Inferring Bacterial Recombination Graphs using BEAST2

Lecturer: Tim Vaughan

Department of Biosystems Science and Enginerring
ETH Zürich, Switzerland

Taming the BEAST, London, July 2017

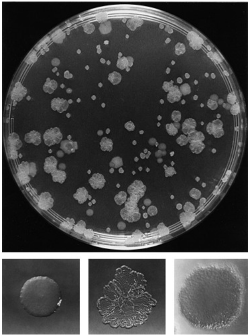
Why study bacterial phylogenetics?

Why study bacterial phylogenetics?

- Bacteria play important roles (both positive and negative) in the health of humans, animals and plants.

Why study bacterial phylogenetics?

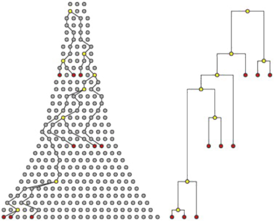
- Bacteria play important roles (both positive and negative) in the health of humans, animals and plants.
- Many bacteria possess interesting and *experimentally accessible* evolutionary dynamics.



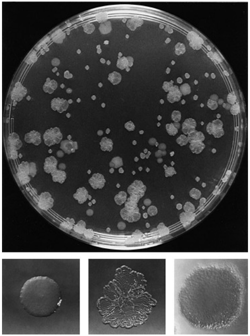
Rainey & Travisano, Nature (1998)

Why study bacterial phylogenetics?

- Bacteria play important roles (both positive and negative) in the health of humans, animals and plants.
- Many bacteria possess interesting and *experimentally accessible* evolutionary dynamics.
- Bacterial genomes are measurably evolving over relatively short study periods.



Drummond & Rambaut, TIEE (2003)



Rainey & Travisano, Nature (1998)

Bacterial Recombination

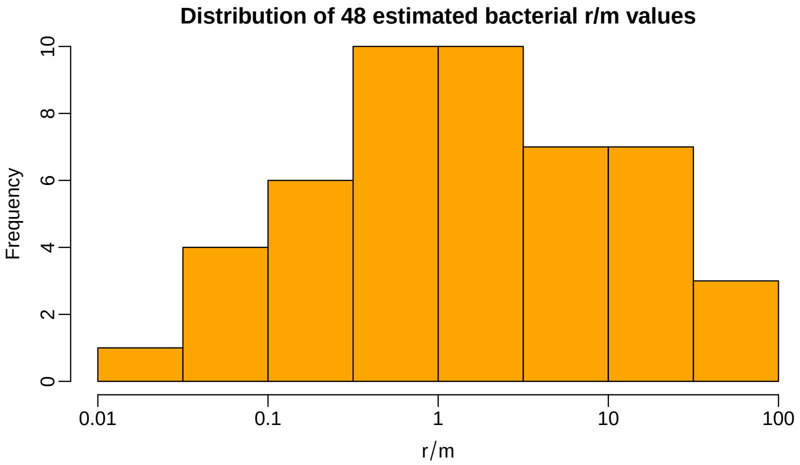
- Bacteria reproduce clonally via binary fission.
- Multiple mechanisms allow for non-vertical transfer of genetic information:
  - Conjugation
  - Natural transformation
  - Phage-mediated transduction
- The frequency at which these events occur depends on the bacterial species (i.e. depends on the genome: a strange loop!)
- The effect of these events can be:
  - Plasmid transfer
  - Insertion
  - Homologous recombination
- Focus solely on homologous recombination: only event which doesn't alter the length of the sequence.

The Problem for Phylogenetic Inference

For many bacteria, the ratio between the recombination rate and the mutation rate is very high.

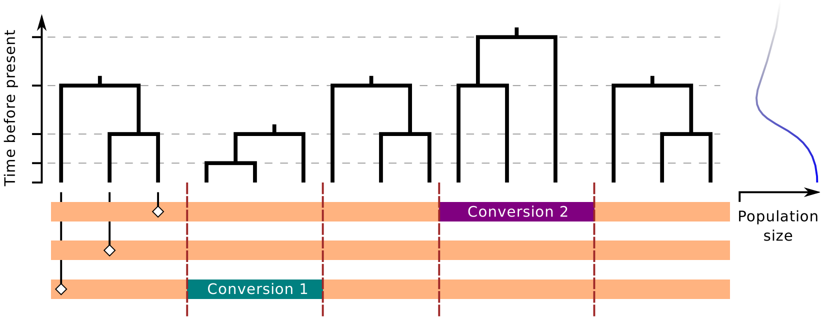
The Problem for Phylogenetic Inference

For many bacteria, the ratio between the recombination rate and the mutation rate is very high.

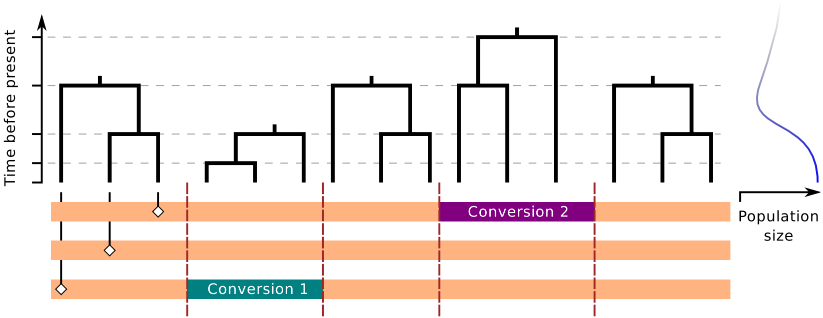


Vos and Didelot (2009)

The Benefit for Demographic Inference



The Benefit for Demographic Inference



- Relationship used by Li and Durbin (2011) to infer human demographic history from pairs of autosomes.

Existing solutions

Existing solutions

- Pre-processing of data to identify and remove non-vertically inherited material. (eg. START: Jolley et al. (2001))

Existing solutions

- Pre-processing of data to identify and remove non-vertically inherited material. (eg. START: Jolley et al. (2001))

Pros	Cons
• Can use standard tools for phylogenetic inference.	• Data is being thrown away. • Ad hoc, may bias results.

