Recombination-aware analysis of bacterial sequence data using BEAST 2

Tim Vaughan Alexei Drummond Nigel French

Infectious Disease Research Centre Mini-Symposium
Massey University, Wellington, 9th–10th September, 2014

ALLAN WILSON CENTRE

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

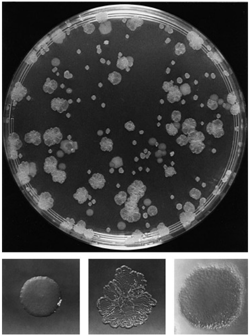
Why study bacterial phylogenetics?

Why study bacterial phylogenetics?

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

Why study bacterial phylogenetics?

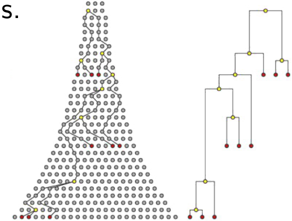
- Bacteria play important roles (both positive and negative) in the health of humans, animals and plants.
- Many bacteria possess interesting and *experimentally accessible* evolutionary dynamics.



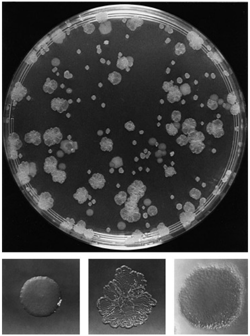
Rainey & Travisano, Nature (1998)

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)

The Apparent Problem

While bacteria reproduce asexually, from a genetic standpoint they are in fact *highly* promiscuous.

The Apparent Problem

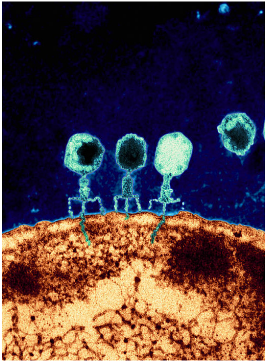
While bacteria reproduce asexually, from a genetic standpoint they are in fact *highly* promiscuous.

Conjugation

The Apparent Problem

While bacteria reproduce asexually, from a genetic standpoint they are in fact *highly* promiscuous.

Conjugation

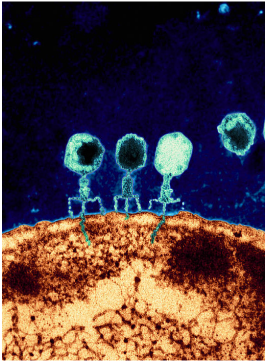


Transduction

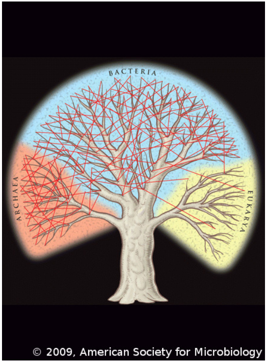
The Apparent Problem

While bacteria reproduce asexually, from a genetic standpoint they are in fact *highly* promiscuous.

Conjugation



Transduction



© 2009, American Society for Microbiology

Network of Life

So what?

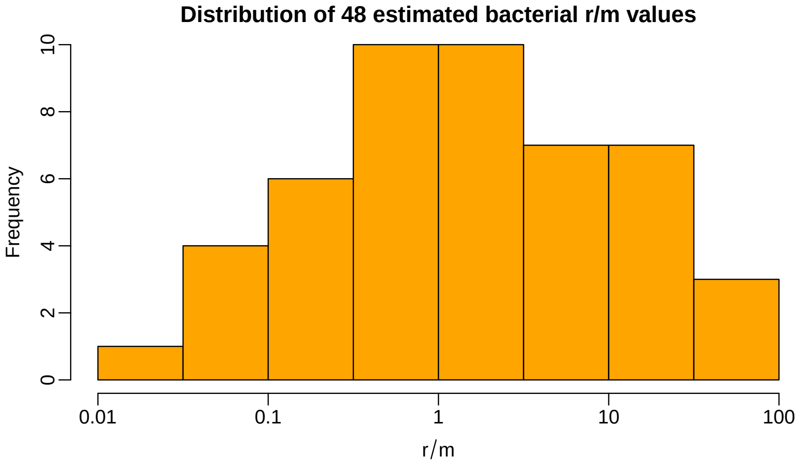
Other organisms (even viruses) employ non-vertical inheritance; why is this a show-stopper for bacterial phylogenetics?

The REAL Problem

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

The REAL Problem

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



Vos and Didelot, The ISME Journal (2009)

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 et al., 2007, 2010)

Existing solutions

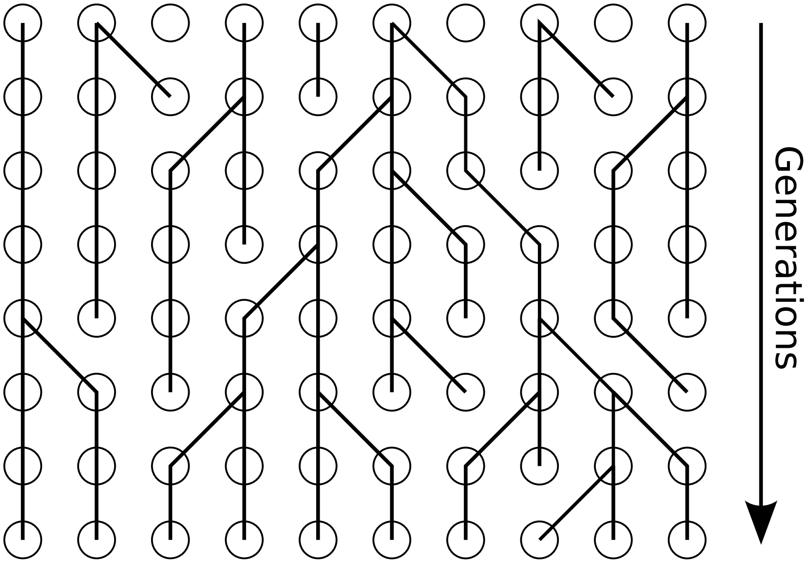
- Pre-processing of data to identify and remove non-vertically inherited material. (eg. START: Jolley et al., 2001)

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 et al., 2007, 2010)

