Bayesian Molecular Epidemiology

Alexei Drummond, alexei@cs.auckland.ac.nz
Computational Evolution Group
University of Auckland, Auckland, New Zealand

May 21-25, MCEB, Montpellier 2013

Bayesian phylogenetics

Phylodynamics

Coalescent models

Birth-death-sampling models

Sampling ancestors

Structured tree models

Perspective

Acknowledgements

Phylodynamics@ Computational Evolution Group

- Denise Kuhnert (PhD student)
- Jessie Wu (PhD student)
- Sasha Gavryuskina (PhD student)
- Tim Vaughan (postdoc)
- Remco Bouckaert (postdoc)
- Walter Xie (research programmer)

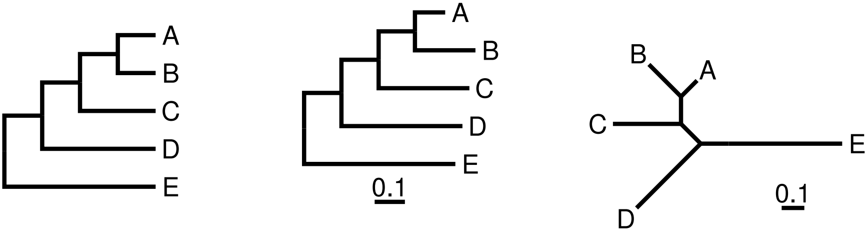
Collaborators

- Sebastian Bonhoeffer, ETH Zurich
- Tanja Stadler, ETH Zurich
- Andrew Rambaut, Edinburgh, UK
- Marc Suchard, UCLA
- David Welch, Auckland, New Zealand
- Peter Drummond, SUT Melbourne

Types of phylogenies and representations

rooted trees

unrooted tree



(a) cladogram

(b) phylogram

(c) unrooted tree

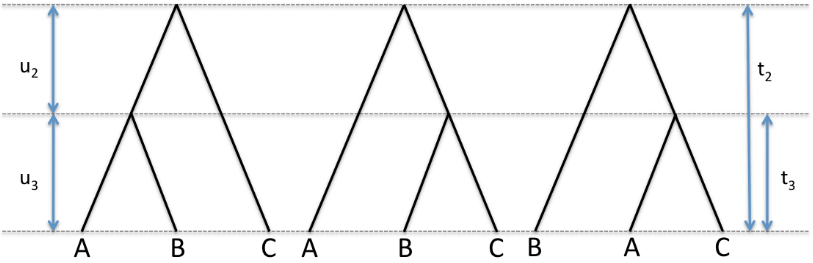
((((A, B), C), D), E);

((((A:0.1, B:0.2):0.12, C:0.3):0.123, D:0.4):0.1234, E:0.5);

branches (edges) and their lengths, nodes, tips (leaves)

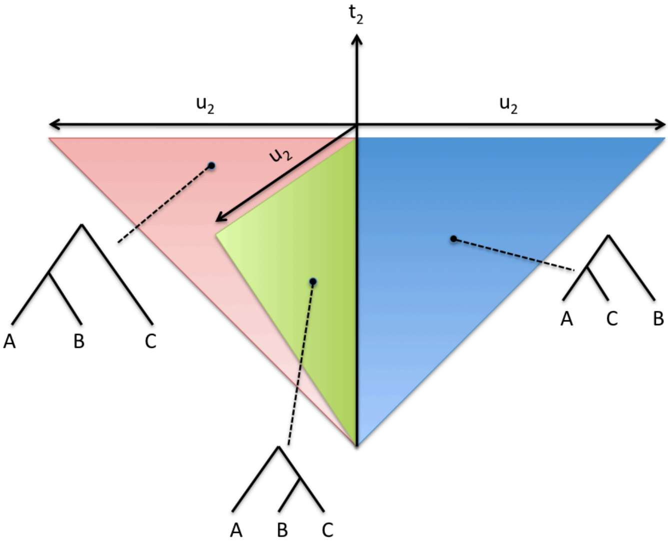
The tip-labeled time-tree

A tip-labeled time-tree is described by a *tip-labeled ranked topology* of size k and *coalescent intervals*, $\mathbf{u} = \{u_2, \ldots, u_k\}$.

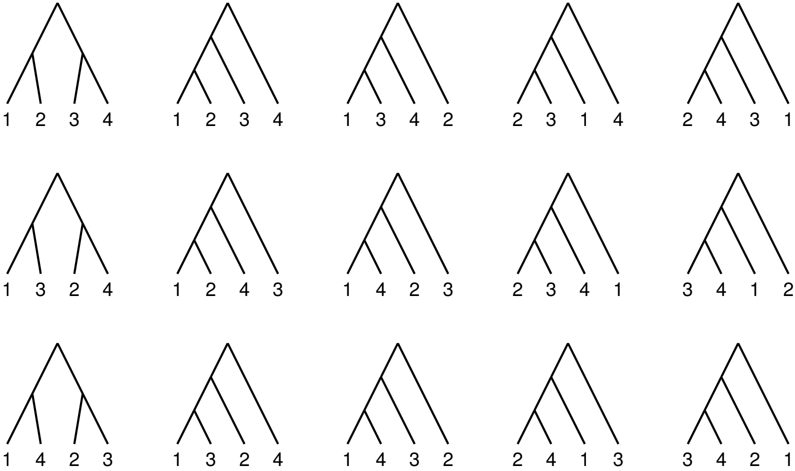


These time-trees of size 3 can be interpreted as describing the possible alternative evolutionary histories or (uniparental) ancestries of the three individuals represented by the labeled tips.

The space of tip-labeled time-trees of size 3



Unranked tree topologies of size 4



How many trees are there?

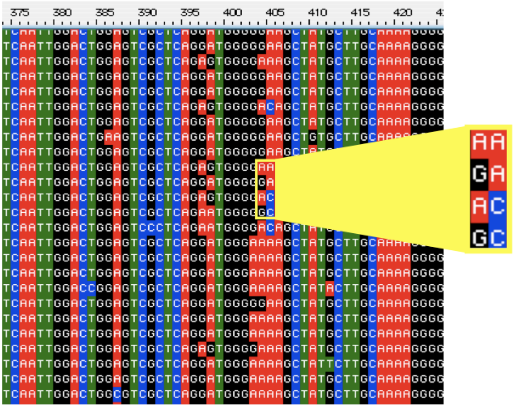
For *n* species there are

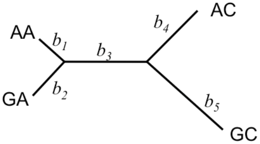
$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \tfrac{(2n-3)!}{(n-2)!2^{n-2}}$$

rooted, tip-labelled binary trees:

n	#trees	
4	15	enumerable by hand
5	105	enumerable by hand on a rainy day
6	945	enumerable by computer
7	10395	still searchable very quickly on computer
8	135135	about the number of hairs on your head
9	2027025	greater than the population of Auckland
10	34459425	$\approx$ upper limit for exhaustive search
20	8.20×10^{21}	$\approx$ upper limit of branch-and-bound searching
48	3.21×10^{70}	$\approx$ the number of particles in the Universe
136	2.11×10^{267}	number of trees to choose from in the “Out of Africa” data (Vigilant *et al*. 1991)

Felsenstein’s likelihood (1981)





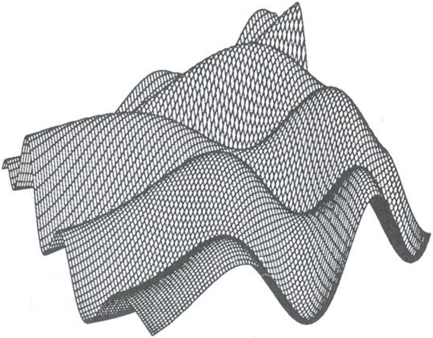
$$L(T) = Pr\{D|T, Q\}$$

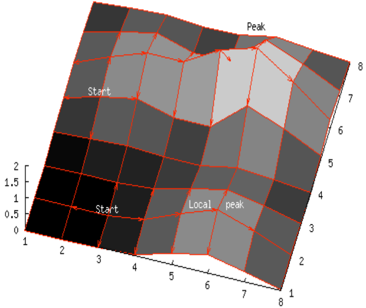
The probability of the data, $Pr\{D|T, Q\}$ can be efficiently calculated given a phylogenetic tree (T), and a **probabilistic model** of molecular evolution (Q).

In statistical phylogenetics, branch lengths are traditionally unconstrained.

Tree space as a hilly landscape

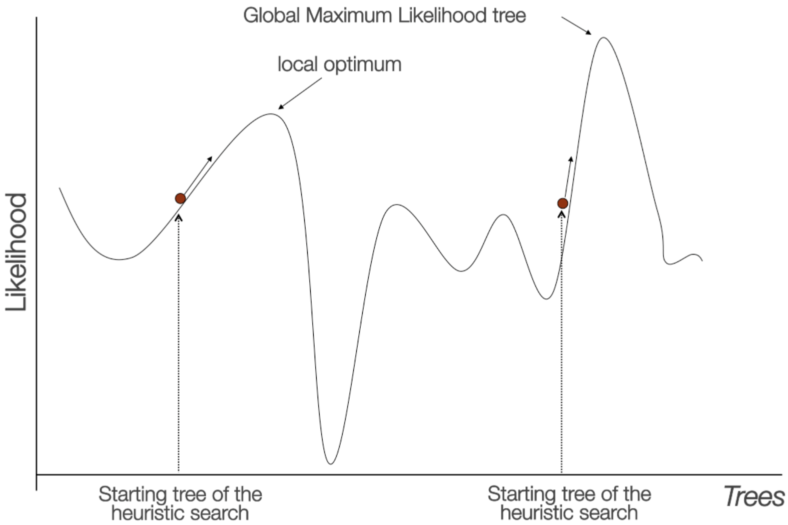
The space of all possible trees can be visualized as a hilly landscape. Nearby points in this landscape represent similar trees, and the height of the landscape is the probability of the tree at that point.





- This space can be **sampled** in a Bayesian analysis with MCMC
- The peak can be identified by a **search algorithm** in the context of maximum likelihoods

Local tree search and multiple optima


Global Maximum Likelihood tree
local optimum
Likelihood
Starting tree of the heuristic search
Starting tree of the heuristic search
Trees

Bayes rule in statistics

$$Pr(\theta|D) = \frac{Pr(D|\theta)Pr(\theta)}{Pr(D)}$$

where

- $P(D|\theta)$ is the likelihood,
- $Pr(\theta)$ is the prior distribution and
- $Pr(\theta|D)$ is the posterior distribution.
- $Pr(D)$ is the marginal likelihood of the data.

Bayes rule in phylogenetics

$$p(T, Q|D) = \frac{Pr\{D|T, Q\}p(T)p(Q)}{Pr\{D\}}$$

where

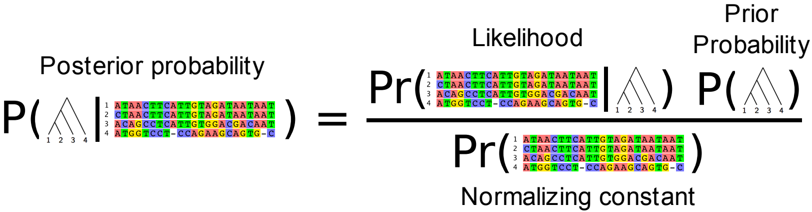
- $Pr(D|T, Q)$ is Felsenstein's likelihood,
- $p(T)$ is the prior distribution on phylogenetic trees,
- $p(Q)$ is the prior distribution on the model of evolution and
- $p(T, Q|D)$ is the posterior distribution
- $Pr(D)$ is the marginal likelihood of the data.

Bayesian reconstruction of phylogenetic trees

Yang & Rannala (1997), Mau, Newton & Larget (1998)

In the context of Bayesian phylogenetics, what we want to compute is the **probability of the tree** given the data.

