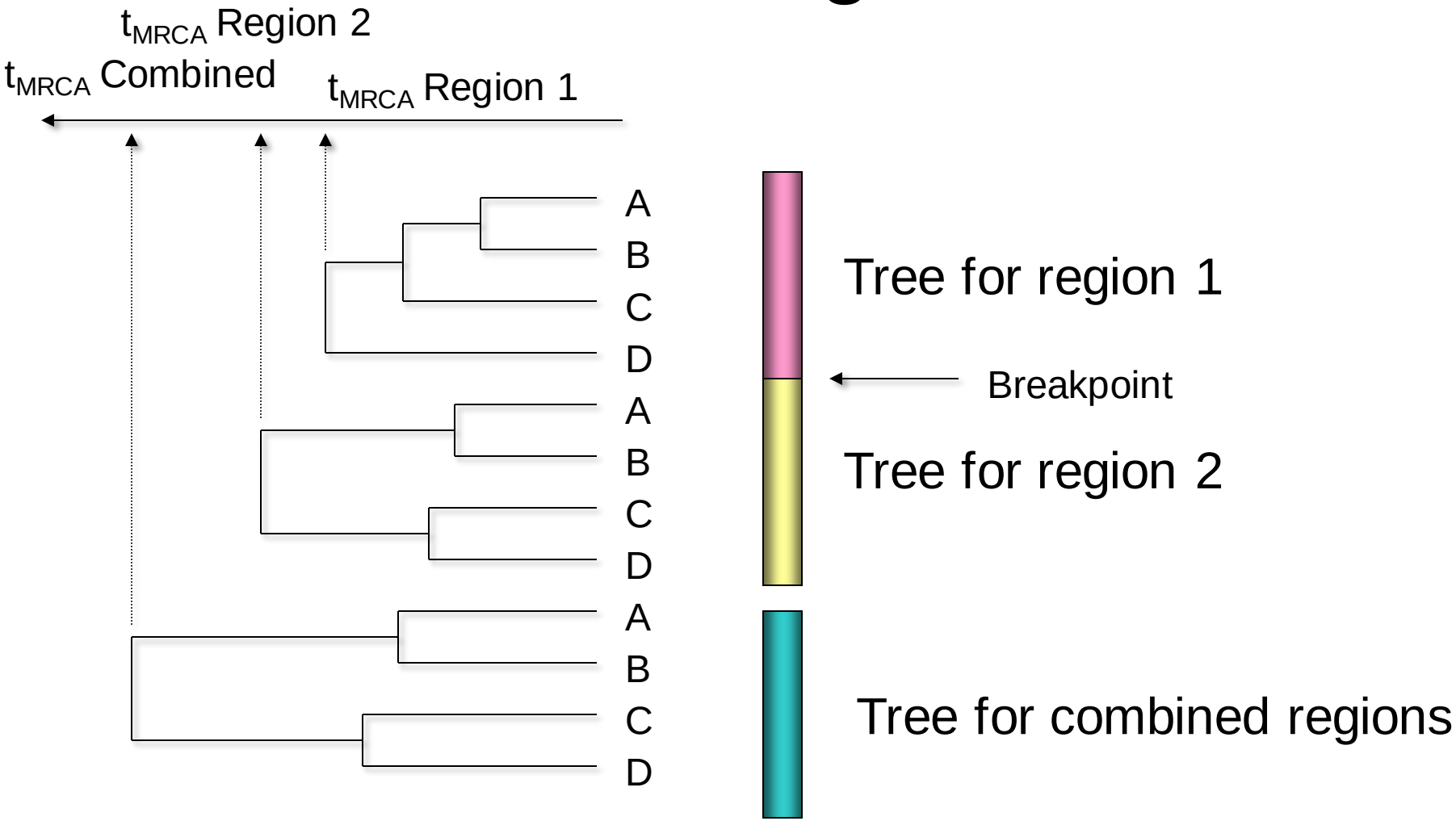
It is important in phylogenetic studies

Recombination allows genomic regions to have different evolutionary histories – i.e. **no single phylogenetic tree can describe the ancestry of recombining sequences**.

This complicates/prevents effective use of phylogenies in tracing **routes of disease transmission/migration**, determining **molecular clock rates**, estimating **mutation bias** and **rate heterogeneity**, and identifying **sites under positive selection**.

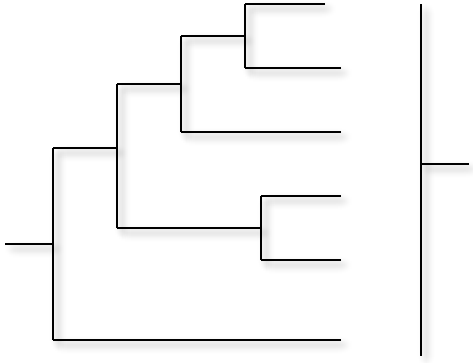
Recombination may compromise guide tree based alignment methods.

Effect of recombination on branch lengths



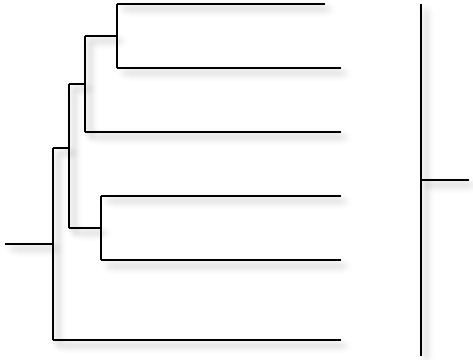
From Awadalla, 2003

Effect of recombination On interpretation of tree shapes



Terminal branches ~ the same length as internal branches.

This tree would be expected if most mutations were neutral and the **population size was constant.** e.g. Papilloma viruses

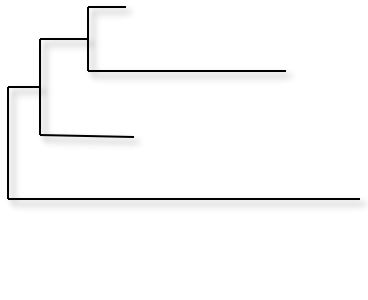


Terminal branches longer than internal branches.

