NC STATE UNIVERSITY

Non-tree like evolution: Detecting and accounting for recombination

Molecular Epidemiology of Infectious Diseases
Lecture 6

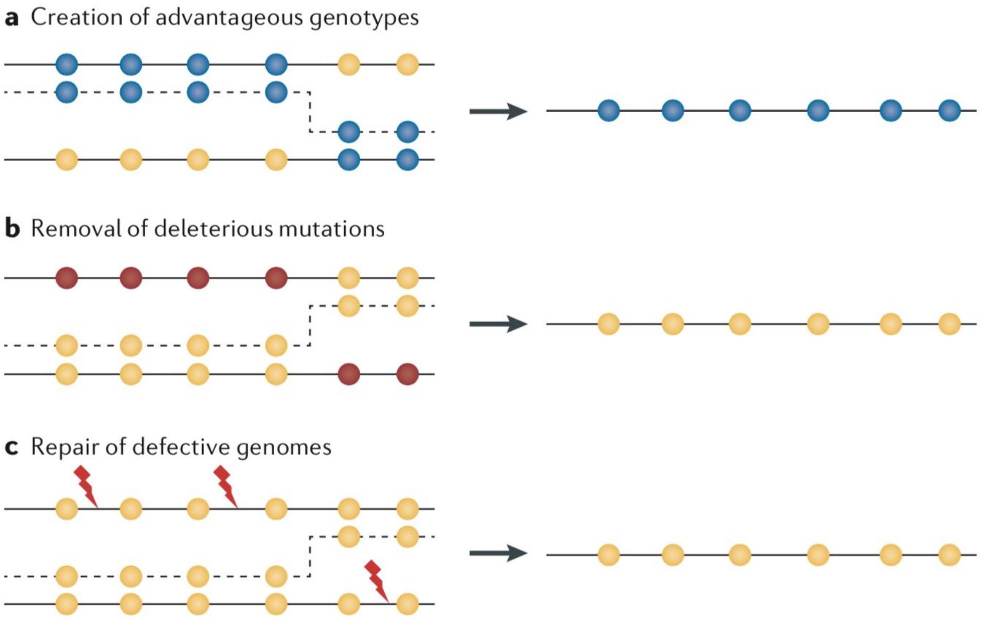
February 19th, 2024

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

The advantages of recombination

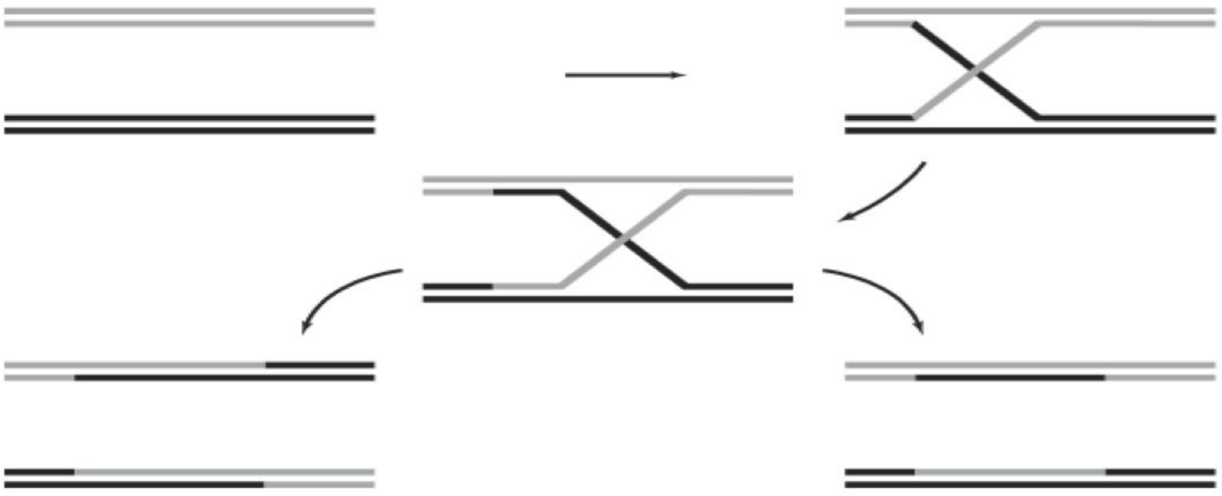
Similar to sexual reproduction, recombination can shuffle parental genetic material to:

- Combine beneficial mutations
- Purge deleterious mutations
- Repair defective genomes


a Creation of advantageous genotypes
b Removal of deleterious mutations
c Repair of defective genomes

Simon-Loriere and Holmes (Nat. Rev. Micro., 2011)

Mechanisms of recombination



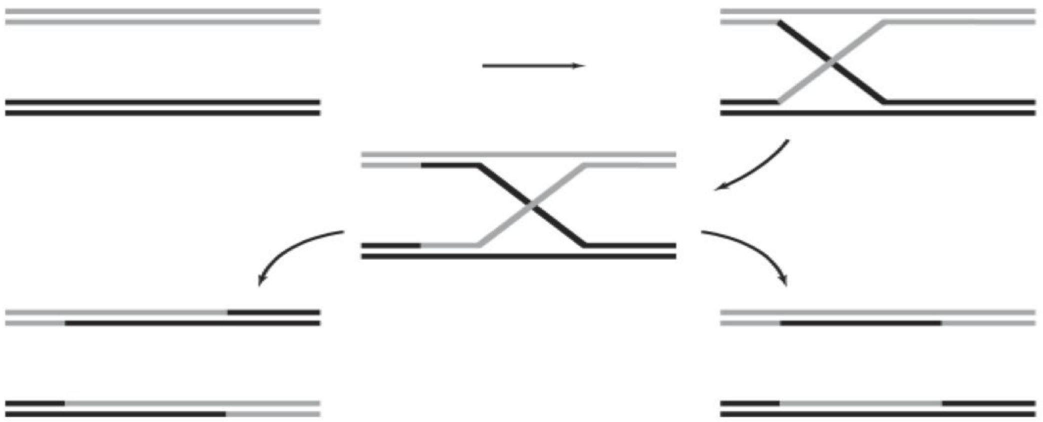
Recombination with crossovers

Recombination via gene conversion

Hein *et al.* (2004)

Mechanisms of recombination

In eukaryotes, recombination is typically due to crossover events



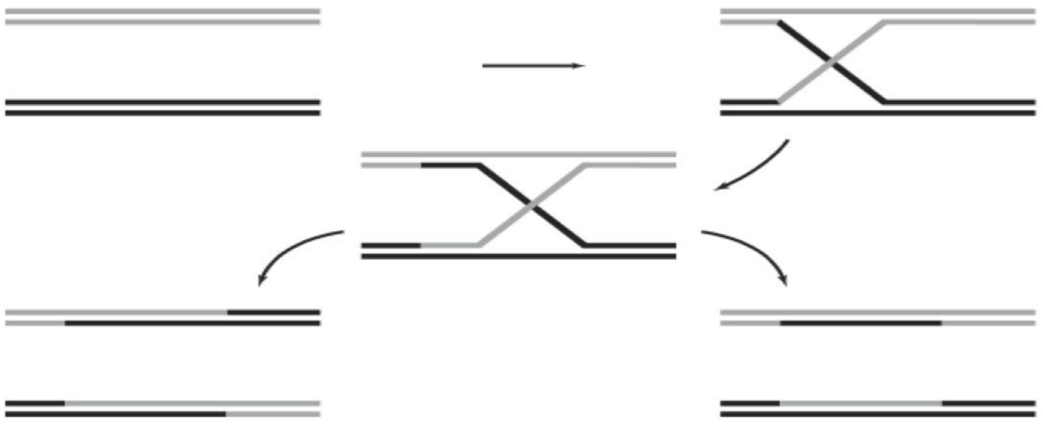
Recombination with crossovers

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

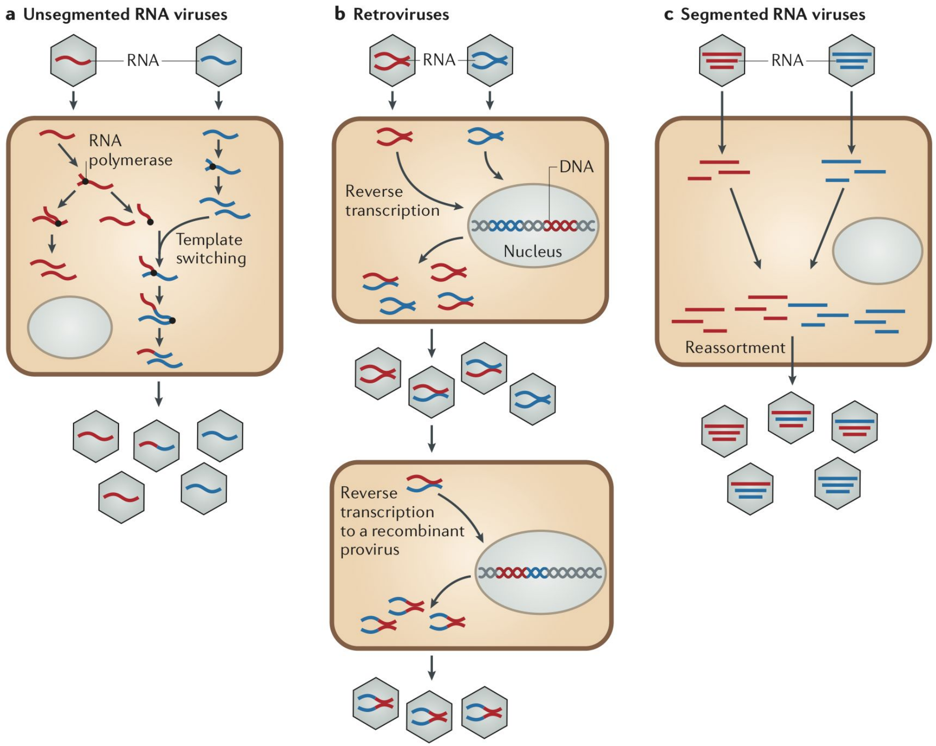
Recombination via gene conversion

Hein *et al.* (2004)

Mechanisms of viral recombination

Co-infection of a cell by genetically distinct viral strains can lead to the generation of recombinant viruses.

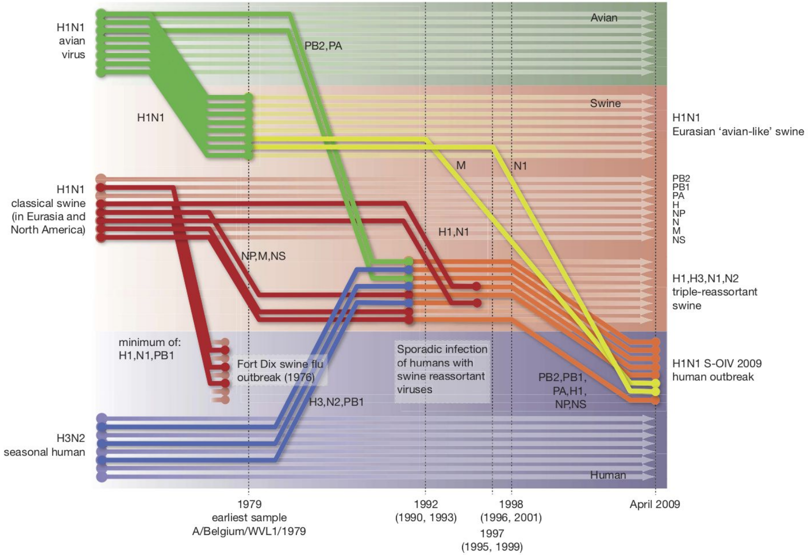
End result: progeny inherit genetic material from both parents.



