Non-tree like evolution: Recombination, ancestral recombination graphs and clonal frames

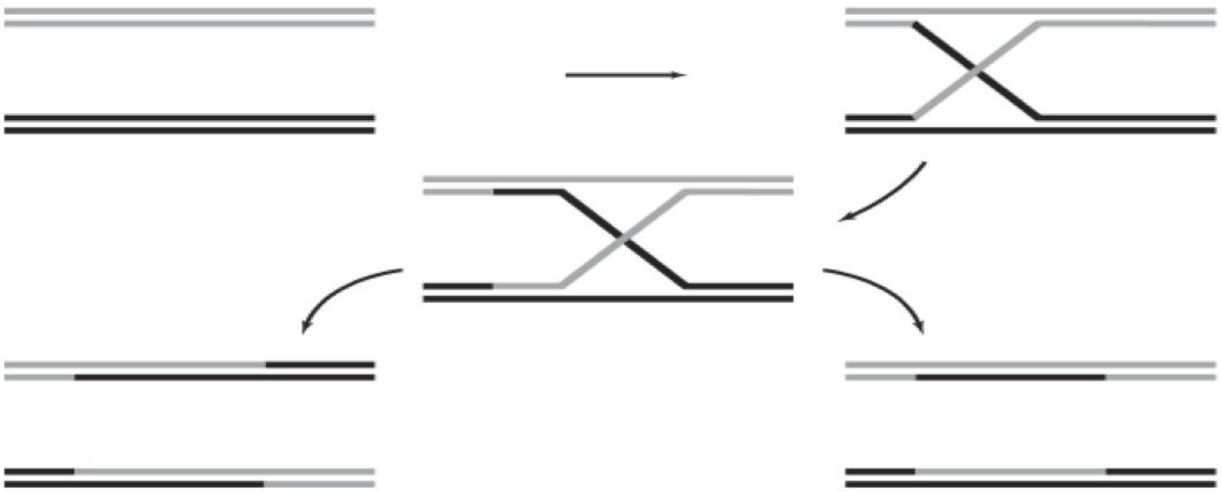
Molecular Epidemiology of Infectious Diseases

Lecture 6

February 24th, 2020

Recombination is a major force shaping the evolution of nearly all microbial pathogens

Mechanisms of recombination



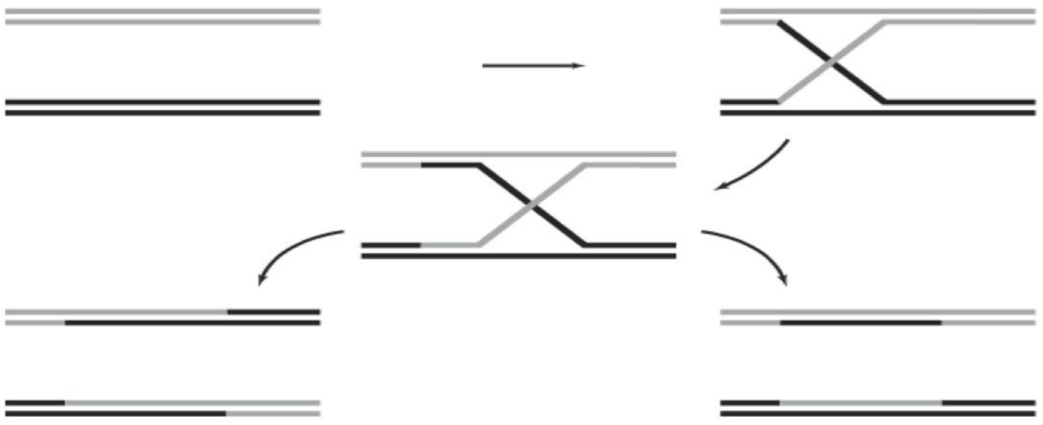
Recombination with crossovers

Recombination via gene conversion

Hein *et al.* (2004)

Mechanisms of recombination

In eukaryotes, recombination is typically due to crossover events



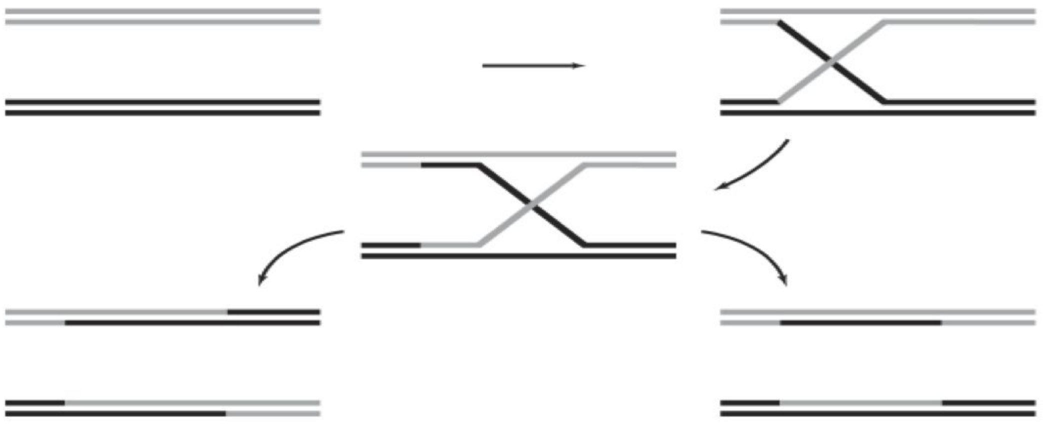
Recombination with crossovers

Recombination via gene conversion

Hein *et al.* (2004)

Mechanisms of recombination

In bacteria, recombination it typically due to gene conversion — the substitution of a small fragment of DNA from one chromosome to another.



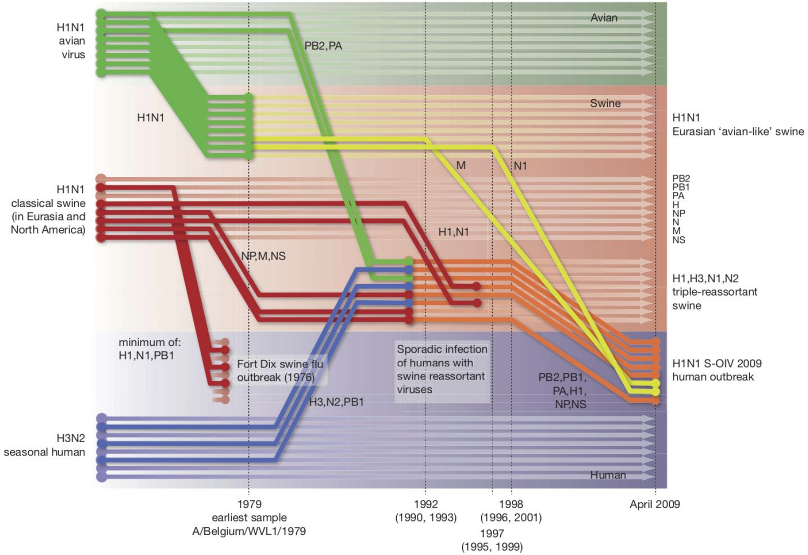
Recombination with crossovers

Recombination via gene conversion

Hein *et al.* (2004)

Mechanisms of recombination

Segmented viruses also undergo reassortment — reshuffling of segments between different progeny viruses



Smith *et al.* (Nature, 2009)

Recombination creates mosaic ancestry

What is of interest is **the ancestry of individual nucleotides in the daughter molecules with respect to the parent nucleotides**

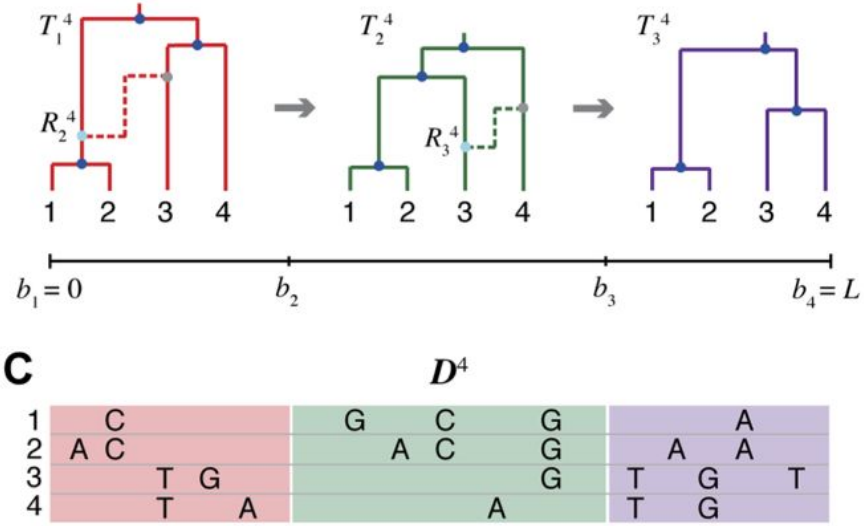
Without any recombination, the entire genome of an individual will share the same ancestry (i.e. phylogenetic history)

With recombination, genomes become mosaics where different segments descend from different ancestors

No single tree can therefore describe the ancestry of a sample of recombining sequences