This tree would be expected if either the **population size was expanding** or **recombination was rampant.** e.g. HIV and Foot and mouth disease virus

Effect of recombination on the inference of mutation rates

P1 AAAAAAAAAAAAAAA
Re AAAAAAAAAACCCCC
X ACAGCGACCAGACAA
P2 CCCCCCCCCCCCCCC



Homoplasies will be detected at these sites
i.e. these sites will look like they have been subject to **recurrent/convergent mutations**
They will appear to be more mutable than they really are.

Mutation-rate heterogeneity over the entire sequence will be overestimated.
Molecular clock hypothesis might be falsely rejected.
Timing of events underestimated due to overestimation of terminal branch lengths.
Selection may be incorrectly inferred

Recombination detection

>30 methods currently described, most with associated software that can be obtained from:
http://www.bioinf.man.ac.uk/~robertson/recombination/

These methods give different kinds of information:
(1) Yes/No

Most methods but not eg **SIMPLOT**,
BOOTSCAN and **RAT**

Recombination detection

>30 methods currently described, most with associated software that can be obtained from:
http://www.bioinf.man.ac.uk/~robertson/recombination/

These methods give different kinds of information:
(1) Yes/No

(2) Breakpoint positions

1/2 of the methods eg **RECPARS, DSS, BARCE, DAMBE, BOOTSCAN SIMPLOT**, **RAT**, **RDP**, **RIP** and **GENECONV**

Recombination detection

>30 methods currently described, most with associated software that can be obtained from:
http://www.bioinf.man.ac.uk/~robertson/recombination/

These methods give different kinds of information:
(1) Yes/No

(2) Breakpoint positions

(3) Recombinants

1/3 methods eg **RDP**, **RIP, RAT, BOOTSCAN/SIMPLOT** and **GENECONV**

Recombination detection

>30 methods currently described, most with associated software that can be obtained from:
http://www.bioinf.man.ac.uk/~robertson/recombination/

These methods give different kinds of information:
(1) Yes/No

(2) Breakpoint positions

(3) Recombinants

(4) Population recombination rates

Very few methods only **DNASP, LDHAT, SITES, INFS/FINS** and **RECOMBINE**

Recombination detection

>30 methods currently described, most with associated software that can be obtained from:
http://www.bioinf.man.ac.uk/~robertson/recombination/