Simon-Loriere and Holmes (Nat. Rev. Micro., 2011)

Mechanisms of recombination

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



Smith *et al.* (Nature, 2009)

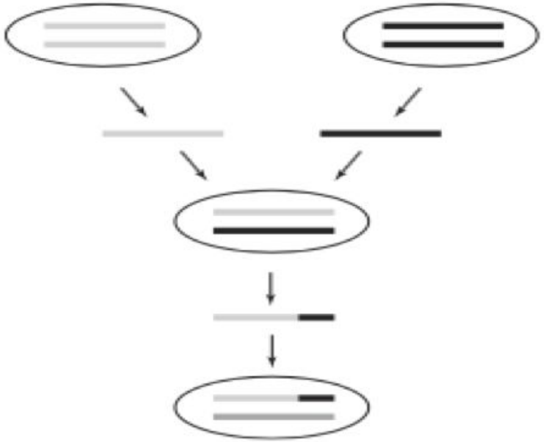
If recombination is so good for pathogens, why is it so bad for phylogenetics?

Recombination creates mosaic ancestry

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

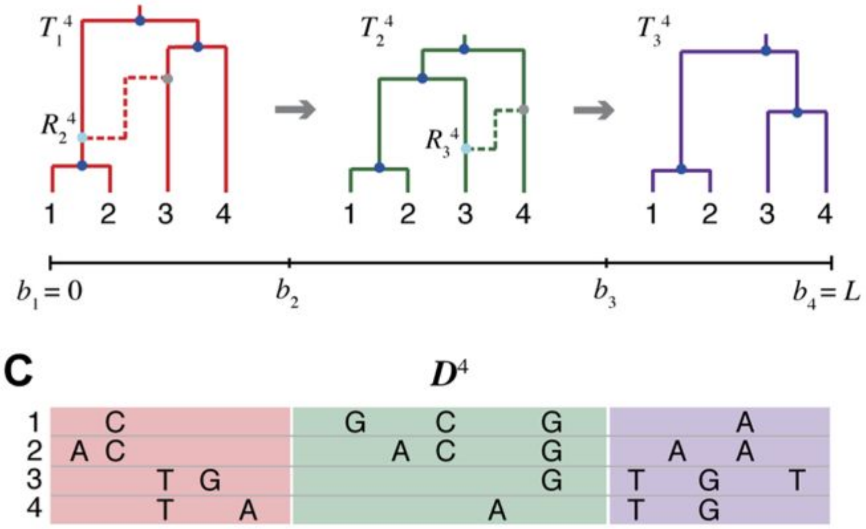
With recombination, genomes become mosaics where different segments descend from different ancestors.

No single phylogenetic tree can therefore describe the genetic 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)

Effect of a single recombination event

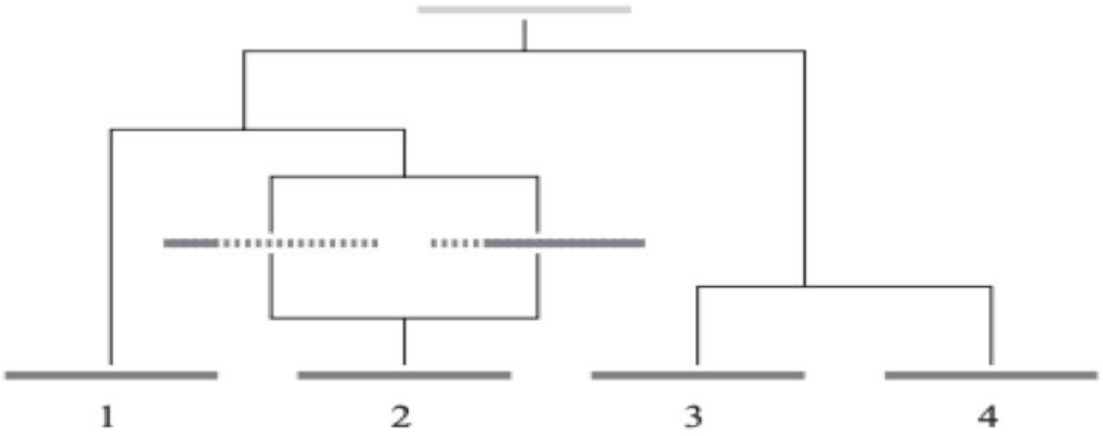
A single recombination event between two sampled lineages will have one of three possible effects on the phylogeny:

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

Hein *et al.* (2004)

Effect of a single recombination event

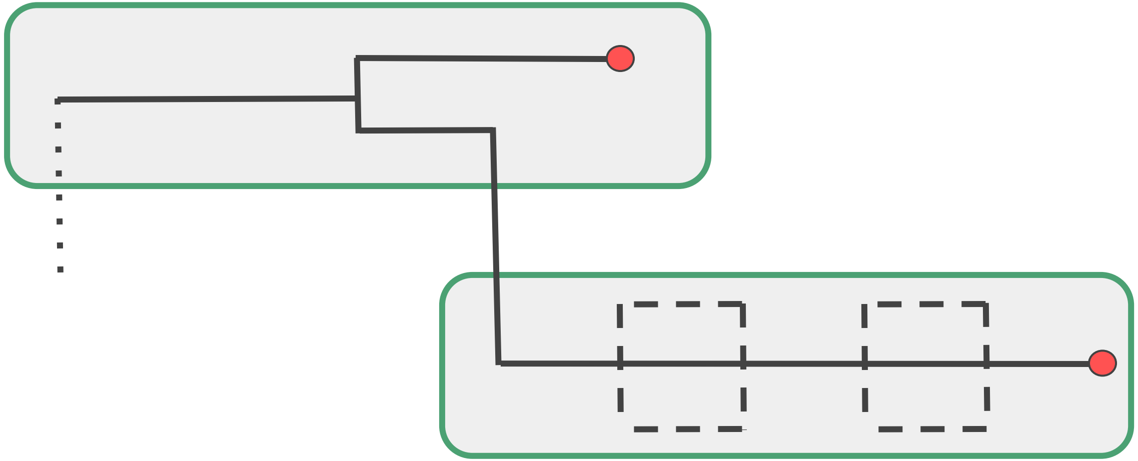
If two recombinant sequences coalesce before they coalesce with any other lineage, the recombination event will have **no effect** on the phylogeny.



Hein *et al.* (2004)

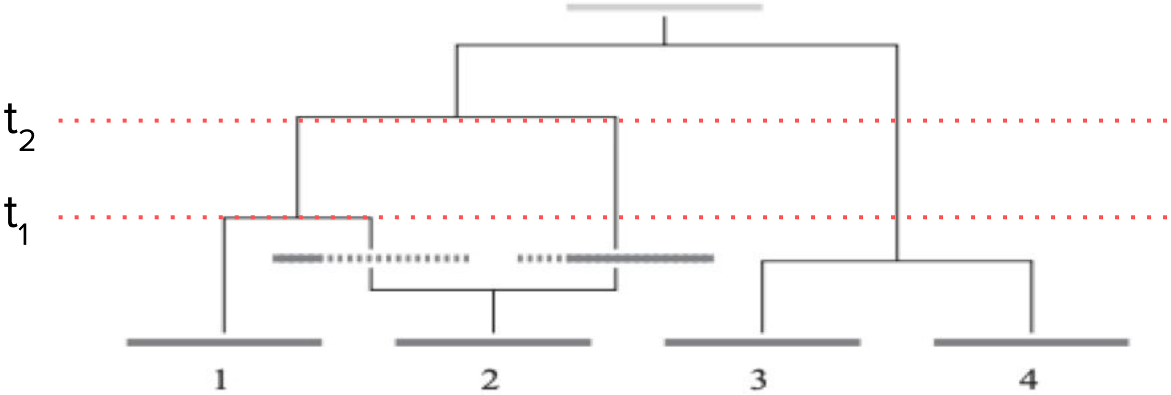
Effect of a single recombination event

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



Effect of a single recombination event

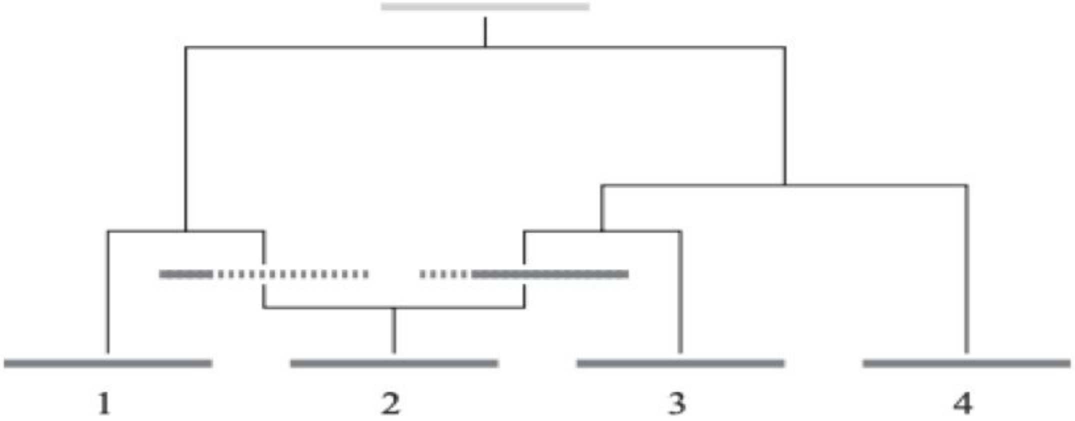
Only **branch lengths will change** if one of two recombining sequences merges with another sequence before coalescing with the other recombining sequence again.