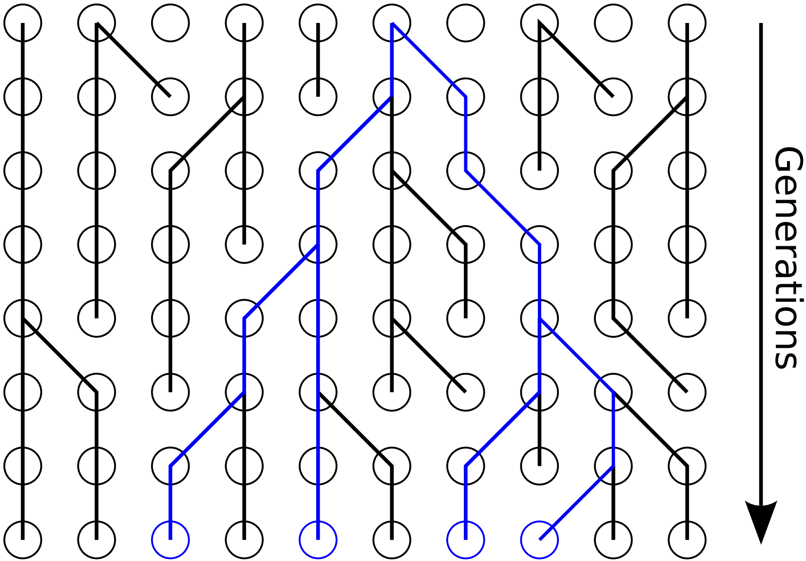
Pros	Cons
• Can make use of all data. • Can infer additional parameters such as recombination rates. • May yield increased confidence in estimates	• Models can be complex, with many parameters. • Both computationally and statistically challenging. • Existing implementations are too restrictive.

The coalescent with gene conversion



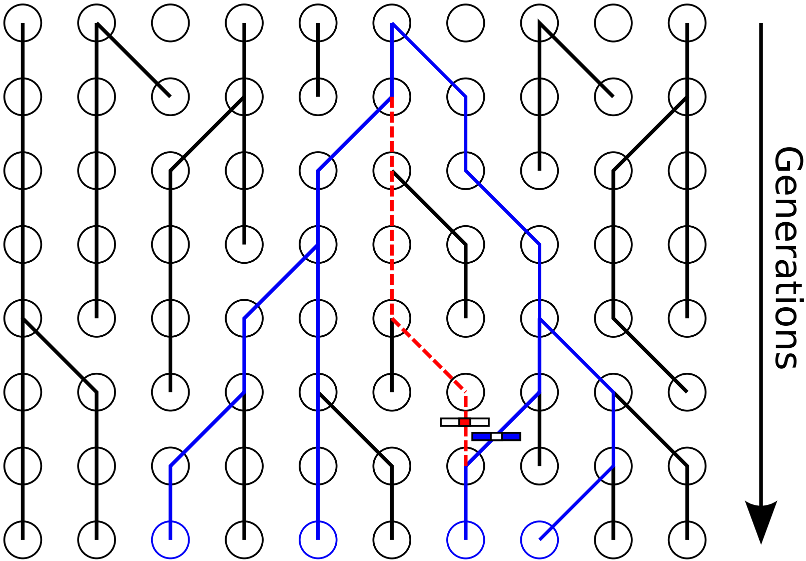
Wiuf, 1999; Wiuf and Hein, 2000

The coalescent with gene conversion



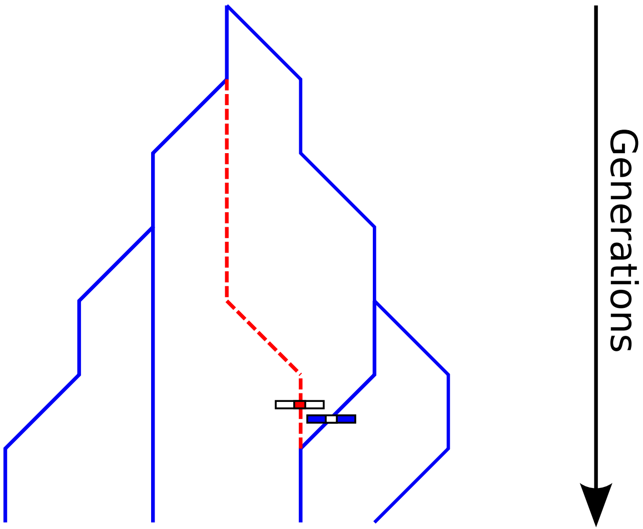
Wiuf, 1999; Wiuf and Hein, 2000

The coalescent with gene conversion



Wiuf, 1999; Wiuf and Hein, 2000

The coalescent with gene conversion

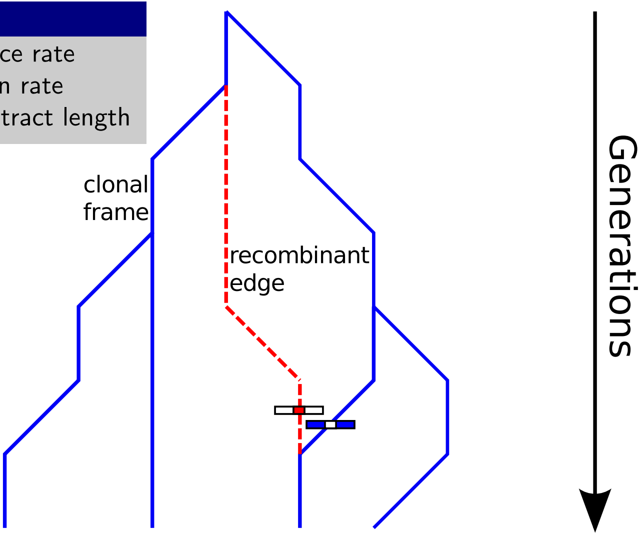


Wiuf, 1999; Wiuf and Hein, 2000

The coalescent with gene conversion

Parameters

$\theta(t)$	Coalescence rate
ρ	Conversion rate
δ	Expected tract length



Wiuf, 1999; Wiuf and Hein, 2000

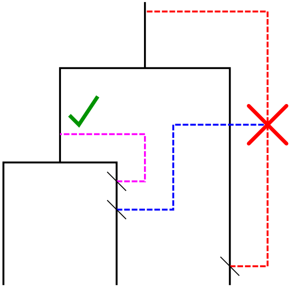
Approximate model

We make the following simplifying assumptions:

Approximate model

We make the following simplifying assumptions:

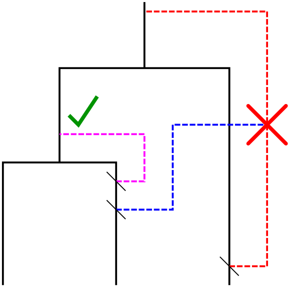
- Following Didelot et al., 2010, we exclude coalescent events between recombinant edges/lineages.