These methods give different kinds of information:
(1) Yes/No

(2) Breakpoint positions

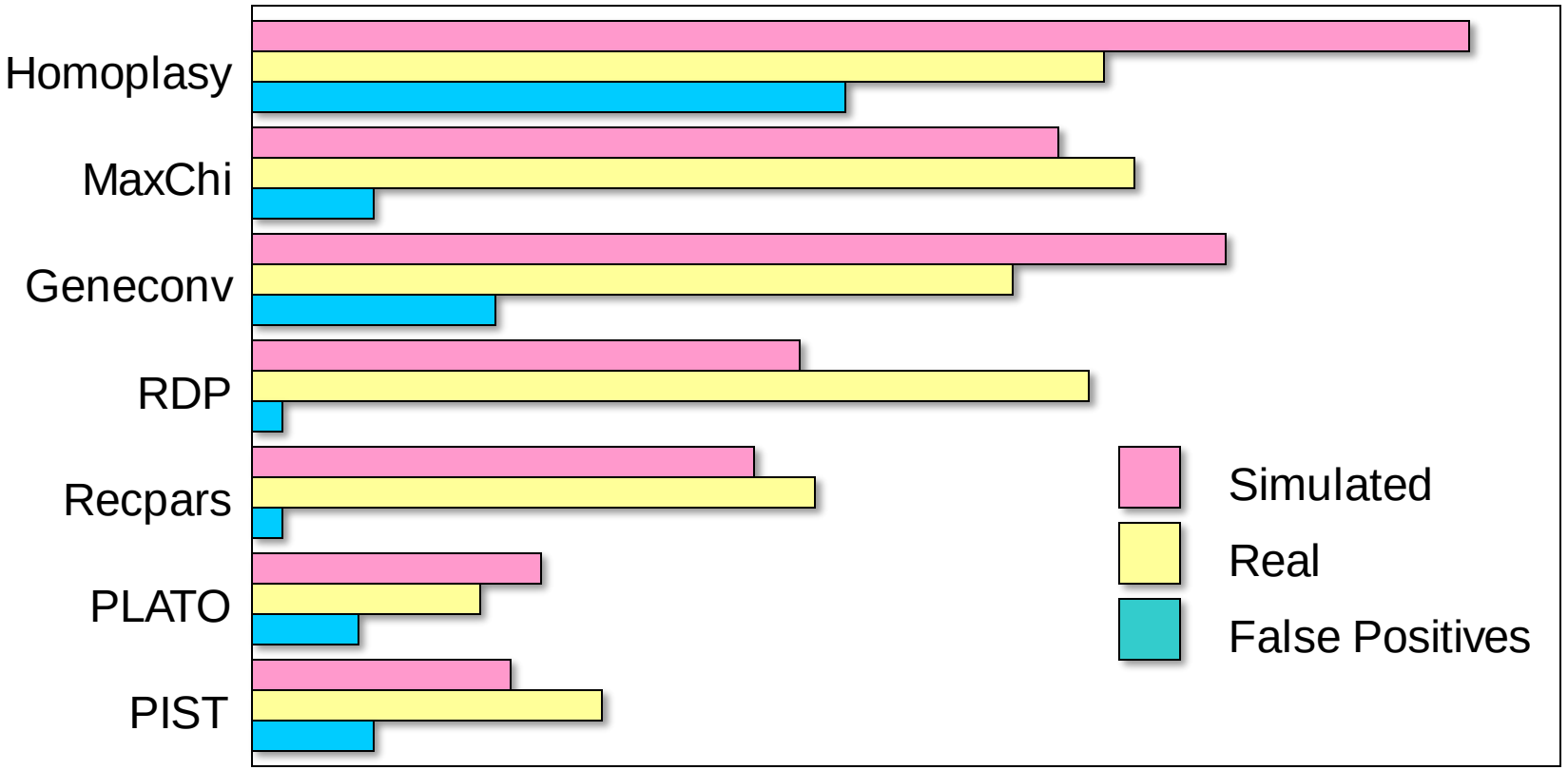
(3) Recombinants

(4) Population recombination rates

Very few methods only **DNASP, LDHAT, SITES, INFS/FINS** and **RECOMBINE**

Method performance

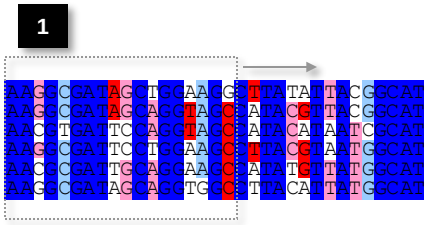
Power



After Posada and Crandall 2001 and Posada 2002

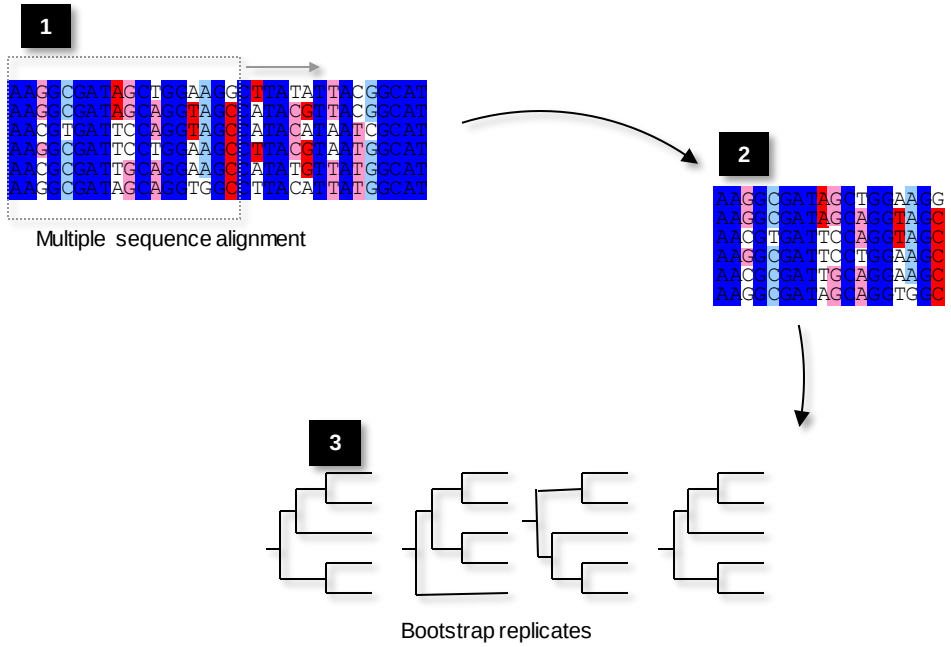
Accuracy of Yes/No answers

How do these methods work?



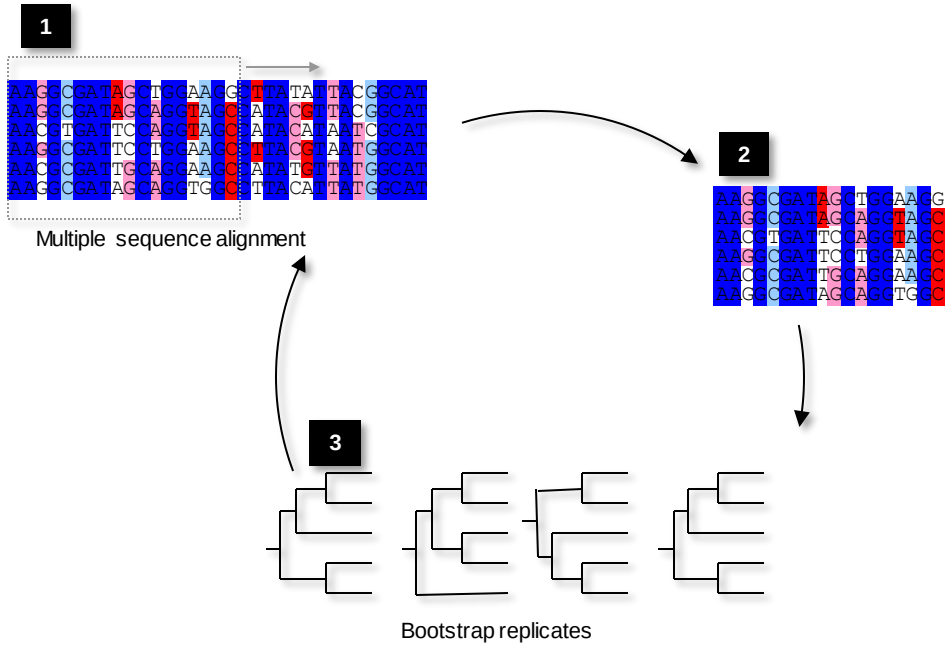
Select the portion of an alignment that falls within a specified window

How do these methods work?



