bayesian coalescent analysis (of viruses)

Alexei Drummond

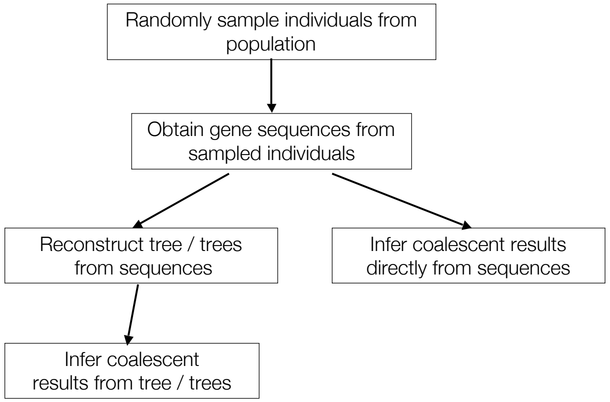
Overview

- Introduction to the Coalescent
- Phylodynamics and the Bayesian skyline plot
- Molecular population genetics for viruses
- Testing model assumptions

The coalescent

- The coalescent is a model of the **ancestral relationships** of a small sample of individuals taken from a large background population.
- The coalescent describes a probability distribution on ancestral genealogies (trees) given a population history.
  - Therefore the coalescent can convert information from ancestral genealogies into information about population history and vice versa.
- The coalescent is a model of ancestral genealogies, not sequences, and its simplest form assumes **neutral evolution**.

Coalescent inference



Randomly sample individuals from population

Obtain gene sequences from sampled individuals

Reconstruct tree / trees from sequences

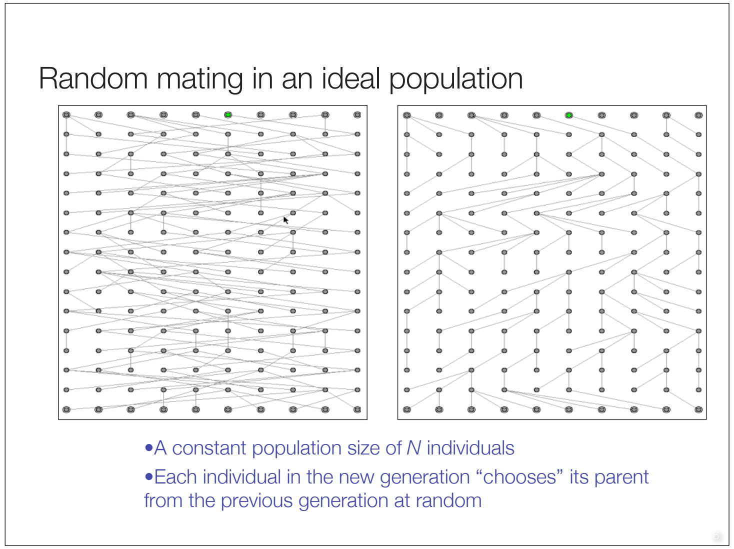
Infer coalescent results directly from sequences

Infer coalescent results from tree / trees

Demographic history

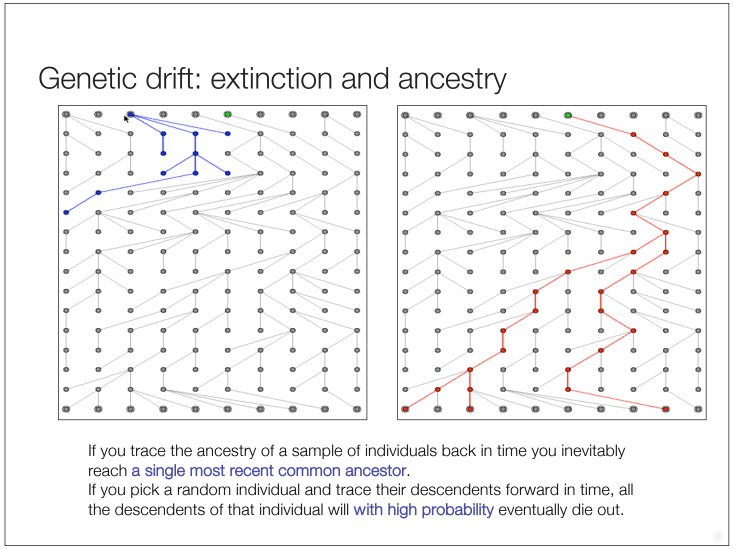
- Change in population size through time

- Applications include
  - Reconstructing infectious disease epidemics
  - Investigating viral dynamics within hosts
  - Identifying bottlenecks