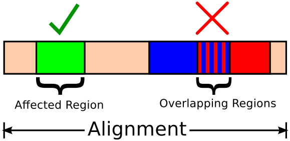
Approximate model

We make the following simplifying assumptions:

- Following Didelot et al., 2010, we exclude coalescent events between recombinant edges/lineages.



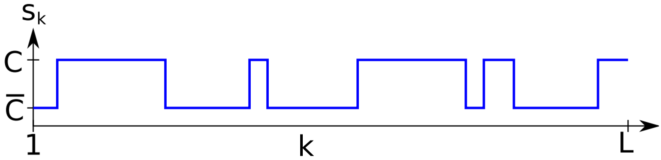
- In addition, we do not permit a site to be affected by more than one conversion at a time.



Approximate model

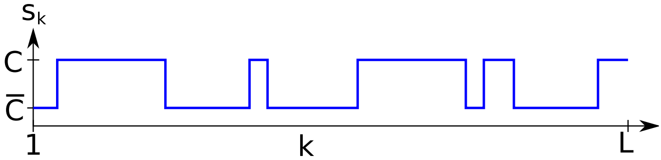
Approximate model

These assumptions allow the distribution of converted sites to be treated as a Markov chain, similar to the Sequentially Markovian Coalescent (McVean and Cardin, 2005).



Approximate model

These assumptions allow the distribution of converted sites to be treated as a Markov chain, similar to the Sequentially Markovian Coalescent (McVean and Cardin, 2005).



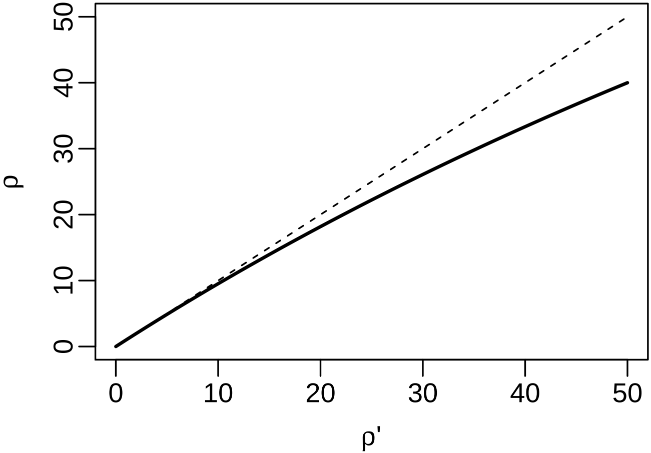
The probability $P(s_k|s_1)$ evolves accoarding to

$$\begin{bmatrix} P(s_{k+1}=C|s_1) \\ P(s_{k+1}=\bar{C}|s_1) \end{bmatrix} = \begin{bmatrix} (1-\delta^{-1}) & \frac{\rho'\lambda_T}{2} \\ \delta^{-1} & 1-\frac{\rho'\lambda_T}{2} \end{bmatrix} \begin{bmatrix} P(s_k=C|s_1) \\ P(s_k=\bar{C}|s_1) \end{bmatrix}$$

where λ_T is the total edge length of the clonal frame T, δ is the expected tract length and ρ' is a conversion rate parameter.

Approximate model

For a given number of expected conversions, the value of the conversion rate parameter ρ in Didelot et al.'s model is always lower than that of the rate parameter ρ' in our model.



Here $\lambda_T = 1$ and $\delta/L = 0.1$.

Bayesian inference framework

We aim to perform inference by using an MCMC algorithm to sample from the posterior

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

where

- A is the sequence alignment,
- μ are the substitution model parameters, and
- G is the full sample genealogy including clonal frame T, recombinant edges R, infected region map M.

Bayesian inference framework

We aim to perform inference by using an MCMC algorithm to sample from the posterior

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

where

- A is the sequence alignment,
- μ are the substitution model parameters, and
- G is the full sample genealogy including clonal frame T, recombinant edges R, infected region map M.

The genealogy density under approximate coalescent with gene conversion can be expanded

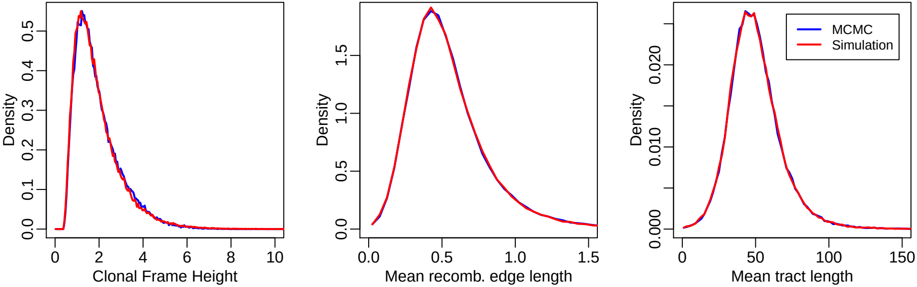
$$f_{CGC}(G | \rho', \delta, \theta) = f(R | T, M, \theta) P(M | T, \rho', \delta) f_C(T | \theta)$$

Implementation and validation

- We have implemented the MCMC algorithm as a BEAST 2 package. http://www.github.com/CompEvol/BACTER

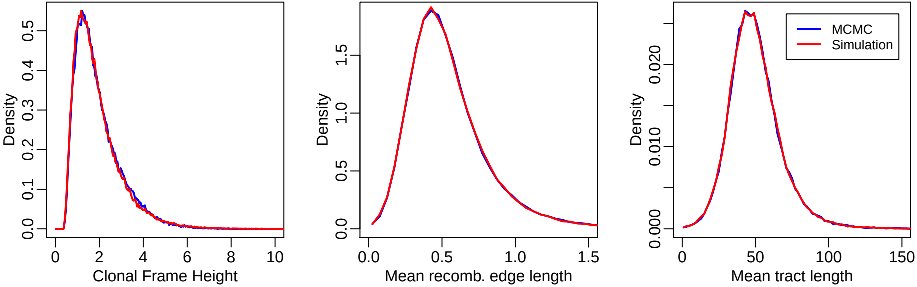
Implementation and validation

- We have implemented the MCMC algorithm as a BEAST 2 package. http://www.github.com/CompEvol/BACTER
- Primary validation involves comparing distributions of summary statistics calculated from simulated ARGs with those sampled via MCMC from the ARG prior $f_{CGC}(G|\rho', \delta, \theta)$.