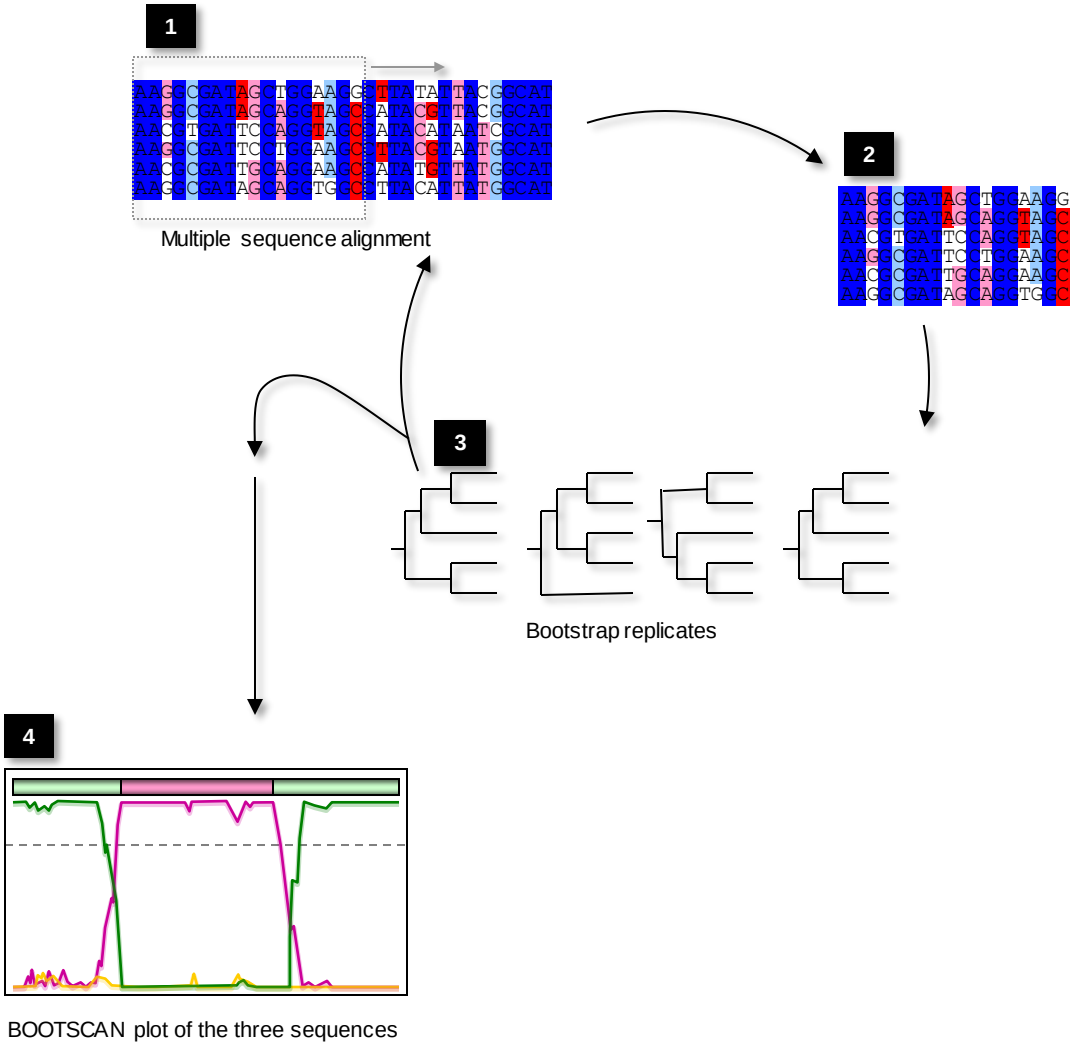
Make bootstrap replicates, calculate distance matrices and construct midpoint rooted neighbour joining/UPGMA trees

How do these methods work?



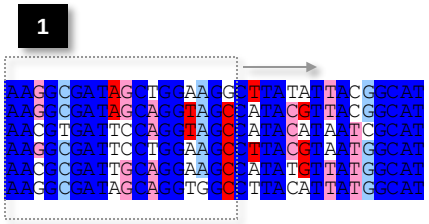
Move window along a specified number of nucleotides and repeat

How do these methods work?

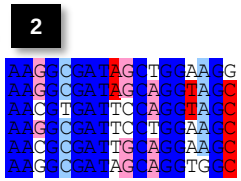


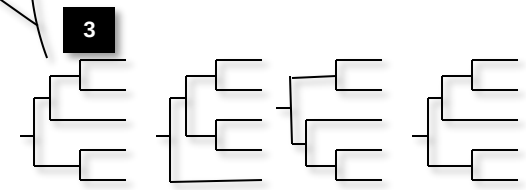
After the last window is examined move on to the detection phase. Select three sequences and retrieve BOOTSCAN plots for the three sequences from stored tree positions/distance matrices

How do these methods work?



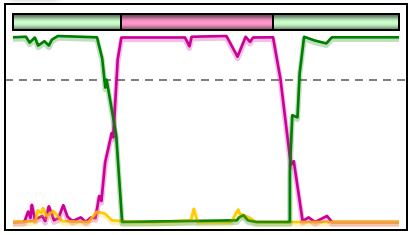
Multiple sequence alignment





Bootstrap replicates

Determine the significance of potential recombination signals using binomial p-value



BOOTSCAN plot of the three sequences

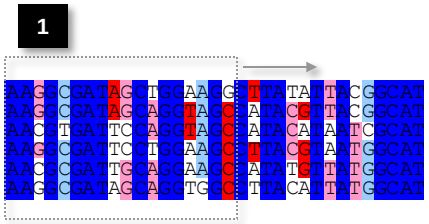
5

$$p\text{-value} = G \times \frac{L}{N} \times \sum_{m=M}^{N} \left(\frac{N!}{m!(N-m)!} \right) p^{N} \times (1-p)^{N-m}$$

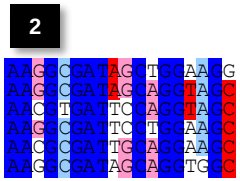
Approximate significance of signals w here:

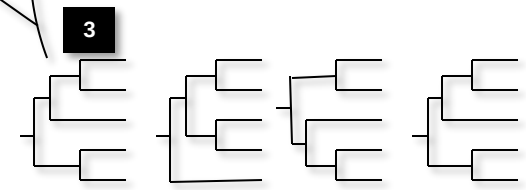
- G is the total number of possible sequence triplets
- L is the length of the sequence
- N is the length of the potential recombinant region
- M is the proportion of nucleotides in common betw een the potential recombinant and parental sequences in the recombinant region
- p is the proportion of nucleotides in common betw een the potential recombinant and parental sequences in the entire sequence

How do these methods work?