Existing solutions

- Pre-processing of data to identify and remove non-vertically inherited material. (eg. START: Jolley et al. (2001))

Pros	Cons
• Can use standard tools for phylogenetic inference.	• Data is being thrown away. • Ad hoc, may bias results.

- Explicit modelling of bacterial recombination. (eg. ClonalFrame and ClonalOrigin: Didelot and Falush (2007); Didelot et al. (2010))

Existing solutions

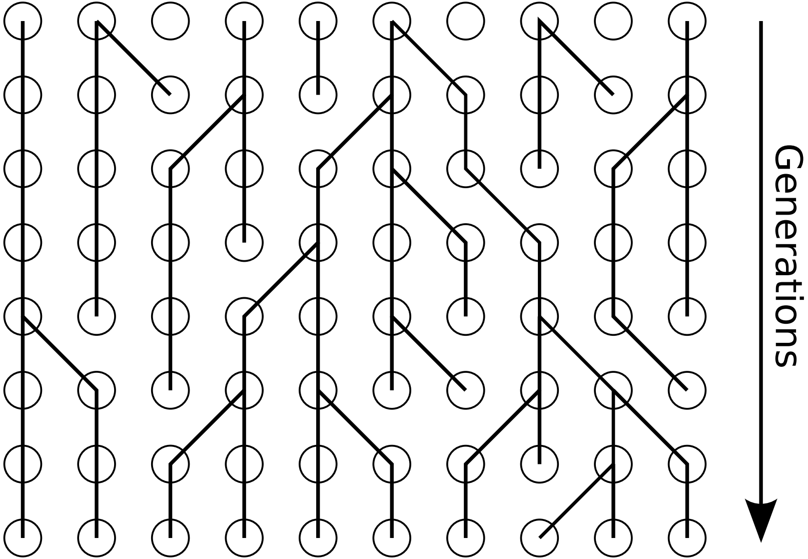
- Pre-processing of data to identify and remove non-vertically inherited material. (eg. START: Jolley et al. (2001))

Pros	Cons
• Can use standard tools for phylogenetic inference.	• Data is being thrown away. • Ad hoc, may bias results.

- Explicit modelling of bacterial recombination. (eg. ClonalFrame and ClonalOrigin: Didelot and Falush (2007); Didelot et al. (2010))

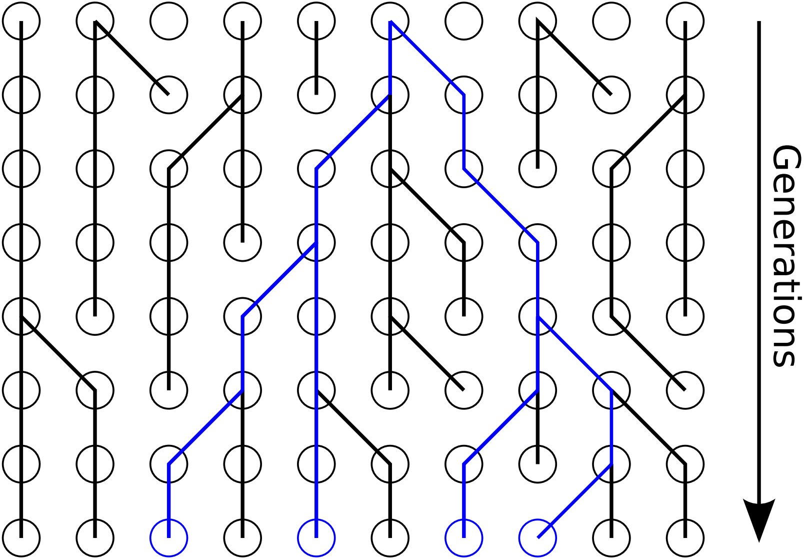
Pros	Cons
• Can make use of all data. • Can infer additional parameters such as recombination rates. • May yield increased confidence in estimates	• Models can be complex, with many parameters. • Both computationally and statistically challenging. • Existing implementations are too restrictive.

The coalescent with gene conversion



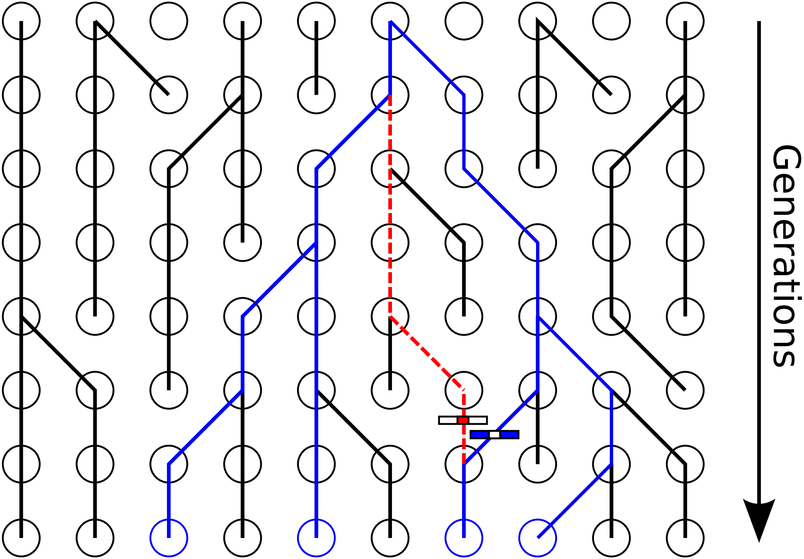
Wiuf (2000); Wiuf and Hein (2000)