Implementation and validation

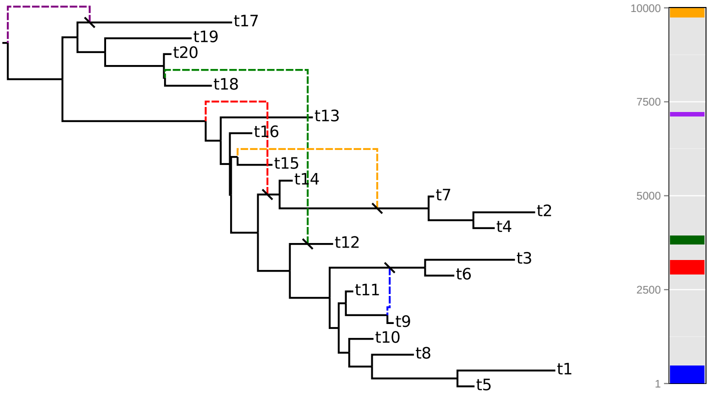
- We have implemented the MCMC algorithm as a BEAST 2 package. `http://www.github.com/CompEvol/BACTER`
- Primary validation involves comparing distributions of summary statistics calculated from simulated ARGs with those sampled via MCMC from the ARG prior $f_{CGC}(G|\rho', \delta, \theta)$.



- In this example we have used 5 heterochronous leaf times, $L = 10^4$, $\rho' = 5$, $\delta = 50$ and $\theta = 1$.

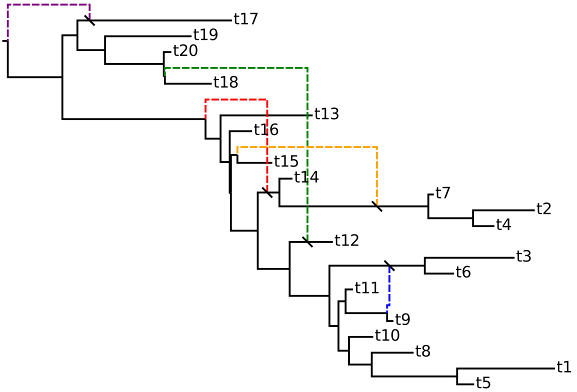
Producing simulated sequence data

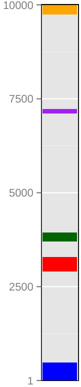
- The following network and conversion map were simulated assuming $\rho' = 200$, $\delta = 500$ and $\theta = 0.01$.



Producing simulated sequence data

- The following network and conversion map were simulated assuming $\rho' = 200$, $\delta = 500$ and $\theta = 0.01$.

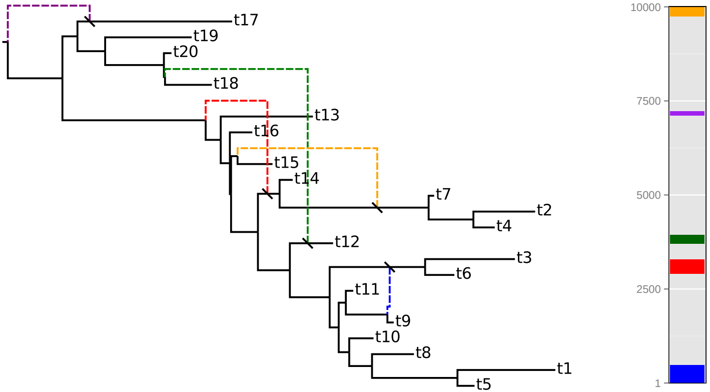




- A 10kb alignment was generated by simulating evolution down this network under Jukes-Cantor with clock rate $\mu = 10$.

Producing simulated sequence data

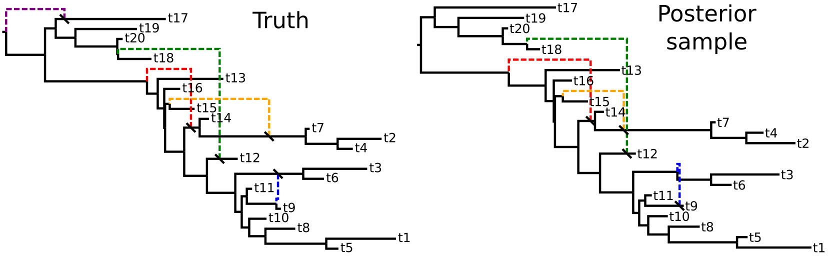
- The following network and conversion map were simulated assuming $\rho' = 200$, $\delta = 500$ and $\theta = 0.01$.



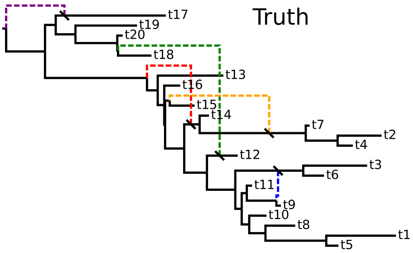
- A 10kb alignment was generated by simulating evolution down this network under Jukes-Cantor with clock rate $\mu = 10$.

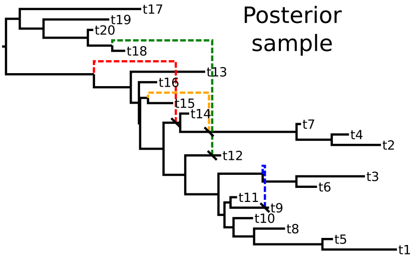
Visualization of network generated automatically from Extended Newick (Cardona et al., BMC Bioinf., 2008) representation using IcyTree (tgvaughan.github.io/icytree).