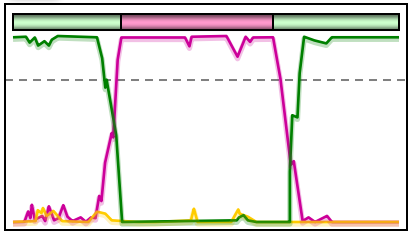
Multiple sequence alignment





Bootstrap replicates

Select another three sequences



BOOTSCAN plot of the three sequences

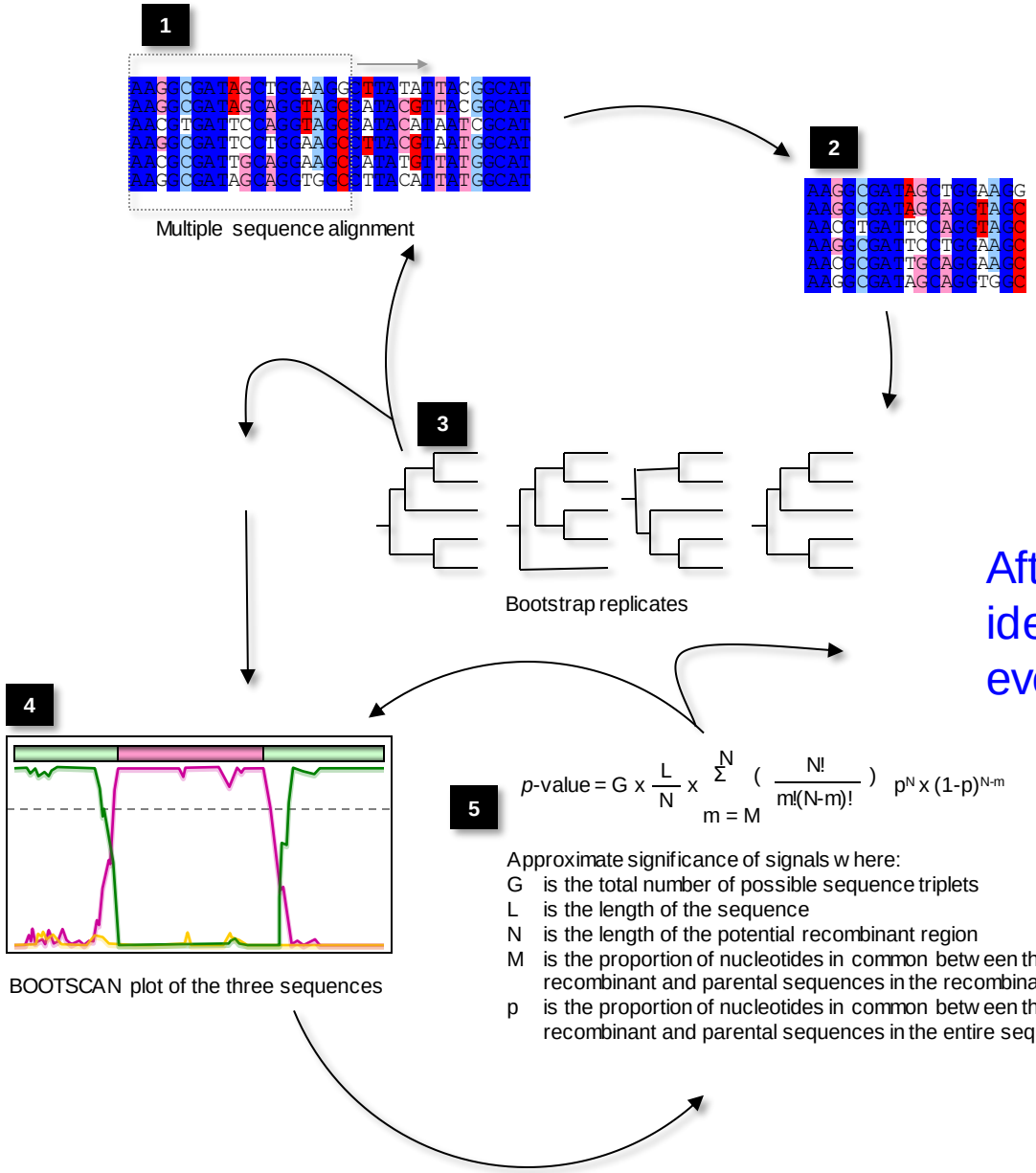
5

$$p\text{-value} = G \times \frac{L}{N} \times \sum_{m=M}^{N} \left(\frac{N!}{m!(N-m)!} \right) p^{N} \times (1-p)^{N-m}$$

Approximate significance of signals w here:

- G is the total number of possible sequence triplets
- L is the length of the sequence
- N is the length of the potential recombinant region
- M is the proportion of nucleotides in common betw een the potential recombinant and parental sequences in the recombinant region
- p is the proportion of nucleotides in common betw een the potential recombinant and parental sequences in the entire sequence

How do these methods work?



After the last triplet is examined, identify unique recombination events

Recombination analysis

Know what you are interested in (recombination rates, breakpoint positions or recombinant identification).

Recombination analysis

Know what you are interested in (recombination rates, breakpoint positions or recombinant identification).

Choose a method with proven power and reasonably low false positive rate – **Remember that it is unknown how methods compare WRT breakpoint or recombinant identification.**

Recombination analysis

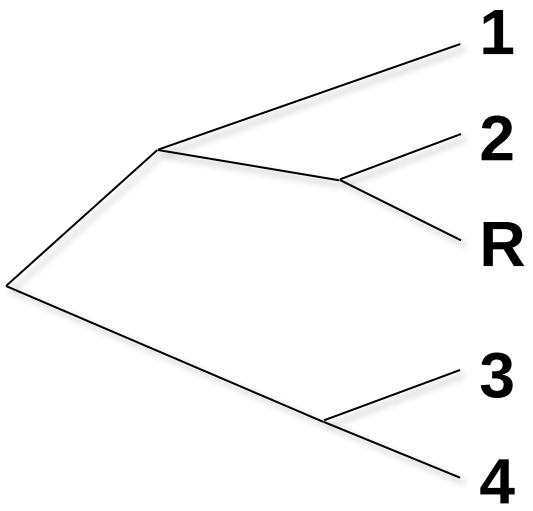
Know what you are interested in (recombination rates, breakpoint positions or recombinant identification).

Choose a method with proven power and reasonably low false positive rate – **Remember that it is unknown how methods compare WRT breakpoint or recombinant identification.**

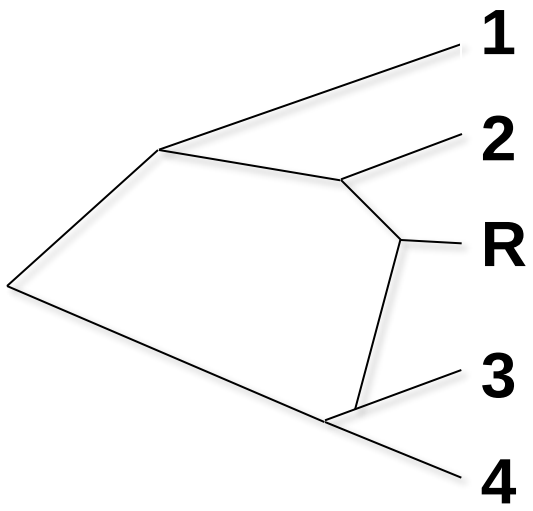
Not all methods work equally well under all conditions - Combinations of methods are preferable.