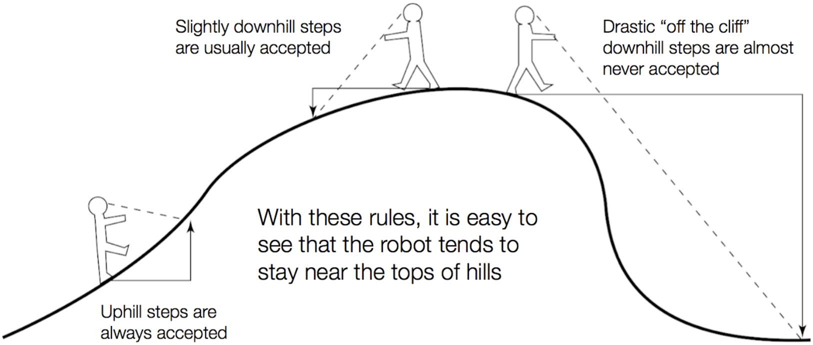
We can compute that from the **likelihood** using **Bayes Theorem**:


Posterior probability
Likelihood
Prior Probability
Normalizing constant
P(
Pr(
1 ATAACTTCATTGTAGATAATAAT
2 CTAACTTCATTGTAGATAATAAT
3 ACAGCCTCATTGTGGACGACAAT
4 ATGGTCCT-CCAGAAGCAGTG-C
1 2 3 4

This is known as the **Posterior probability** of the tree. Another method of reconstructing the evolutionary history is then to find the tree that has the **Maximum Posterior probability**.

Markov chain Monte Carlo (MCMC) robot

[courtesy of Paul O Lewis]



Slightly downhill steps are usually accepted

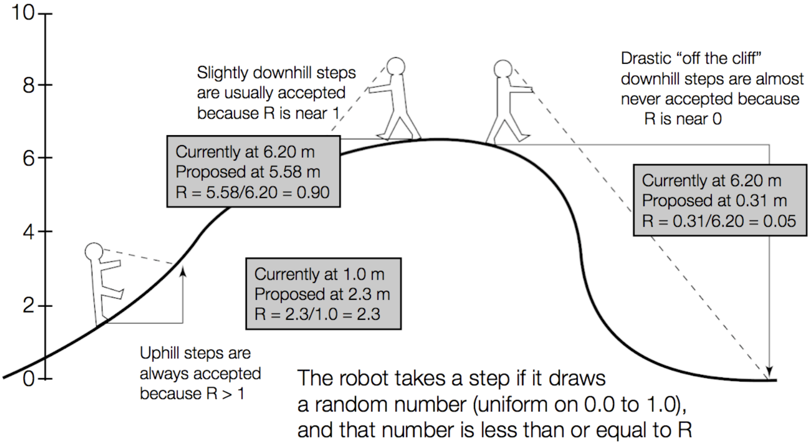
Drastic “off the cliff” downhill steps are almost never accepted

With these rules, it is easy to see that the robot tends to stay near the tops of hills

Uphill steps are always accepted

Markov chain Monte Carlo (MCMC) robot

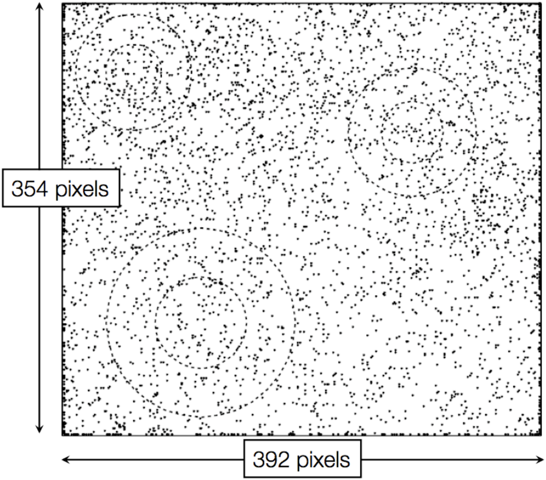
[courtesy of Paul O Lewis]


10
8
6
4
2
0
Slightly downhill steps are usually accepted because R is near 1
Drastic “off the cliff” downhill steps are almost never accepted because R is near 0
Currently at 6.20 m
Proposed at 5.58 m
R = 5.58/6.20 = 0.90
Currently at 6.20 m
Proposed at 0.31 m
R = 0.31/6.20 = 0.05
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
Uphill steps are always accepted because R > 1

The robot takes a step if it draws a random number (uniform on 0.0 to 1.0), and that number is less than or equal to R

Pure Random Walk

[courtesy of Paul O Lewis]



Proposal scheme:

- random direction
- gamma-distributed step length (mean 45 pixels, s.d. 40 pixels)
- reflection at edges

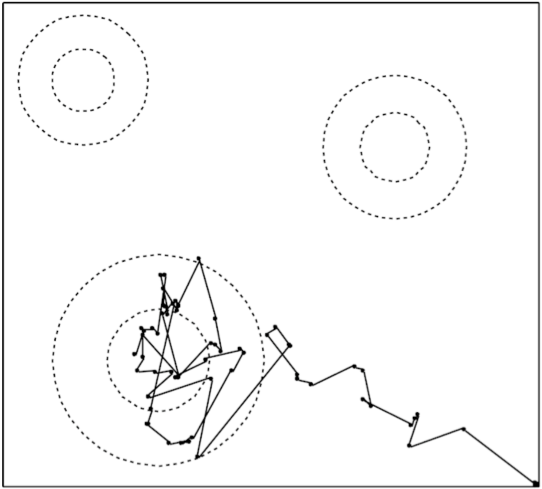
Target distribution:

- equal mixture of 3 bivariate normal hills
- inner contours: 50%
- outer contours: 95%

In this case the robot is accepting every step and 5000 steps are shown

Burn In

[courtesy of Paul O Lewis]



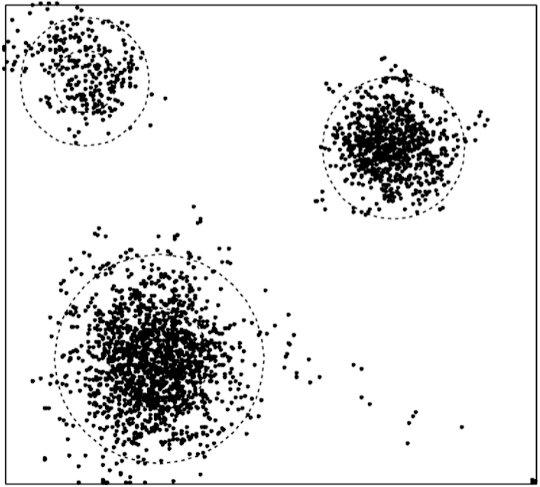
Robot is now following the rules and thus quickly finds one of the three hills.

Note that first few steps are not at all representative of the distribution.

100 steps taken from starting point

Target Distribution Approximation

[courtesy of Paul O Lewis]



How good is the MCMC approximation?

- 51.2% of points are inside inner contours (cf. 50% actual)
- 93.6% of points are inside outer contours (cf. 95% actual)

Approximation gets better the longer the chain is allowed to run.

5000 steps taken

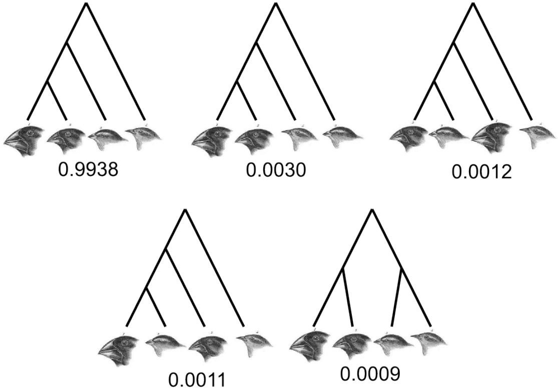
Target distribution *versus* proposal distribution

- The target distribution is the posterior distribution of interest
- The proposal distribution is used to decide which point to try next
  - you have much flexibility here, and the choice affects only the efficiency of the MCMC algorithm
  - MCMC using a symmetric proposal distribution is the Metropolis algorithm (Metropolis et al. 1953)
  - Use of an asymmetric proposal distribution requires a modification proposed by Hastings (1970), and is known as the Metropolis-Hastings algorithm

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1092.

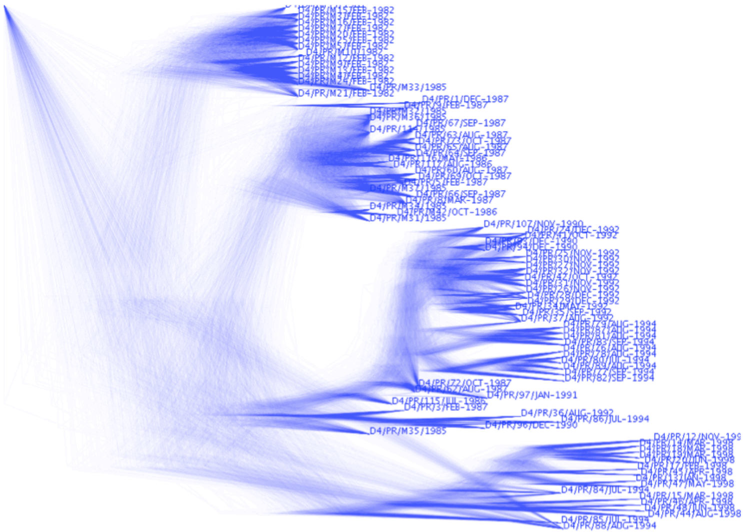
The Posterior Distribution on Darwin’s Finches

```
1 ATAACTTCATTGTAGATAATAAT
2 CTAACTTCATTGTAGATAATAAT
3 ACAGCCTCATTGTGGACGACAAT
4 ATGGTCCT-CCAGAAGCAGTG-C
```



This posterior probability distribution was computed using **Markov chain Monte Carlo** implemented in the BEAST software package (Drummond & Rambaut, 2007).

The posterior distribution for a moderately large time tree



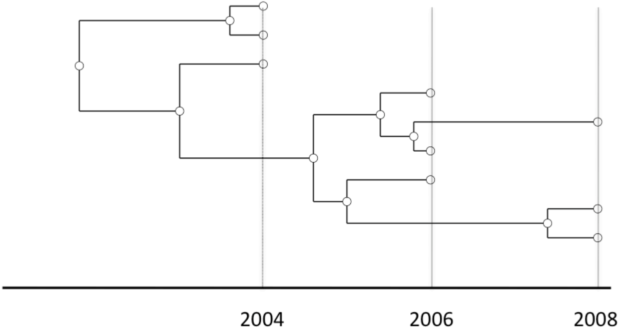
Summary of BEAST 2 capabilities

Analysis	Estimate phylogenies from alignments	
	Estimate dates of most recent common ancestors	
	Estimate gene and species trees	
	Infer population histories	
	Estimate substitution rates	
	Phylogeography	
	Path sampling	
	Simulation studies	
Models	Trees	Gene trees, species trees, structured coalescent, serially sampled trees
	Tree Likelihood	Felsenstein, threaded, Beagle
		Continuous, Ancestral reconstruction
		SNAPP
		Auto Partition
	Substitution Models	JC96, HKY, TN93, GTR
		Covarion, Stochastic Dollo
		RB, substBMA
		Blosum62, CPREV, Dayhoff, JTT, MTREV, WAG
	Frequency models	Fixed, estimated, empirical
	Sitemodels	Gamma site model, Mixture site model
	Tree Priors	Coalescent constant, exponential, skyline
		Birth Death Yule, Birth Death Sampling Skyline
		Yule with callibration correction
		Multi species coalescent
	Clock Models	Strict, Relaxed, Random
	Prior distributions	Uniform, 1/X, Normal, LogNormal, Gamma, Beta, etc.

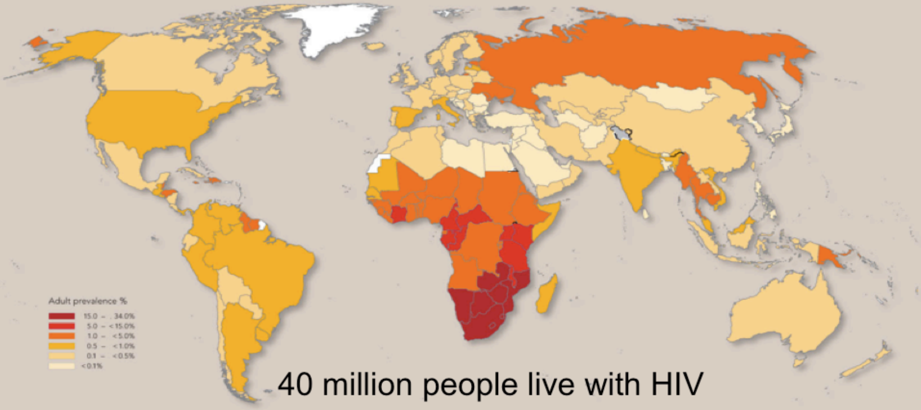
Evolution is happening right now!