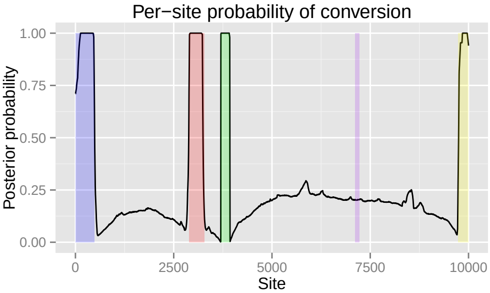
Network inference from simulated data

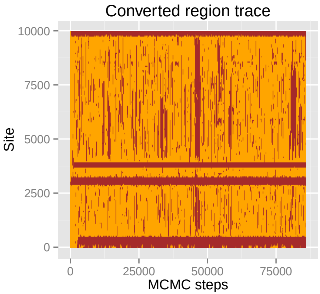


Network inference from simulated data

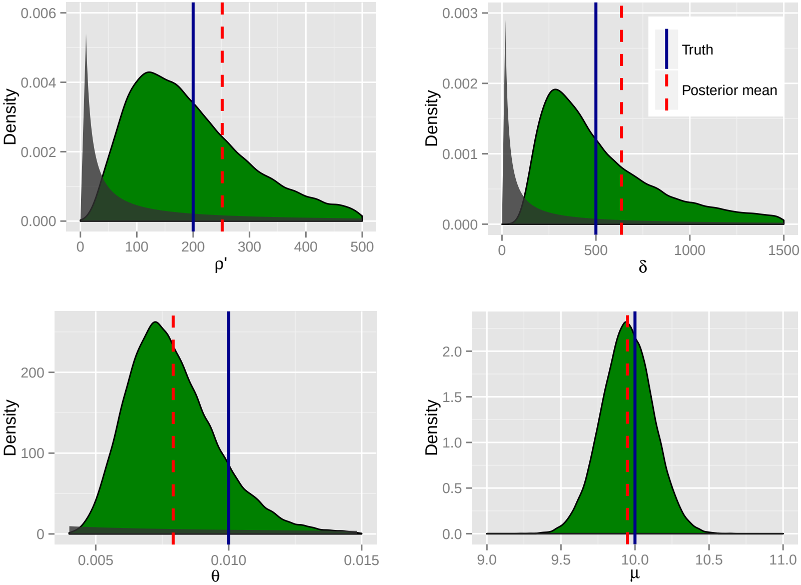




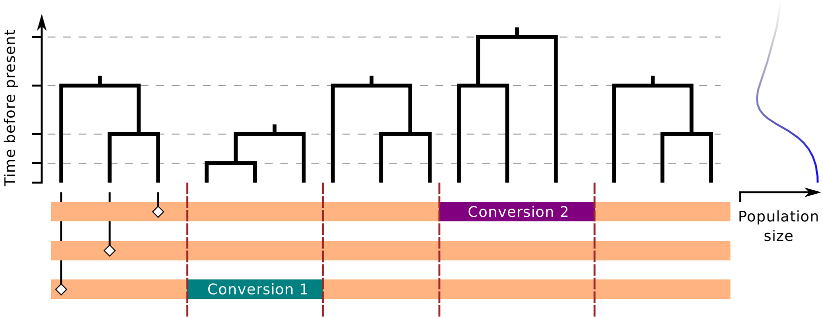




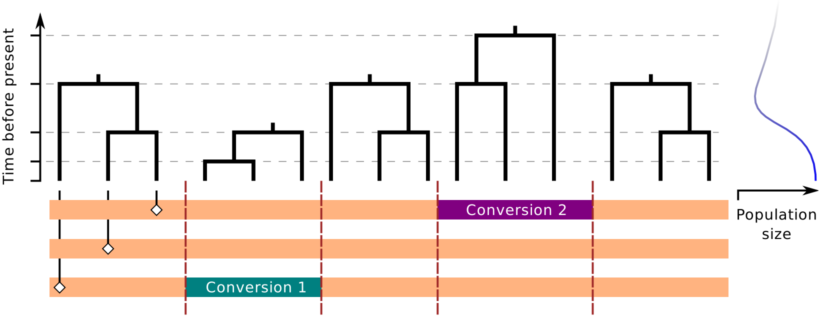
Parametric inference from simulated data



Benefit to the inference of demographic parameters



Benefit to the inference of demographic parameters



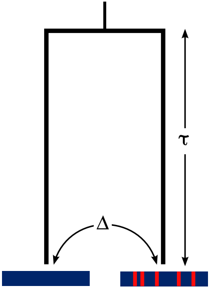
- Relationship used by Li and Durbin, Nature (2011) to infer human demographic history from pairs of autosomes.

How much inference power can we gain?

Consider an alignment of two sequences of length L. With complete linkage, the probability for the number of segregating sites under the Jukes-Cantor substitution model is

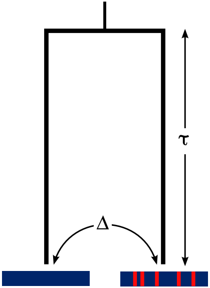
$$P(\Delta|\tau) = \frac{1}{4^L}\left(\frac{2}{3}\mu\tau\right)^{\Delta} e^{-2(L-\Delta)\mu\tau}$$

in the limit $\tau \ll 1/\mu$.



How much inference power can we gain?

Consider an alignment of two sequences of length L. With complete linkage, the probability for the number of segregating sites under the Jukes-Cantor substitution model is



$$P(\Delta|\tau) = \frac{1}{4^L}\left(\frac{2}{3}\mu\tau\right)^{\Delta} e^{-2(L-\Delta)\mu\tau}$$

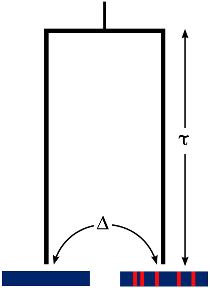
in the limit $\tau \ll 1/\mu$.
The density of τ under the coalescent with population size parameter θ is

$$P(\tau|\theta) = \frac{1}{\theta}e^{-\frac{\tau}{\theta}}$$

How much inference power can we gain?