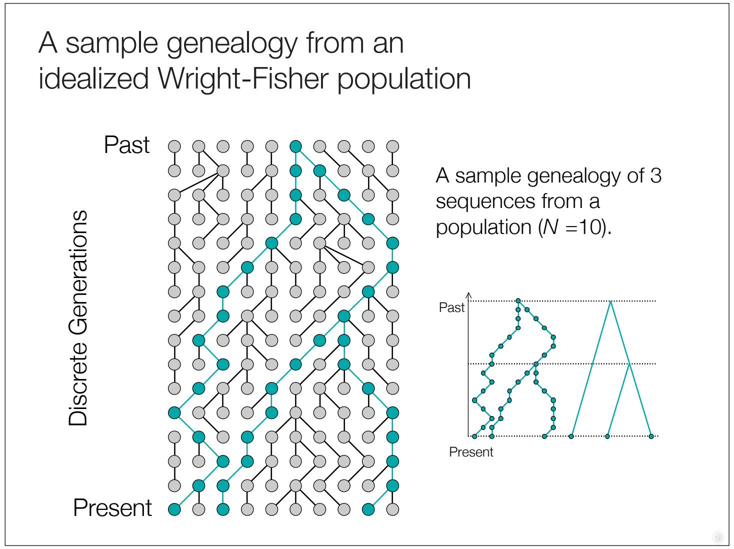
Random mating in an ideal population

- A constant population size of *N* individuals
- Each individual in the new generation “chooses” its parent from the previous generation at random



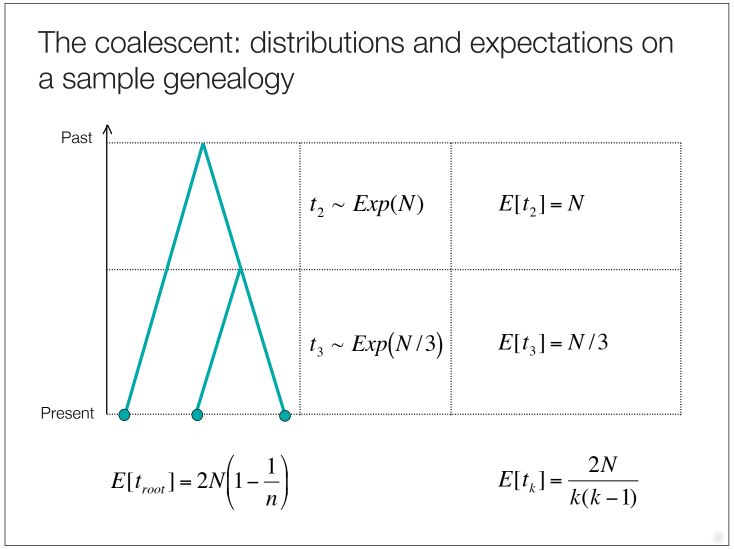
Genetic drift: extinction and ancestry

If you trace the ancestry of a sample of individuals back in time you inevitably reach a single most recent common ancestor.
If you pick a random individual and trace their descendents forward in time, all the descendents of that individual will with high probability eventually die out.



A sample genealogy from an idealized Wright-Fisher population

A sample genealogy of 3 sequences from a population (*N* =10).