Rodrigo and Felsenstein, 1999; Drummond *et al*, 2002

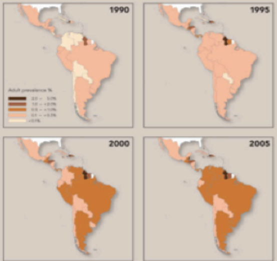
Many pathogens, such as HIV, Hepatitis C and Influenza A, evolve very rapidly, so that samples of the virus population from different times directly reveal evolutionary change.

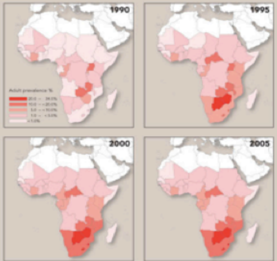


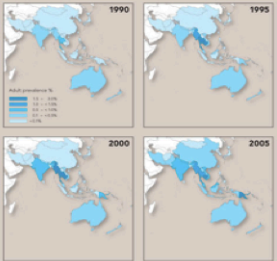
In fact it becomes possible to **calibrate** the tree and thus place the tree on a time scale - by constraining the tips to known sampling times



40 million people live with HIV

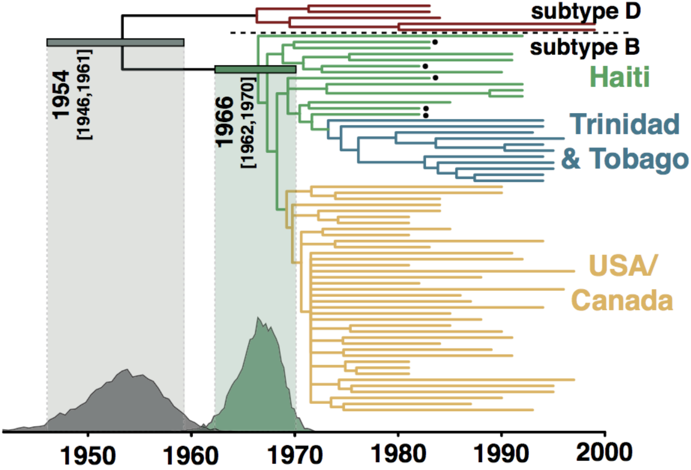






A calibrated phylogenetic inference

Origin of HIV Epidemic in the Americas, Gilbert *et al* (2007)



A phylogenetic reconstruction of samples of HIV-1 virus. Each degree one node represents a single infected individual from whom a blood sample has been taken.

Phylodynamics

- The intersection of **phylogenetics** and **mathematical epidemiology**
- Includes estimation of epidemiological parameters from phylogenetic data
- In a Bayesian setting, this has the familiar flavor of a hierarchical tree prior
- The hyperparameters of the tree prior become dynamical parameters of the epidemiological model
- The most common approach is to leverage coalescent theory, by using coalescent machinery augmented with deterministic models of effective population size parametrized by R_0 or its epidemiological constituents (net infection rate *et cetera*).

Coalescent models

Bayesian coalescent inference

- Kingman's coalescent is a **mathematical theory describing a genealogy of a small random sample** from a large background population.
- Provides a probability distribution over tree space given a population size history: $P(\mathcal{G}|N)$
- Old coalescent trees come from large populations
- Star-like coalescent trees come from exponentially growing populations
- In a Bayesian framework the coalescent is a hierarchical prior on tree space.
- Backwards in time model
- Applied to both within-host and between-host population dynamics

The coalescent with serial samples

Many epidemiological agents evolve very rapidly, so that the effect of sampling the population at different times becomes important.

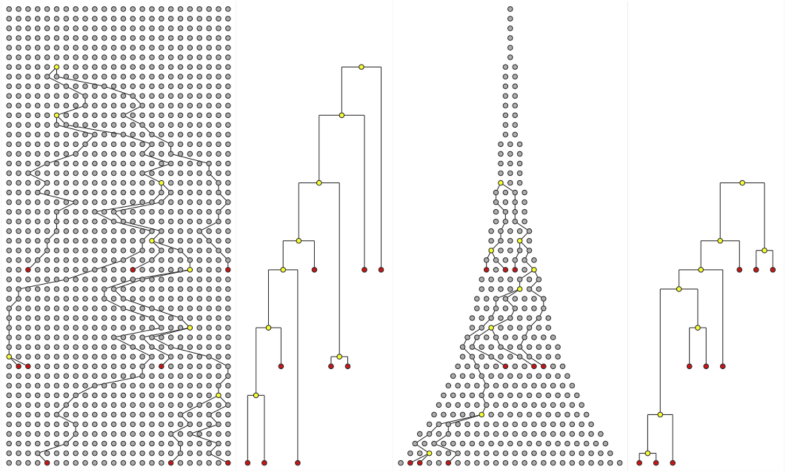
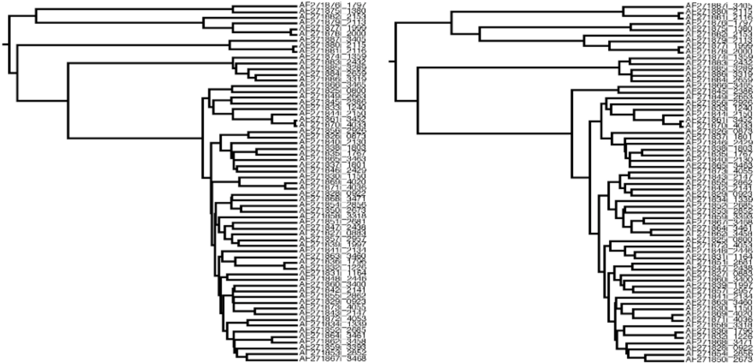


Fig. 3. The underlying Wright–Fisher population and serially-sampled genealogies from two populations. The first population has a constant population size over the history of the genealogy, while the second population has been exponentially growing. The coalescent likelihood calculates the probability of a genealogy given a particular background population history (e.g., constant or exponentially growing) and can therefore be employed to estimate the population history that best reflects the shape of the co-estimated phylogeny.

Bayesian integration of uncertainty in genealogies

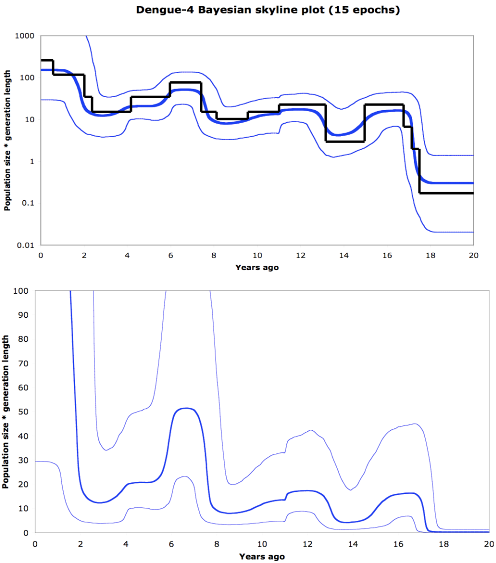


How similar are these two trees? Both of them are plausible given the data. We can use Bayesian Markov chain Monte Carlo to average the coalescent over all plausible trees.

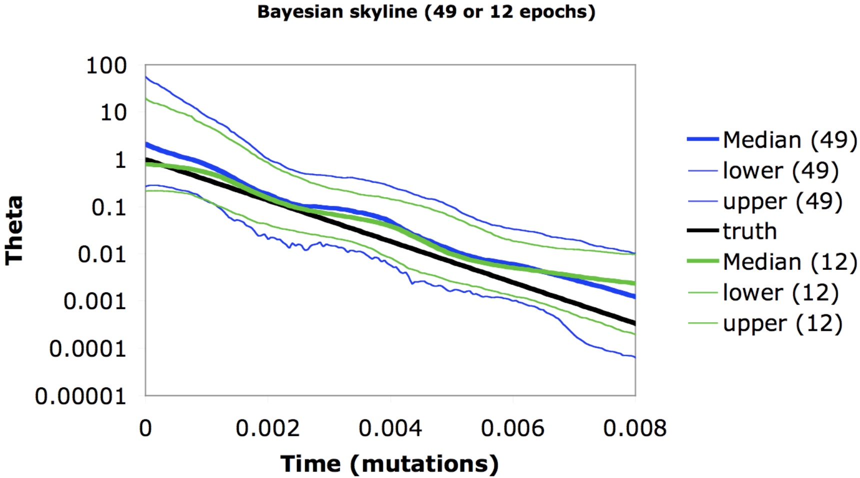
The Bayesian skyline plot

Drummond *et al* (2005), *Molecular Biology and Evolution*

The Bayesian skyline plot estimates a demographic function that has a certain fixed number of steps (in this example 15) and then integrates over all possible positions of the break points, and population sizes within each epoch.

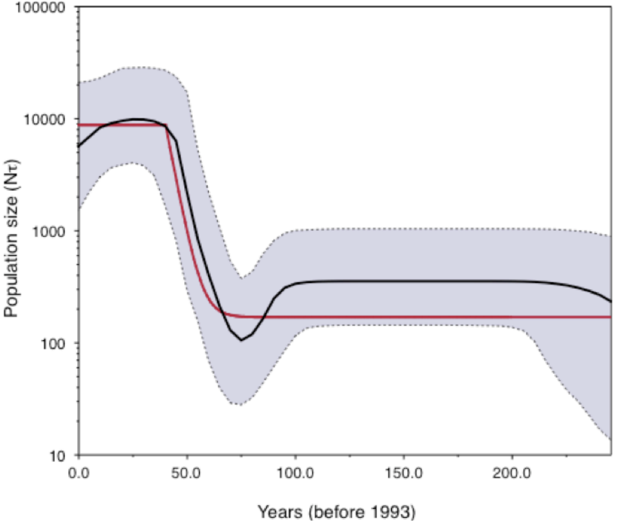


Validating the Bayesian skyline plot



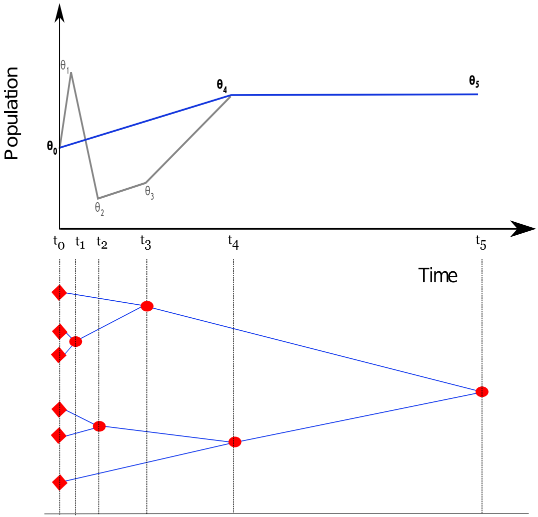
Comparison of BSP to parametric coalescent model

Hepatitis C in Egypt

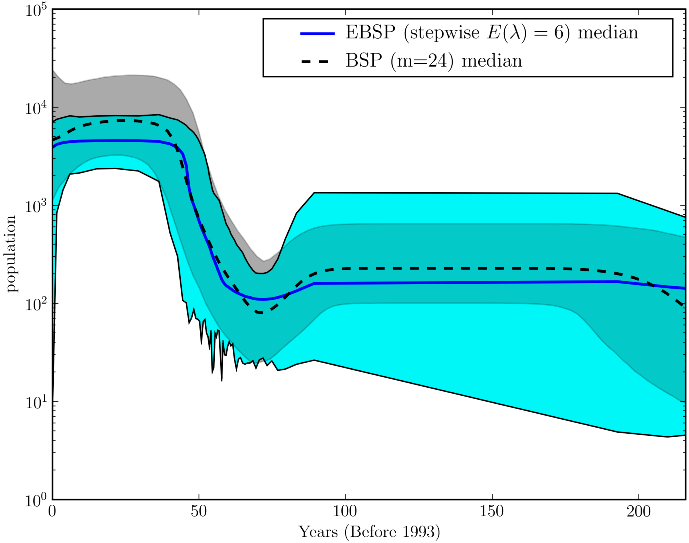


Extending the BSP with Stochastic Variable Selection

Heled and Drummond (2008), *Molecular Biology and Evolution*

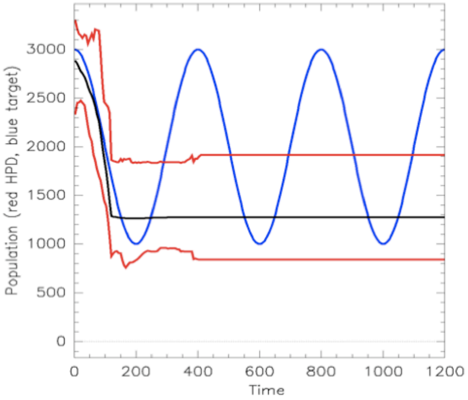


Comparison of EBSP to BSP on Egypt Hepatitis C



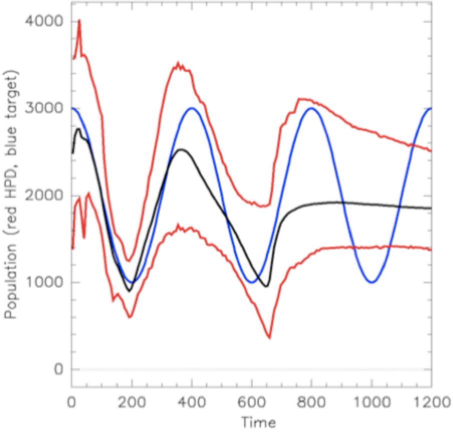
Detecting evolutionary bottlenecks using EBSP