Consider an alignment of two sequences of length L. With complete linkage, the probability for the number of segregating sites under the Jukes-Cantor substitution model is

$$P(\Delta|\tau) = \frac{1}{4^L}\left(\frac{2}{3}\mu\tau\right)^{\Delta} e^{-2(L-\Delta)\mu\tau}$$



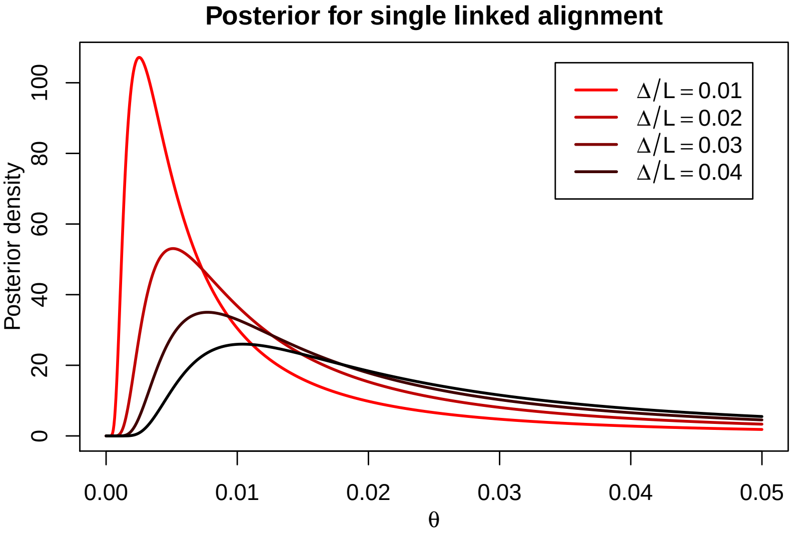
in the limit $\tau \ll 1/\mu$.
The density of τ under the coalescent with population size parameter θ is

$$P(\tau|\theta) = \frac{1}{\theta}e^{-\frac{\tau}{\theta}}$$

Using the Jeffreys prior for θ, the posterior density becomes

$$P(\theta|\Delta) = \frac{\Delta(2(L-\Delta)\mu)^{\Delta}\theta^{\Delta-1}}{(2(L-\Delta)\mu\theta+1)^{\Delta+1}}$$

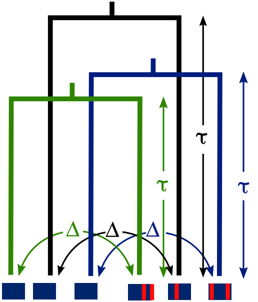
How much inference power can we gain?



How much inference power can we gain?

Now assume the sequence is divided into n loci each of length L/n, and with its own τ_i and Δ_i. The posterior density then becomes

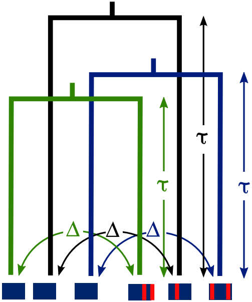
$$P(\theta|\vec{\Delta}) = \frac{\theta^{-1}}{Z} \prod_{i=1}^{n} \frac{(\frac{2}{3}\mu)^{\Delta_i}(\Delta_i!)\theta^{\Delta_i}}{(2(L/n - \Delta_i)\mu\theta + 1)^{\Delta_i + 1}}$$



How much inference power can we gain?

Now assume the sequence is divided into n loci each of length L/n, and with its own τ_i and Δ_i. The posterior density then becomes

$$P(\theta|\vec{\Delta}) = \frac{\theta^{-1}}{Z} \prod_{i=1}^{n} \frac{(\frac{2}{3}\mu)^{\Delta_i}(\Delta_i!)\theta^{\Delta_i}}{(2(L/n-\Delta_i)\mu\theta+1)^{\Delta_i+1}}$$

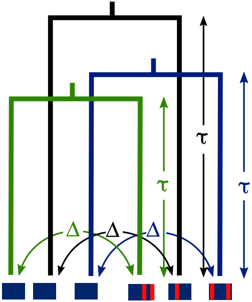


Unlike the single locus case, the normalizing constant Z and hence the density itself must be evaluated numerically.

How much inference power can we gain?

Now assume the sequence is divided into n loci each of length L/n, and with its own τ_i and Δ_i. The posterior density then becomes

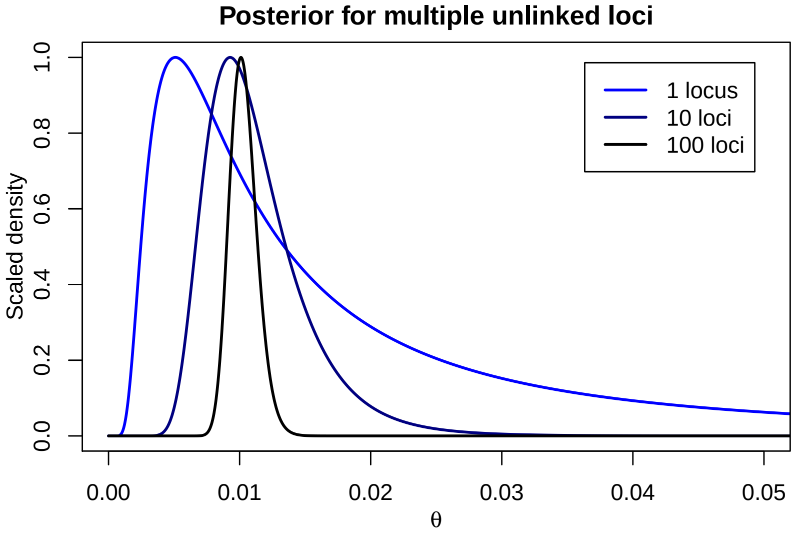
$$P(\theta|\vec{\Delta}) = \frac{\theta^{-1}}{Z} \prod_{i=1}^{n} \frac{(\frac{2}{3}\mu)^{\Delta_i} (\Delta_i!) \theta^{\Delta_i}}{(2(L/n - \Delta_i)\mu\theta + 1)^{\Delta_i + 1}}$$