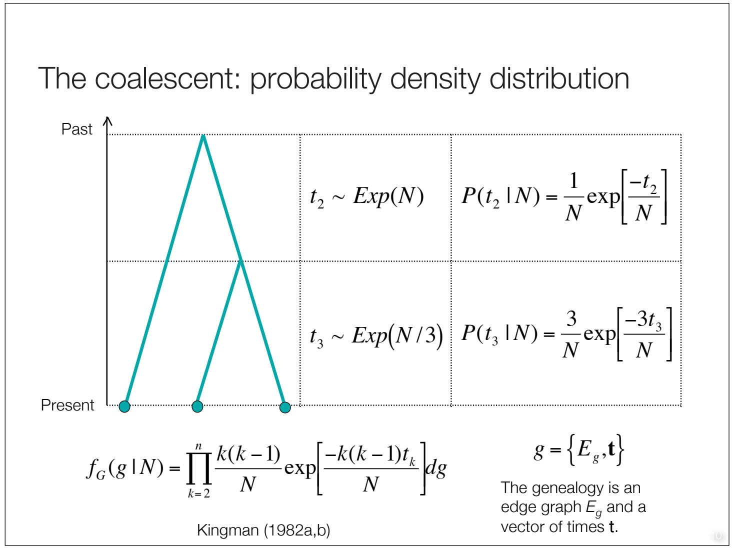


The coalescent: distributions and expectations on a sample genealogy

$t_2 \sim Exp(N)$	$E[t_2] = N$
$t_3 \sim Exp(N/3)$	$E[t_3] = N/3$

$$E[t_{root}] = 2N\left(1 - \frac{1}{n}\right)$$

$$E[t_k] = \frac{2N}{k(k-1)}$$



The coalescent: probability density distribution

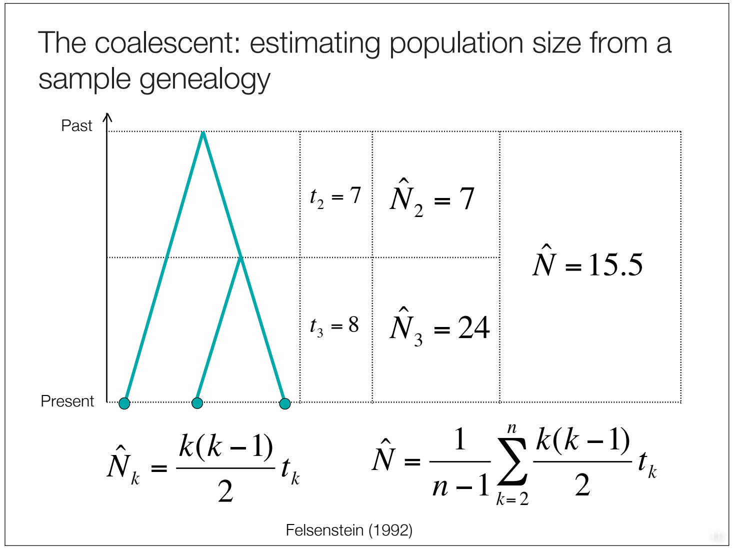
$t_2 \sim Exp(N)$	$P(t_2 \mid N) = \frac{1}{N}\exp\left[\frac{-t_2}{N}\right]$
$t_3 \sim Exp(N/3)$	$P(t_3 \mid N) = \frac{3}{N}\exp\left[\frac{-3t_3}{N}\right]$

$$f_G(g \mid N) = \prod_{k=2}^{n} \frac{k(k-1)}{N}\exp\left[\frac{-k(k-1)t_k}{N}\right]dg$$

Kingman (1982a,b)

$$g = \{E_g, \mathbf{t}\}$$

The genealogy is an edge graph E_g and a vector of times **t**.

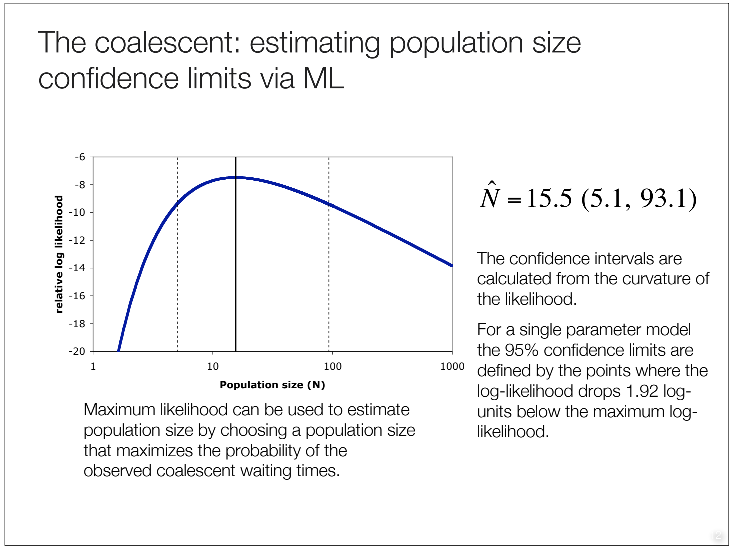


The coalescent: estimating population size from a sample genealogy

$t_2 = 7$	$\hat{N}_2 = 7$	$\hat{N} = 15.5$
$t_3 = 8$	$\hat{N}_3 = 24$	

$$\hat{N}_k = \frac{k(k-1)}{2} t_k \qquad \hat{N} = \frac{1}{n-1}\sum_{k=2}^{n} \frac{k(k-1)}{2} t_k$$

Felsenstein (1992)