Recombination creates mosaic ancestry

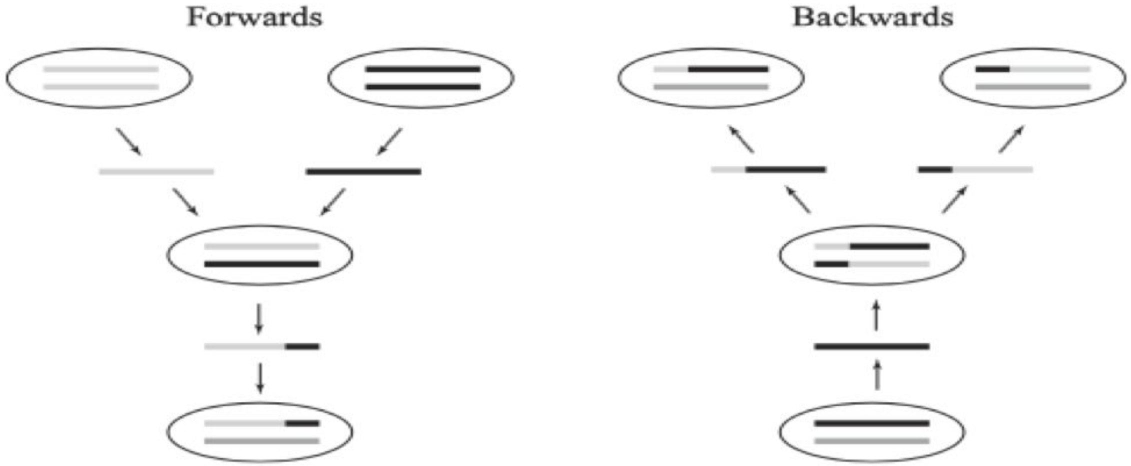
Different regions of the genome will have different phylogenetic histories:



Rasmussen *et al.* (PLoS Gen, 2014)

Recombination in phylogenies

In a sense, recombination events are the opposite of coalescent events in that genetic material is split among two different ancestors



Hein *et al.* (2004)

Effect of a single recombination event

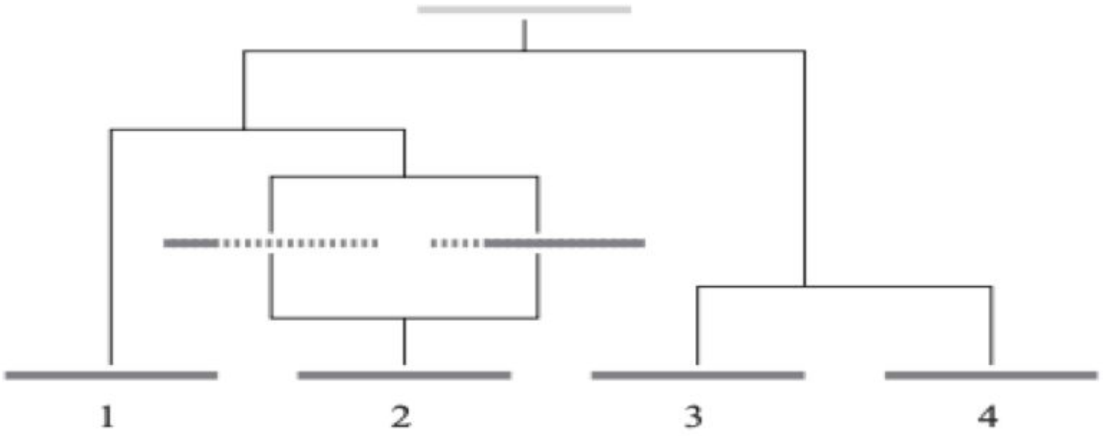
A single recombination event between two sampled lineages will have one of three possible effects on the phylogeny:

- No effect
- Effect only the branch lengths
- Effect the tree topology

Hein *et al.* (2004)

Effect of a single recombination event

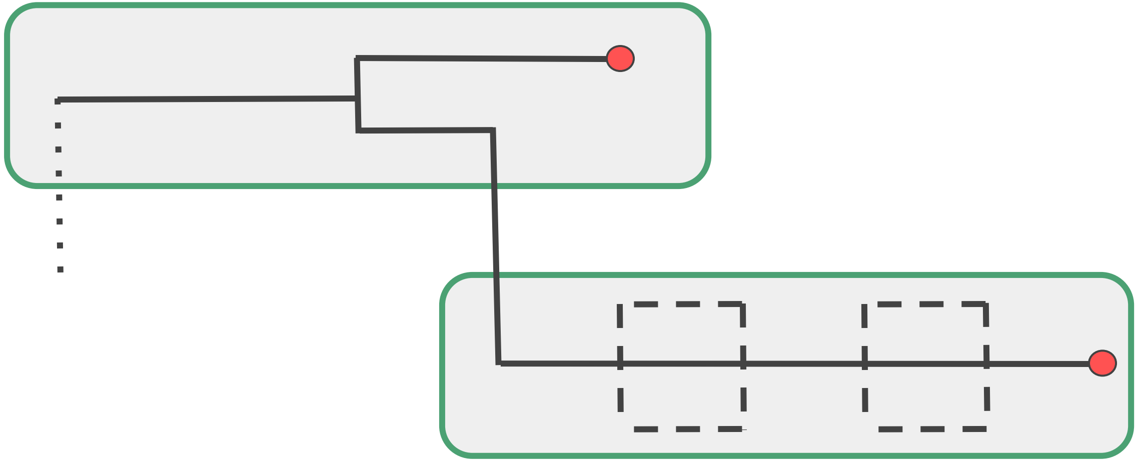
If two recombinant sequences coalesce before they coalesce with any other lineage, the recombination event will have **no effect** on the phylogeny.



Hein *et al.* (2004)

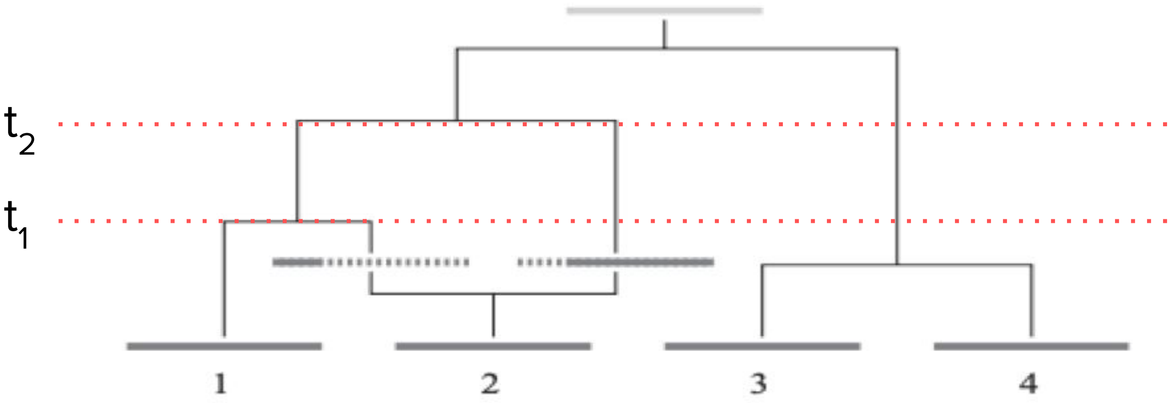
Effect of a single recombination event

Recombination events within individual hosts will generally have no impact on the overall pathogen phylogeny



Effect of a single recombination event

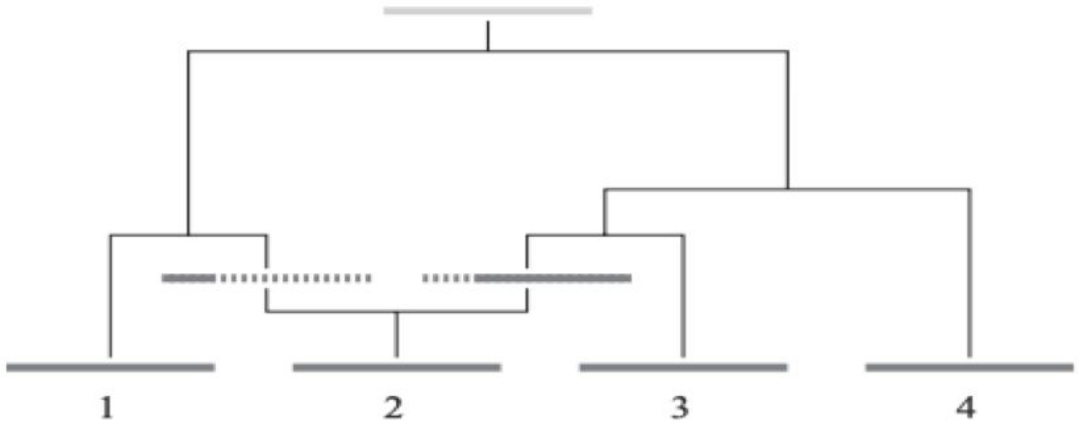
Only **branch lengths will change** if one of two recombining sequences merges with another sequence before coalescing with the other recombining sequence again.