Hein *et al.* (2004)

Effect of a single recombination event

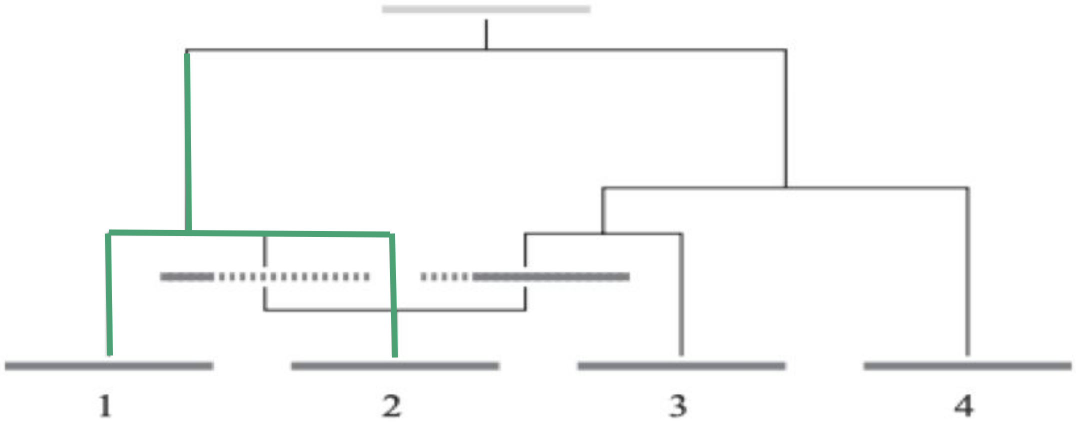
The **tree topology will change** if the two recombining sequences coalesce with other sequences before the two recombining sequences coalesce.



Hein *et al.* (2004)

Effect of a single recombination event

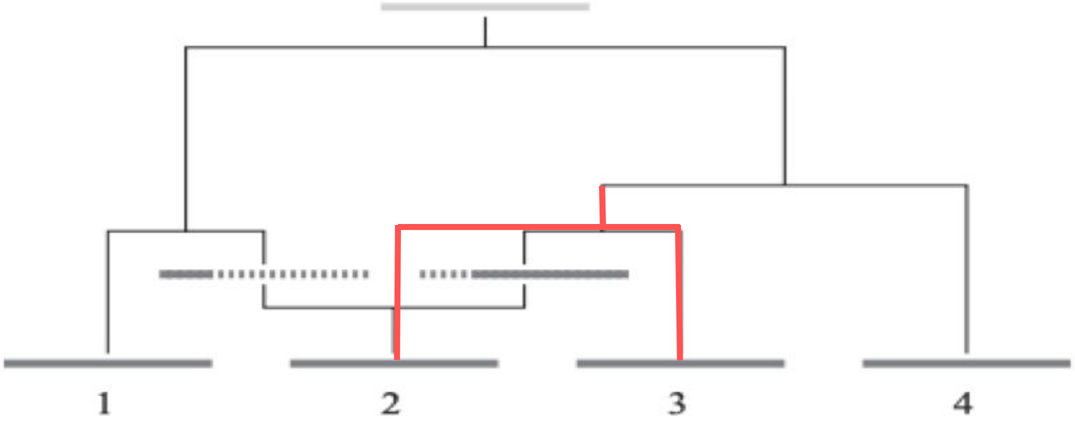
The **tree topology will change** if the two recombining sequences coalesce with other sequences before the two recombining sequences coalesce.



Hein *et al.* (2004)

Effect of a single recombination event

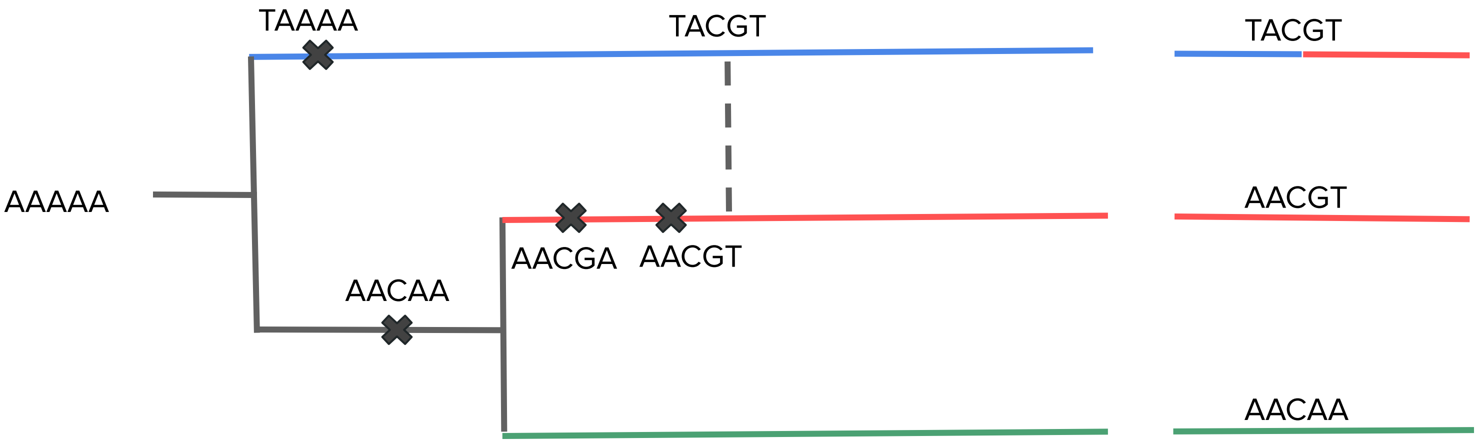
The **tree topology will change** if the two recombining sequences coalesce with other sequences before the two recombining sequences coalesce.



Hein *et al.* (2004)

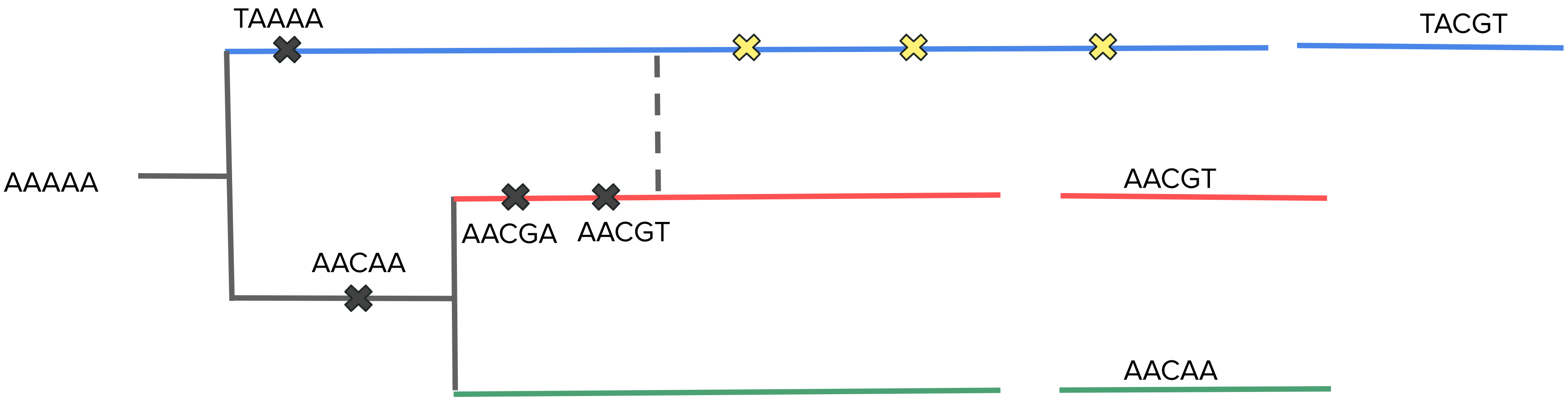
Effect of a single recombination event

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

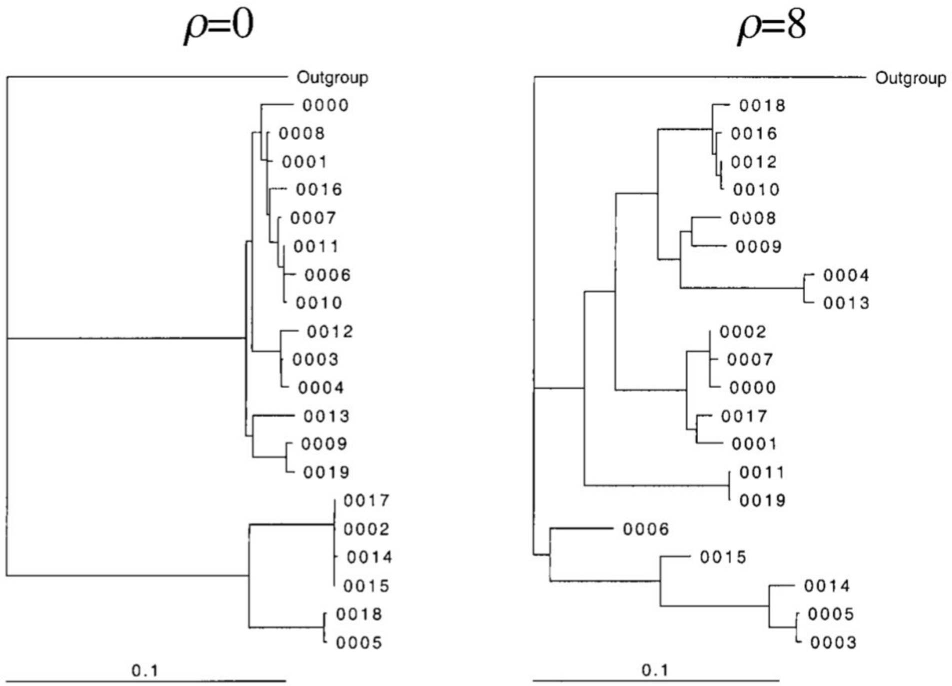
In the presence of multiple recombination events, phylogenies:

- Have longer terminal branches
- Tree shape become more star-like
- Mutations accumulate in a less clock-like manner***

*** Wreaks havoc on estimating the molecular clock rate

Schierup and Hein (2000)

Effect of many recombination events



Schierup and Hein (2000)

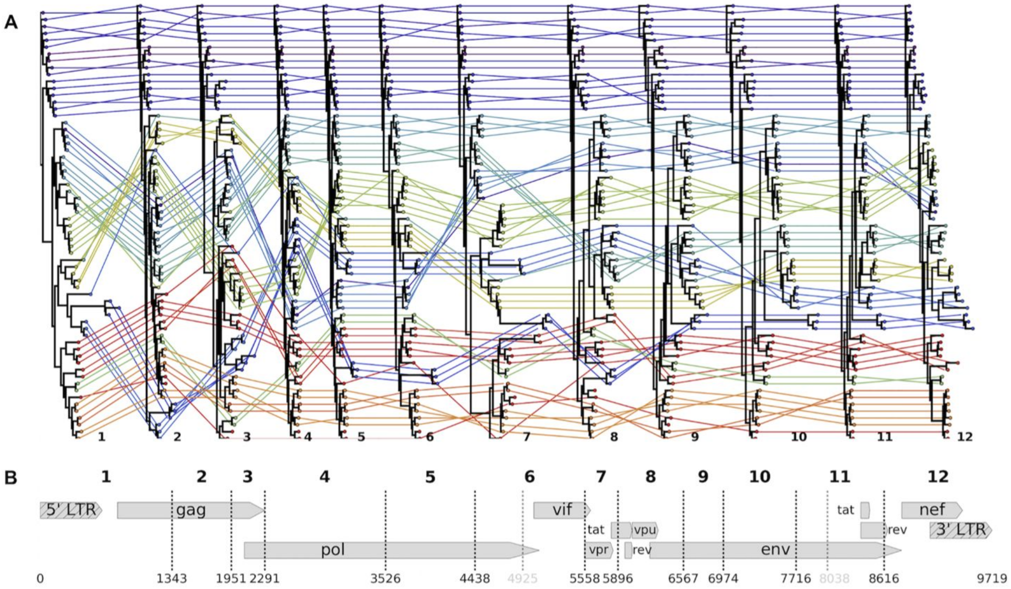
We therefore need to be able to detect and/or account for recombination in phylogenetic analyses

How do we detect recombination?