The coalescent: estimating population size confidence limits via ML

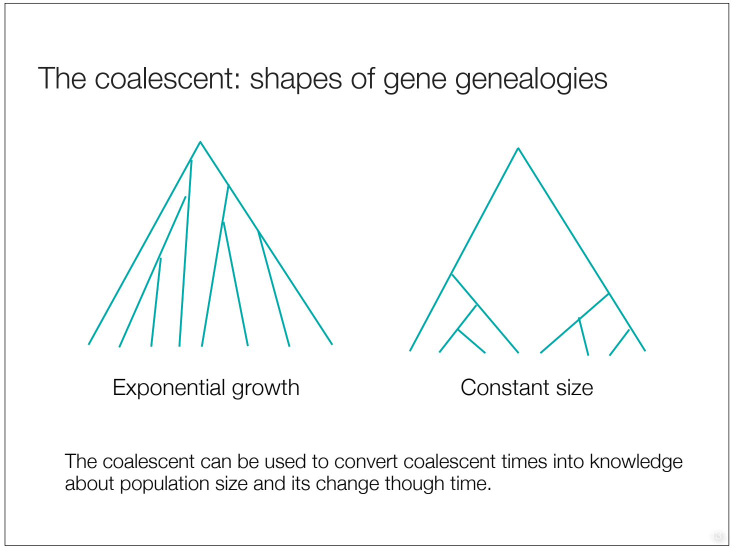
Maximum likelihood can be used to estimate population size by choosing a population size that maximizes the probability of the observed coalescent waiting times.

$$\hat{N} = 15.5\ (5.1,\ 93.1)$$

The confidence intervals are calculated from the curvature of the likelihood.

For a single parameter model the 95% confidence limits are defined by the points where the log-likelihood drops 1.92 log-units below the maximum log-likelihood.

The coalescent: shapes of gene genealogies

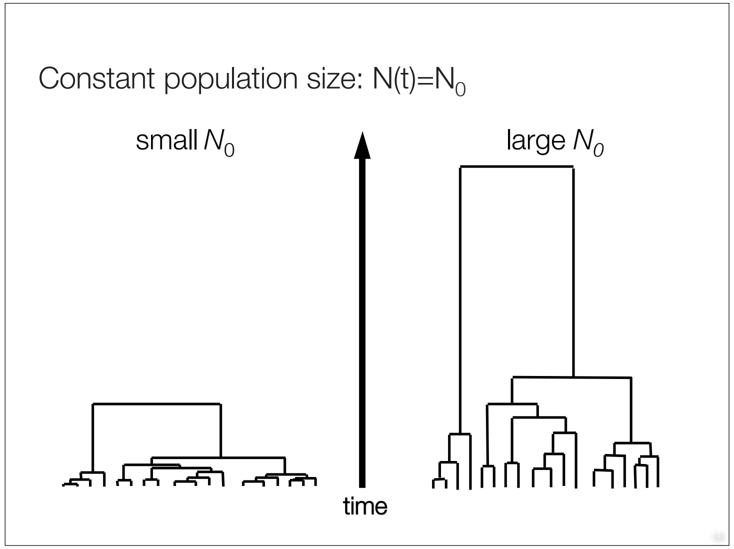


Exponential growth

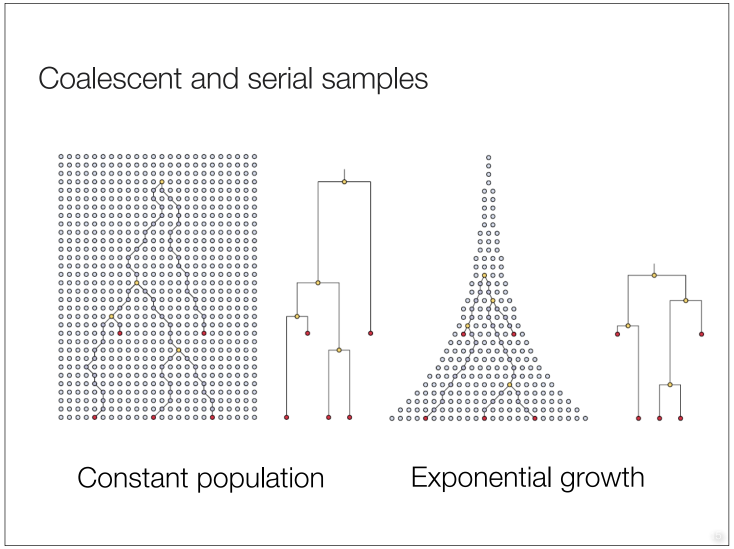
Constant size

The coalescent can be used to convert coalescent times into knowledge about population size and its change though time.