The coalescent with gene conversion



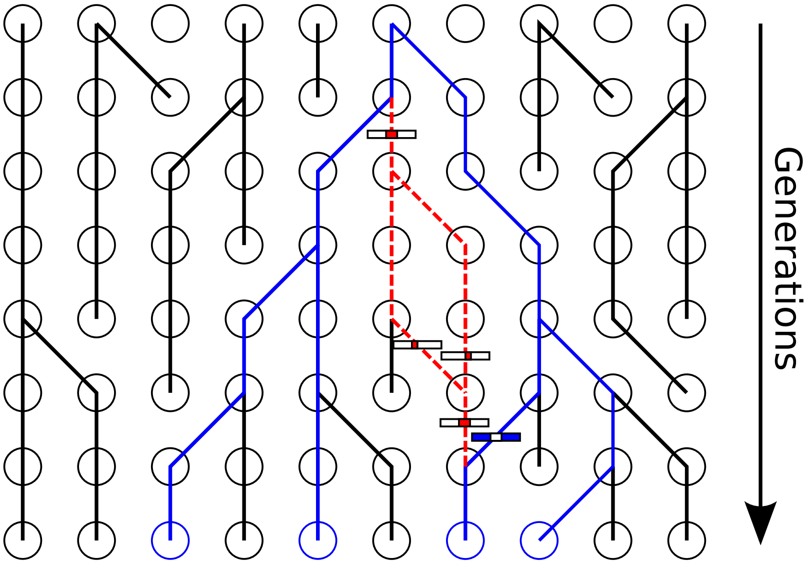
Wiuf (2000); Wiuf and Hein (2000)

The coalescent with gene conversion



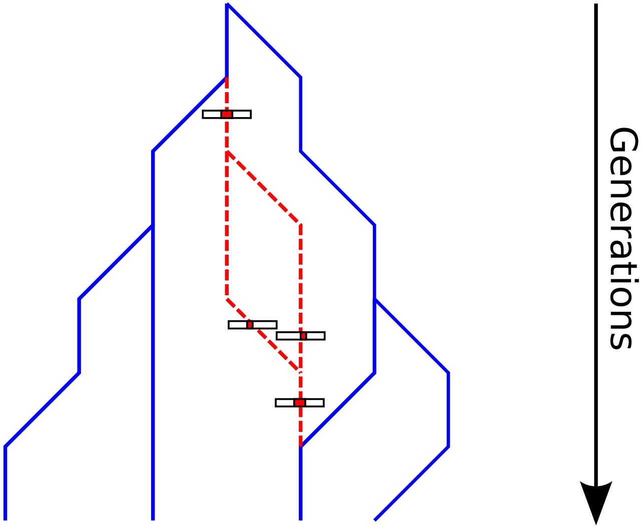
Wiuf (2000); Wiuf and Hein (2000)

The coalescent with gene conversion



Wiuf (2000); Wiuf and Hein (2000)

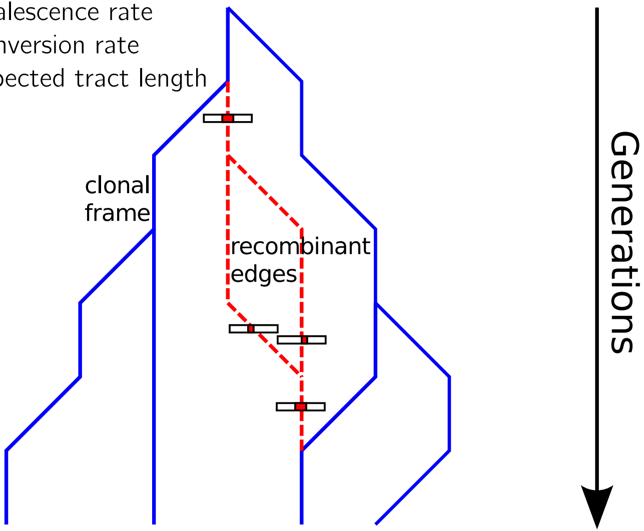
The coalescent with gene conversion



Wiuf (2000); Wiuf and Hein (2000)

The coalescent with gene conversion

$(N_e(t)g)^{-1}$	Coalescence rate
ρ_s	Conversion rate
δ	Expected tract length

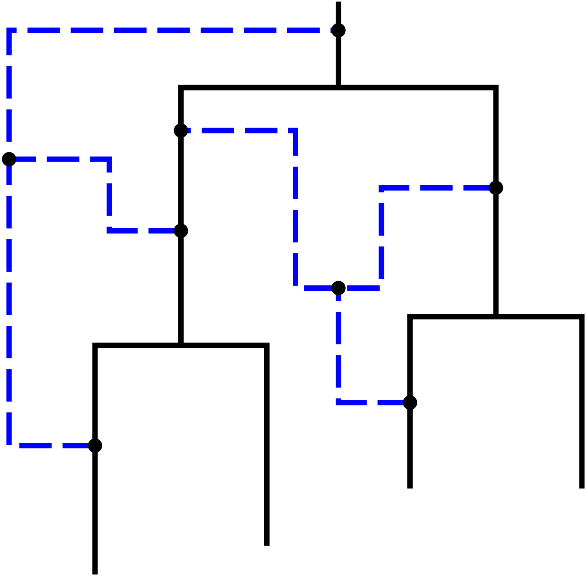


Wiuf (2000); Wiuf and Hein (2000)

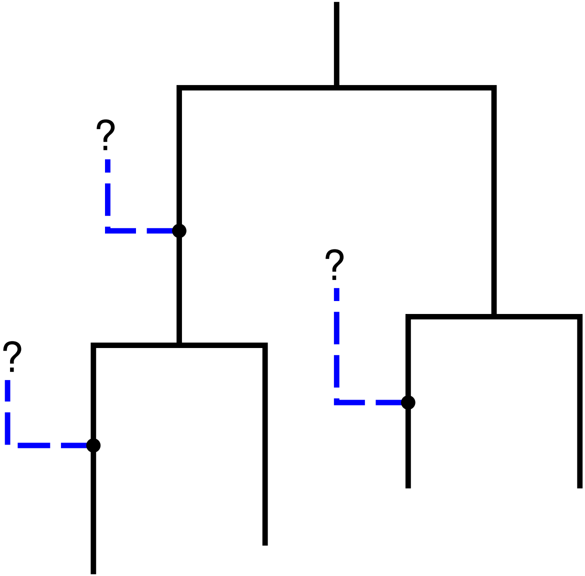
Problem

The space of possible ancestral recombination graphs is extremely large. Even two samples have infinitely many distinct ancestries!