Hein *et al.* (2004)

Effect of a single recombination event

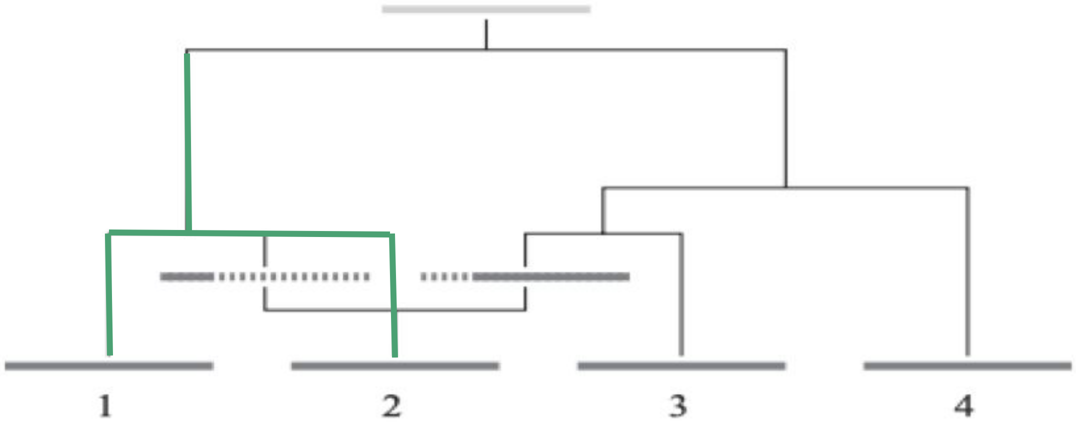
The **tree topology will change** if the two recombining sequences coalesce with other sequences before the two recombining sequences coalesce.



Hein *et al.* (2004)

Effect of a single recombination event

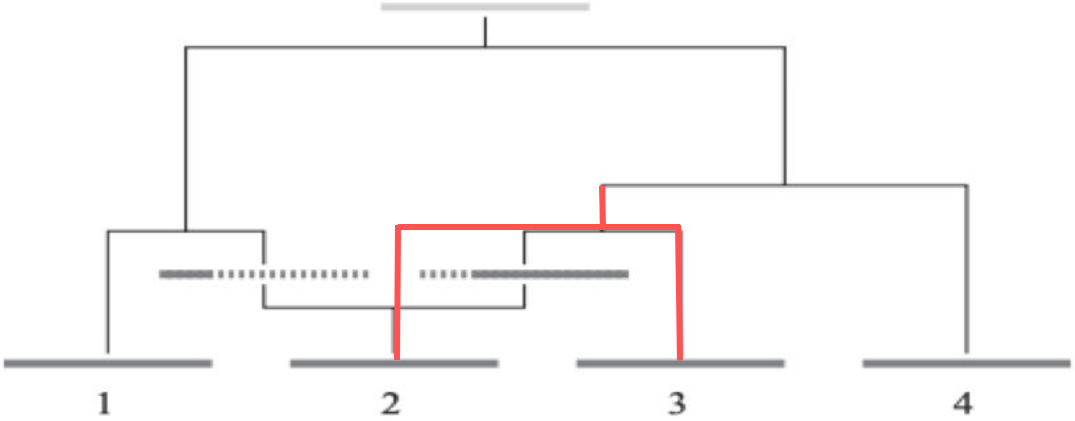
The tree topology will change if the two recombining sequences coalesce with other sequences before the two recombining sequences coalesce.



Hein *et al.* (2004)

Effect of a single recombination event

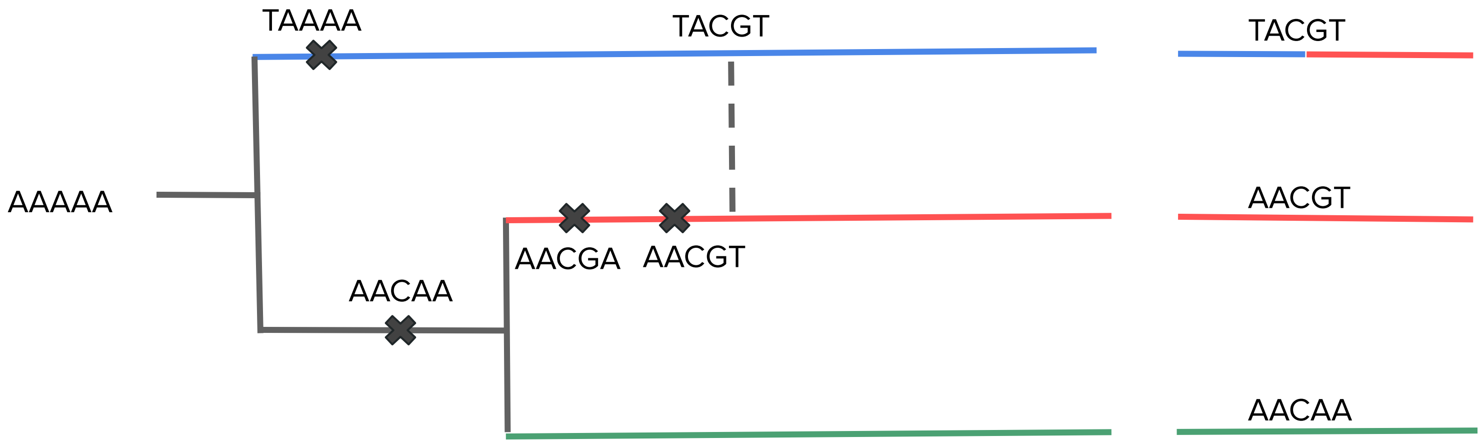
The tree topology will change if the two recombining sequences coalesce with other sequences before the two recombining sequences coalesce.



Hein *et al.* (2004)

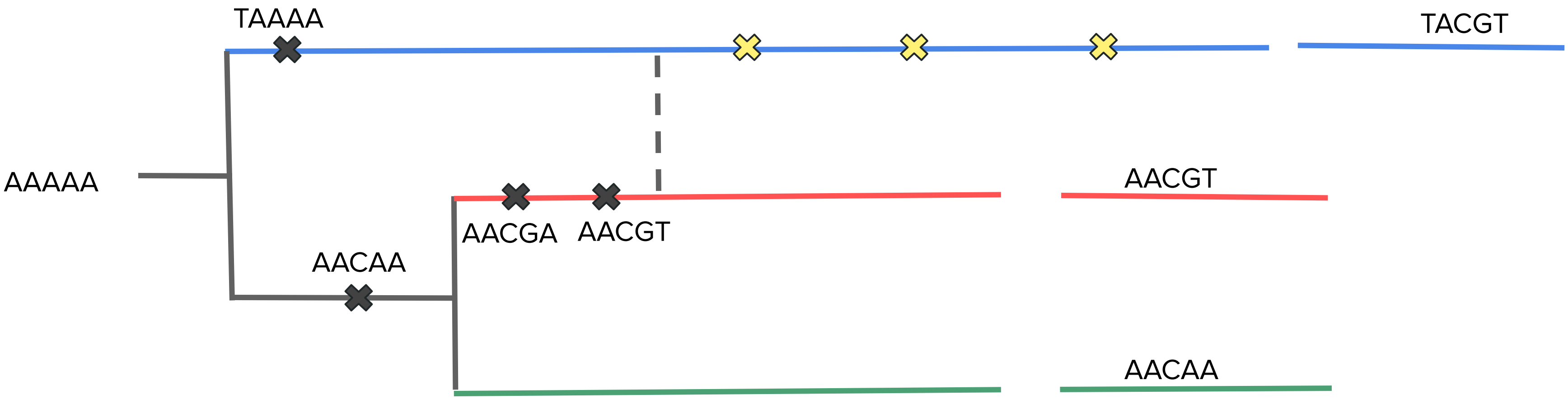
Effect of a single recombination event

A recombination event between two sequences can generate recombinant sequences that are quite genetically divergent from the parent sequences.



Effect of a single recombination event

This will result in abnormally long branches leading to recombinant sequences if recombination is ignored when reconstructing the phylogeny.