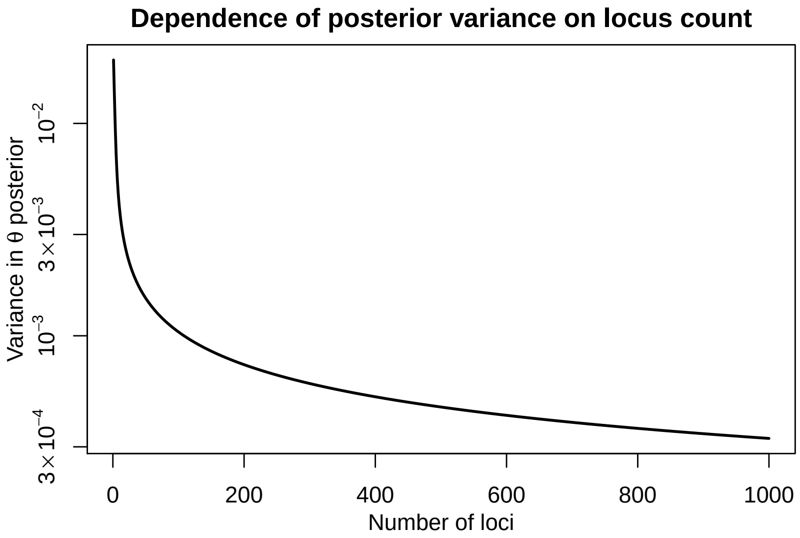
Unlike the single locus case, the normalizing constant Z and hence the density itself must be evaluated numerically.

- Can get a *rough* idea of the effect of increasing homologous conversion rate by fixing $\Delta_i = \Delta/n$ and varying n.

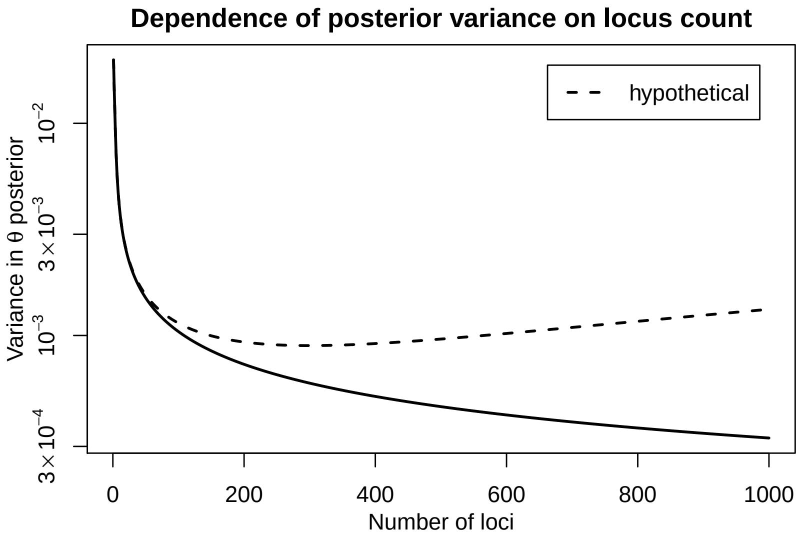
How much inference power can we gain?



How much inference power can we gain?

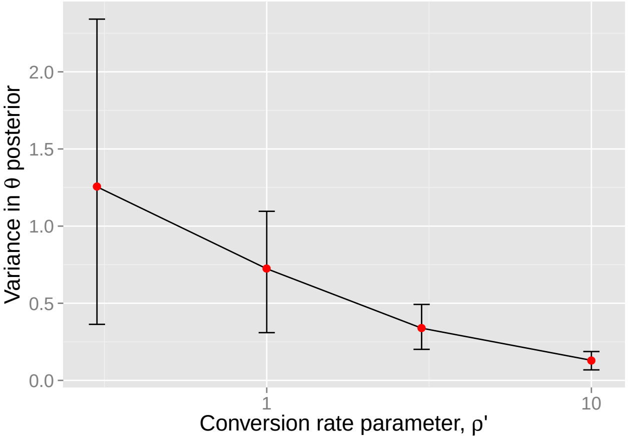


How much inference power can we gain?



Simulation study

- Performed joint inference of ARG and θ from 5 datasets for 4 distinct values of the conversion rate parameter ρ'.



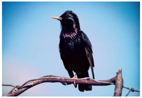
$\Longrightarrow$ Increased conversion can improve demographic inference.

Campylobacter genomic data

- Genus of spiral-shaped bacteria responsible for the majority of gastroenteritis in the developed world.

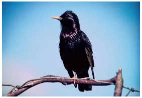
Campylobacter genomic data

- Genus of spiral-shaped bacteria responsible for the majority of gastroenteritis in the developed world.
- Often isolated from feces of domestic farm animals and environmental sources in NZ.



Campylobacter genomic data

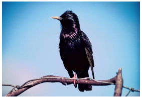
- Genus of spiral-shaped bacteria responsible for the majority of gastroenteritis in the developed world.
- Often isolated from feces of domestic farm animals and environmental sources in NZ.



- Full genomes sequenced from 60 *C. coli* and *C. jejuni* isolates sampled from a variety of sources in NZ between 2005 and 2009.

Campylobacter genomic data

- Genus of spiral-shaped bacteria responsible for the majority of gastroenteritis in the developed world.
- Often isolated from feces of domestic farm animals and environmental sources in NZ.



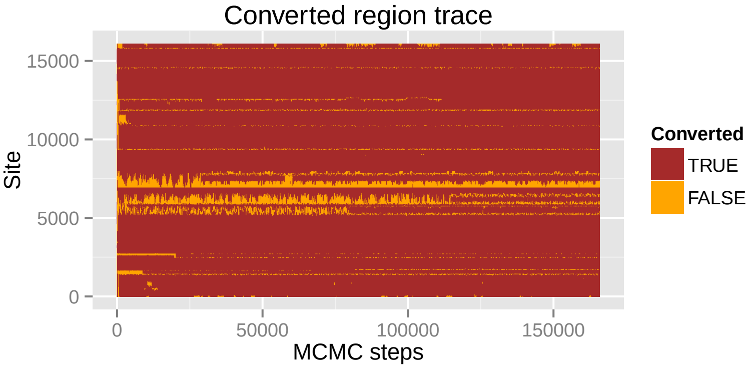
- Full genomes sequenced from 60 *C. coli* and *C. jejuni* isolates sampled from a variety of sources in NZ between 2005 and 2009.
- Inference performed on alignment of contiguous 16kb region between the genes aspA and uncA (inclusive).