Full ARG

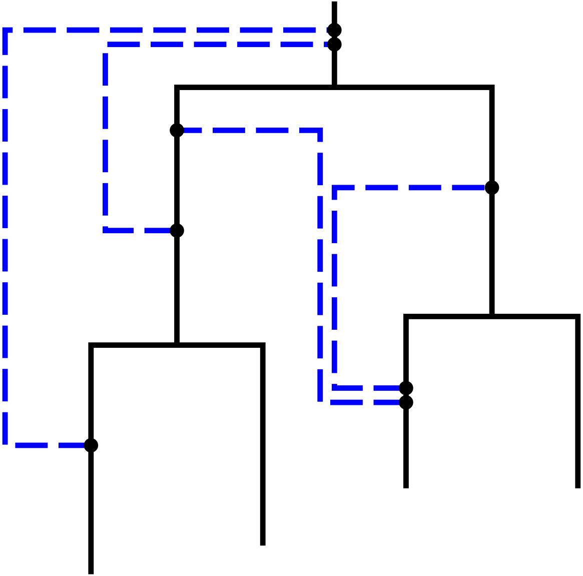


Approximation 1: ClonalFrame



Didelot and Falush (2007)

Approximation 2: ClonalOrigin



Didelot et al. (2010)

Inference under the ClonalOrigin model

Inference follows the standard Bayesian phylogenetic tradition:

$$f(G, N, \mu, \rho, \delta | \mathcal{A}) \propto P_F(\mathcal{A} | G, \mu) f_{CGC}(G | N, \rho, \delta) f_{\text{prior}}(N, \mu, \rho, \delta)$$

where

$\mathcal{A}$ is the sequence alignment,

μ are the substitution model parameters, and

G is the full sample genealogy including clonal frame T and M conversions $\{C_i\}_{i \in [1 \dots M]}$.

The genealogy density under ClonalOrigin model can be expanded

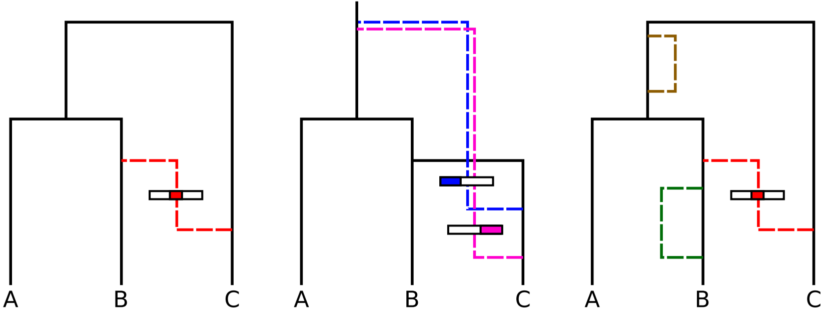
$$f_{CGC}(G | \rho', \delta, N) = \left(\prod_{i=1}^{M} f(C_i | T, N, \delta) \right) P(M | T, \rho) f_C(T | N)$$

Identifiability

Despite using a simplified model, an infinite number of ARGs still possess the same likelihood given a sequence alignment.

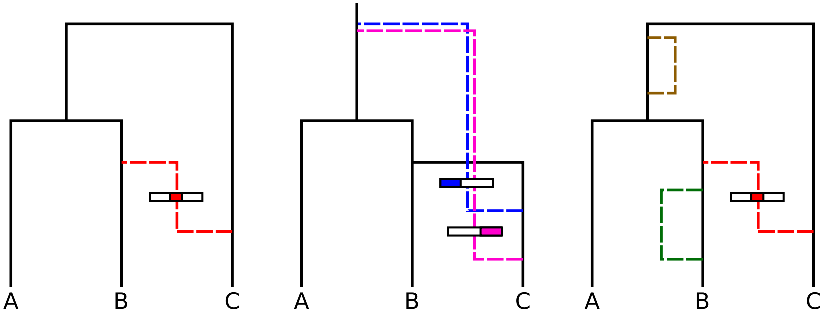
Identifiability

Despite using a simplified model, an infinite number of ARGs still possess the same likelihood given a sequence alignment.



Identifiability

Despite using a simplified model, an infinite number of ARGs still possess the same likelihood given a sequence alignment.



Very important for an MCMC algorithm to propose state changes which minimize effect on likelihood.

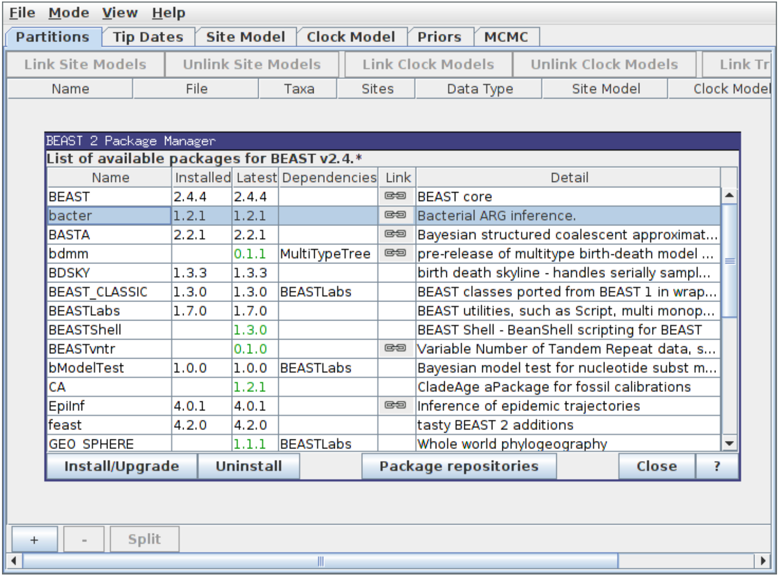
ClonalOrigin in BEAST

bacter

- BEAST package that performs inference under the ClonalOrigin model.
- Joint inference of clonal frame and recombinant edges.
- Can be combined with usual variety of substitution models and parametric population models.
- Straight-forward usage via BEAUti.