- Phylogenetic discordance between loci
- Linkage disequilibrium maps
- Substitution distribution/mosaic tests

Phylogenetic discordance

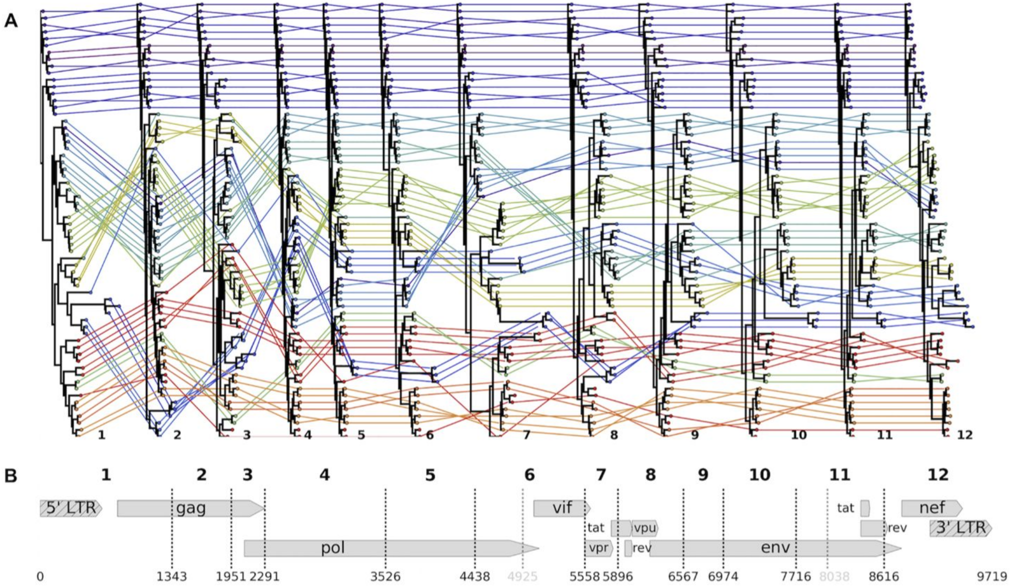
Phylogenetic discordance between ‘local’ trees can be used to detect recombination but may also arise due to errors in reconstruction.



Bell and Bedford (PLoS Pathogens, 2017)

Phylogenetic discordance

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



Phylogenetic recombination detection methods like **GARD** (Pond *et al*., 2006) allow for statistical tests of discordance.

Bell and Bedford (PLoS Pathogens, 2017)

How do we detect recombination?

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

Linkage disequilibrium

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

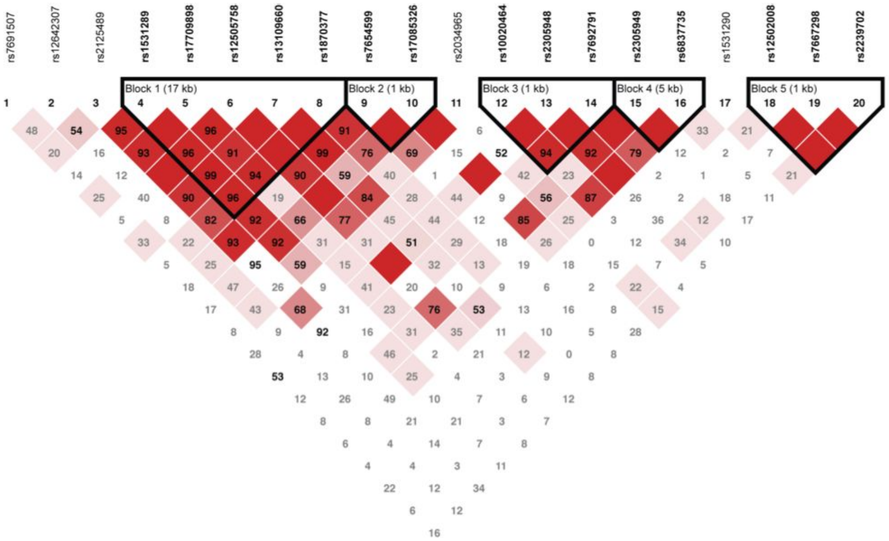
LD at the population level may arise due to alleles being physically linked into haplotypes.

LD can be quantified by looking at correlations in the presence/absence of alleles between different sites.

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

Linkage disequilibrium maps

Sharp changes in linkage disequilibrium -- correlations in the presence/absence of alleles -- can indicate recombination in the history of the sample



Linkage disequilibrium: correlations between sites in the presence or absence of alleles.

Fang et al. (2009)

How do we detect recombination?

- Phylogenetic discordance between loci
- Linkage disequilibrium maps
- Substitution distribution/mosaic tests

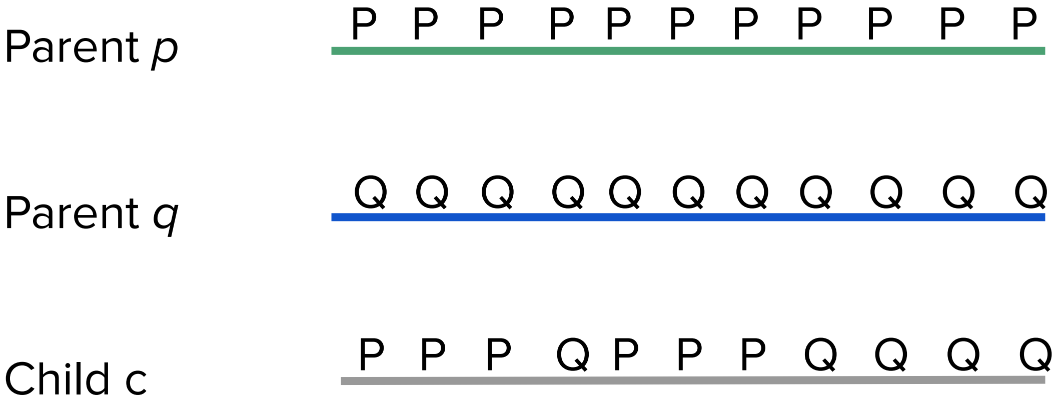
How do we detect recombination?

Many substitution methods test for clustering/mosaicism of mutations in configurations that are unlikely to have evolved by mutation alone.

Pairs or triplets of sequences are compared, one is assumed to be a potential child sequence that could have arisen by the other “parent” sequences recombining.

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

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.

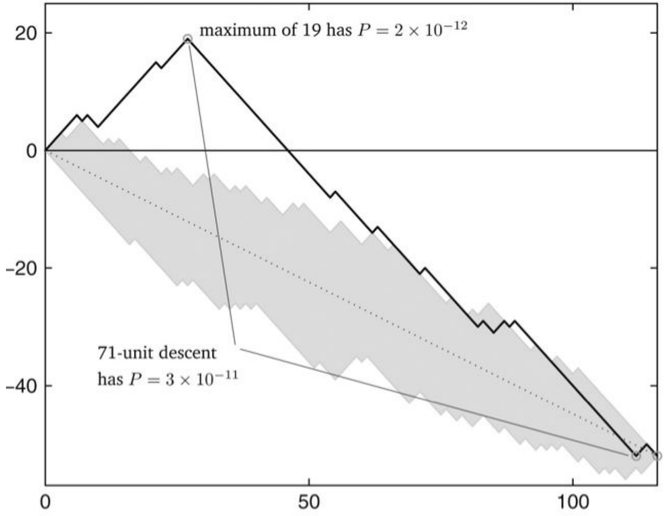
Let the *P*'s be thought of as up steps in the random walk.

And the *Q*'s as down steps.

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

The 3SEQ test for *Neisseria*

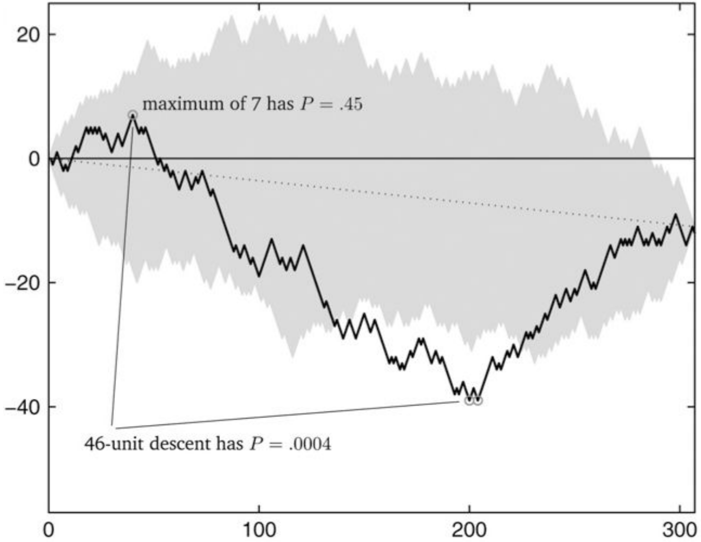
A recombinant will have a statistically improbable heights with its up steps clustered towards one end and down steps clustered towards the other end.



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)

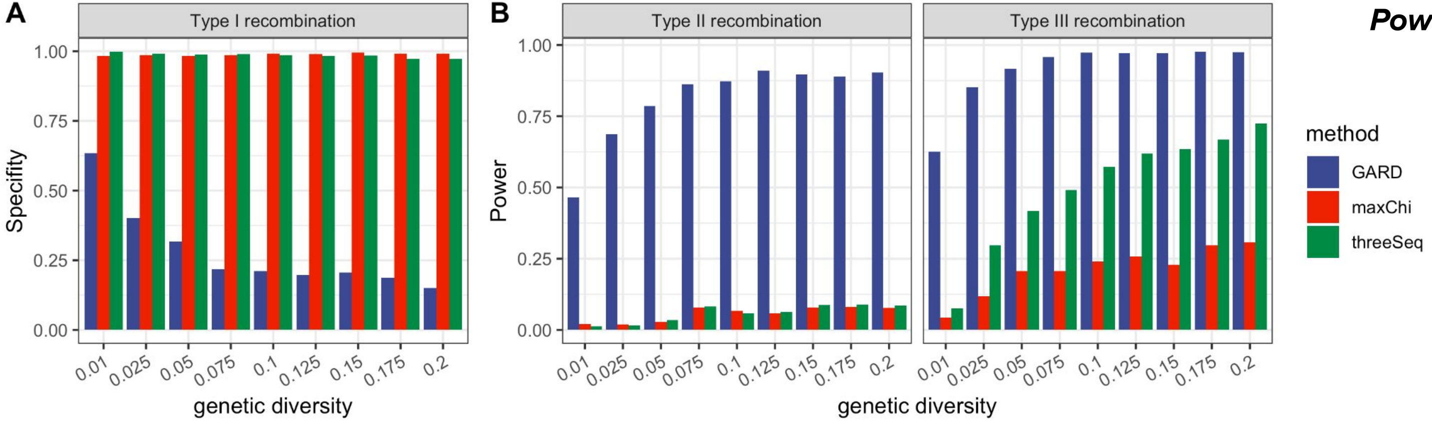
But which methods work best for detecting and localizing recombination breakpoints?