480 contemporaneous samples from a single locus



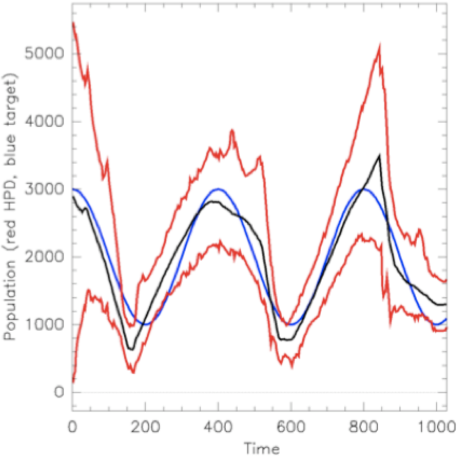
Detecting evolutionary bottlenecks using EBSP

16 contemporaneous samples from each of 32 loci



Detecting evolutionary bottlenecks using EBSP

480 samples sampled through time from a single locus



The population dynamics of genetic diversity in Influenza A

Rambaut *et al* (2008) *Nature* **453**:615-620

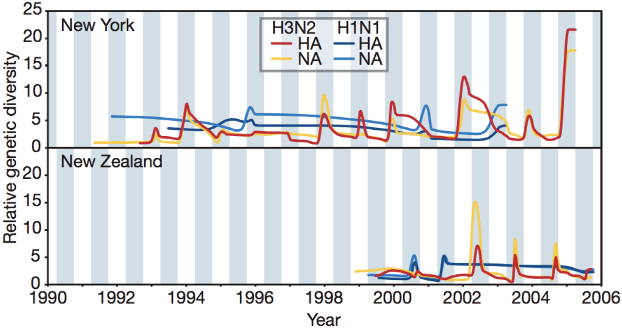
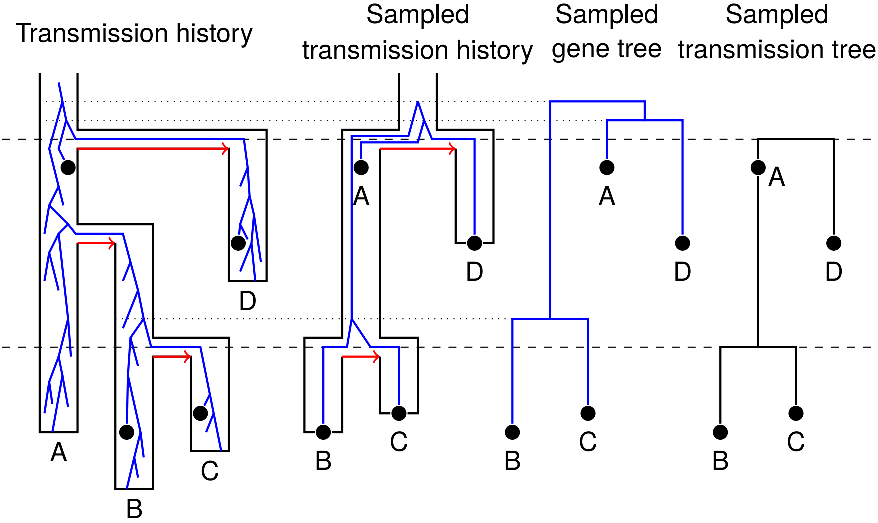


Figure 1 | Population dynamics of genetic diversity in influenza A virus. Bayesian skyline plots of the HA and NA segments for the A/H3N2 and A/H1N1 subtypes in New York state (top) and New Zealand (bottom). The horizontal shaded blocks represent the winter seasons. The *y*-axes represent a measure of relative genetic diversity (see Methods for details). The shorter timescale of New Zealand skyline plot is due to the shorter sampling period.

Birth-death models



Transmission history

Sampled transmission history

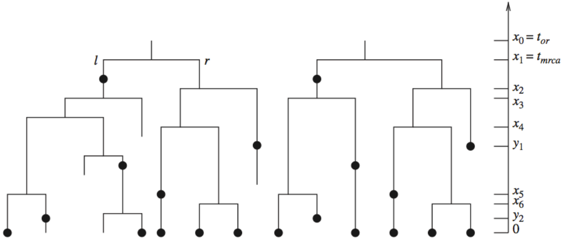
Sampled gene tree

Sampled transmission tree

An oriented transmission tree and the embedded un-oriented viral gene tree.

Birth-death-serial-sampled (BDSS) tree prior

Stadler, 2010



The per-lineage dynamics are captured by a simple set of rate equations:

$$I \xrightarrow{\lambda} 2I \qquad I \xrightarrow{\mu} R \qquad I \xrightarrow{r\psi} U \tag{1}$$

R_0 is the expected number of secondary infections per infected individual:

$$R_0 = \frac{\lambda}{\mu + r\psi} \tag{2}$$

Where r is the probability that sampling removes the lineage from infectious category.

Connecting coalescent growth rates and epidemic models

There is a simple relationship between R_0 and growth rate g at the start of the epidemic:

$$R_0 = 1 + \frac{g}{d} \tag{3}$$

where d is total death rate (Wallinga & Lipsitch, 2007). Taking:

$$d = \mu + r\psi \tag{4}$$

$$g = \lambda - d \tag{5}$$

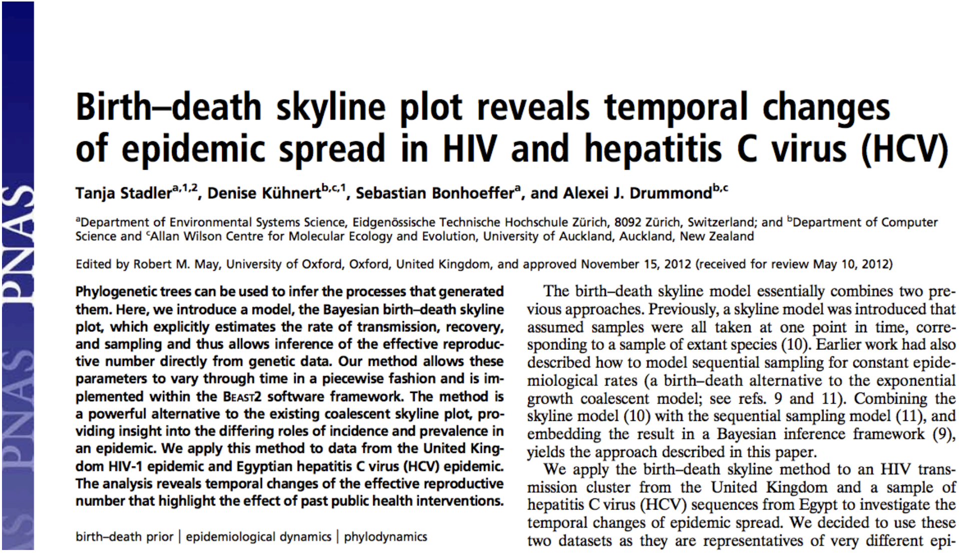
it is easy to show this R_0 is the same as for BDSS model, so coalescent-estimated g is also an estimate of $\lambda - \mu - r\psi$.

Can we still estimate g accurately with exponential coalescent?

Estimating growth rate based on coalescent approach

Table : The measure of accuracy of estimating growth rate g in exponential growth tree prior, where true value $g = \lambda - \mu - r\psi = 4.24 \times 10^{-4}$

BDSS	1 tree	2 trees	5 trees
mean of median	0.0004488872	0.0004460723	0.0004396722
relative error	0.1705658	0.1316335	0.07757073
relative bias	0.05869633	0.05205729	0.03696277
HPD interval width	0.0003617696	0.0002531587	0.0001581470
95% HPD accuracy	95%	96%	93%
Coalescent	1 tree	2 trees	5 trees
mean of median	0.0004845768	0.0004319822	0.0004147897
relative error	0.2701248	0.1972525	0.1244552
relative bias	0.1428698	0.01882604	-0.02172247
HPD interval width	0.0001942935	0.0001265572	7.699674×10^{-5}
95% HPD accuracy	48%	46%	46%



Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)

Tanja Stadler[a,1,2], Denise Kühnert[b,c,1], Sebastian Bonhoeffer[a], and Alexei J. Drummond[b,c]

[a]Department of Environmental Systems Science, Eidgenössische Technische Hochschule Zürich, 8092 Zürich, Switzerland; and [b]Department of Computer Science and [c]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

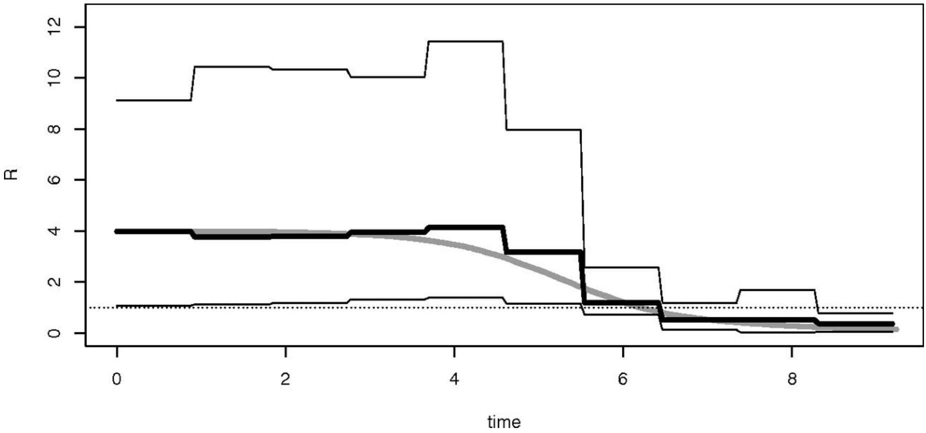
Edited by Robert M. May, University of Oxford, Oxford, United Kingdom, and approved November 15, 2012 (received for review May 10, 2012)

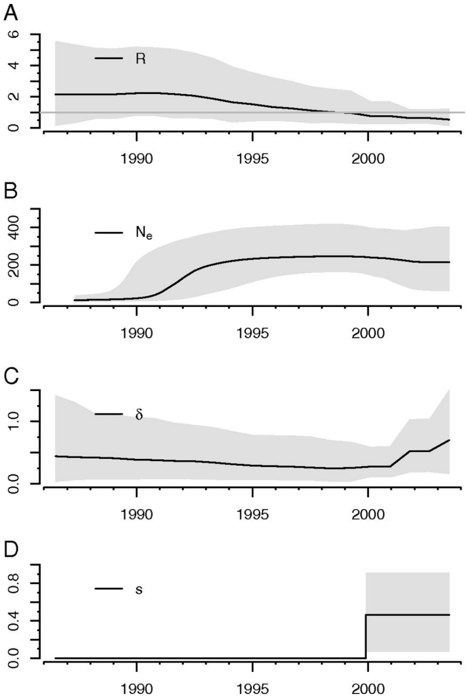
Phylogenetic trees can be used to infer the processes that generated them. Here, we introduce a model, the Bayesian birth–death skyline plot, which explicitly estimates the rate of transmission, recovery, and sampling and thus allows inference of the effective reproductive number directly from genetic data. Our method allows these parameters to vary through time in a piecewise fashion and is implemented within the BEAST2 software framework. The method is a powerful alternative to the existing coalescent skyline plot, providing insight into the differing roles of incidence and prevalence in an epidemic. We apply this method to data from the United Kingdom HIV-1 epidemic and Egyptian hepatitis C virus (HCV) epidemic. The analysis reveals temporal changes of the effective reproductive number that highlight the effect of past public health interventions.

birth–death prior | epidemiological dynamics | phylodynamics

The birth–death skyline model essentially combines two previous approaches. Previously, a skyline model was introduced that assumed samples were all taken at one point in time, corresponding to a sample of extant species (10). Earlier work had also described how to model sequential sampling for constant epidemiological rates (a birth–death alternative to the exponential growth coalescent model; see refs. 9 and 11). Combining the skyline model (10) with the sequential sampling model (11), and embedding the result in a Bayesian inference framework (9), yields the approach described in this paper.