Accounting for recombination

Phylogenetic Inferences:
(1) Present network graphs



Standard bifurcating tree

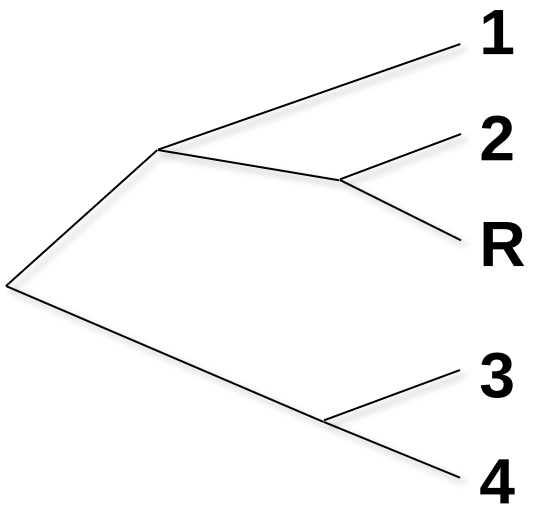


Network graph indicating the dual ancestry of sequence R

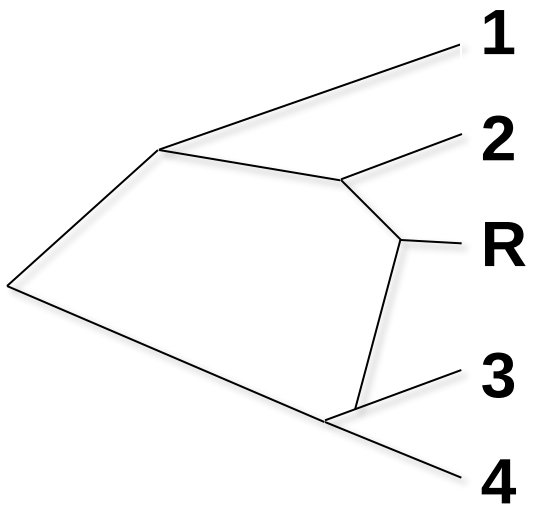
Programs like **SplitsTree** accept sequence alignments and produce network graphs rather than bifurcating trees

Accounting for recombination

Phylogenetic Inferences:
(1) Present network graphs



Standard bifurcating tree

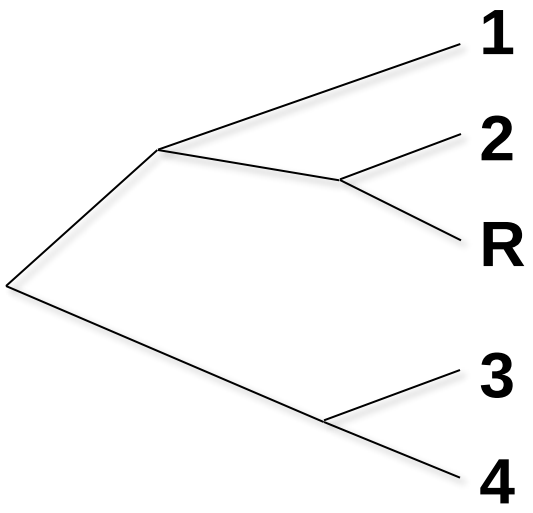


Network graph indicating the dual ancestry of sequence R

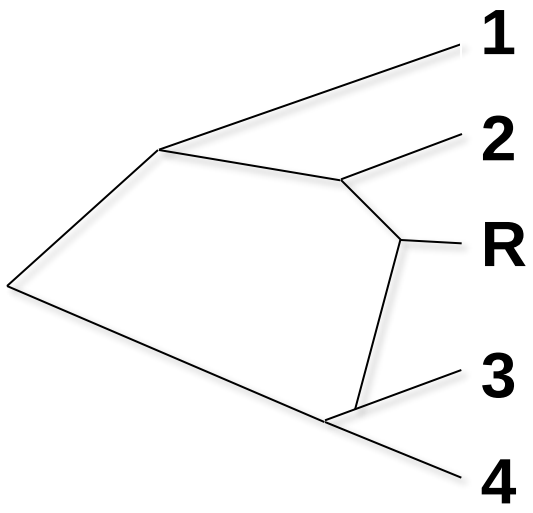
Note that such network graphs are not, strictly speaking, phylogenetic trees that represent recombination

Accounting for recombination

Phylogenetic Inferences:
(1) Present network graphs



Standard bifurcating tree

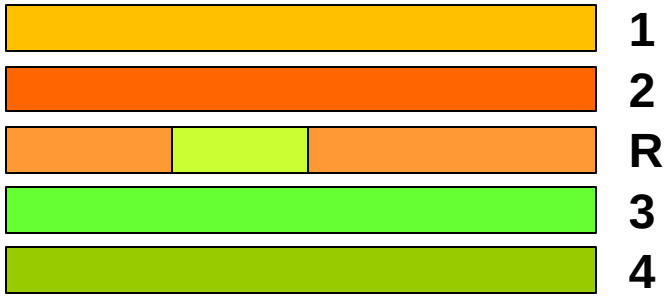


Network graph indicating the dual ancestry of sequence R

Their branch-lengths are not proportional to mutation numbers and their topologies reflect non-tree-like evolution which can have causes other than recombination