Constant population size: N(t)=N_0



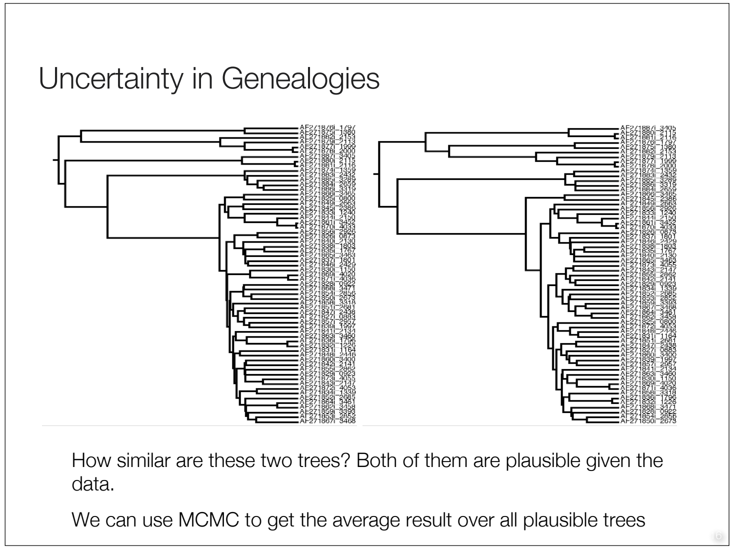
Coalescent and serial samples



Constant population

Exponential growth

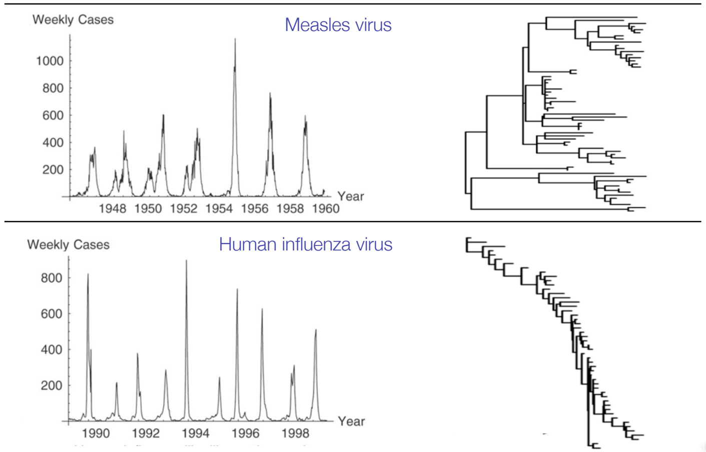
Uncertainty in Genealogies



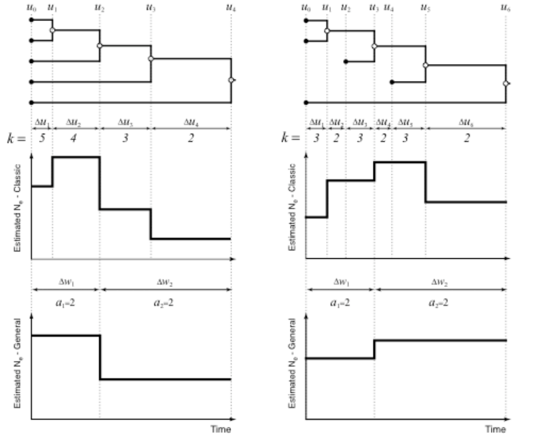
How similar are these two trees? Both of them are plausible given the data.

We can use MCMC to get the average result over all plausible trees

Virus population dynamics



Population size changes


u_0 u_1 u_2 u_3 u_4
Δu_1 Δu_2 Δu_3 Δu_4
$k=$ 5 4 3 2
Estimated N_e - Classic
Δw_1 Δw_2
a_1=2 a_2=2
Estimated N_e - General
Time
u_0 u_1 u_2 u_3 u_4 u_5 u_6
Δu_1 Δu_2 Δu_3 Δu_4 Δu_5 Δu_6
$k=$ 3 2 3 2 3 2
Estimated N_e - Classic
Δw_1 Δw_2
a_1=2 a_2=2
Estimated N_e - General
Time

The generalized skyline plot

- Visual framework for exploring the demographic history of sampled DNA sequences
- Input: a single estimated ancestral genealogy (a tree)
- Output: nonparametric plot of the population size through time

  Groups adjacent coalescent intervals

  Converts information within these intervals to estimates of population size

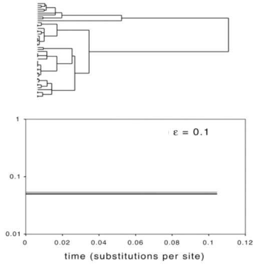
$$\hat{N}_k = \frac{k(k-1)}{2} t_k \qquad \hat{N}_{k,l} = \frac{k(k-l)}{2l} \sum_{i=k-l+1}^{k} t_i$$

Estimate of population size from single coalescent interval

Estimate of population size from *l* adjacent coalescent intervals.