Sensitivity versus specificity

Detection power increases with genetic diversity but there is a tradeoff between power (sensitivity) and specificity.

Specificity = True negative rate

Power = True positive rate



Shi Cen

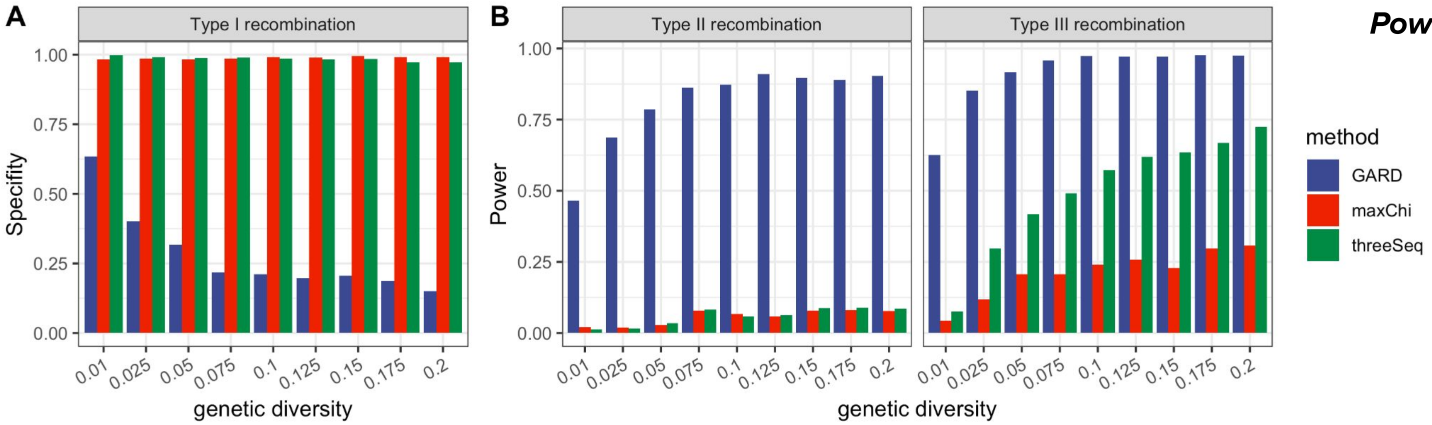
Sensitivity versus specificity

Phylogenetic discordance methods have higher power but low specificity.

Substitution methods like 3SEQ have lower power but very high specificity.

Specificity = True negative rate

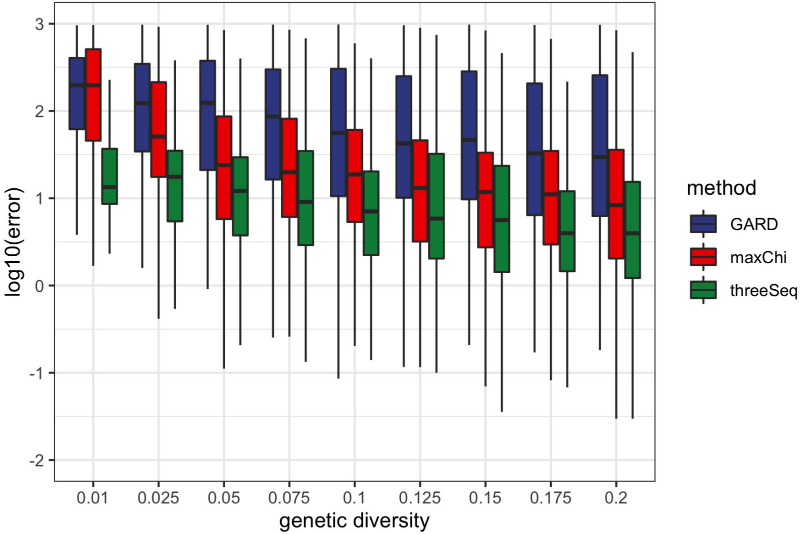
Power = True positive rate

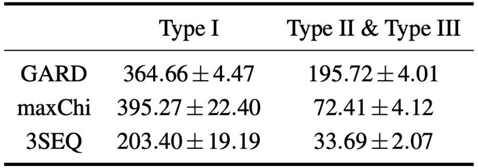


Cen & Rasmussen (bioRxiv, 2023)

Breakpoint location accuracy

3SEQ performs best in accurately locating breakpoints but is still highly dependent on patterns of genetic polymorphisms in the sequences.





	Type I	Type II & Type III
GARD	364.66 ± 4.47	195.72 ± 4.01
maxChi	395.27 ± 22.40	72.41 ± 4.12
3SEQ	203.40 ± 19.19	33.69 ± 2.07

Localization accuracy (mean±SEM) of three detection methods

Cen & Rasmussen (bioRxiv, 2023)

Phylogenetic methods that account for recombination

Some potential options

Remove recombinant sequences from alignments.

Remove recombinant genomic regions and reconstruct local trees from recombination-free blocks.

Assume evolution is mostly tree-like and reconstruct a clonal frame

Reconstruct a full ancestral recombination graph

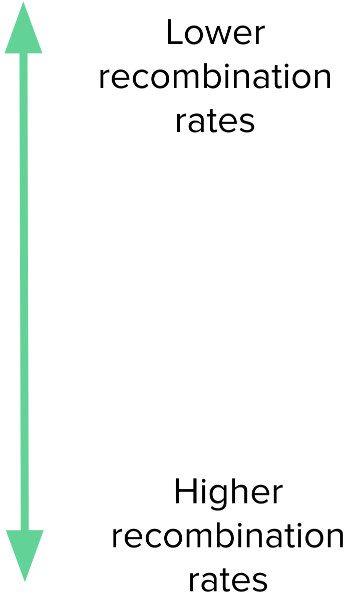
Some potential options

Remove recombinant sequences from alignments.

Remove recombinant genomic regions and reconstruct local trees from recombination-free blocks.

Assume evolution is mostly tree-like and reconstruct a clonal frame

Reconstruct a full ancestral recombination graph



Lower recombination rates

Higher recombination rates

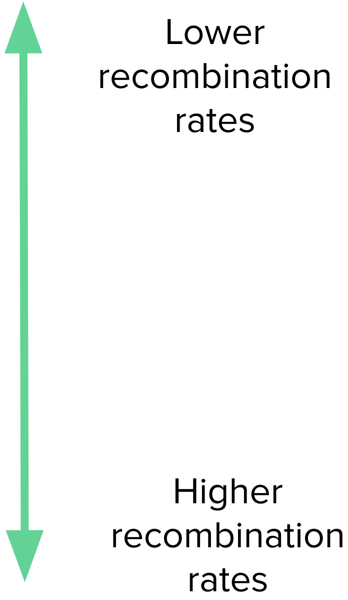
Some potential options

Remove recombinant sequences from alignments.

Remove recombinant genomic regions and reconstruct local trees from recombination-free blocks.

Assume evolution is mostly tree-like and reconstruct a clonal frame

Reconstruct a full ancestral recombination graph

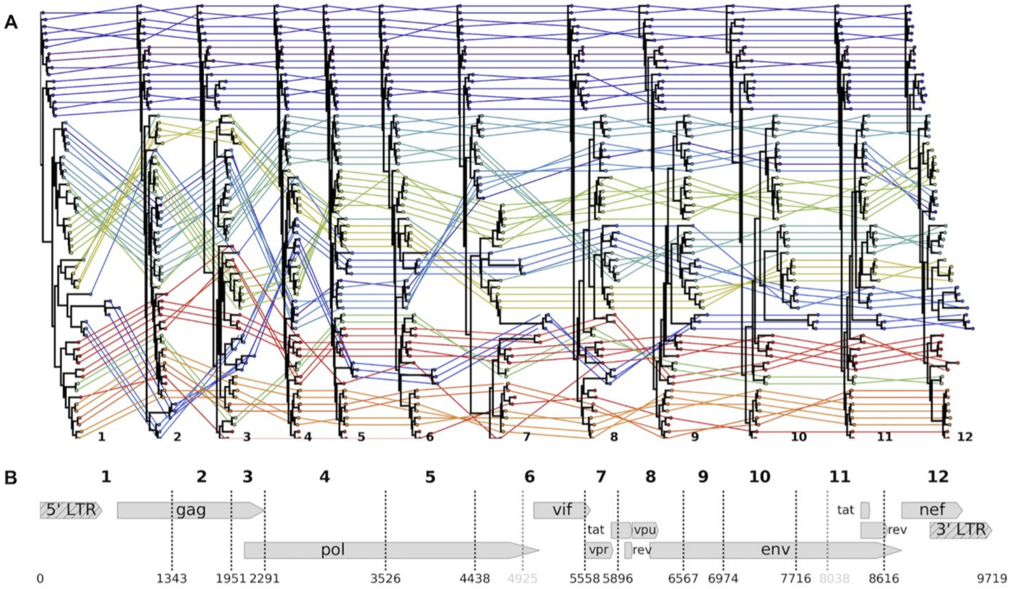


Lower recombination rates

Higher recombination rates

Inferring local trees

Local trees can be reconstructed for each non-recombinant region between detected breakpoints if there is sufficient genetic diversity between breakpoints..