Accounting for recombination

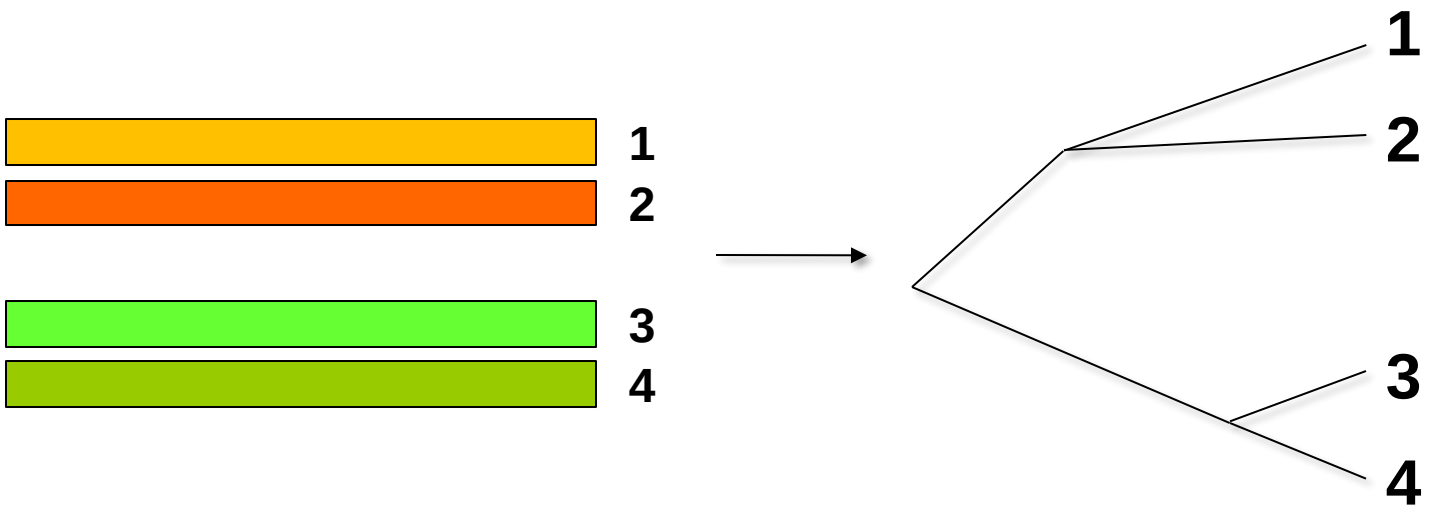
Phylogenetic Inferences:
(1) Present network graphs
(2) Remove recombinants



Identify sequence “R” as the recombinant using a computer program like RDP or VisRD

Accounting for recombination

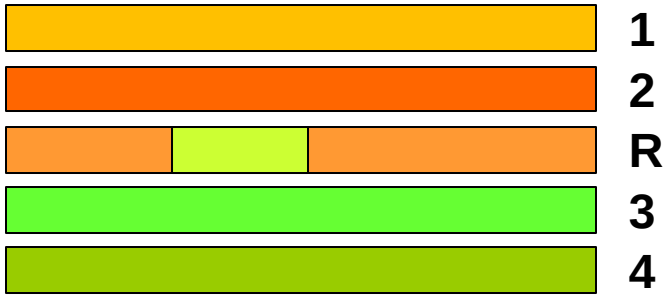
Phylogenetic Inferences:
(1) Present network graphs
(2) Remove recombinants



Remove the recombinant sequence and draw the tree

Accounting for recombination

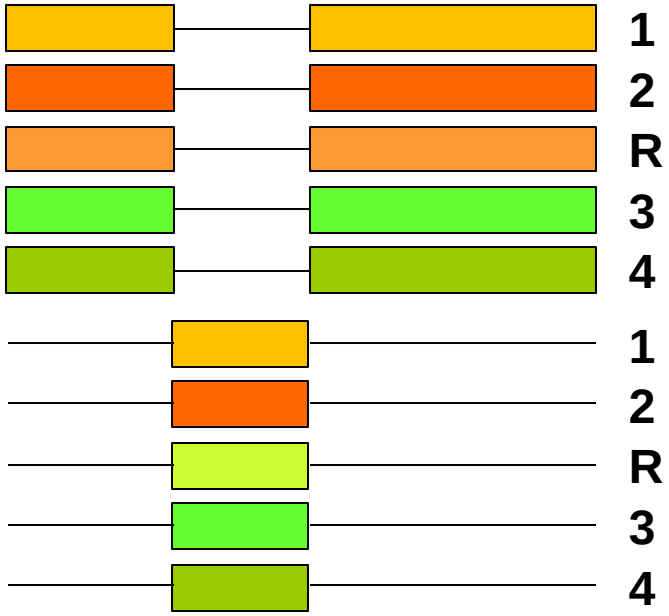
Phylogenetic Inferences:
(1) Present network
(2) Remove recombinants
(3) Split alignment at recombination breakpoints



Identify the positions of recombination breakpoints…..

Accounting for recombination

Phylogenetic Inferences:
(1) Present network
(2) Remove recombinants
(3) Split alignment at recombination breakpoints

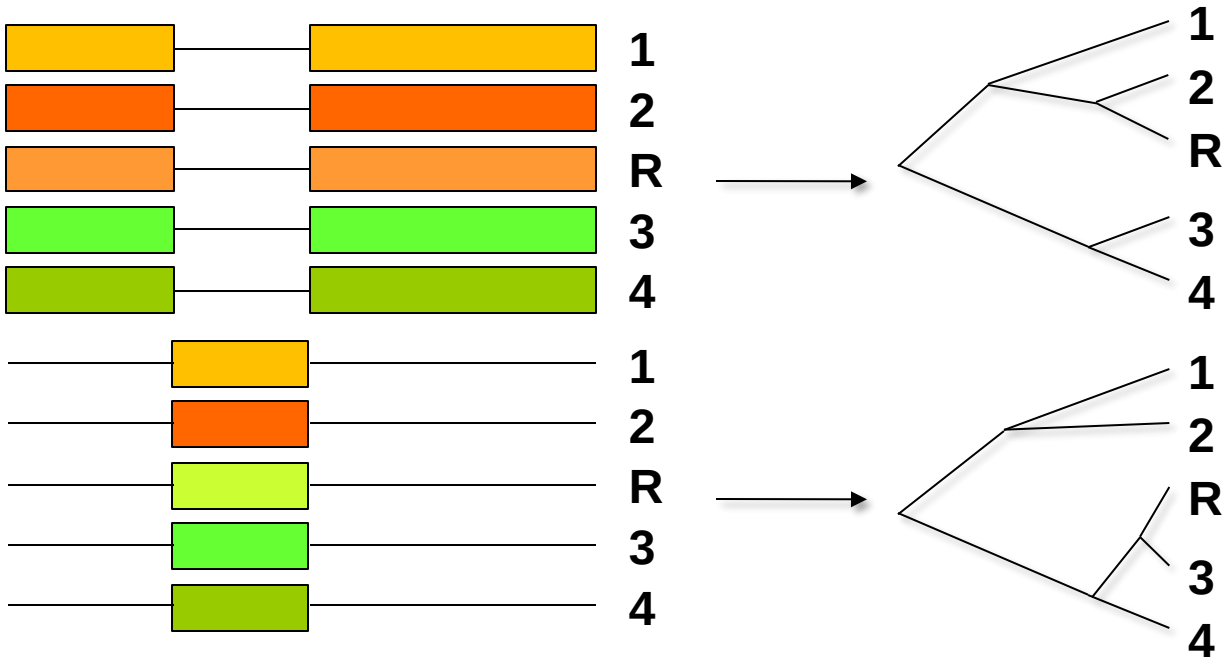


= bits of the alignments treated as missing data

…… split the alignment into two separate parts…..