`http://tgvaughan.github.io/bacter`

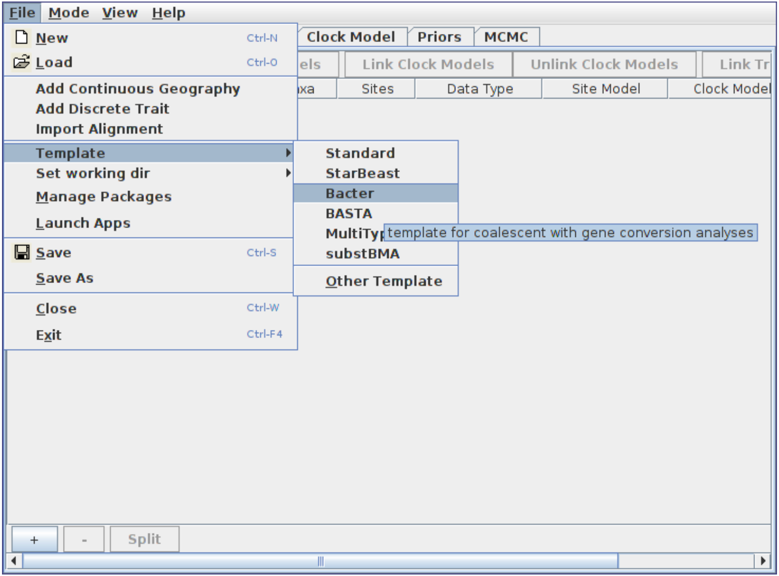
Setting up a Bacter analysis



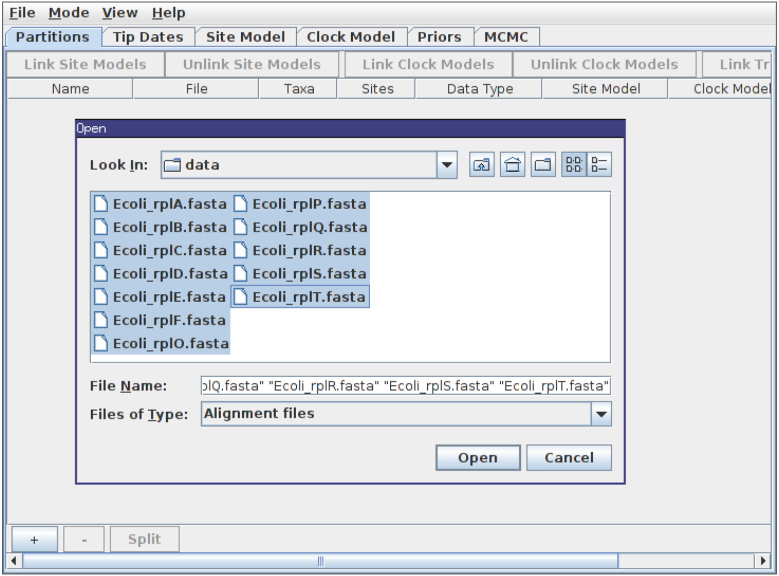
List of available packages for BEAST v2.4.*

Name	Installed	Latest	Dependencies	Link	Detail
BEAST	2.4.4	2.4.4		🔗	BEAST core
bacter	1.2.1	1.2.1		🔗	Bacterial ARG inference.
BASTA	2.2.1	2.2.1		🔗	Bayesian structured coalescent approximat...
bdmm		0.1.1	MultiTypeTree	🔗	pre-release of multitype birth-death model ...
BDSKY	1.3.3	1.3.3			birth death skyline - handles serially sampl...
BEAST_CLASSIC	1.3.0	1.3.0	BEASTLabs		BEAST classes ported from BEAST 1 in wrap...
BEASTLabs	1.7.0	1.7.0			BEAST utilities, such as Script, multi monop...
BEASTShell		1.3.0			BEAST Shell - BeanShell scripting for BEAST
BEASTvntr		0.1.0		🔗	Variable Number of Tandem Repeat data, s...
bModelTest	1.0.0	1.0.0	BEASTLabs		Bayesian model test for nucleotide subst m...
CA		1.2.1			CladeAge aPackage for fossil calibrations
EpiInf	4.0.1	4.0.1		🔗	Inference of epidemic trajectories
feast	4.2.0	4.2.0			tasty BEAST 2 additions
GEO_SPHERE		1.1.1	BEASTLabs		Whole world phylogeography

Setting up a Bacter analysis



Setting up a Bacter analysis



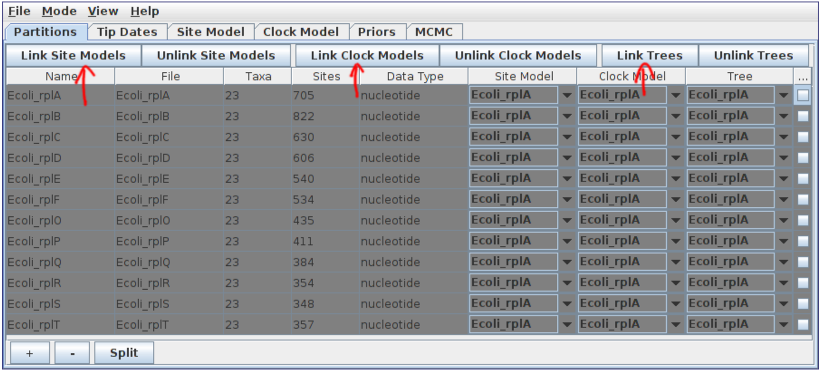
Look In: data

Ecoli_rplA.fasta
Ecoli_rplB.fasta
Ecoli_rplC.fasta
Ecoli_rplD.fasta
Ecoli_rplE.fasta
Ecoli_rplF.fasta
Ecoli_rplO.fasta
Ecoli_rplP.fasta
Ecoli_rplQ.fasta
Ecoli_rplR.fasta
Ecoli_rplS.fasta
Ecoli_rplT.fasta

File Name: plQ.fasta" "Ecoli_rplR.fasta" "Ecoli_rplS.fasta" "Ecoli_rplT.fasta"