Bell and Bedford (PLoS Pathogens, 2017)

Recombination vs. mutation rates

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

If *r/m* << 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

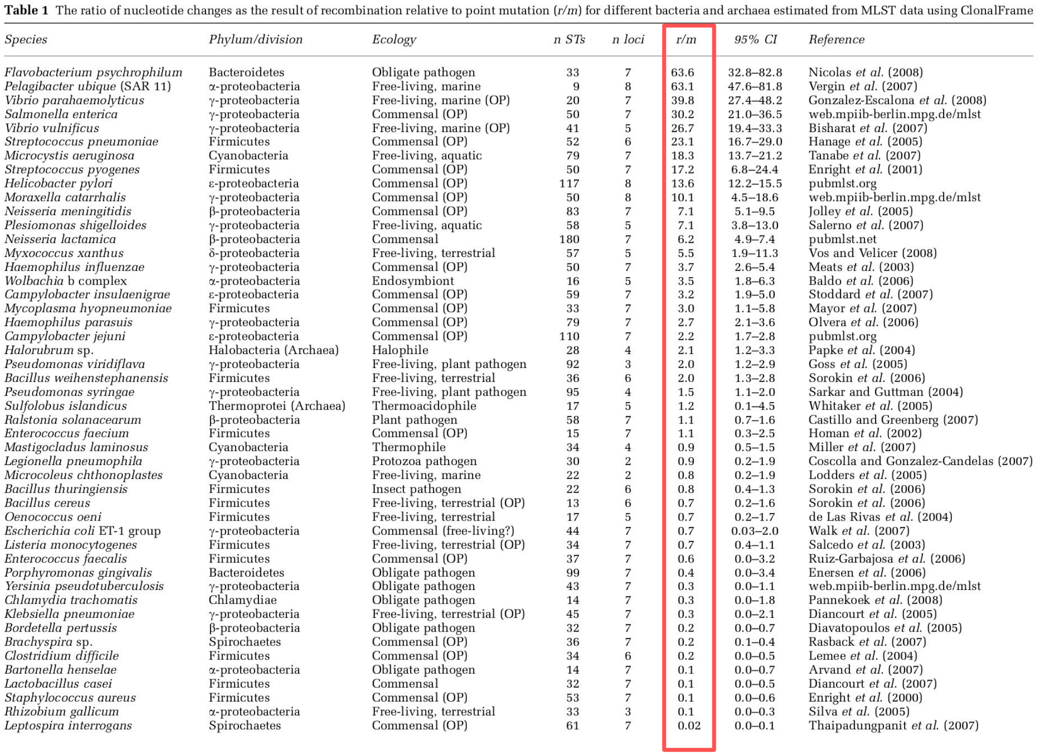


Table 1 The ratio of nucleotide changes as the result of recombination relative to point mutation (*r*/*m*) for different bacteria and archaea estimated from MLST data using ClonalFrame

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)

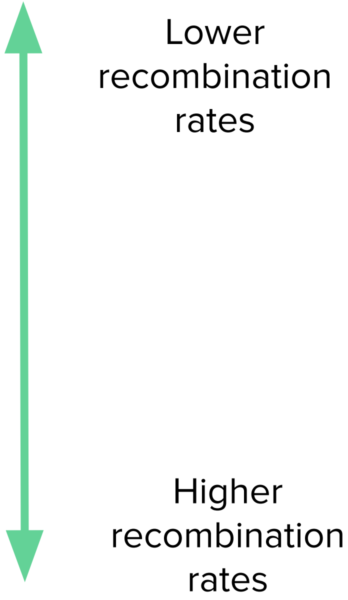
Some potential options

Remove recombinant sequences from alignments.

Remove recombinant genomic regions and reconstruct local trees from recombination-free blocks.

Assume evolution is mostly tree-like and reconstruct a clonal frame

Reconstruct a full ancestral recombination graph



Lower recombination rates

Higher recombination rates

Clonal frames

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

Assumes the majority of the genome is inherited clonally (vertically) 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 (i.e. the core genome) but gene conversion and other horizontal transfers 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

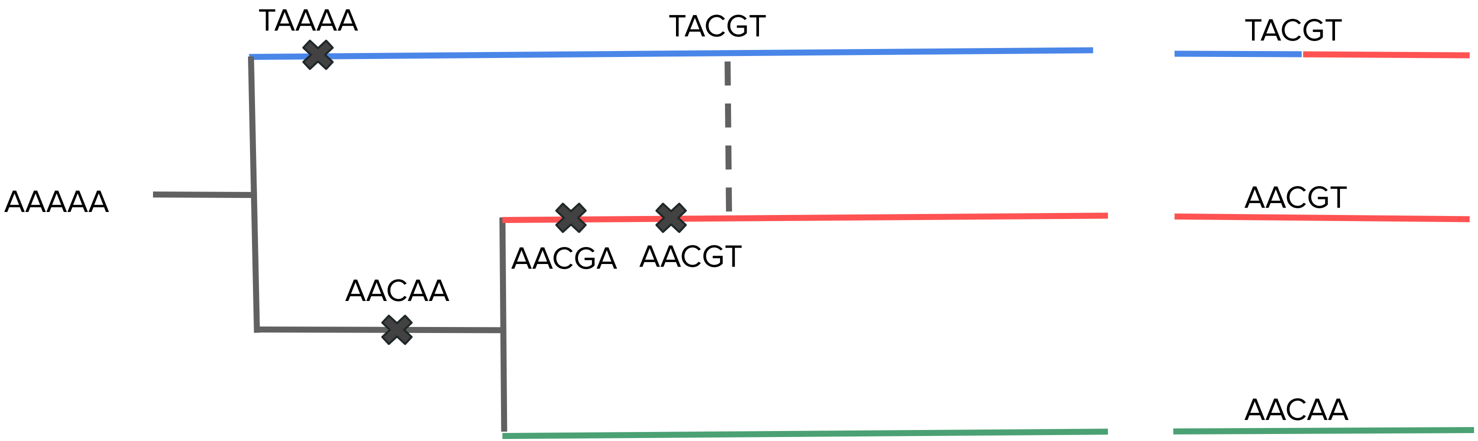
The genomic location of *insertions* caused by recombination are estimated along each branch of the tree using a Hidden Markov Model.