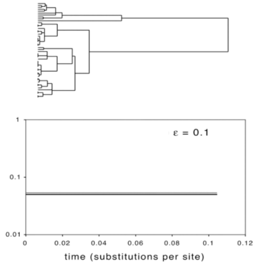
Examples

- I: Constant population size
- N(t)=N(0)

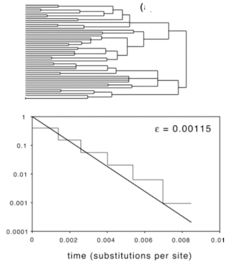


Skyline Plot

- I: Constant population size
- N(t)=N(0)



II: Exponential growth

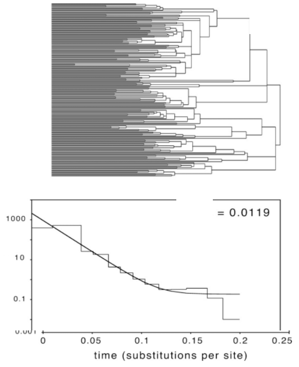


Skyline Plot

- III: HIV-1 group M
- (tree estimated in Yusim *et al* (2001) Phil. Trans. Roy. Soc. Lond. B 356: 855-866)

Black curve is a parametric estimate obtained from the same data under the “expansion model”

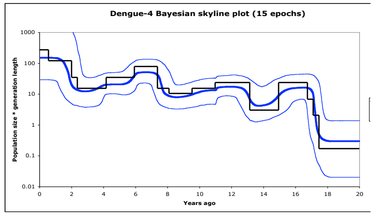
Results follow accepted demographic pattern for the HIV pandemic

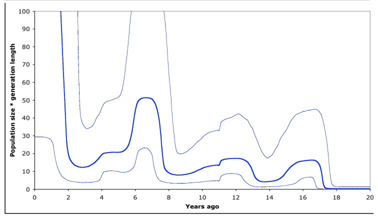


The Bayesian skyline plot

Estimate a demographic function that has a certain fixed number of steps (in this example 15) and then integrate over all possible positions of the break points.

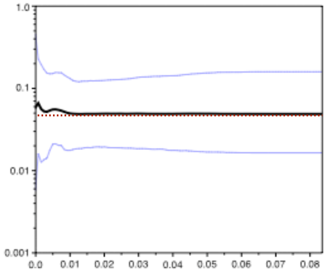
Explains the Dengue data quite well (test of neutrality do not reject the data if we use the Bayesian skyline plot to describe the demographic history.



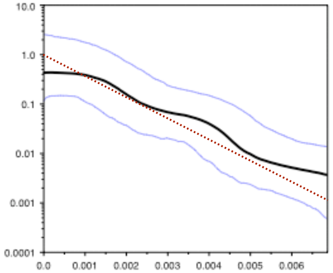


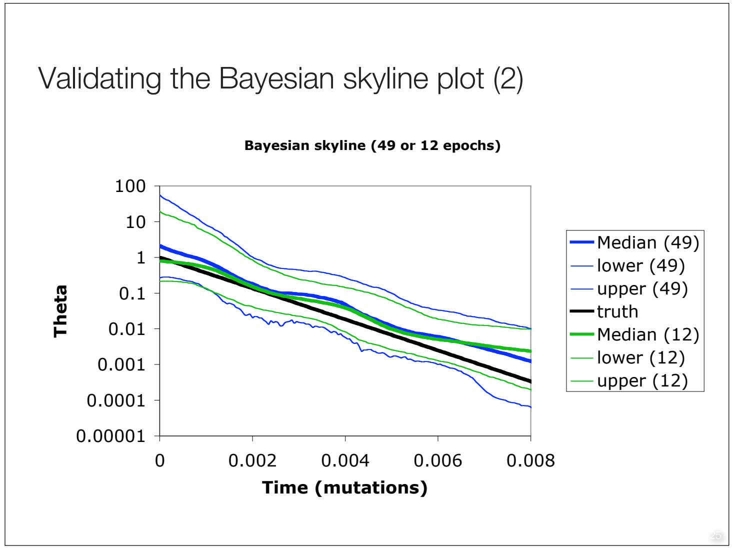
Validating the Bayesian skyline plot (1)