Accounting for recombination

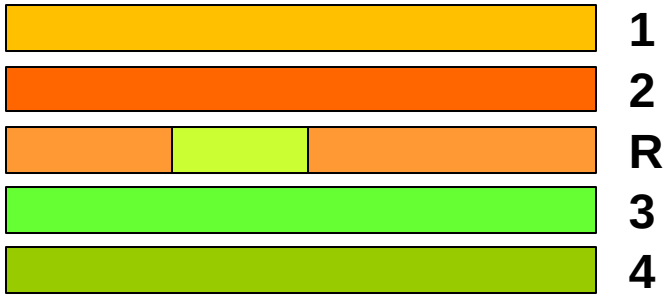
Phylogenetic Inferences:
(1) Present network
(2) Remove recombinants
(3) Split alignment at recombination breakpoints



…… and make a separate tree for each part

Accounting for recombination

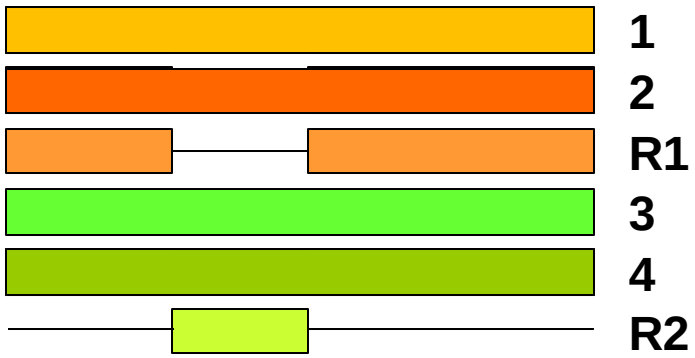
Phylogenetic Inferences:
(1) Present network
(2) Remove recombinants
(3) Split alignment at recombination breakpoints
(4) Split only the recombinant sequences



Use a program such as RDP to identify the recombinant sequences and their breakpoints……

Accounting for recombination

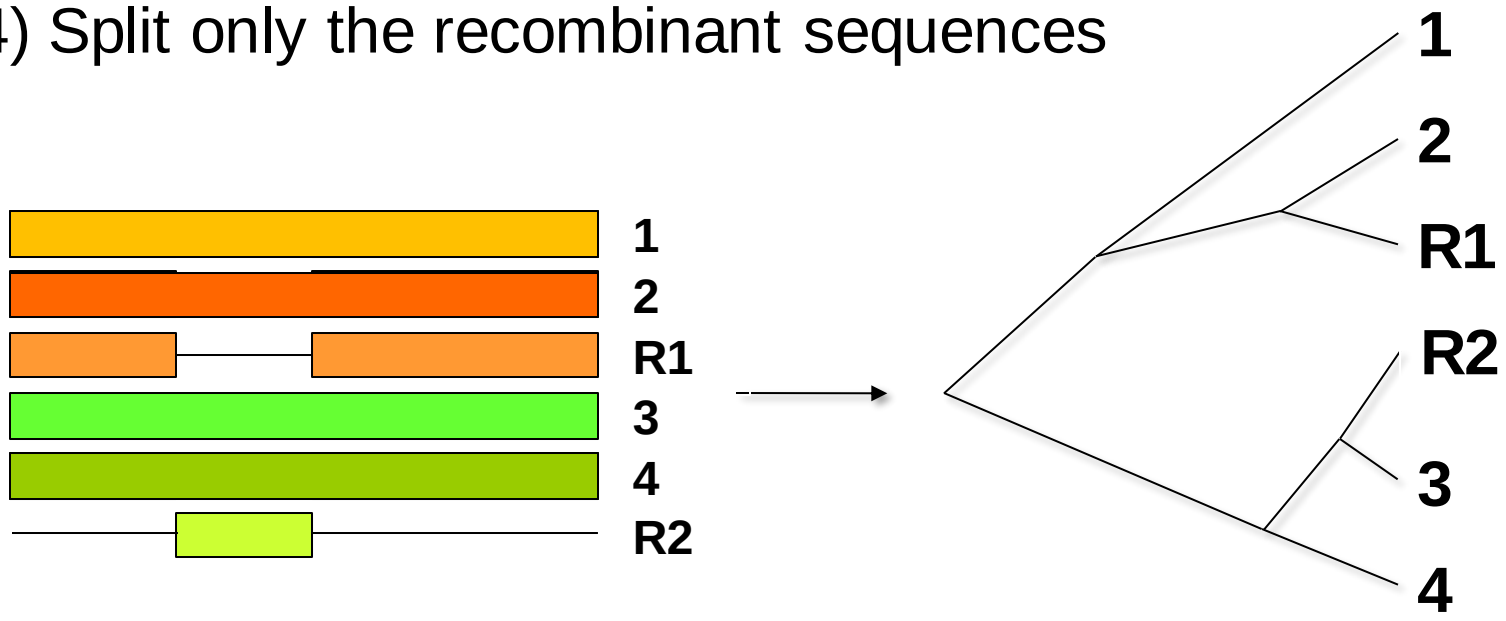
Phylogenetic Inferences:
(1) Present network
(2) Remove recombinants
(3) Split alignment at recombination breakpoints
(4) Split only the recombinant sequencesz



…… split the recombinants into their component parts…….

Accounting for recombination

Phylogenetic Inferences:
(1) Present network
(2) Remove recombinants
(3) Split alignment at recombination breakpoints
(4) Split only the recombinant sequences



…… and draw a single tree but with some sequences represented more than once (programs like RDP will do this for you).

Accounting for recombination

Population Genetic Inferences:

(1) Estimate the population recombination rate from the data and include this rate in models.

Programs like DNASP can account for recombination (if given estimated recombination rates) during inference of population genetic parameters

Accounting for recombination

Population Genetic Inferences:

(1) Estimate the population recombination rate from the data and include this rate in models.

(2) Carry on as if recombination didn't exist.

This is unfortunately the solution most often used