Effect of many recombination events

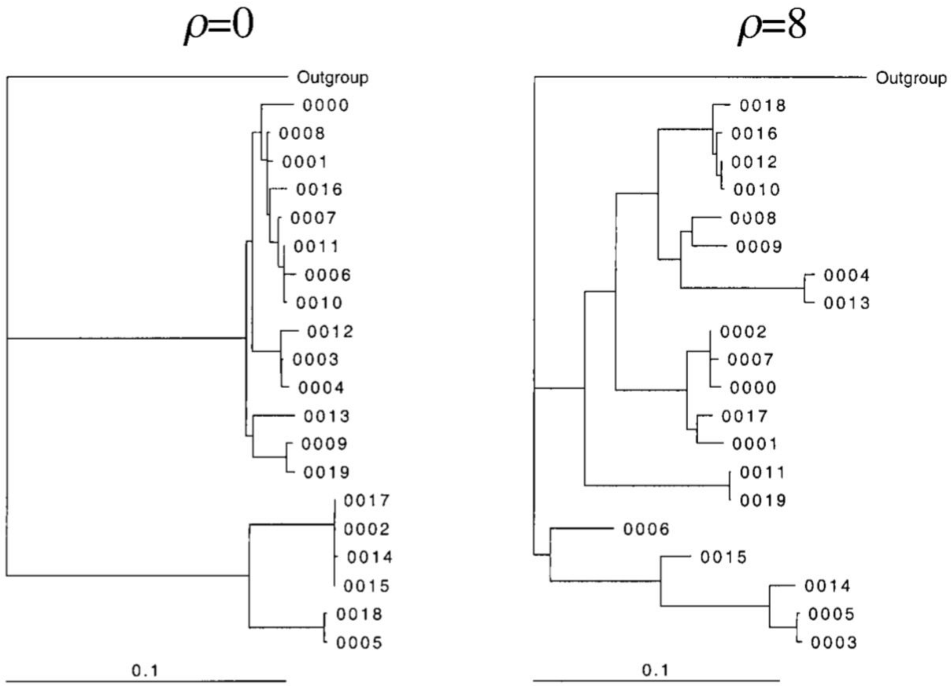
In the presence of multiple recombination events, phylogenies:

- Have longer terminal branches
- Become more star-like
- Behave less clock-like***

*** Wreaks havoc on estimating the molecular clock rate

Schierup and Hein (2000)

Effect of many recombination events



Schierup and Hein (2000)

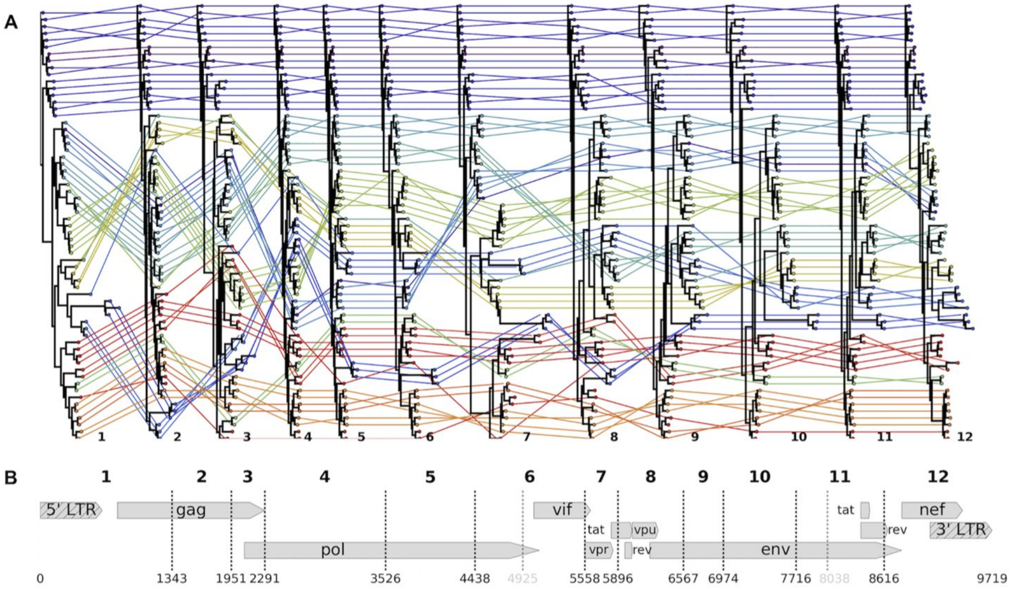
We therefore need to be able to detect and/or account for recombination in phylogenetic analyses

How do we detect recombination?

- Phylogenetic discordance between loci
- Linkage disequilibrium maps
- Triplet sequence tests

Phylogenetic discordance

Phylogenetic discordance between ‘local’ trees can be used to detect recombination but may also arise due to errors in reconstruction.



Bell and Bedford (PLoS Pathogens, 2017)

Linkage disequilibrium

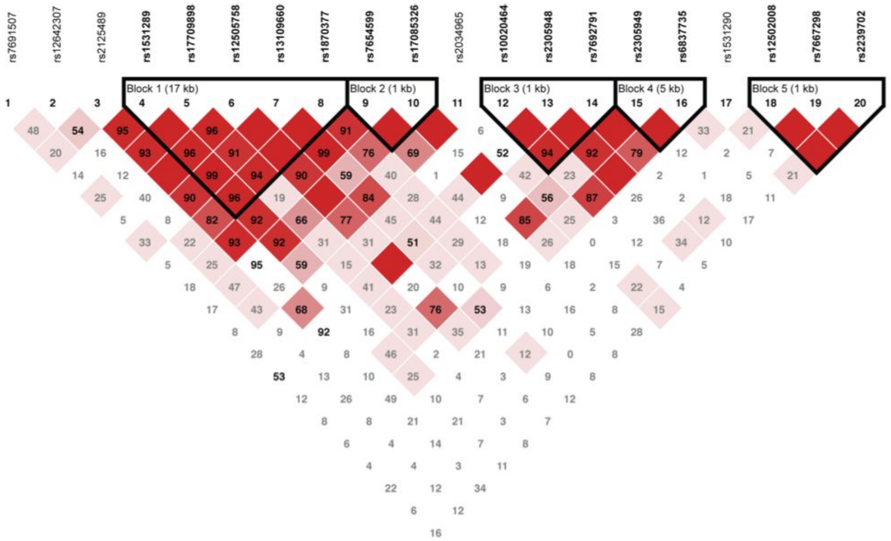
Linkage disequilibrium is the non-random association of alleles at different loci in a given population

LD at the population level may arise due to alleles being physically linked into haplotypes

LD is expected to decay over long distances in the genome due to recombination

Linkage disequilibrium maps

Sharp changes in linkage disequilibrium can indicate recombination in the history of the sample



Fang et al. (2009)

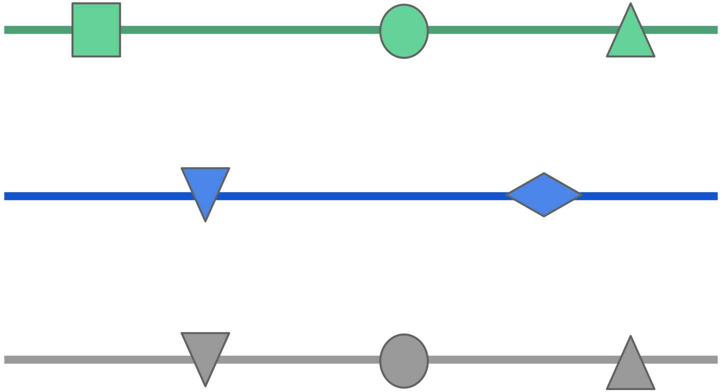
How do we detect recombination?

Many statistical tests of recombination employ a **triplet test**

Three sequences are compared, positing one as a potential child sequence that could have arisen by the two other “parent” sequences recombining.

We’ll consider the 3SEQ test of Boni *et al*. (Genetics, 2007)