We apply the birth–death skyline method to an HIV transmission cluster from the United Kingdom and a sample of hepatitis C virus (HCV) sequences from Egypt to investigate the temporal changes of epidemic spread. We decided to use these two datasets as they are representatives of very different epi-


R
0
2
4
6
8
10
12
time
0
2
4
6
8


A
R
0
2
4
6
1990
1995
2000
B
Ne
0
200
400
1990
1995
2000
C
δ
0.0
1.0
1990
1995
2000
D
s
0.0
0.4
0.8
1990
1995
2000

A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics

Timothy G. Vaughan*[1,2] and Alexei J. Drummond[1,3]

[1]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

[2]Institute of Veterinary Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

[3]Department of Computer Science, University of Auckland, Auckland, New Zealand

***Corresponding author:** E-mail: t.g.vaughan@massey.ac.nz.

Associate editor: Asger Hobolth

Abstract

In this article, we present a versatile new software tool for the simulation and analysis of stochastic models of population phylodynamics and chemical kinetics. Models are specified via an expressive and human-readable XML format and can be used as the basis for generating either single population histories or large ensembles of such histories. Importantly, phylogenetic trees or networks can be generated alongside the histories they correspond to, enabling investigations into the interplay between genealogies and population dynamics. Summary statistics such as means and variances can be recorded in place of the full ensemble, allowing for a reduction in the amount of memory used—an important consideration for models including large numbers of individual subpopulations or demes. In the case of population size histories, the resulting simulation output is written to disk in the flexible JSON format, which is easily read into numerical analysis environments such as R for visualization or further processing. Simulated phylogenetic trees can be recorded using the standard Newick or NEXUS formats, with extensions to these formats used for non-tree-like inheritance relationships.

Key words: **stochastic simulation, population genetics, phylogenetic trees, chemical kinetics simulation, epidemic modeling.**

```
<beast version='2.0' namespace='beast.core.parameter:master.beast'>

 <run spec='Trajectory' simulationTime='50'>

  <model spec='Model'>

   <!-- Compartment populations in model-->
   <population spec='Population' populationName='S' id='S'/>
   <population spec='Population' populationName='I' id='I'/>
   <population spec='Population' populationName='R' id='R'/>

   <!-- Reactions giving rise to stochastic dynamics -->
   <reaction spec='Reaction' reactionName='Infection' rate='0.001'>
     S + I -> 2I
   </reaction>
   <reaction spec='Reaction' reactionName='Recovery' rate='0.2'>
     I -> R
   </reaction>

  </model>

  <!-- Initial compartment occupancies -->
  <initialState spec='InitState'>
   <populationSize spec='PopulationSize' population='@S' size='999'/>
   <populationSize spec='PopulationSize' population='@I' size='1'/>
  </initialState>

  <!-- Output file specification -->
  <output spec='JsonOutput' fileName='SIR_output.json'/>

 </run>
</beast>
```

FIG. 1. MASTER input file specifying a single fixed time length simulation of a stochastic SIR model.

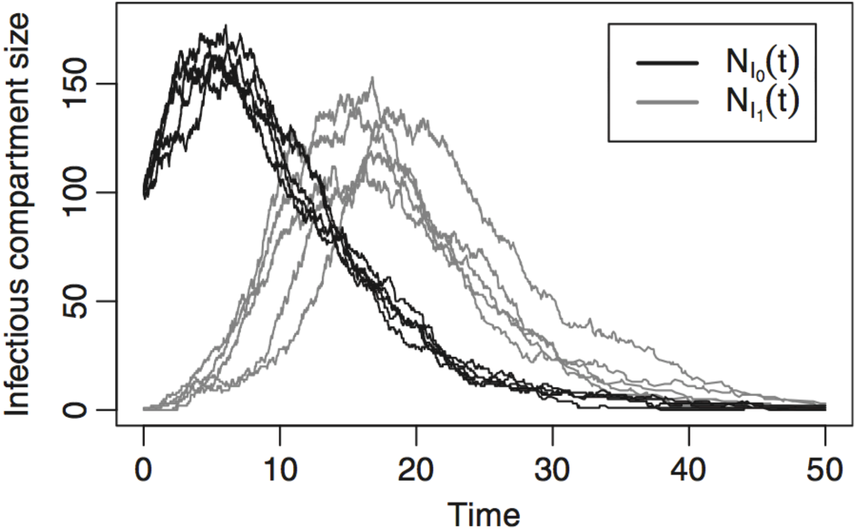


FIG. 3. Histories generated using the two-deme structured SIR model. Note the clear delay between peak infection in deme 0 and peak infection in deme 1.

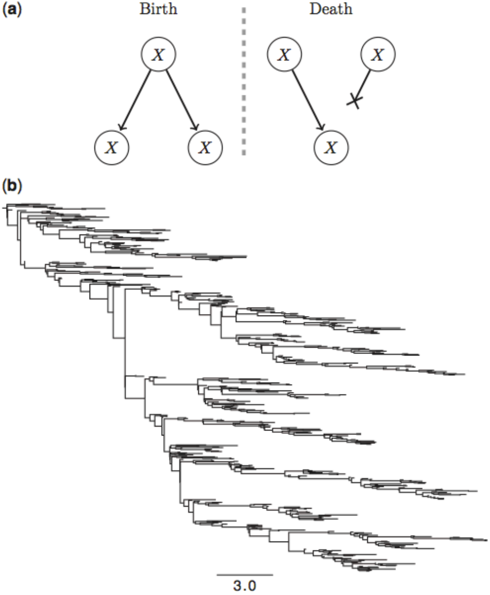
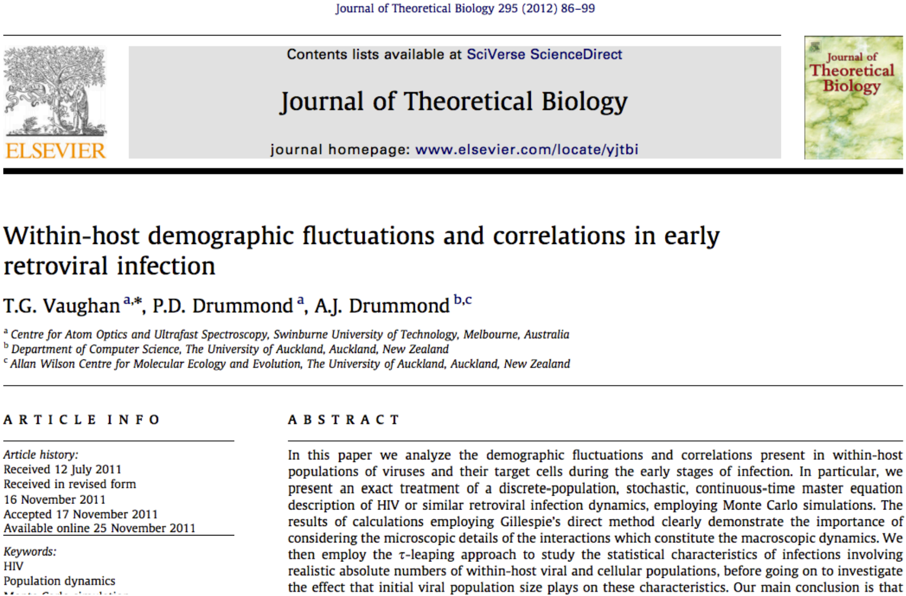


Fig. 5. A stochastic logistic model with inheritance tracking. (*a*) Inheritance relationships between reactants (top) and products (bottom). (*b*) A typical tree produced by MASTER.



Journal of Theoretical Biology 295 (2012) 86–99

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Within-host demographic fluctuations and correlations in early retroviral infection

T.G. Vaughan [a,*], P.D. Drummond [a], A.J. Drummond [b,c]

[a] Centre for Atom Optics and Ultrafast Spectroscopy, Swinburne University of Technology, Melbourne, Australia
[b] Department of Computer Science, The University of Auckland, Auckland, New Zealand
[c] Allan Wilson Centre for Molecular Ecology and Evolution, The University of Auckland, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form
16 November 2011
Accepted 17 November 2011
Available online 25 November 2011

Keywords:
HIV
Population dynamics
Monte Carlo simulation

ABSTRACT

In this paper we analyze the demographic fluctuations and correlations present in within-host populations of viruses and their target cells during the early stages of infection. In particular, we present an exact treatment of a discrete-population, stochastic, continuous-time master equation description of HIV or similar retroviral infection dynamics, employing Monte Carlo simulations. The results of calculations employing Gillespie's direct method clearly demonstrate the importance of considering the microscopic details of the interactions which constitute the macroscopic dynamics. We then employ the τ-leaping approach to study the statistical characteristics of infections involving realistic absolute numbers of within-host viral and cellular populations, before going on to investigate the effect that initial viral population size plays on these characteristics. Our main conclusion is that

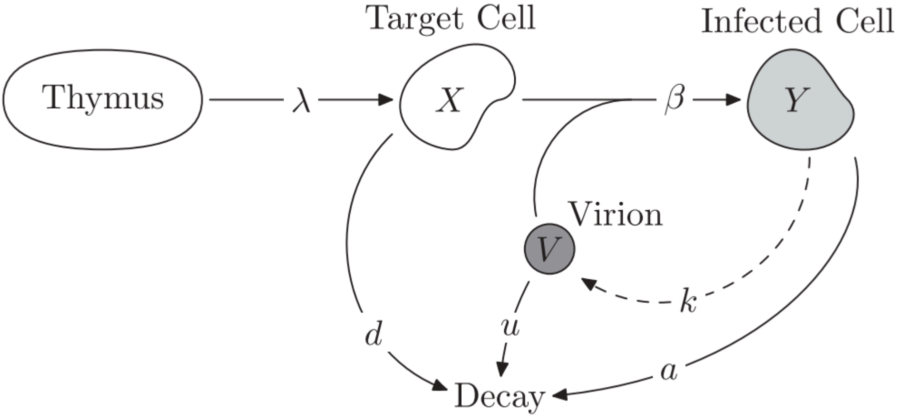


Fig. 1. Schematic of the model used in this paper, detailing the microscopic processes involving target cells (*X*), infected cells (*Y*) and virions (*V*). Each arrow represents a single process occurring at the rate given by its label, with its tail(s) indicating the one or more bodies which instigate the process and the head indicating the product. The dashed line indicates that infected cells are not consumed in the production of virions.

$$0 \xrightarrow{\lambda} X$$

$$X + V \xrightarrow{\beta} Y$$

$$Y \xrightarrow{k} Y + V$$

$$X \xrightarrow{d} 0$$

$$Y \xrightarrow{a} 0$$

$$V \xrightarrow{u} 0$$

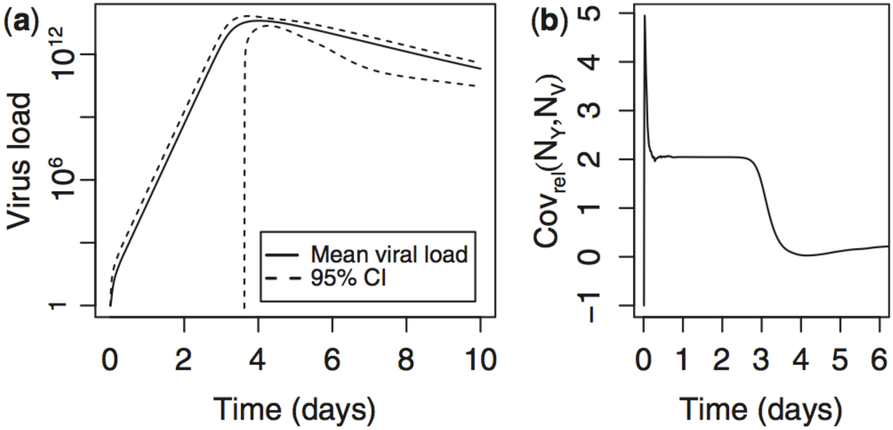
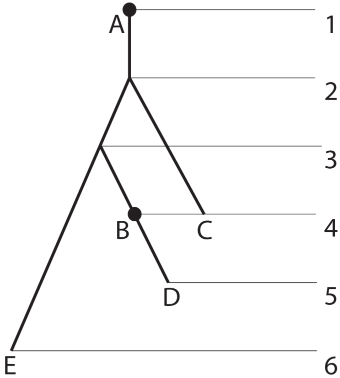


FIG. 4. Using MASTER to perform within-host infection dynamics simulations. (*a*) Expected viral load conditional on chronic infection. (*b*) Relative covariance between infected cell and virion within-host populations.