Simulated data: Constant population

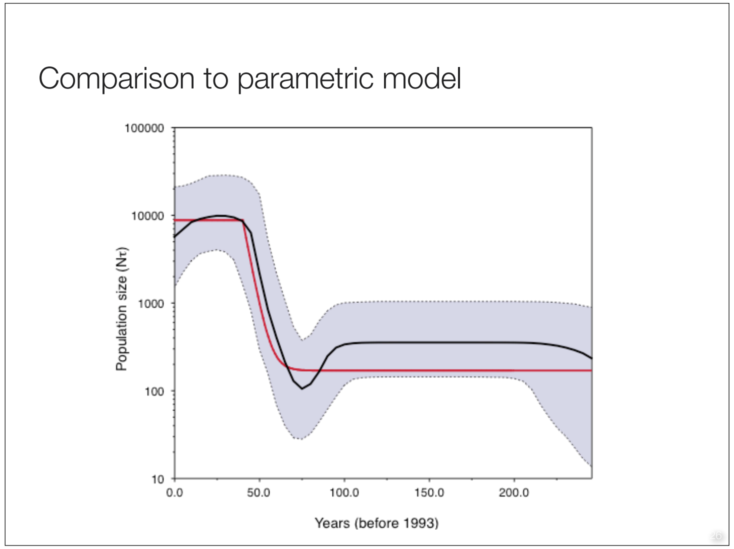


Simulated data: Exponential growth

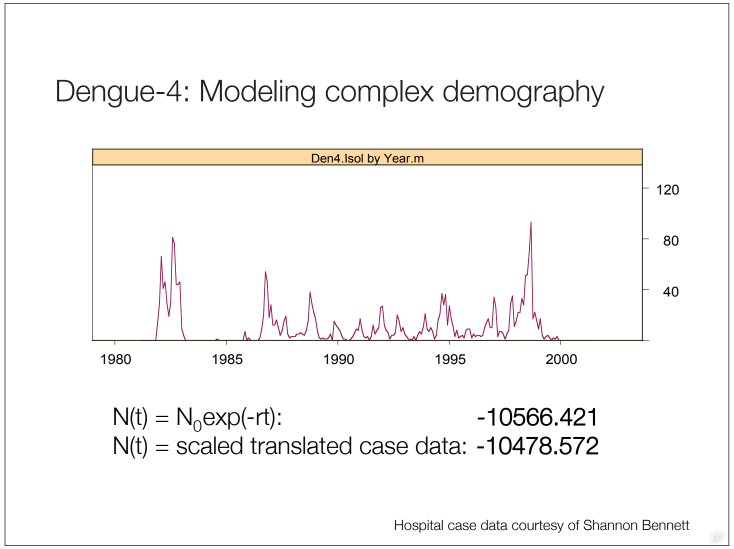




Validating the Bayesian skyline plot (2)