The 3SEQ triplet test



Parent *p*

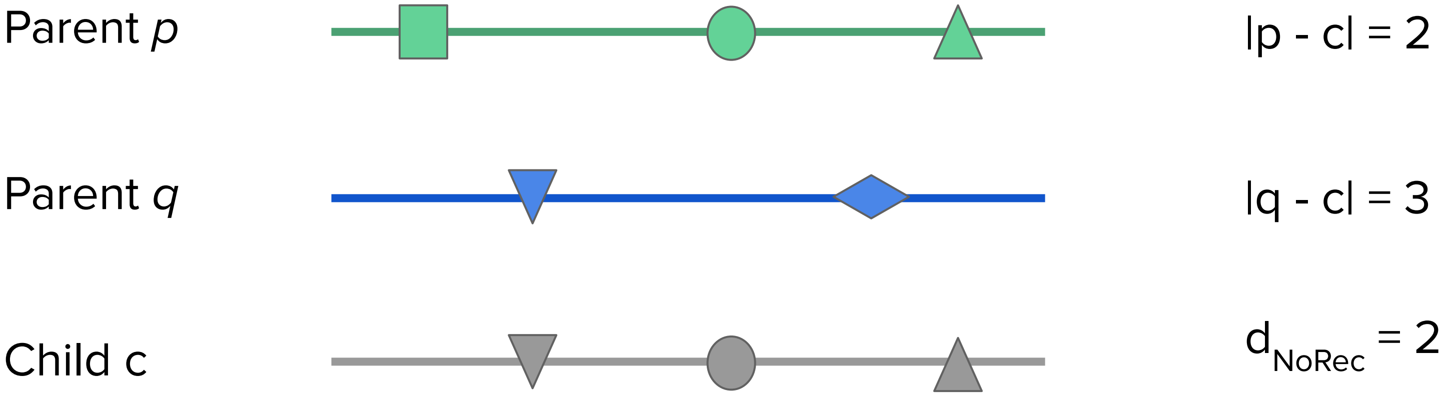
Parent *q*

Child c

Here, |p - q| = 5

Let |p - q| represent the number of mutations separating sequences p and q

The 3SEQ triplet test



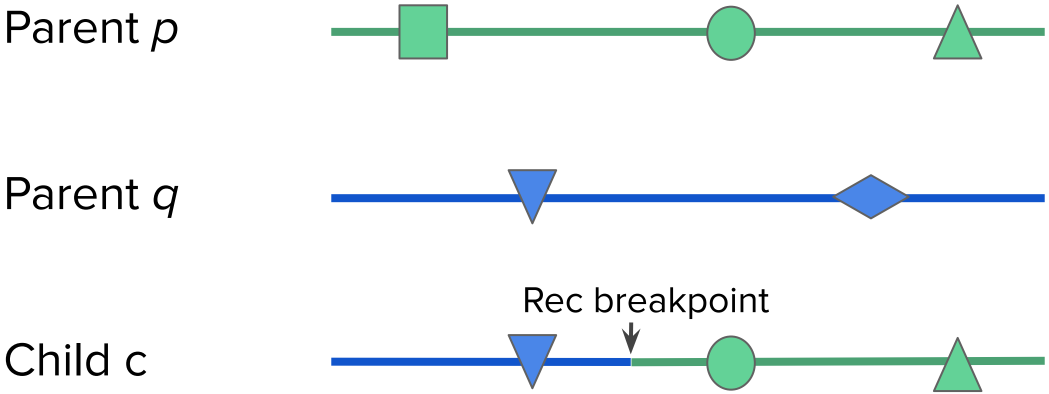
Parent *p* $|p - c| = 2$

Parent *q* $|q - c| = 3$

Child c $d_{NoRec} = 2$

Let d_{NoRec} be the minimum distance from the child to either parent. d_{NoRec} is the number of mutations the child would need to undergo if it descended from one of the parents without recombination.

The 3SEQ triplet test



Parent *p*

Parent *q*

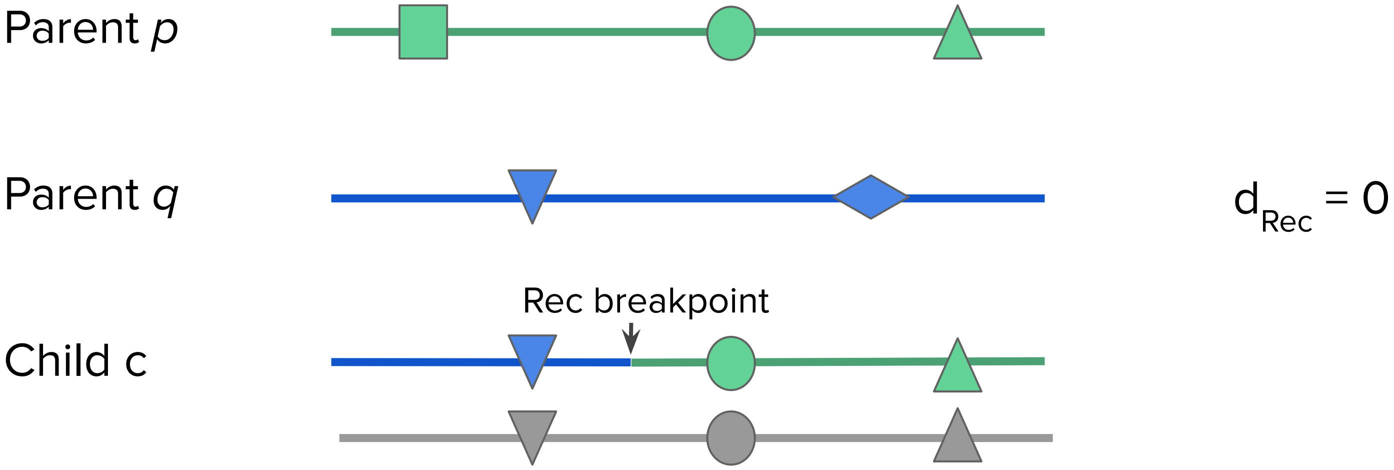
Rec breakpoint

Child c

$$d_{Rec} = \min_{0 \leq l \leq < L} \left(|(pq)_l - c| \right)$$

Let d_{Rec} be the minimum distance between the child and the *optimal* recombinant sequence we can create from parents *p* and *q*.

The 3SEQ triplet test



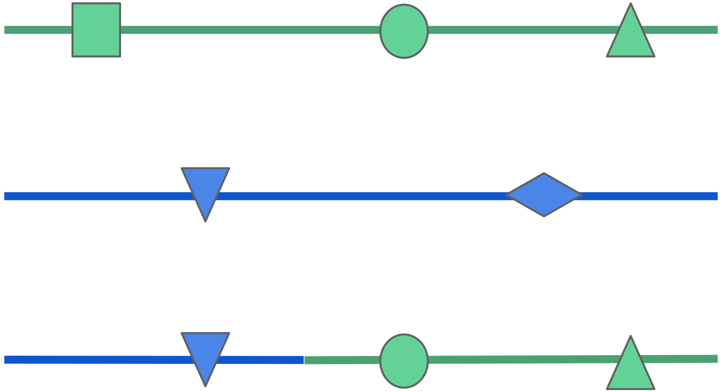
Here the best recombinant we can create has all the same mutations as child c

The 3SEQ triplet test

Parent *p*

Parent *q*

Child c



$d_{NoRec} = 2$

$d_{Rec} = 0$

$\Delta = 2$

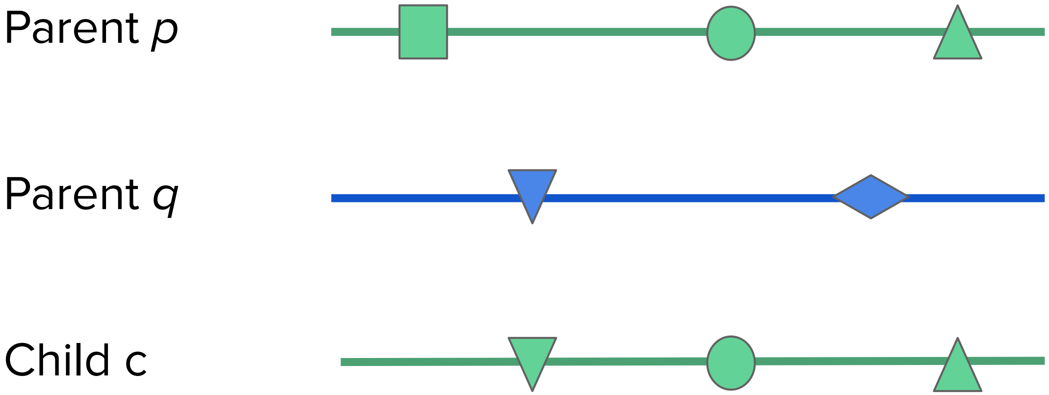
Let $\Delta = d_{NoRec} - d_{Rec}$, the number of mutations that can be “explained” away by recombination.

The 3SEQ triplet test

For any sequence triplet, the larger Δ is, the more evidence there is for recombination.

The problem: a particular sequence triplet could randomly have a large Δ if one side of the child sequence appeared to be closer to parent *p* and the other side appeared closer to parent *q* by chance.

The 3SEQ triplet test



For example, child *c* could have descended from parent *p* but the upside down triangle mutation could have occurred by chance.

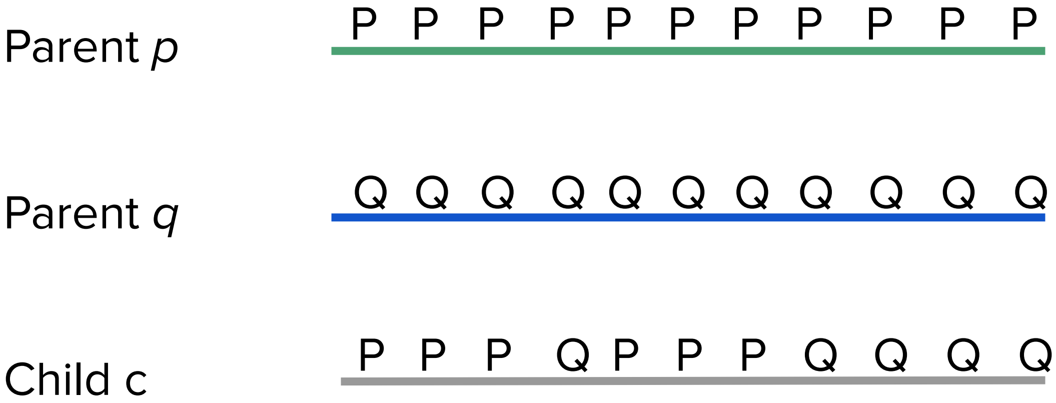
The 3SEQ triplet test

For any sequence triplet, the larger Δ is, the more evidence there is for recombination.

The problem: a particular sequence triplet could have a large Δ by chance if the the one side of the child sequence appeared to be closer to parent *p* and the other side appeared closer to parent *q*.

We therefore need to test whether the **order of mutations** in the child is highly nonrandom or can be explained by chance.

The 3SEQ triplet test



Parent *p* P P P P P P P P P P P

Parent *q* Q Q Q Q Q Q Q Q Q Q Q

Child c P P P Q P P P Q Q Q Q

Let the *P*'s be mutations that the child shares in common with parent p and the *Q*'s be mutations the child shares with parent q

The 3SEQ triplet test

We can think of the mutations as up and down steps in a discrete random walk.