Sampling ancestors

Fully ranked tree with sampled internal nodes



How many trees with sampled internal nodes are there?

We can recursively count these trees using equations:

$$S(n) = R(n) = \frac{n!(n-1)!}{2^{n-1}}$$

$$S(n_1, \dots, n_m) = \sum_{i=1}^{n_m} \sum_{j=0}^{min\{i, n_{m-1}\}} \binom{i}{j} \binom{n_{m-1}}{j}$$

$$\times \frac{R(n_m)}{R(i)} S(n_1, \dots, n_{m-1} + i - j)$$

Time complexity is $O(mn^2)$, where n is the number of sampled individuals and m is the number of sampling times.

Birth-death-sampling-through-time model with sampled ancestors

- birth rate λ
- death rate μ
- sampling rate ψ
- become noninfectious probability *r*

This model produces only trees in which each sampled node has distinct rank.

Bayesian MCMC analysis with BEAST 2

Since this model produces trees which are not necessarily bifurcating we need to extend Bayesian MCMC methods and adapt BEAST 2 for dealing with a new type of tree.

- Prior distribution
- Proposal mechanism
- Likelihood (peeling algorithm)

Prior distribution

Stadler at el.:

$$f[\mathcal{T}|\lambda, \mu, \psi, r, t_{or} = x_0] = \lambda^{m-1}(\psi(1-r))^k$$

$$\times \prod_{i=0}^{m-1} \frac{1}{q(x_i)} \prod_{i=1}^{m} \psi(r + (1-r)p_0(y_i))q(y_i)$$

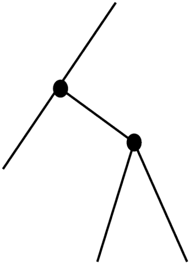
Proposal mechanism

An extension of Wilson-Balding operator

- Choose an edge e_i that terminates at node i.
- Choose an edge e_j such that at least one end of e_j is above i or a leaf j which is above i excluding the edges adjacent to e_i.
- Prune the edge e_i together with the descendant subtree and attach it to the edge e_j or to the leaf j.

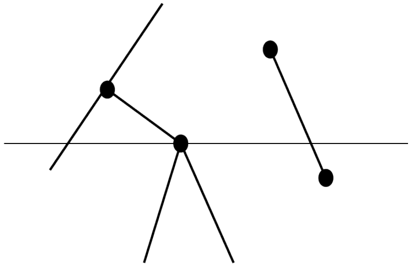
Proposal mechanism

Wilson-Balding operator



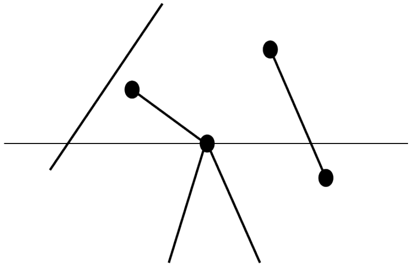
Proposal mechanism

Wilson-Balding operator



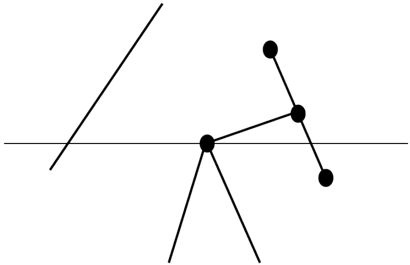
Proposal mechanism

Wilson-Balding operator



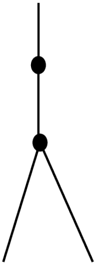
Proposal mechanism

Wilson-Balding operator



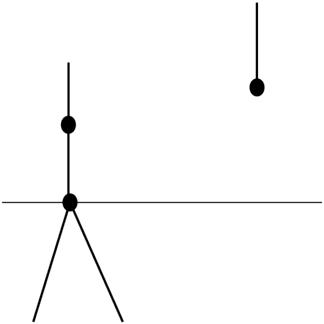
Proposal mechanism

Extension of Wilson-Balding operator



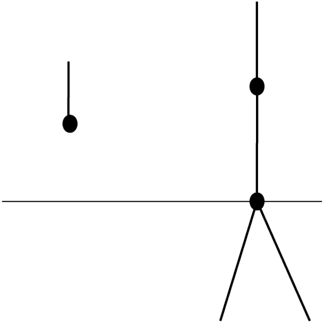
Proposal mechanism

Extension of Wilson-Balding operator



Proposal mechanism

Extension of Wilson-Balding operator



Proposal mechanism

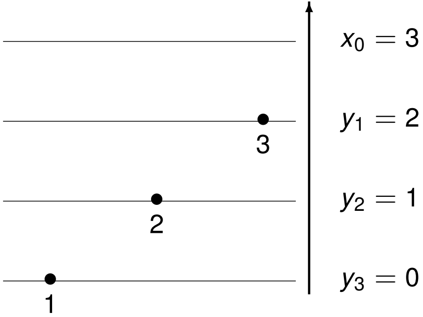
Every tree is reachable with finite number of moves.

Hastings ratio is as follows:

attaching to	internal	leaf	root
removing from	branch		branch
internal branch	$\frac{\lvert l_j \rvert}{\lvert l_i \rvert}$	$\frac{D}{(D-1)} \frac{1}{\lvert l_i \rvert}$	$\frac{e^{\lvert x_j \rvert}}{\lvert l_i \rvert}$
node	$\frac{D}{(D+1)} \lvert l_j \rvert$	1	$\frac{D}{(D+1)} e^{\lvert x_j \rvert}$
root branch	$\frac{\lvert l_j \rvert}{e^{\lvert x_i \rvert}}$	$\frac{D}{(D-1)} \frac{1}{e^{\lvert x_i \rvert}}$	-

Sampling from prior BEAST 2

$\lambda = 2$, $\mu = 1$, $\psi = 1$, and $r = 0.5$.



Sampling from prior BEAST 2

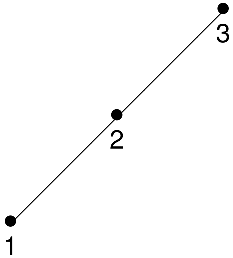
Chain length of 100000 and log every 100.
Thus, 1000 trees were sampled. ESS is 1000.

Count	Percent	Topology
263	26.30	((1)2)3
242	24.20	((1,2))3
238	23.80	((1,2),3)
193	19.30	((1)2,3)
26	2.60	(1,(2)3)
20	2.00	((1)3,2)
14	1.40	(1,(2,3))
4	0.40	((1,3),2)

74.4% of trees have sampled internal nodes.

Sampling from prior BEAST 2

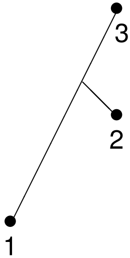
26.3%



((1)2)3

Sampling from prior BEAST 2

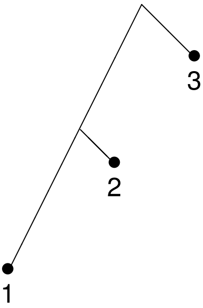
24.2%



((1,2))3

Sampling from prior BEAST 2

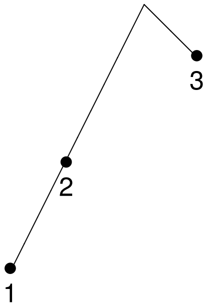
23.8%



((1,2),3)

Sampling from prior BEAST 2

19.3%



((1)2,3)

Sampling from prior BEAST 2

$A[B_1, \ldots, B_k]$ is a subtree with the root at sampled node A and $B_1, \ldots, B_k$ are all the sampled node under nodes A that occurs in this subtree.

Count	Percent	Clade
1000	100.00	1[]
544	54.40	2[]
505	50.50	3[1, 2]
456	45.60	2[1]
449	44.90	3
26	2.60	3[2]
20	2.00	3[1]

Sampling from prior BEAST 2

$A[B]$ means that sampled node A is an ancestor of sampled node B.

Count	Percent	Pair
531	53.10	3[2]
525	52.50	3[1]
456	45.60	2[1]

Sampling from prior BEAST 2

$r = 0.9$
$ESS = 995.77$

Count	Percent	Topology
708	70.80	((1,2),3)
106	10.60	((1)2,3)
97	9.70	((1,2))3
32	3.20	(1,(2,3))
31	3.10	((1,3),2)
13	1.30	((1)2)3
8	0.80	(1,(2)3)
5	0.50	((1)3,2)

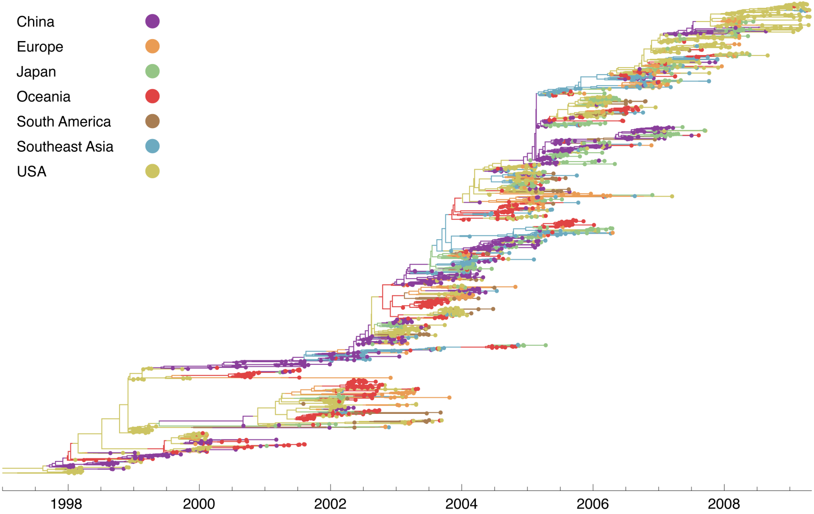
22.9 % of trees have sampled internal nodes.

Further work:

- Likelihood
- More operators
- Using other models, i.e. skyline model

Structured tree models

Structured trees



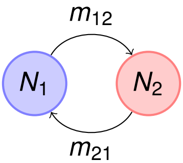
Structured Coalescent

- Accommodates subdivision (demes) in the population
- Initially described by Tajima (1989) and Hudson (1990)
- Implemented in Migrate (Beerli and Felsenstein, 1999; 2001)
  - Estimates subpopulation sizes and migration rates in both ML and Bayesian framework

More recent Extensions

- Serial sampling of data (Ewing *et al.*, 2004)
- Number demes change over time (Ewing and Rodrigo, 2006a)
- Ghost demes - demes that are hidden/not sampled (but you know they are there; Beerli, 2004; Ewing and Rodrigo, 2006b)

Two-deme Wright-Fisher model



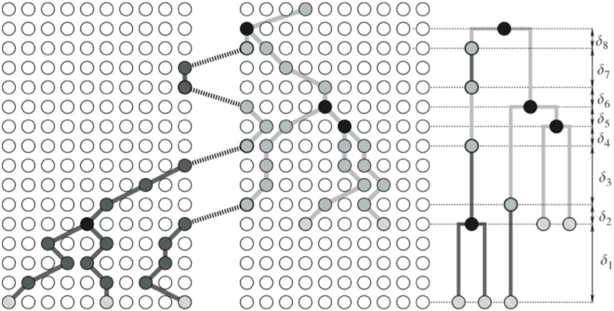
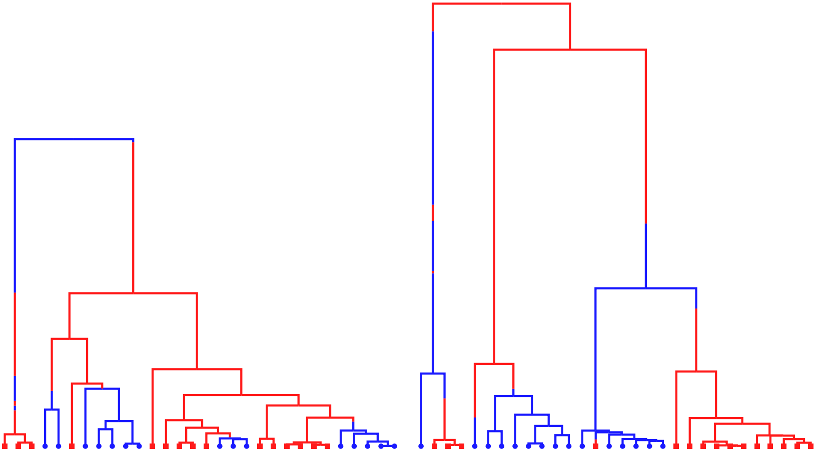


FIG. 2.5. A simplified view of Fisher–Wright subpopulations with migration. Migration events, shown as dashed lines between subpopulations, are explicitly placed on the genealogy (right), as bold circles. The δs signify intervals between migration nodes, coalescent nodes, and leaf nodes.