Campylobacter dataset analysis results

- Analyzed alignment assuming a strict clock, a GTR$+\gamma$ substitution model and $\rho'/\mu = 3$ (motivated by Fearnhead et al., JME, 2012).

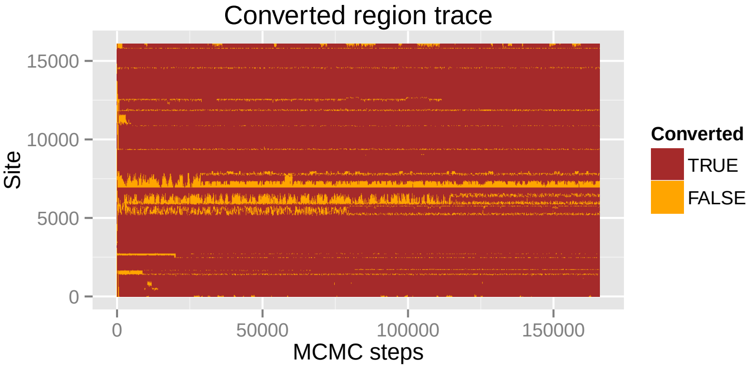
Campylobacter dataset analysis results

- Analyzed alignment assuming a strict clock, a GTR+γ substitution model and $\rho'/\mu = 3$ (motivated by Fearnhead et al., JME, 2012).



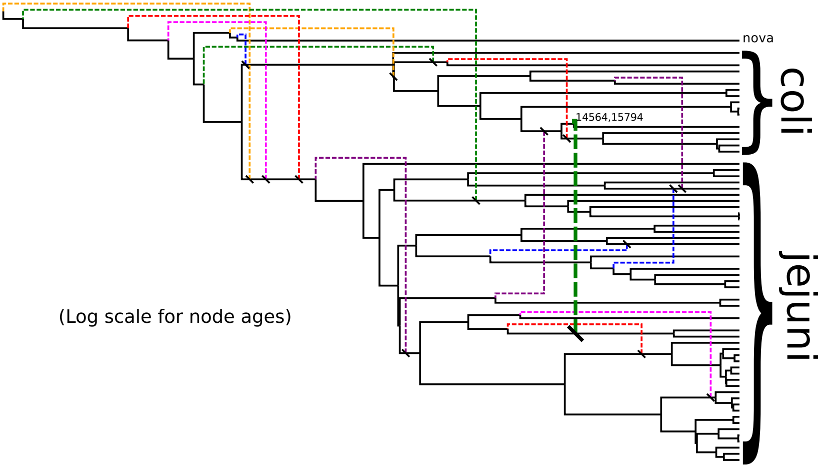
Campylobacter dataset analysis results

- Analyzed alignment assuming a strict clock, a GTR+γ substitution model and $\rho'/\mu = 3$ (motivated by Fearnhead et al., JME, 2012).



- Campylobacter dataset in obvious danger of violating the “no overlap” assumption of the model.

Campylobacter dataset analysis results



(Log scale for node ages)

- Known gene conversion recovered: incorporation of C. coli uncA gene by ST61 C. jejuni strain (and a close relative) as described by Wilson et al., MBE, 2009.

Summary

- Have implemented a scheme for Bayesian inference under an approximate coalescent with gene conversion model (inspired by Didelot et al., 2010) as a BEAST 2 package.

Summary

- Have implemented a scheme for Bayesian inference under an approximate coalescent with gene conversion model (inspired by Didelot et al., 2010) as a BEAST 2 package.
- Scheme is capable of recovering model parameters (ρ', δ, θ, and μ) from sequence alignments, as well as jointly inferring the sites affected by conversion and the ARG.

Summary

- Have implemented a scheme for Bayesian inference under an approximate coalescent with gene conversion model (inspired by Didelot et al., 2010) as a BEAST 2 package.
- Scheme is capable of recovering model parameters (ρ', δ, θ, and μ) from sequence alignments, as well as jointly inferring the sites affected by conversion and the ARG.
- Simulated data analyses confirm that the ability of our scheme to estimate population size improves with increasing conversion rates.

Summary

- Have implemented a scheme for Bayesian inference under an approximate coalescent with gene conversion model (inspired by Didelot et al., 2010) as a BEAST 2 package.
- Scheme is capable of recovering model parameters (ρ', δ, θ, and μ) from sequence alignments, as well as jointly inferring the sites affected by conversion and the ARG.
- Simulated data analyses confirm that the ability of our scheme to estimate population size improves with increasing conversion rates.
- The assumption that each site is affected by at most one conversion seems to be violated in the case of the available Campylobacter data.

Future goals

- Relax the non-overlapping conversion assumption.
  (In progress.)

Future goals

- Relax the non-overlapping conversion assumption. **(In progress.)**
- Perform a full assessment of the ability the inference scheme to correctly infer unknowns in the presence of model misspecification. **(In progress.)**

Future goals

- Relax the non-overlapping conversion assumption. **(In progress.)**
- Perform a full assessment of the ability the inference scheme to correctly infer unknowns in the presence of model misspecification. **(In progress.)**
- Investigate parametric and non-parametric inference of demographic history dynamics.

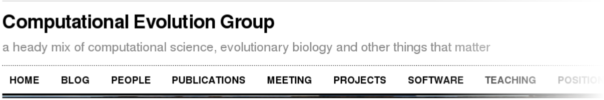
Future goals

- Relax the non-overlapping conversion assumption. **(In progress.)**
- Perform a full assessment of the ability the inference scheme to correctly infer unknowns in the presence of model misspecification. **(In progress.)**
- Investigate parametric and non-parametric inference of demographic history dynamics.

BEAST 2 package source code

The BEAST 2 package is still in development, but the source code is available at `http://www.github.com/CompEvol/BACTER`.

Acknowledgements



Computational Evolution Group

a heady mix of computational science, evolutionary biology and other things that matter

- David Welch

- Patrick Biggs

NeSI
New Zealand eScience Infrastructure

ALLAN WILSON CENTRE