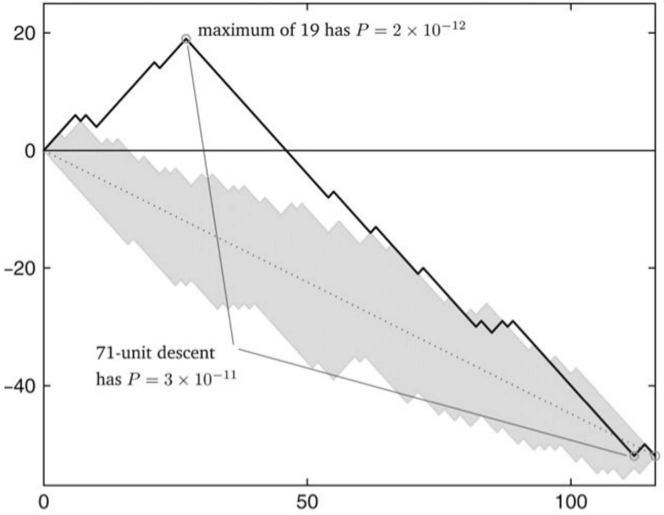
Let the *P*'s be thought of as up steps in the random walk.

And the *Q*'s as down steps.

A hypergeometric random walk model can be used test whether the distribution of *P*'s and *Q*'s is nonrandom based on the height of the random walk.

The 3SEQ test for *Neisseria*

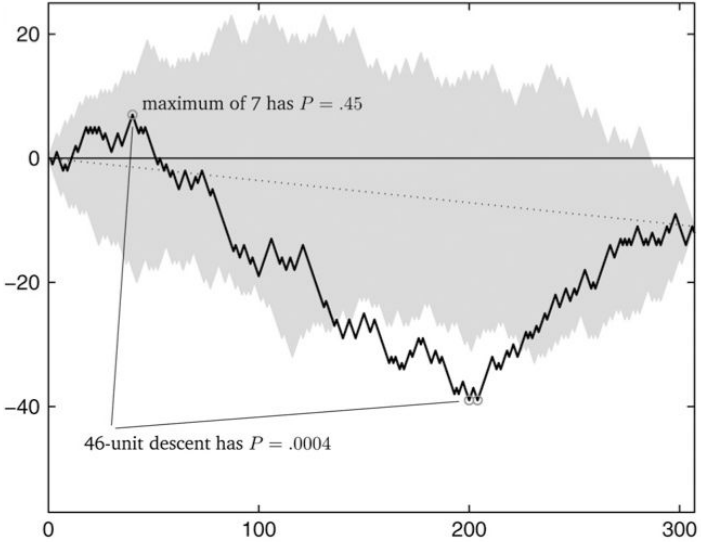
A recombinant will have a statistically improbable heights with its up steps clustered towards one end and down steps clustered towards the other end.



Boni *et al.* (2007)

The 3SEQ test for 1918 Spanish influenza

Small deviations from plausible random walks provide weak evidence for recombination



Boni *et al.* (2007)

Phylogenetic methods that account for recombination

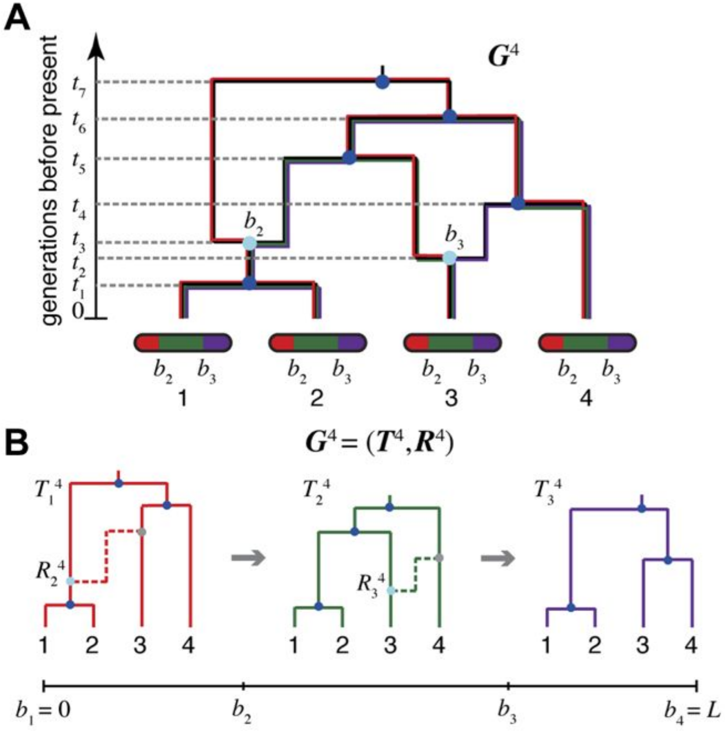
Ancestral recombination graphs

ARGs provide a complete record of the ancestry of all sequences as a graph/network.

This graph includes all recombination and coalescent events in the history of the sample as well as information about the location of recombination breakpoints.

The local phylogeny at each genomic position is embedded in the full ARG

A hypothetical ARG



Rasmussen *et al.* (PLoS Gen, 2014)

Ancestral recombination graphs

ARGs are in theory the ideal way to represent the history of sequences with recombination.

However, even state-of-the-art methods like *ARGweaver* (Rasmussen et al., 2014) that employ very efficient HMM methods work with at most dozens of sequences.

Notoriously difficult to infer full ARGs and generally computationally impossible.

Clonal frames

A **clonal frame** attempts to describe the true ancestral relationships among sampled sequences as a single tree.

Assumes the majority of the genome is inherited clonally while accounting for recombination within certain regions of the genome

Clonal frames are a popular choice for bacteria where the majority of the genome is assumed to be inherited clonally but gene conversion overwrites small portions of the genome.

The ClonalFrameML approach

A ML phylogeny is reconstructed from a multiple genome alignment which is taken to represent the initial clonal frame

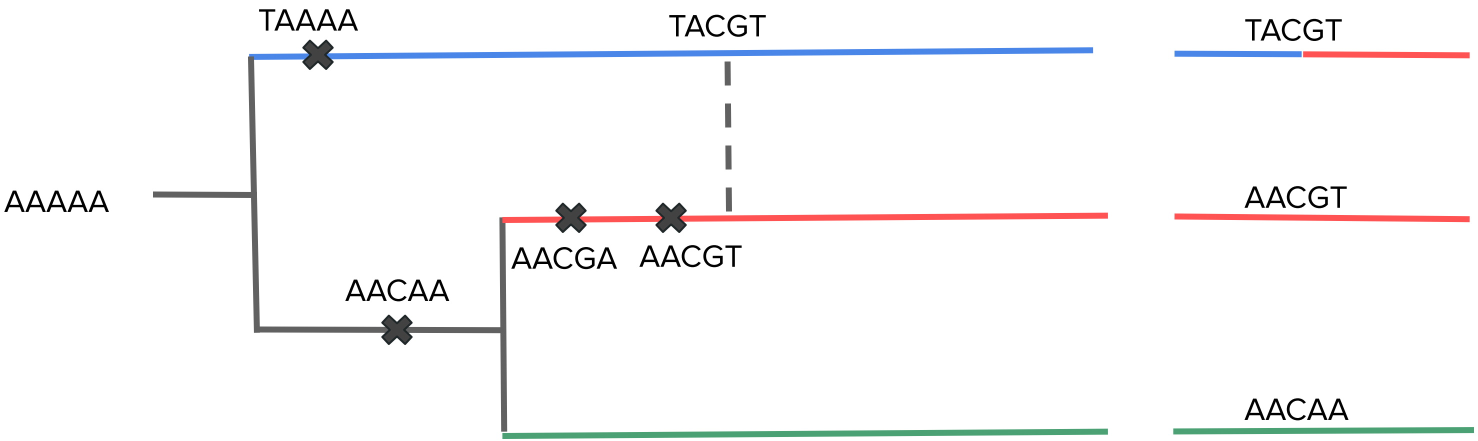
The genomic location of *insertions* caused by recombination are estimated along each branch of the tree using a Hidden Markov Model.

Recombination events are identified and initial ML phylogeny can be refined by ignoring recombinant regions of the genome.