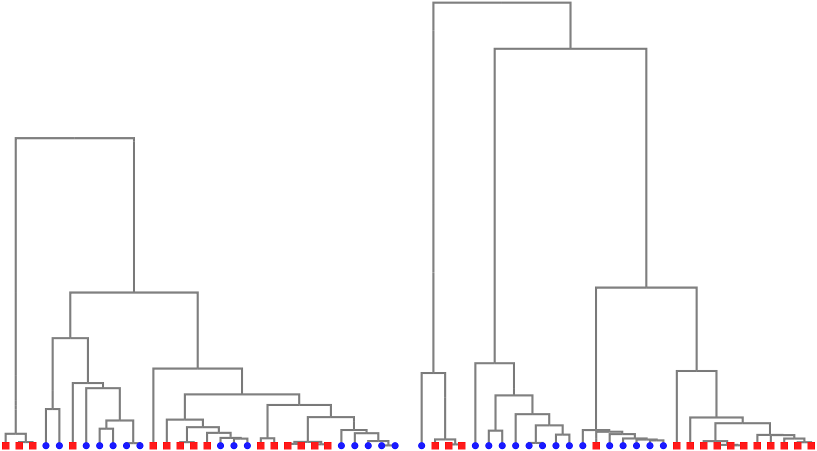
- In general, N_i is the population size of population (deme) i.
- m_{ij} is the probability that an individual in population i was produced from a parent in population j.

Two-deme structured coalescent trees



$N_1 = N_2 = 1000, m_{12} = m_{21} = 0.0008$. There are 15 samples from each deme, all sampled at the same time.

Two-deme structured coalescent trees

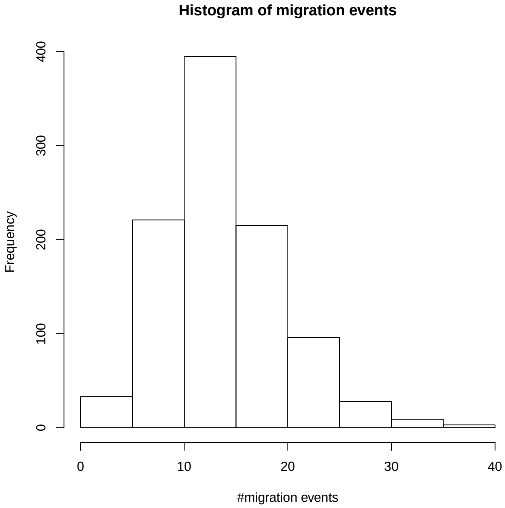


A standard phylogenetic inference method would infer just the tree.
Here we show the true trees, tips annotated with known demes.

Structured coalescent likelihood

The structured coalescent likelihood can be expressed as a product over time intervals from the tips to the root of the ancestral genealogy.

In the standard panmictic coalescent, the number of intervals is known, but in the structured coalescent its an unknown random variable.



Prior distribution of the number of migration events in the two deme, 30 sample example.

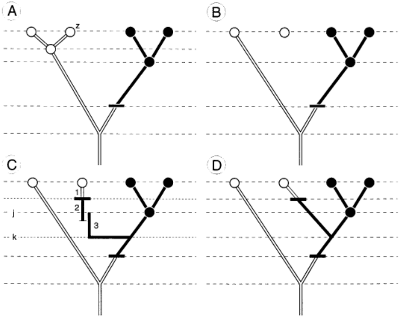
Bayesian MCMC of structured coalescent

In a non-structured Bayesian coalescent analysis, the tree topology and coalescent times are sampled in a Markov chain of correlated states. The size of the discrete structure is fixed to $n - 1$ coalescent events. Operators involve modifying the ancestral relationship tree topology or altering the times of the coalescent events.

For the structured coalescent we have to introduce new “operators” that can add or remove migration events to the ancestral history. When a migration event is added it must be given a time and location on the tree. This increases the state space and thus the computational demand on the inference.

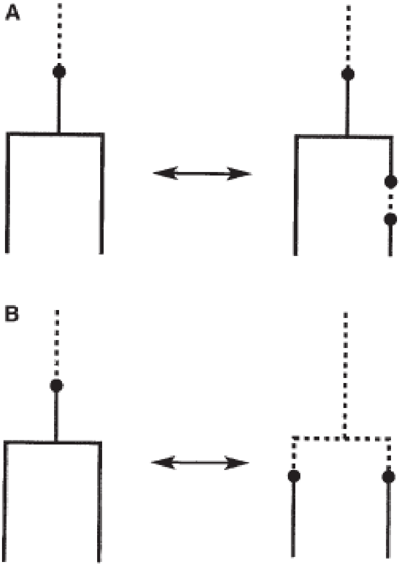
Beerli and Felsenstein (2001) proposal distribution

1. Tree proposals based on “dissolving” part of the tree and then redrawing from the (conditional) prior.
2. Good at sampling from the prior
3. Bad when the sequence data is informative about the tree, because random coalescent subtrees won’t fit the sequence data well.



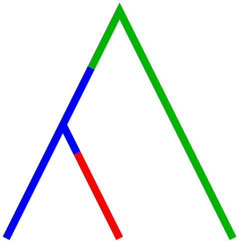
Ewing, Nicholls and Rodrigo (2004) proposal distribution

1. "Standard" tree state proposals, rejecting when inconsistent typed tree generated.
2. Type-specific operators
    - A. Migration-pair birth/death move
    - B. Migration merge-split move
3. Relatively poor mixing.



Operator design strategy

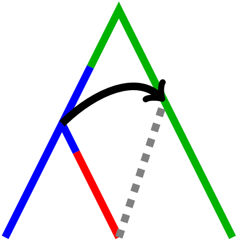
With some exceptions, we take the following general approach to operator development.



Operator design strategy

With some exceptions, we take the following general approach to operator development.

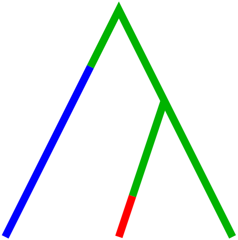
- Apply a standard tree move paying no attention to types.



Operator design strategy

With some exceptions, we take the following general approach to operator development.

- Apply a standard tree move paying no attention to types.
- Type-changes along altered branches are regenerated.
- Regeneration is accomplished by drawing new migration paths from a continuous time Markov process generated by the current rate matrix conditional on types at each end of the branch.



Uniformization method (Fearnhead and Sherlock, 2006)

Method for drawing trajectories from a continuous time Markovian jump process conditioned on the beginning and end states:

$$\frac{\partial}{\partial t} P_i(t) = \sum_j m_{ij} P_j(t) \tag{6}$$

The uniformized process has a state independent intensity $\rho = \max_i(-m_{ii})$ and a discrete-time transition matrix

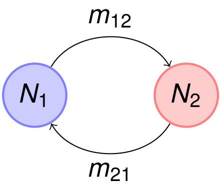
$$U = \frac{1}{\rho} m + I. \tag{7}$$

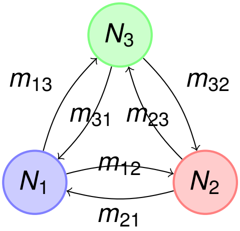
Method

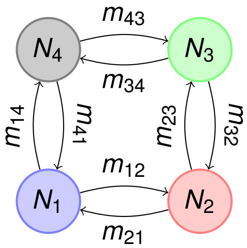
1. Generate event times according to Poisson process with rate ρ.
2. Use standard forward-backward algorithm to determine transitions at these event times conditional on end states.

Comparison with ENR04-style sampler

- For comparison, we have taken the operators used by Ewing, Nicholls and Rodrigo (2004) in their multi-type tree sampler and re-implemented them in BEAST 2.
- The benefit of their operators is that they are computationally *simple* and hence achieve reasonable mixing despite being “small” moves.
- The results were compared on three sets of simulated data. Simulated on 2-demes, 3-demes, 4-demes respectively:



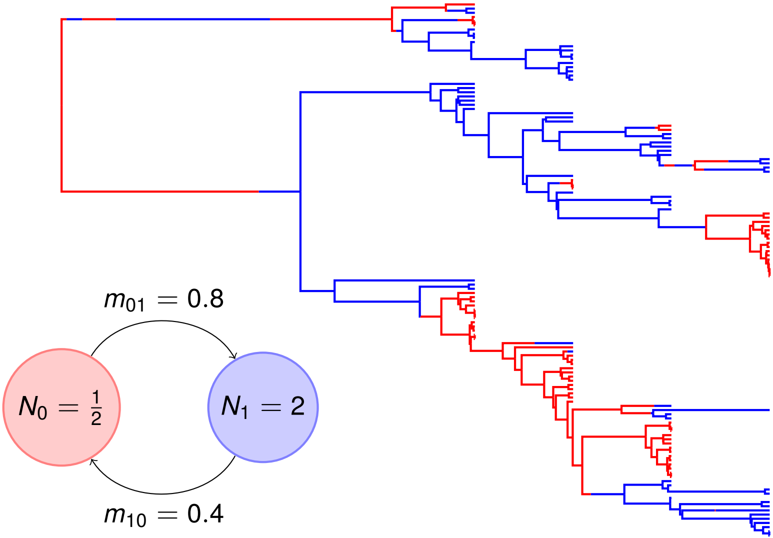




Proposal kernel weights

	Kernel weights	
Operator	ENR04	VD13
Scale(**m**)	1	1
Scale(**N**)	1	1
Scale(μ)	1	1
Scale(κ)	1	1
DeltaExchange(π)	1	1
UpDown(**N**,$\langle\mu,$ **m**$\rangle$)	1	1
MultiTypeUniform	10	10
UpDown($\langle$Tree, **N**$\rangle$, $\langle\mu,$ **m**$\rangle$)	10	10
Scale(Tree)	10	10
TypeSubtreeExchangeEasy	10	-
TypeWilsonBaldingEasy	10	-
TypePairBirthDeath	10	-
TypeMergeSplitExtended	10	-
TypeBirthDeath	10	-
TypeSubtreeExchange	-	10
TypeWilsonBalding	-	10
NodeShiftRetype(root)	-	10
NodeShiftRetype(rest)	-	10

Two-deme: true tree example

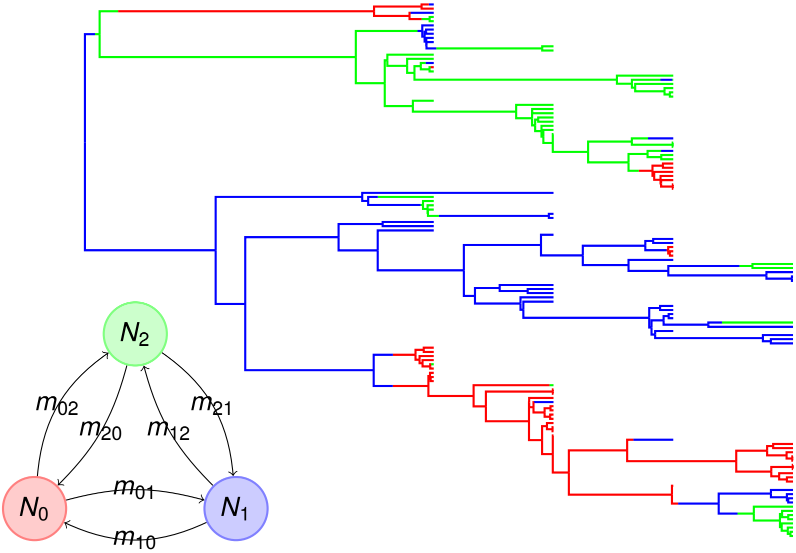


Two-deme performance comparison

	95% HPD coverage		mean ESS		seconds/eff. sample	
Parameter	VD13	ENR04	VD13	ENR04	VD13	ENR04
N_0	0.96	0.96	3337	1517	12	21
N_1	0.96	0.96	4827	1632	9	19
$m_{0,1}$	0.93	0.93	5918	2296	7	14
$m_{1,0}$	0.90	0.90	5927	1945	7	16
μ	0.94	0.94	2112	388	20	82
Tree Height	0.93	0.92	3274	206	13	154
Tree Length	-	-	1319	235	32	135

VD13 is 2 to 12 times faster depending on the summary statistic. Tree length is a good central statistic.