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

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

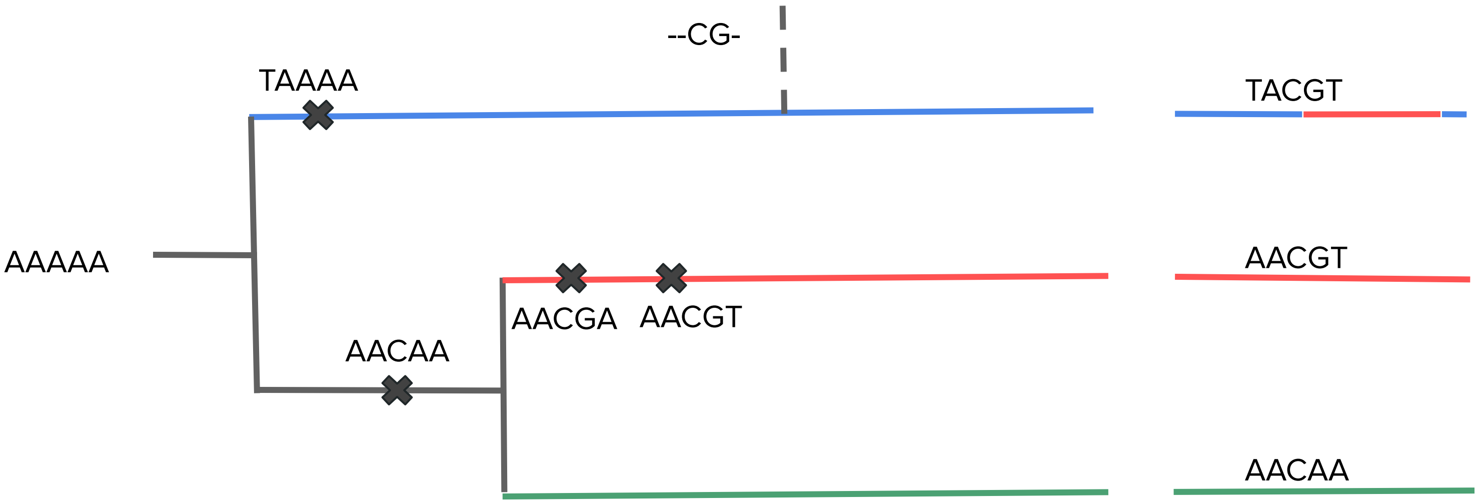
The ClonalFrame model of recombination

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

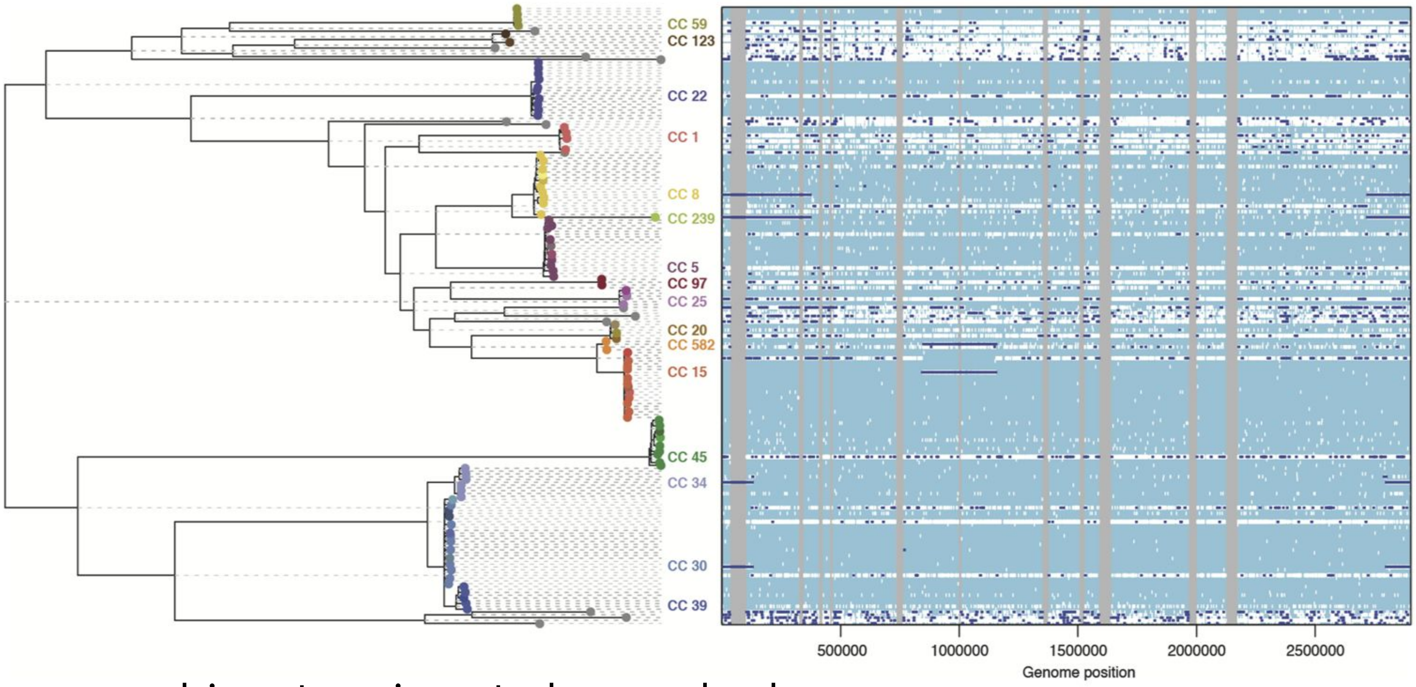


The ClonalFrame model of recombination

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



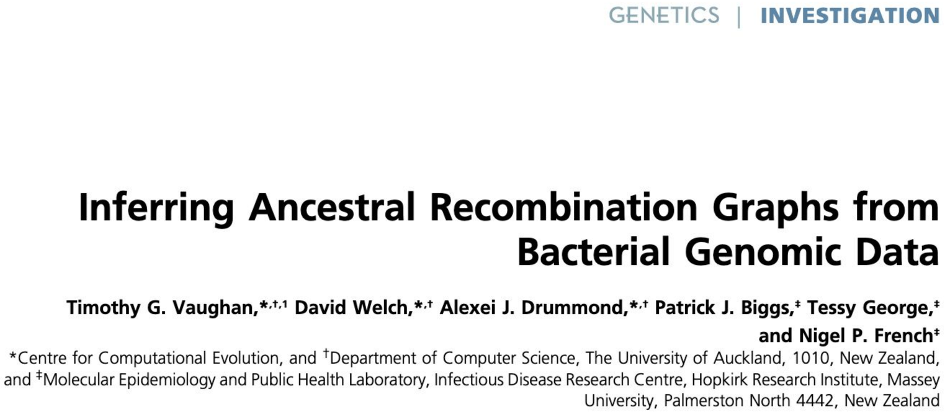
ClonalFrame of *Staphylococcus aureus*



Dark blue = recombinant regions to be masked

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

Bacter: Clonal frames in BEAST 2

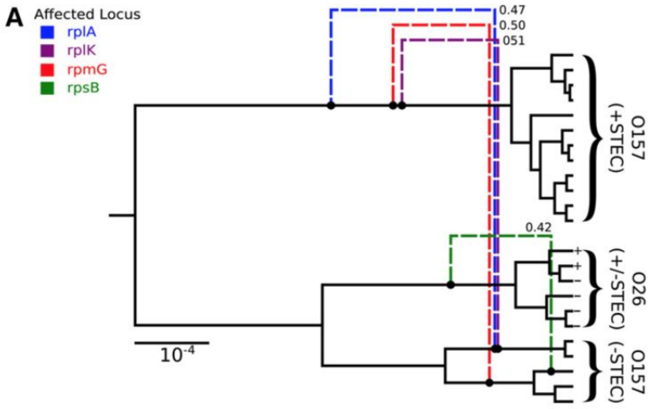


GENETICS | INVESTIGATION

Inferring Ancestral Recombination Graphs from Bacterial Genomic Data

Timothy G. Vaughan,[*,†,1] David Welch,[*,†] Alexei J. Drummond,[*,†] Patrick J. Biggs,[‡] Tessy George,[‡] and Nigel P. French[‡]

[*]Centre for Computational Evolution, and [†]Department of Computer Science, The University of Auckland, 1010, New Zealand, and [‡]Molecular Epidemiology and Public Health Laboratory, Infectious Disease Research Centre, Hopkirk Research Institute, Massey University, Palmerston North 4442, New Zealand



https://taming-the-beast.org/tutorials/Bacter-Tutorial/

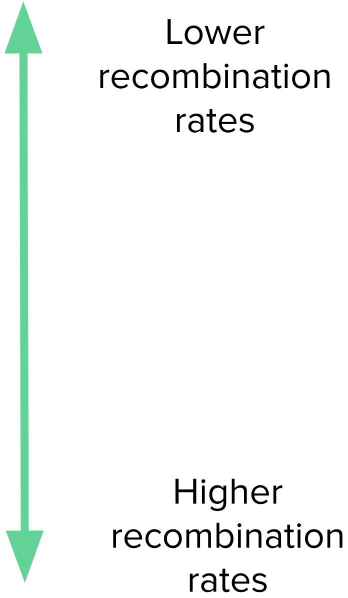
Some potential options

Remove recombinant sequences from alignments.

Remove recombinant genomic regions and reconstruct local trees from recombination-free blocks.

Assume evolution is mostly tree-like and reconstruct a clonal frame

Reconstruct a full ancestral recombination graph



Lower recombination rates

Higher recombination rates

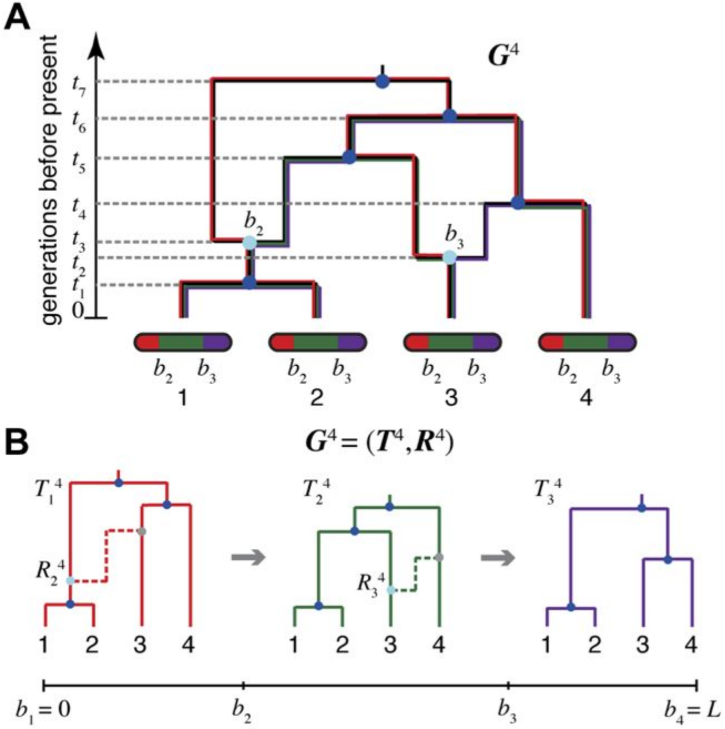
Ancestral recombination graphs

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)

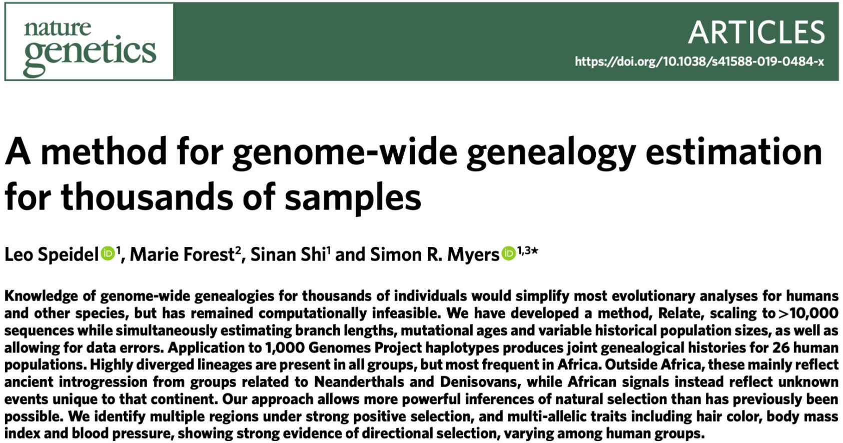
Ancestral recombination graphs

ARGs are in theory the ideal way to represent the full ancestral 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, but in recent years several methods have allowed for much faster inference by approximating ARGs as a sequence of correlated local trees.

Faster approximate ARG methods



nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0484-x

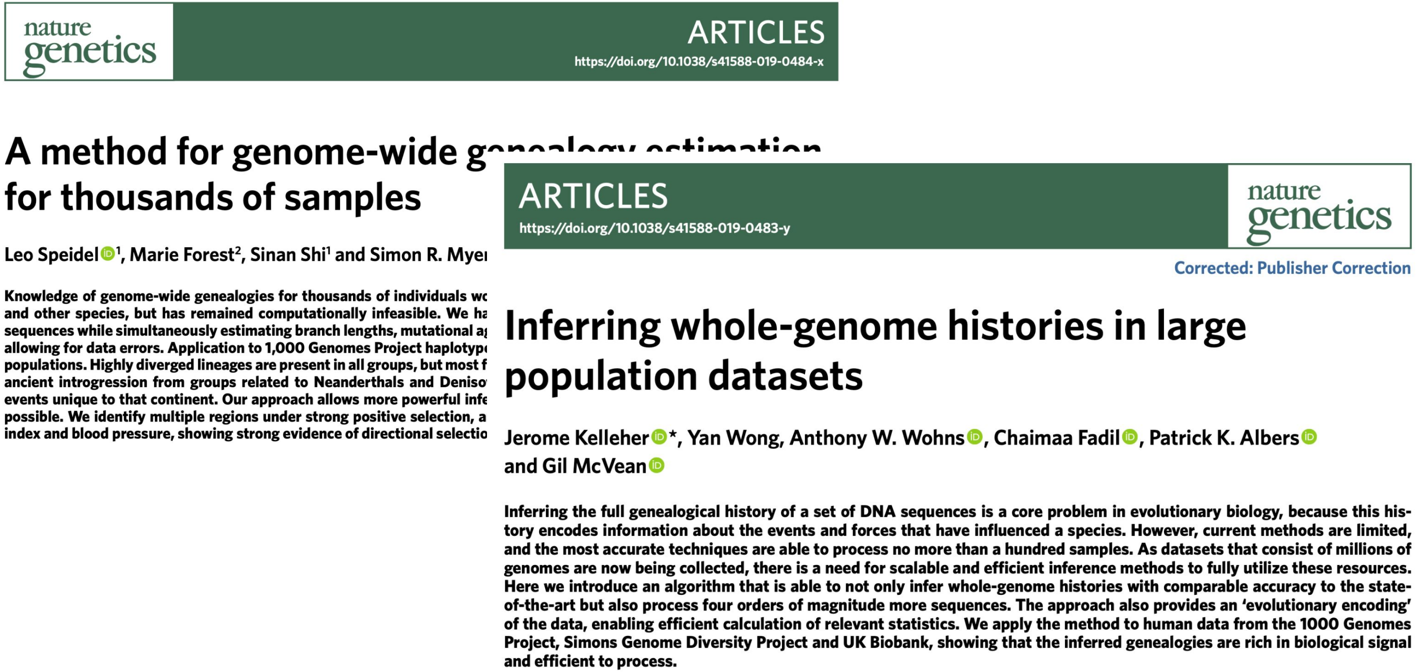
A method for genome-wide genealogy estimation for thousands of samples

Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myers[1,3]*

Knowledge of genome-wide genealogies for thousands of individuals would simplify most evolutionary analyses for humans and other species, but has remained computationally infeasible. We have developed a method, Relate, scaling to >10,000 sequences while simultaneously estimating branch lengths, mutational ages and variable historical population sizes, as well as allowing for data errors. Application to 1,000 Genomes Project haplotypes produces joint genealogical histories for 26 human populations. Highly diverged lineages are present in all groups, but most frequent in Africa. Outside Africa, these mainly reflect ancient introgression from groups related to Neanderthals and Denisovans, while African signals instead reflect unknown events unique to that continent. Our approach allows more powerful inferences of natural selection than has previously been possible. We identify multiple regions under strong positive selection, and multi-allelic traits including hair color, body mass index and blood pressure, showing strong evidence of directional selection, varying among human groups.