Didelot *et al.* (PLoS Comp Bio, 2015)

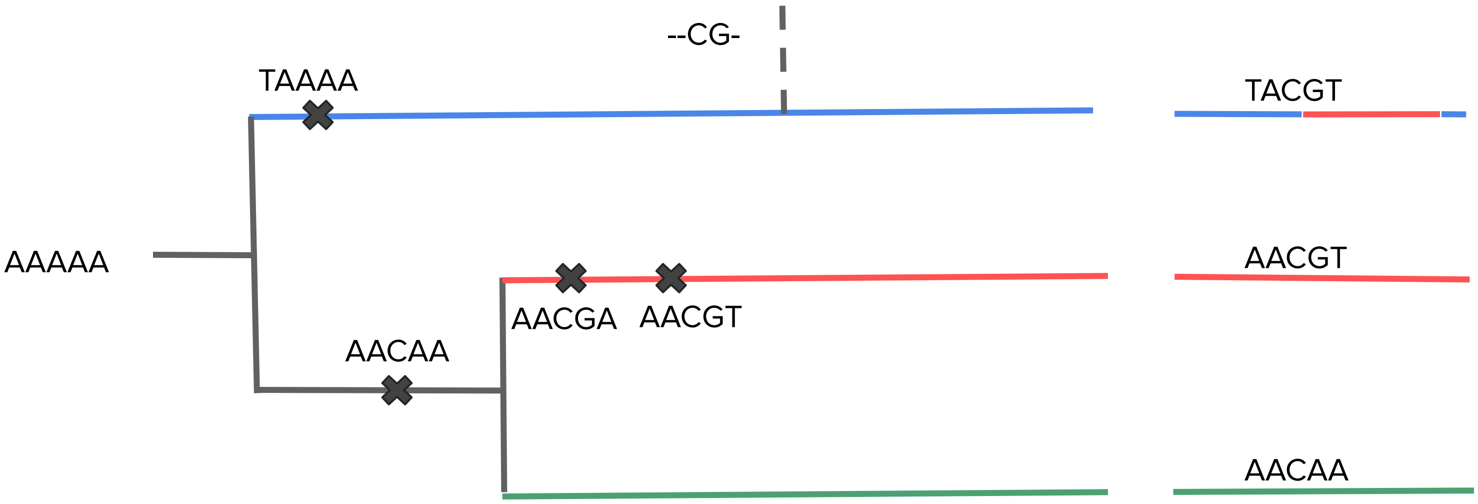
The ClonalFrame model of recombination

The ClonalFrame model of recombination does not consider recombination events between sampled lineages in the phylogeny.

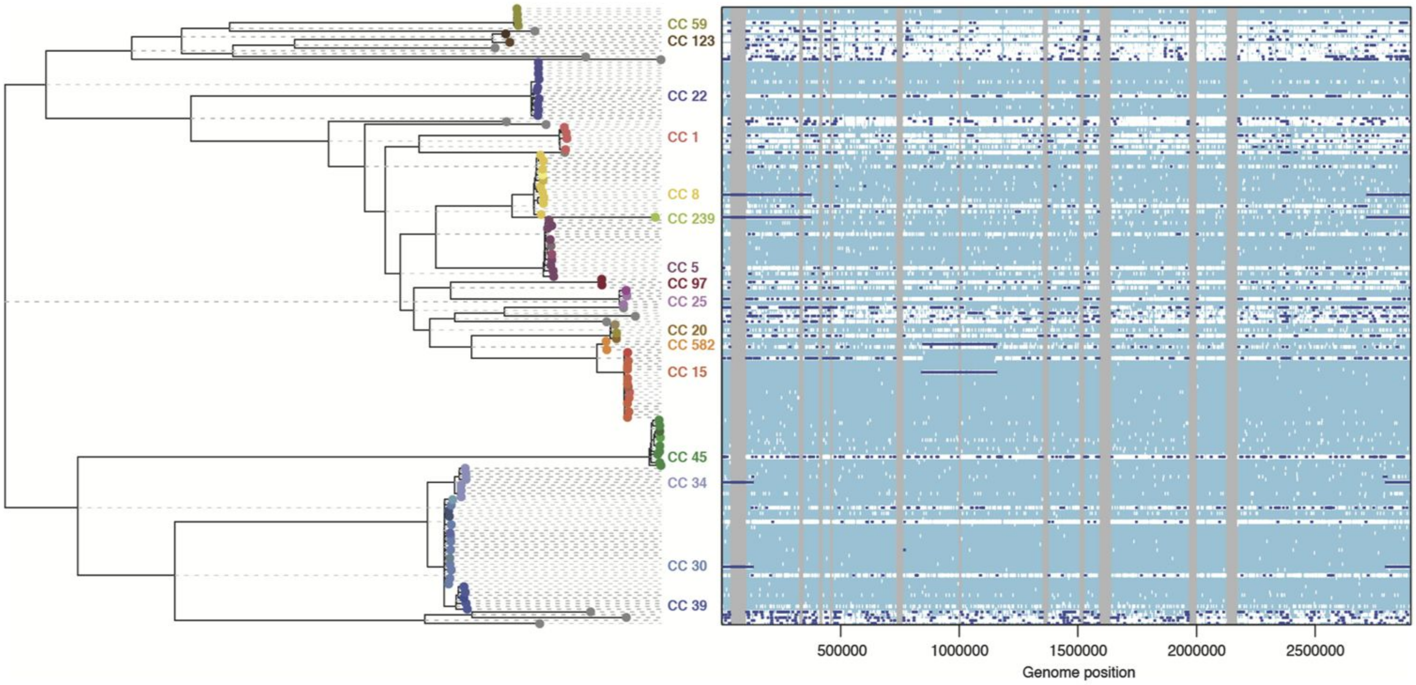


The ClonalFrame model of recombination

Rather the model assumes recombination events overwrite short sequences by inserting genetic material that is **external** to the sampled sequences.



ClonalFrame of *Staphylococcus aureus*



Didelot *et al.* (PLoS Comp Bio, 2015)

Some practical remedies

If reconstructing the full phylogenetic history of the sequences is not the ultimate goal, we can also:

- Infer local phylogenies for different loci or non-recombinant blocks
- Remove potential recombinant sequences if recombinants are rare
- Strip genomic regions affected by recombination from sequence alignments

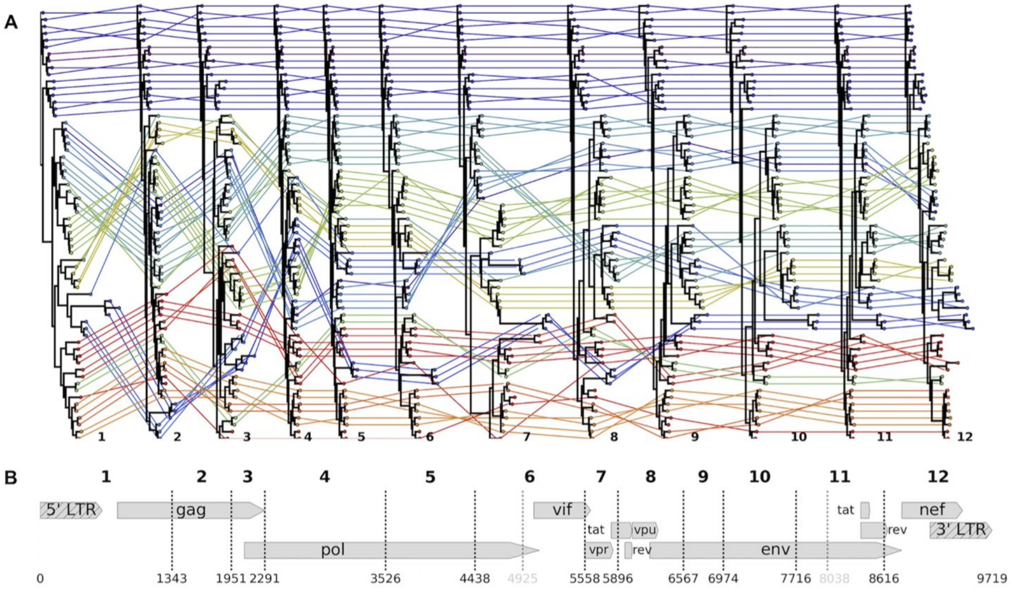
Some practical remedies

If reconstructing the full phylogenetic history of the sequences is not the ultimate goal, we can also:

- Infer local phylogenies for different loci or non-recombinant blocks
- Remove potential recombinant sequences if recombinants are rare
- Strip genomic regions affected by recombination from sequence alignments

Inferring local trees

Local phylogenenies reconstructed from different regions of the genome represent different, albeit correlated, realizations of the evolutionary process.



Bell and Bedford (PLoS Pathogens, 2017)

Recombination vs. mutation rates

Whether or not it is possible to infer phylogenies ultimately depends of the ratio of the recombination rate *r* to the mutation rate *m*.

If $r/m \ll 1$, most changes in the genome occur due to mutation and it will generally be possible to infer local phylogenies within non-recombining regions.

If $r/m > 1$, most changes occur by recombination and there will not be enough mutations between recombination breakpoints to reliably reconstruct phylogenies.