Three-deme: true tree example

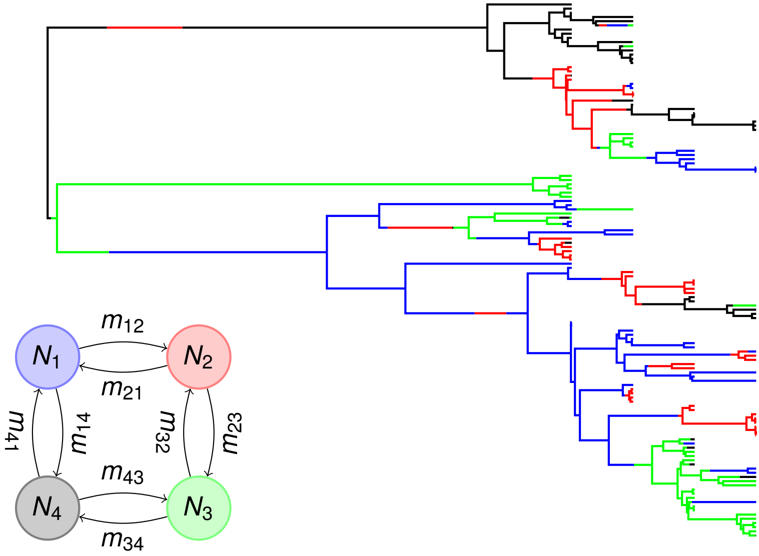


Three-deme performance comparison

	95% HPD coverage		mean ESS		seconds/eff. sample	
Parameter	VD13	ENR04	VD13	ENR04	VD13	ENR04
N_0	0.96	0.95	3119	1620	16	21
N_1	0.98	0.98	4064	1767	12	20
N_2	0.97	0.97	3244	1563	16	22
$m_{0,1}$	0.93	0.93	1499	954	34	37
$m_{1,0}$	0.93	0.93	1279	800	41	44
$m_{0,2}$	0.93	0.93	1477	981	35	36
$m_{2,0}$	0.95	0.95	1385	866	37	40
$m_{1,2}$	0.94	0.92	1205	746	43	47
$m_{2,1}$	0.90	0.93	1489	941	35	37
μ	0.99	0.98	1411	266	37	132
Tree Height	0.98	0.97	1874	125	28	284
Tree Length	-	-	896	159	59	223

VD13 is 1 to 10 times faster depending on the summary statistic. Tree length is a good central statistic.

Four-deme: true tree example



Four-deme performance comparison

	95% HPD coverage		mean ESS		seconds/eff. sample	
Parameter	VD13	ENR04	VD13	ENR04	VD13	ENR04
N_0	0.96	0.96	2560	1578	23	25
N_1	0.94	0.94	2743	1486	21	27
N_2	0.91	0.91	2458	1467	24	27
N_3	0.95	0.95	2699	1569	22	25
$m_{0,3}$	0.80	0.81	935	739	64	55
$m_{0,1}$	0.93	0.93	811	652	74	62
$m_{2,1}$	0.91	0.92	913	693	65	58
$m_{3,0}$	0.95	0.95	898	726	66	56
$m_{1,0}$	0.84	0.85	777	649	77	62
$m_{1,2}$	0.84	0.85	718	542	83	75
$m_{3,2}$	0.88	0.89	901	756	66	53
$m_{2,3}$	0.93	0.93	914	716	65	56
μ	0.94	0.93	956	185	62	220
Tree Height	0.95	0.96	1115	91	53	447

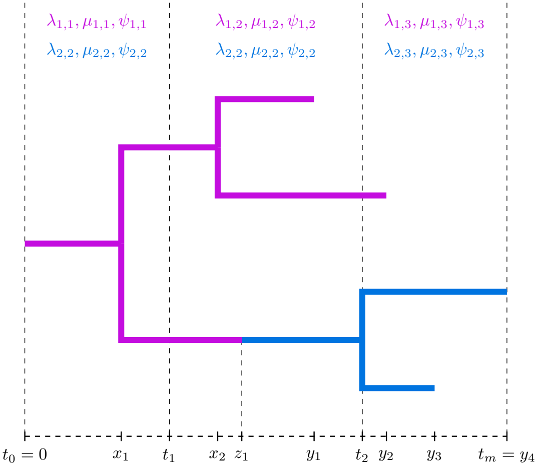
Multi-type birth-death process

Assume that the process is started with one infected individual in deme or of type $i \in \{1 \dots d\}$ at time $t = 0$. With time increasing from the past to the present, in a time step Δt the process can undergo

1. a birth event, so that another infected individual is created in deme i:
$$N_i(t + \Delta t) = N_i(t) + 1,$$
2. a death event, implying the recovery or removal of an infected individual in deme i:
$$N_i(t + \Delta t) = N_i(t) - 1,$$
3. a sampling event, yielding the removal of an infected individual as in 2., but this time the removal is observed, or
4. a migration event, indicating that an individual changes from deme i to deme $j \neq i$:
$$N_i(t + \Delta t) = N_i(t) - 1 \text{ and } N_j(t + \Delta t) = N_j(t) + 1.$$

The process terminates when no infected individuals are left.

Multi-type birth-death process notation


$\lambda_{1,1}, \mu_{1,1}, \psi_{1,1}$
$\lambda_{2,2}, \mu_{2,2}, \psi_{2,2}$
$\lambda_{1,2}, \mu_{1,2}, \psi_{1,2}$
$\lambda_{2,2}, \mu_{2,2}, \psi_{2,2}$
$\lambda_{1,3}, \mu_{1,3}, \psi_{1,3}$
$\lambda_{2,3}, \mu_{2,3}, \psi_{2,3}$
$t_0 = 0$
x_1
t_1
x_2
z_1
y_1
t_2
y_2
y_3
$t_m = y_4$

Notation under the multi-type birth–death model. Birth events are denoted by x_j, sampling events by y_j and the one migration event z_1.

Priors for comparison of BDMM with Structured Coalescent

	Simulations	Multi-type Birth–death	Structured Coalescent
$\mathcal{R}_i$	LogN(0.4,0.6)	LogN(0.5,1)	-
δ	LogN(-1,1)	$\mathcal{N}$(80,20)	-
s	B(1,10)	B(1,100)	-
t_m	-	LogN(2.,1.25)	-
$m_{ij}^{(sc)}$	Exp(0.01)	Exp(0.01)	Exp(0.01)
N_i	-	-	LogN(-2,2)

Table : **Prior distributions** for the simulation study and the phylogeographic analysis of human Influenza H3N2 sequences from Australia and New Zealand. The Beta distribution is denoted by B, the normal distribution by $\mathcal{N}$, $(i, j \in \{1{,}2\})$.

Parameter estimates

	Multi-type Birth–death		Structured Coalescent	
	Median	95% HPD	Median	95% HPD
$\mathcal{R}_1$	1.00	(0.93–1.06)	-	-
$\mathcal{R}_2$	1.01	(0.98–1.05)	-	-
δ_1	71	(26–112)	-	-
δ_2	72	(26–114)	-	-
s_1	0.0035	(0.0001–0.0132)	-	-
s_2	0.0013	(0.0001–0.0066)	-	-
m_{12}	0.024	(0.0002–0.064)	-	-
m_{21}	0.035	(0.0036–0.077)	-	-
Migration events in tree	18	(13–23)	11	(8–13)
Root of the tree (yr)	1997	(1996–1998)	1999	(1997–2000)
Origin of the epidemic (yr)	1978	(1966–1985)	-	-
N_1	-	-	0.85	(0.43–1.37)
N_2	-	-	1.22	(0.77–1.74)
m_{12}^{sc}	-	-	0.024	(9×10^{-7}–0.099)
m_{21}^{sc}	-	-	0.076	(0.004–0.153)

Table : **Phylogeographic analysis of Influenza H3N2.** Posterior median estimates with 95% HPD intervals of Australasian H3 dataset.

Comparison of estimate migration history

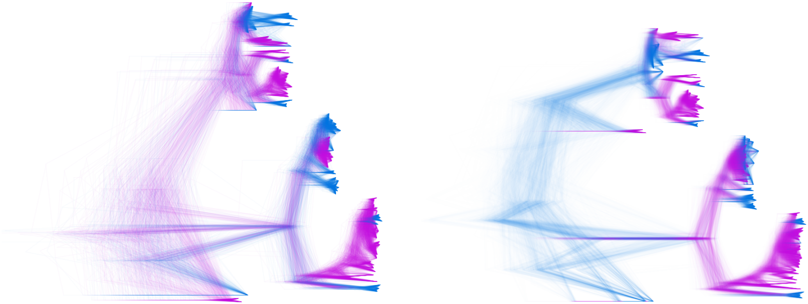


Figure : **H3N2 analysis: Posterior distribution of multi-type phylogenies.** The posterior phylogenies of the multi-type birth–death analysis and Structured coalescent of human influenza virus, with the two sampling locations Australia and New Zealand denoted by blue and purple, respectively, were plotted with the program DensiTree.

Posterior distribution of number of migrations

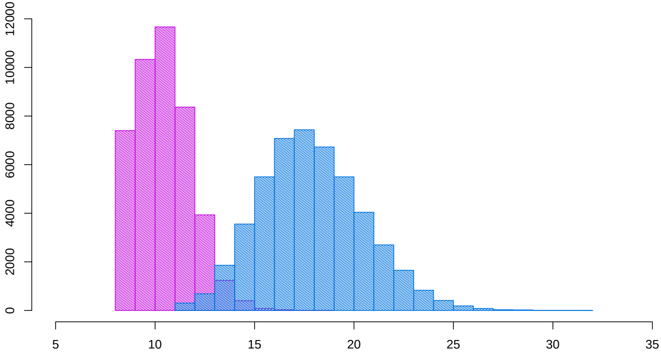


Figure : **H3N2 analysis: Posterior distribution of the number of migrations.** The histograms show the posterior density of the number of migrations in the phylogenetic tree from the analysis under the multi-type birth–death model (blue) and under the structured coalescent (purple).

Integrating population dynamics with population genetics?

Genealogical models

- Focus on genetics of a population, especially neutrality
- Account for stochastic nature of mutation and drift
- Forward simulation and equilibrium solutions
- Powerful inference tools

Population dynamics models

- Focus on coupled interactions between different types, hosts
- Often deterministic, rather than stochastic models
- Forward simulation and equilibrium solutions
- Parameters closely aligned to real measurable quantities

Integrating population dynamics with population genetics?

Genealogical models

- Generally poor at describing dynamics and selection
- Not readily parameterized by “real-life” parameters
- Parameters such as $N_e(t)$ can’t be compared with real measurements in absolute terms

Population dynamics models

- Poor at handling evolution
- Poor at describing genetic variation
- Poor inference tools

Dynamical population genetics

What would a synthesis look like?

- Microscopic descriptions of all processes including
  - Selection, competition
  - Mutation, type switching
  - Birth, death, infection, genetic drift *et cetera*
  - Demographic stochasticity
  - Environmental stochasticity
- Natural modeling of stochastic parts of the process
- Retains non-linear coupling between different types and hosts
- Handles both neutral and selected variation
- Parameters can be readily connected with real measurable quantities
- Simulation, analysis and inference tools