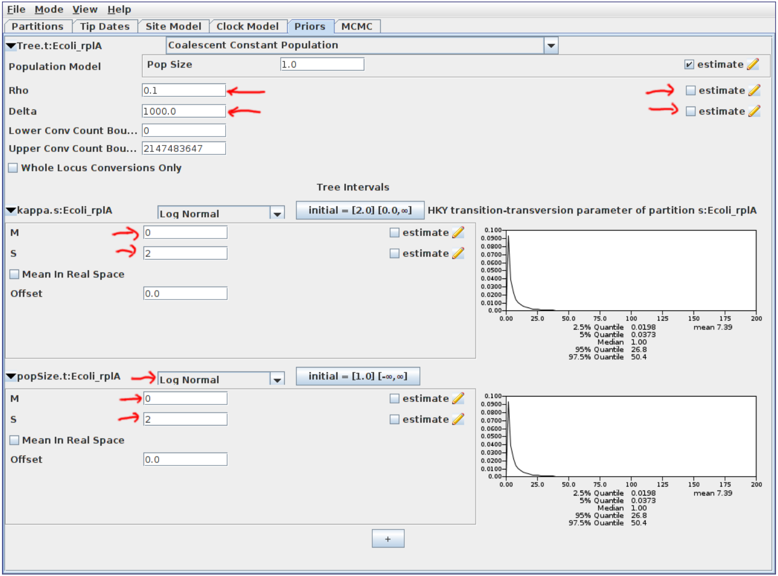
Files of Type: Alignment files

Setting up a Bacter analysis



Name	File	Taxa	Sites	Data Type	Site Model	Clock Model	Tree	...
Ecoli_rplA	Ecoli_rplA	23	705	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplB	Ecoli_rplB	23	822	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplC	Ecoli_rplC	23	630	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplD	Ecoli_rplD	23	606	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplE	Ecoli_rplE	23	540	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplF	Ecoli_rplF	23	534	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplO	Ecoli_rplO	23	435	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplP	Ecoli_rplP	23	411	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplQ	Ecoli_rplQ	23	384	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplR	Ecoli_rplR	23	354	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplS	Ecoli_rplS	23	348	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□
Ecoli_rplT	Ecoli_rplT	23	357	nucleotide	Ecoli_rplA	Ecoli_rplA	Ecoli_rplA	□

Setting up a Bacter analysis



Tree.t:Ecoli_rplA: Coalescent Constant Population

Population Model: Pop Size 1.0 ☑ estimate

Rho 0.1 □ estimate

Delta 1000.0 □ estimate

Lower Conv Count Bou... 0

Upper Conv Count Bou... 2147483647

□ Whole Locus Conversions Only

Tree Intervals

kappa.s:Ecoli_rplA: Log Normal — initial = [2.0] [0.0,∞] — HKY transition-transversion parameter of partition s:Ecoli_rplA

M 0 □ estimate

S 2 □ estimate

□ Mean In Real Space

Offset 0.0

```
  2.5% Quantile  0.0198        mean 7.39
    5% Quantile  0.0373
         Median  1.00
   95% Quantile  26.8
 97.5% Quantile  50.4
```

popSize.t:Ecoli_rplA: Log Normal — initial = [1.0] [-∞,∞]

M 0 □ estimate

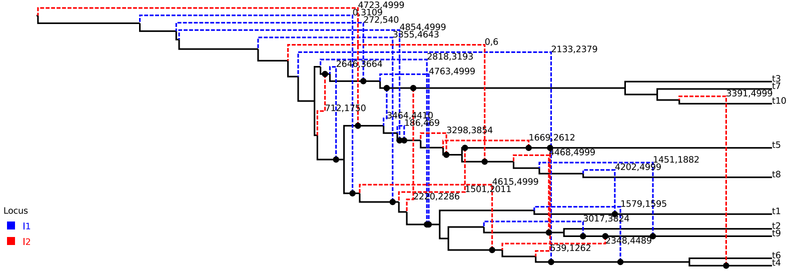
S 2 □ estimate

□ Mean In Real Space

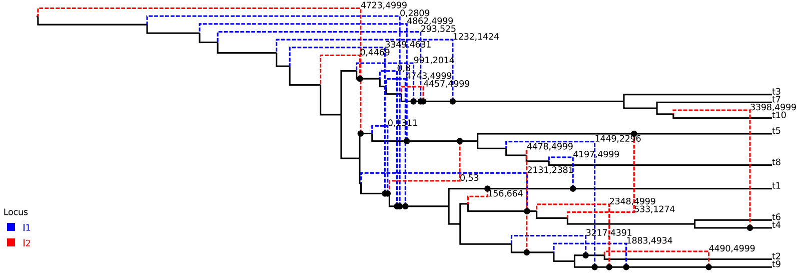
Offset 0.0

```
  2.5% Quantile  0.0198        mean 7.39
    5% Quantile  0.0373
         Median  1.00
   95% Quantile  26.8
 97.5% Quantile  50.4
```

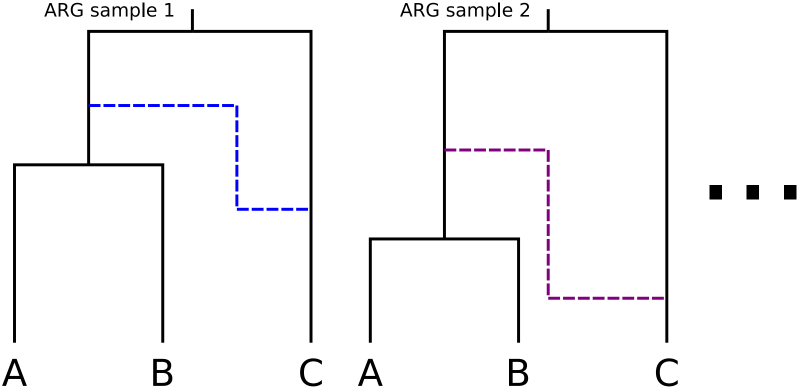
True ARG:



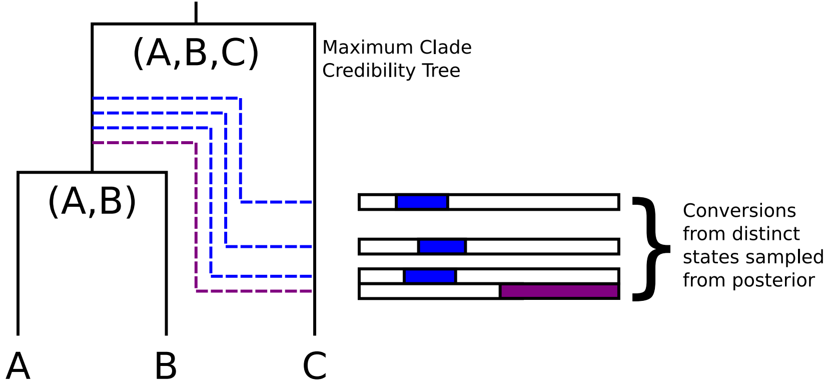
Randomly-selected ARG from MCMC:



Summary networks



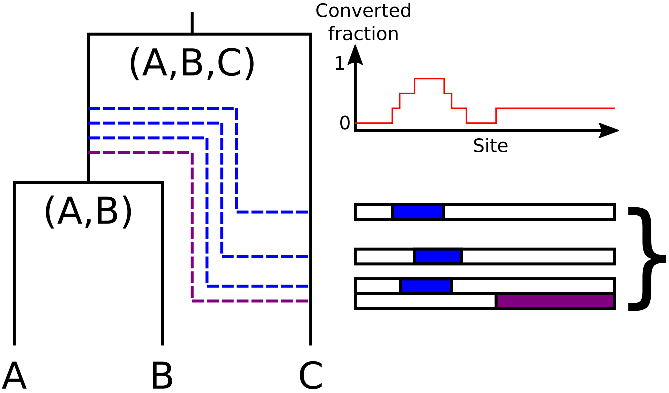
Summary networks



Maximum Clade Credibility Tree

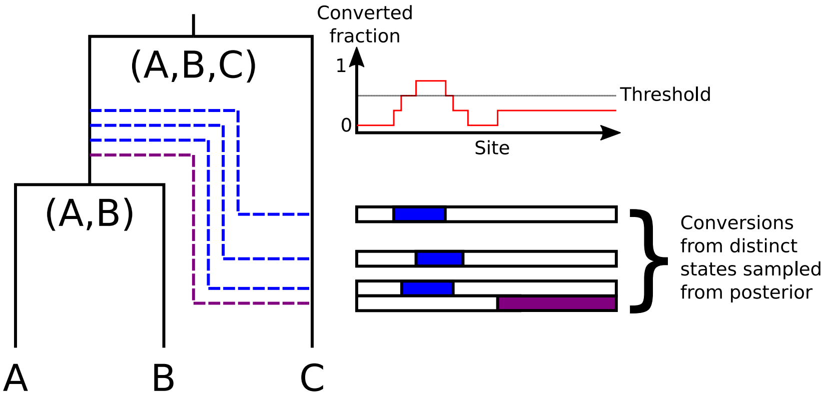
Conversions from distinct states sampled from posterior

Summary networks



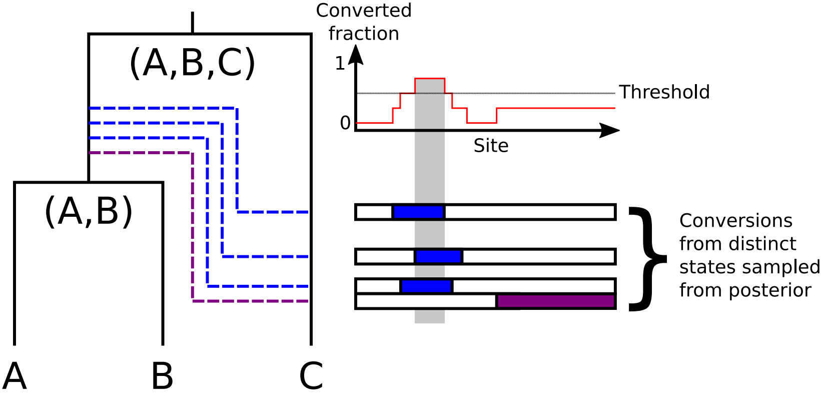
Conversions from distinct states sampled from posterior