Recombination vs. mutation rates

The ratio r/m varies widely among different microbial pathogens

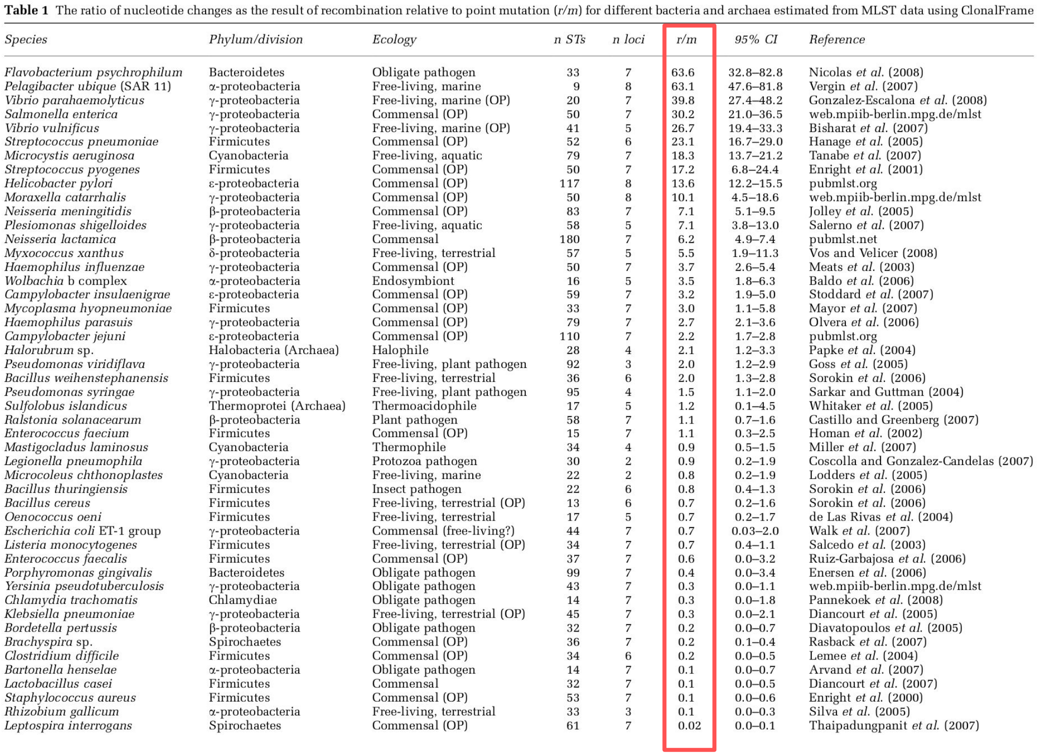


Table 1 The ratio of nucleotide changes as the result of recombination relative to point mutation (*r*/*m*) for different bacteria and archaea estimated from MLST data using ClonalFrame

Species	*Phylum/division*	*Ecology*	*n STs*	*n loci*	*r/m*	*95% CI*	*Reference*
Flavobacterium psychrophilum	Bacteroidetes	Obligate pathogen	33	7	63.6	32.8–82.8	Nicolas *et al.* (2008)
Pelagibacter ubique (SAR 11)	α-proteobacteria	Free-living, marine	9	8	63.1	47.6–81.8	Vergin *et al.* (2007)
Vibrio parahaemolyticus	γ-proteobacteria	Free-living, marine (OP)	20	7	39.8	27.4–48.2	Gonzalez-Escalona *et al.* (2008)
Salmonella enterica	γ-proteobacteria	Commensal (OP)	50	7	30.2	21.0–36.5	web.mpiib-berlin.mpg.de/mlst
Vibrio vulnificus	γ-proteobacteria	Free-living, marine (OP)	41	5	26.7	19.4–33.3	Bisharat *et al.* (2007)
Streptococcus pneumoniae	Firmicutes	Commensal (OP)	52	6	23.1	16.7–29.0	Hanage *et al.* (2005)
Microcystis aeruginosa	Cyanobacteria	Free-living, aquatic	79	7	18.3	13.7–21.2	Tanabe *et al.* (2007)
Streptococcus pyogenes	Firmicutes	Commensal (OP)	50	7	17.2	6.8–24.4	Enright *et al.* (2001)
Helicobacter pylori	ε-proteobacteria	Commensal (OP)	117	8	13.6	12.2–15.5	pubmlst.org
Moraxella catarrhalis	γ-proteobacteria	Commensal (OP)	50	8	10.1	4.5–18.6	web.mpiib-berlin.mpg.de/mlst
Neisseria meningitidis	β-proteobacteria	Commensal (OP)	83	7	7.1	5.1–9.5	Jolley *et al.* (2005)
Plesiomonas shigelloides	γ-proteobacteria	Free-living, aquatic	58	5	7.1	3.8–13.0	Salerno *et al.* (2007)
Neisseria lactamica	β-proteobacteria	Commensal	180	7	6.2	4.9–7.4	pubmlst.net
Myxococcus xanthus	δ-proteobacteria	Free-living, terrestrial	57	5	5.5	1.9–11.3	Vos and Velicer (2008)
Haemophilus influenzae	γ-proteobacteria	Commensal (OP)	50	7	3.7	2.6–5.4	Meats *et al.* (2003)
Wolbachia b complex	α-proteobacteria	Endosymbiont	16	5	3.5	1.8–6.3	Baldo *et al.* (2006)
Campylobacter insulaenigrae	ε-proteobacteria	Commensal (OP)	59	7	3.2	1.9–5.0	Stoddard *et al.* (2007)
Mycoplasma hyopneumoniae	Firmicutes	Commensal (OP)	33	7	3.0	1.1–5.8	Mayor *et al.* (2007)
Haemophilus parasuis	γ-proteobacteria	Commensal (OP)	79	7	2.7	2.1–3.6	Olvera *et al.* (2006)
Campylobacter jejuni	ε-proteobacteria	Commensal (OP)	110	7	2.2	1.7–2.8	pubmlst.org
Halorubrum sp.	Halobacteria (Archaea)	Halophile	28	4	2.1	1.2–3.3	Papke *et al.* (2004)
Pseudomonas viridiflava	γ-proteobacteria	Free-living, plant pathogen	92	3	2.0	1.2–2.9	Goss *et al.* (2005)
Bacillus weihenstephanensis	Firmicutes	Free-living, terrestrial	36	6	2.0	1.3–2.8	Sorokin *et al.* (2006)
Pseudomonas syringae	γ-proteobacteria	Free-living, plant pathogen	95	4	1.5	1.1–2.0	Sarkar and Guttman (2004)
Sulfolobus islandicus	Thermoprotei (Archaea)	Thermoacidophile	17	5	1.2	0.1–4.5	Whitaker *et al.* (2005)
Ralstonia solanacearum	β-proteobacteria	Plant pathogen	58	7	1.1	0.7–1.6	Castillo and Greenberg (2007)
Enterococcus faecium	Firmicutes	Commensal (OP)	15	7	1.1	0.3–2.5	Homan *et al.* (2002)
Mastigocladus laminosus	Cyanobacteria	Thermophile	34	4	0.9	0.5–1.5	Miller *et al.* (2007)
Legionella pneumophila	γ-proteobacteria	Protozoa pathogen	30	2	0.9	0.2–1.9	Coscolla and Gonzalez-Candelas (2007)
Microcoleus chthonoplastes	Cyanobacteria	Free-living, marine	22	2	0.8	0.2–1.9	Lodders *et al.* (2005)
Bacillus thuringiensis	Firmicutes	Insect pathogen	22	6	0.8	0.4–1.3	Sorokin *et al.* (2006)
Bacillus cereus	Firmicutes	Free-living, terrestrial (OP)	13	6	0.7	0.2–1.6	Sorokin *et al.* (2006)
Oenococcus oeni	Firmicutes	Free-living, terrestrial	17	5	0.7	0.2–1.7	de Las Rivas *et al.* (2004)
Escherichia coli ET-1 group	γ-proteobacteria	Commensal (free-living?)	44	7	0.7	0.03–2.0	Walk *et al.* (2007)
Listeria monocytogenes	Firmicutes	Free-living, terrestrial (OP)	34	7	0.7	0.4–1.1	Salcedo *et al.* (2003)
Enterococcus faecalis	Firmicutes	Commensal (OP)	37	7	0.6	0.0–3.2	Ruiz-Garbajosa *et al.* (2006)
Porphyromonas gingivalis	Bacteroidetes	Obligate pathogen	99	7	0.4	0.0–3.4	Enersen *et al.* (2006)
Yersinia pseudotuberculosis	γ-proteobacteria	Obligate pathogen	43	7	0.3	0.0–1.1	web.mpiib-berlin.mpg.de/mlst
Chlamydia trachomatis	Chlamydiae	Obligate pathogen	14	7	0.3	0.0–1.8	Pannekoek *et al.* (2008)
Klebsiella pneumoniae	γ-proteobacteria	Free-living, terrestrial (OP)	45	7	0.3	0.0–2.1	Diancourt *et al.* (2005)
Bordetella pertussis	β-proteobacteria	Obligate pathogen	32	7	0.2	0.0–0.7	Diavatopoulos *et al.* (2005)
Brachyspira sp.	Spirochaetes	Commensal (OP)	36	7	0.2	0.1–0.4	Rasback *et al.* (2007)
Clostridium difficile	Firmicutes	Commensal (OP)	34	6	0.2	0.0–0.5	Lemee *et al.* (2004)
Bartonella henselae	α-proteobacteria	Obligate pathogen	14	7	0.1	0.0–0.7	Arvand *et al.* (2007)
Lactobacillus casei	Firmicutes	Commensal	32	7	0.1	0.0–0.5	Diancourt *et al.* (2007)
Staphylococcus aureus	Firmicutes	Commensal (OP)	53	7	0.1	0.0–0.6	Enright *et al.* (2000)
Rhizobium gallicum	α-proteobacteria	Free-living, terrestrial	33	3	0.1	0.0–0.3	Silva *et al.* (2005)
Leptospira interrogans	Spirochaetes	Commensal (OP)	61	7	0.02	0.0–0.1	Thaipadungpanit *et al.* (2007)

Vos & Didelot (ISME, 2008)

On Wednesday we will look at how to detect recombination using RDP4.