Relate -- Speidel *et al.* (2019)

Faster approximate ARG methods



nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0484-x

A method for genome-wide genealogy estimation
for thousands of samples

Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myei

Knowledge of genome-wide genealogies for thousands of individuals wc
and other species, but has remained computationally infeasible. We ha
sequences while simultaneously estimating branch lengths, mutational a
allowing for data errors. Application to 1,000 Genomes Project haplotyp
populations. Highly diverged lineages are present in all groups, but most f
ancient introgression from groups related to Neanderthals and Deniso
events unique to that continent. Our approach allows more powerful infe
possible. We identify multiple regions under strong positive selection, a
index and blood pressure, showing strong evidence of directional selectio

ARTICLES

https://doi.org/10.1038/s41588-019-0483-y

nature genetics

Corrected: Publisher Correction

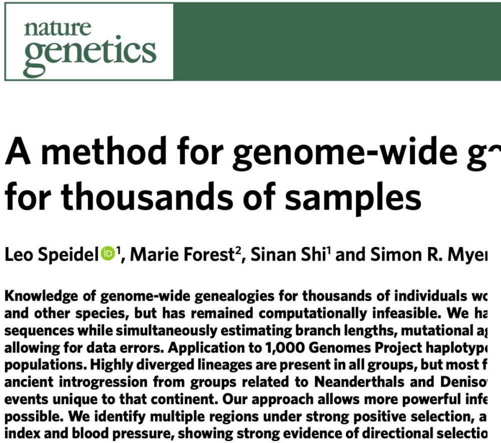
Inferring whole-genome histories in large population datasets

Jerome Kelleher*, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers
and Gil McVean

Inferring the full genealogical history of a set of DNA sequences is a core problem in evolutionary biology, because this history encodes information about the events and forces that have influenced a species. However, current methods are limited, and the most accurate techniques are able to process no more than a hundred samples. As datasets that consist of millions of genomes are now being collected, there is a need for scalable and efficient inference methods to fully utilize these resources. Here we introduce an algorithm that is able to not only infer whole-genome histories with comparable accuracy to the state-of-the-art but also process four orders of magnitude more sequences. The approach also provides an 'evolutionary encoding' of the data, enabling efficient calculation of relevant statistics. We apply the method to human data from the 1000 Genomes Project, Simons Genome Diversity Project and UK Biobank, showing that the inferred genealogies are rich in biological signal and efficient to process.

tsinfer -- Kelleher *et al.* (2019)

Faster approximate ARG methods



nature genetics

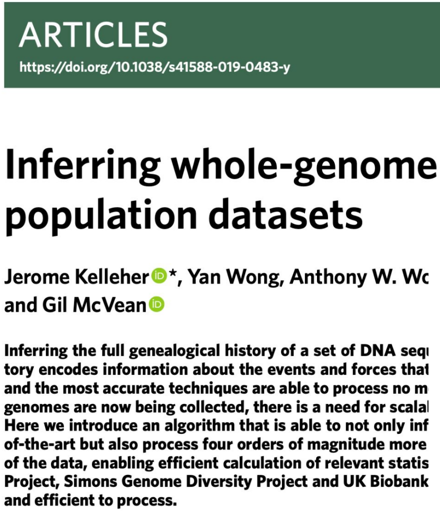
ARTICLES

https://doi.org/10.1038/s41588-019-0484-x

A method for genome-wide genealogy estimation
for thousands of samples

Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myer

Knowledge of genome-wide genealogies for thousands of individuals wo
and other species, but has remained computationally infeasible. We ha
sequences while simultaneously estimating branch lengths, mutational a
allowing for data errors. Application to 1,000 Genomes Project haplotyp
populations. Highly diverged lineages are present in all groups, but most f
ancient introgression from groups related to Neanderthals and Deniso
events unique to that continent. Our approach allows more powerful infe
possible. We identify multiple regions under strong positive selection, a
index and blood pressure, showing strong evidence of directional selectio



ARTICLES

https://doi.org/10.1038/s41588-019-0483-y

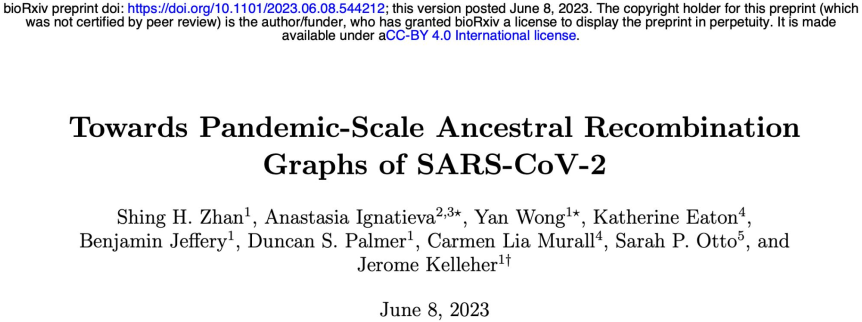
nature genetics

Corrected: Publisher Correction

Inferring whole-genome histories in large
population datasets

Jerome Kelleher*, Yan Wong, Anthony W. Wo
and Gil McVean

Inferring the full genealogical history of a set of DNA sequ
tory encodes information about the events and forces that
and the most accurate techniques are able to process no m
genomes are now being collected, there is a need for scalal
Here we introduce an algorithm that is able to not only inf
of-the-art but also process four orders of magnitude more
of the data, enabling efficient calculation of relevant statis
Project, Simons Genome Diversity Project and UK Biobank
and efficient to process.



bioRxiv preprint doi: https://doi.org/10.1101/2023.06.08.544212; this version posted June 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Towards Pandemic-Scale Ancestral Recombination Graphs of SARS-CoV-2

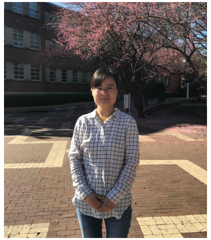
Shing H. Zhan[1], Anastasia Ignatieva[2,3⋆], Yan Wong[1⋆], Katherine Eaton[4], Benjamin Jeffery[1], Duncan S. Palmer[1], Carmen Lia Murall[4], Sarah P. Otto[5], and Jerome Kelleher[1†]

June 8, 2023

sc2ts -- Zhan *et al.* (2023)

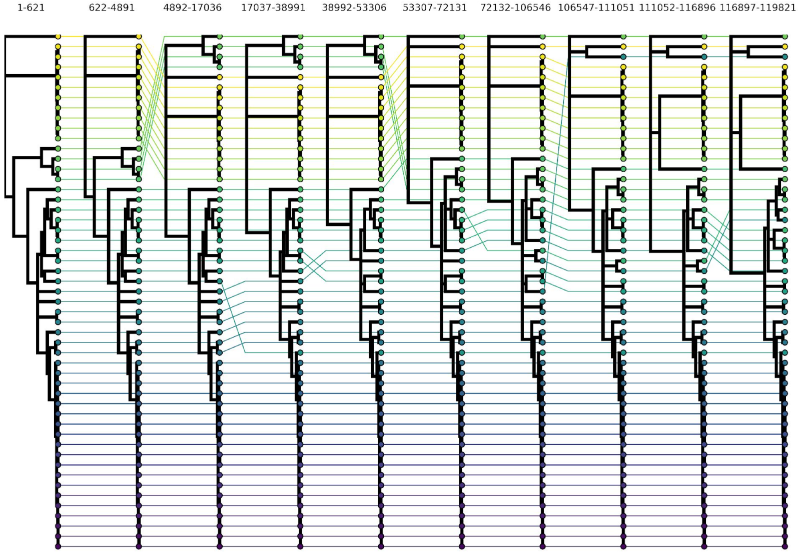
Demographic inference from ARGs

(Structured) coalescent methods can be adapted to ARGs, allowing for demographic inference from many different correlated but different local trees.



Fangfang Guo

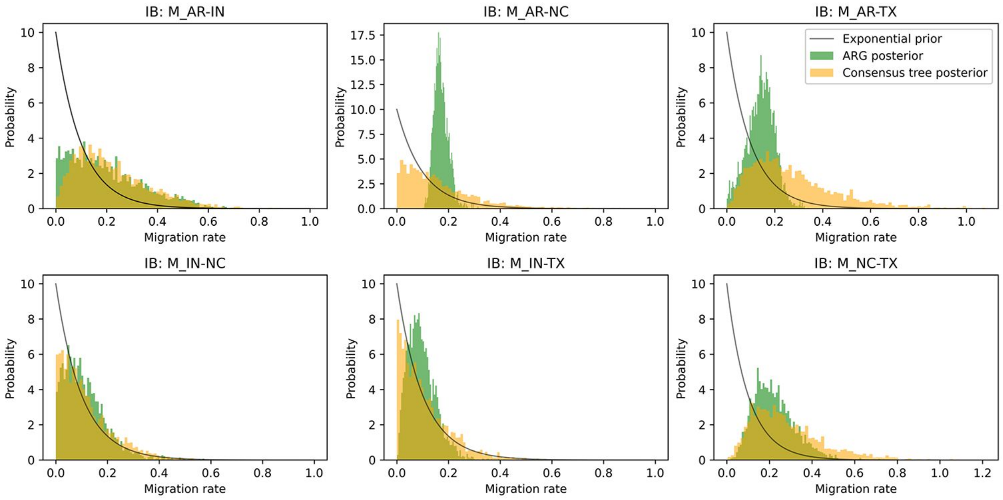
Ignazio Carbone

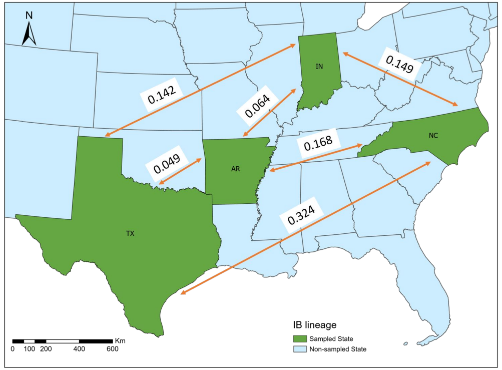


ARG reconstructed from chromosome 3 of the *A. flavus* genome

Demographic inference from ARGs

Because ARGs contain many different trees, there is often way more information about demographic parameters in ARGs than any single phylogenetic tree.





Guo *et al.* (PLoS Comp Bio, 2022)

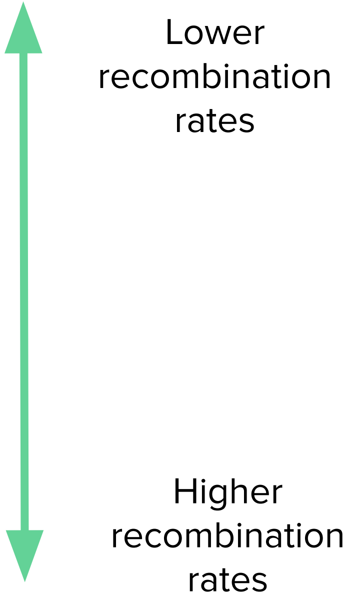
Some potential options

Remove recombinant sequences from alignments.

Remove recombinant genomic regions and reconstruct local trees from recombination-free blocks.

Assume evolution is mostly tree-like and reconstruct a clonal frame

Reconstruct a full ancestral recombination graph



Lower recombination rates

Higher recombination rates

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