Summary networks

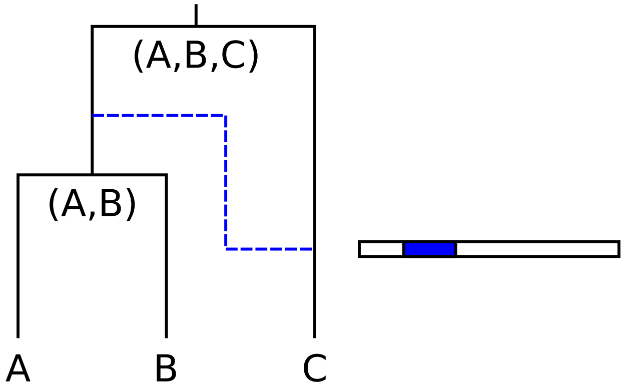


Conversions from distinct states sampled from posterior

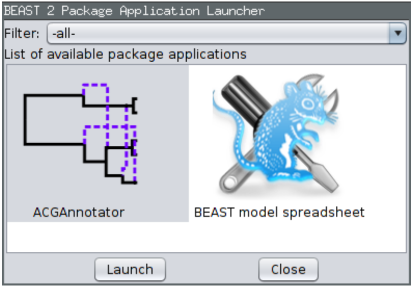
Summary networks



Summary networks



ACGAnnotator

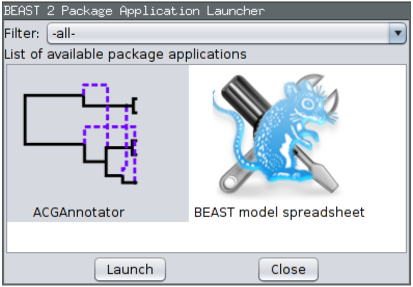


BEAST 2 Package Application Launcher

Filter: -all-

List of available package applications

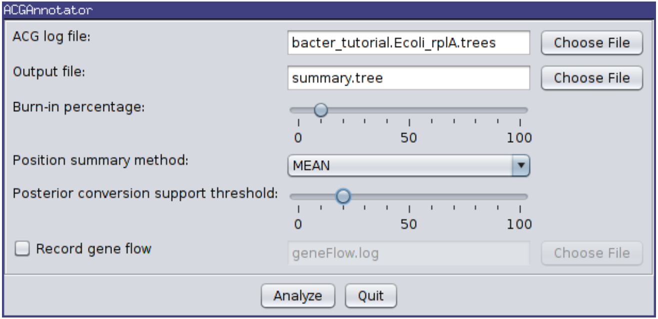
ACGAnnotator



BEAST 2 Package Application Launcher

Filter: -all-

List of available package applications



ACGAnnotator

ACG log file: bacter_tutorial.Ecoli_rplA.trees

Output file: summary.tree

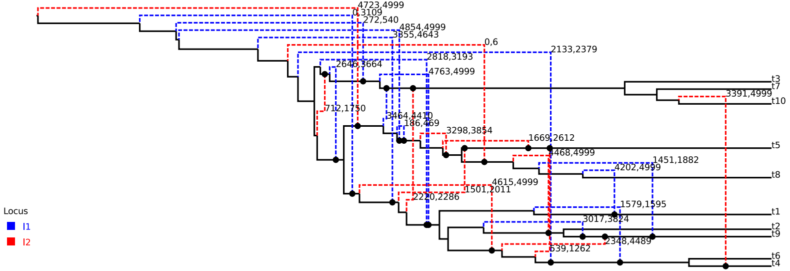
Burn-in percentage:

Position summary method: MEAN

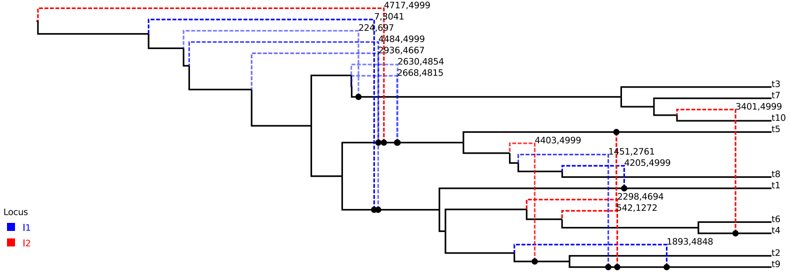
Posterior conversion support threshold:

☐ Record gene flow geneFlow.log

True ARG:

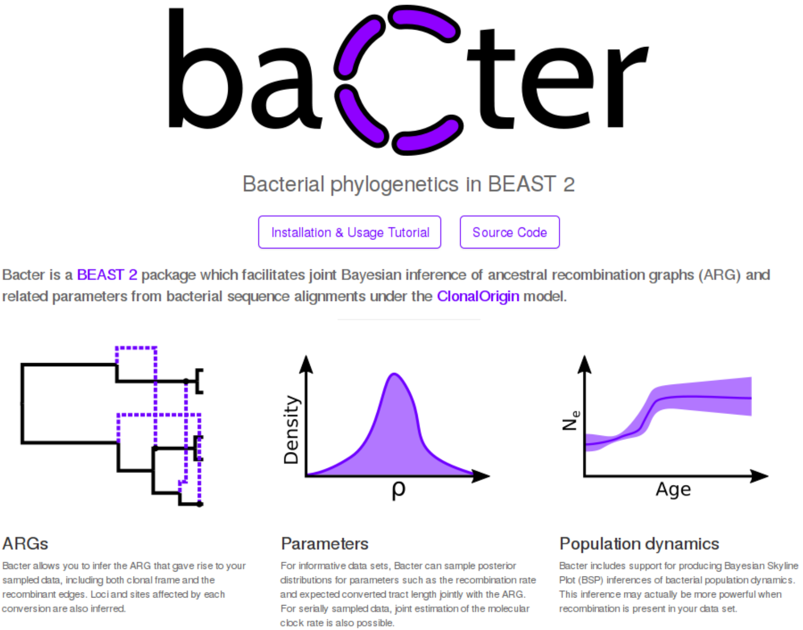


Summary ARG from MCMC:



Bacter limitations

- Computational complexity scales with the number of proposed conversions.
    - This can be huge, even for small sample sets!
    - Can't use BEAGLE to speed things up: doesn't perform well due to peculiarities of ARG likelihood computation.
- MCMC algorithm used does not intelligently locate conversions.
    - Looking at fixing this in the near future.
- Summary algorithm can produce peculiar results.
    - More research needs to be done to find a better algorithm.



bacter

Bacterial phylogenetics in BEAST 2

Installation & Usage Tutorial

Source Code

Bacter is a BEAST 2 package which facilitates joint Bayesian inference of ancestral recombination graphs (ARG) and related parameters from bacterial sequence alignments under the ClonalOrigin model.

ARGs

Bacter allows you to infer the ARG that gave rise to your sampled data, including both clonal frame and the recombinant edges. Loci and sites affected by each conversion are also inferred.

Parameters

For informative data sets, Bacter can sample posterior distributions for parameters such as the recombination rate and expected converted tract length jointly with the ARG. For serially sampled data, joint estimation of the molecular clock rate is also possible.

Population dynamics

Bacter includes support for producing Bayesian Skyline Plot (BSP) inferences of bacterial population dynamics. This inference may actually be more powerful when recombination is present in your data set.

tgvaughan.github.io/bacter

Bacter Tutorial

1. Open Bacter tutorial at taming-the-beast.github.io/tutorials/Bacter-Tutorial
2. Begin the tutorial!

References

Didelot, X. and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, 175:1251.

Didelot, X., Lawson, D., Daarling, A., and Falush, D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, 186:1435.

Jolley, K. A., Feil, E. J., Chan, M. S., and Maiden, M. C. (2001). Sequence type analysis and recombinational tests (start). *Bioinformatics (Oxford, England)*, 17:1230–1231.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

Vos, M. and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J*, 3(2):199–208.

Wiuf, C. (2000). A coalescence approach to gene conversion. *Theor Popul Biol*, 57(4):357–367.

Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics*, 155(1):451–462.