Comparison to parametric model

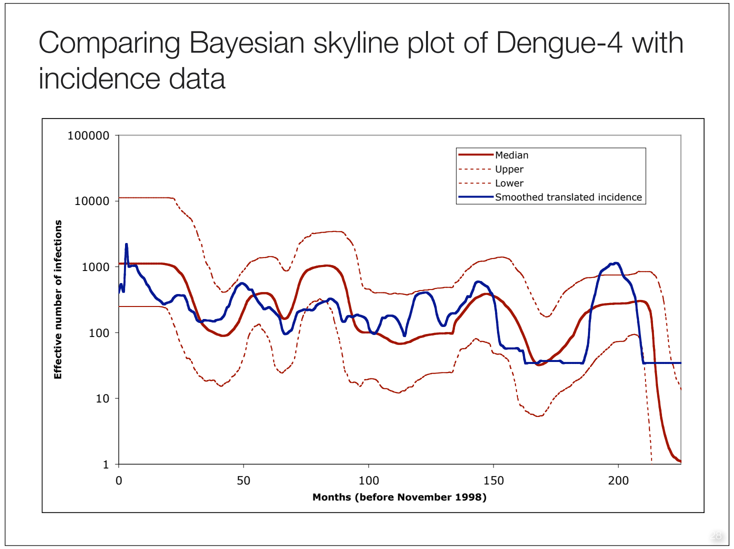


Dengue-4: Modeling complex demography

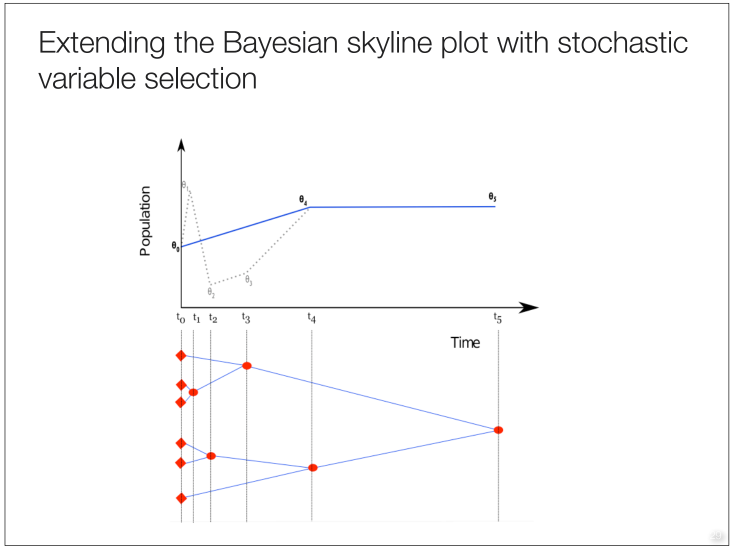
$N(t) = N_0 exp(-rt)$: -10566.421

N(t) = scaled translated case data: -10478.572

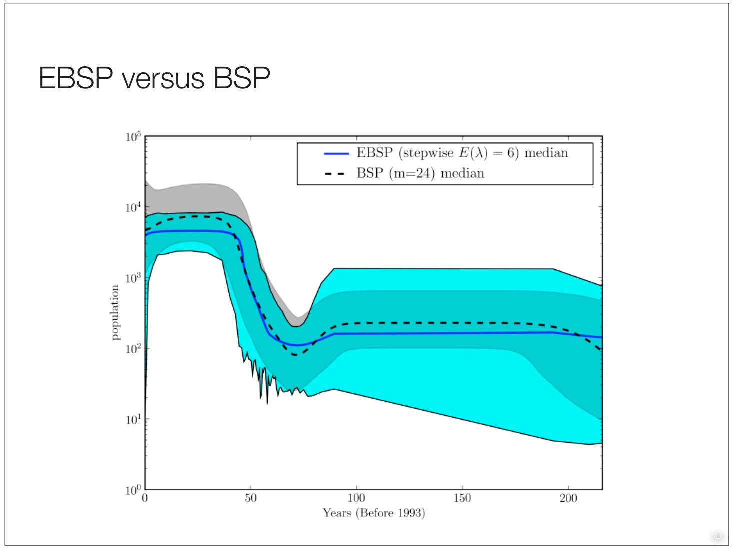
Hospital case data courtesy of Shannon Bennett



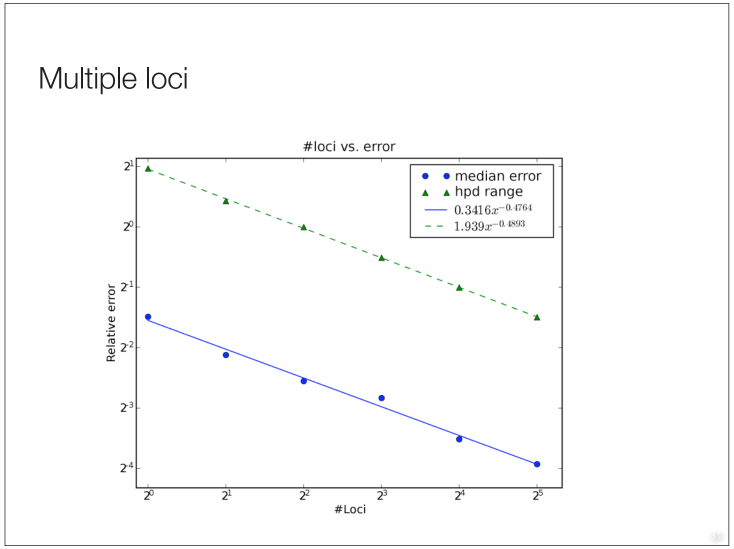
Comparing Bayesian skyline plot of Dengue-4 with incidence data



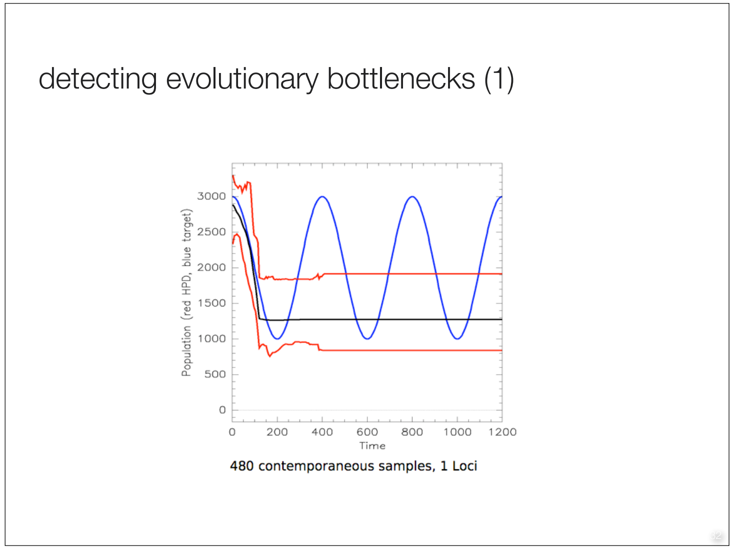
Extending the Bayesian skyline plot with stochastic variable selection



EBSP versus BSP



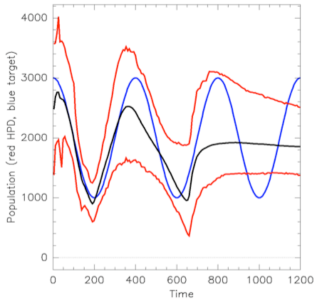
Multiple loci



detecting evolutionary bottlenecks (1)

480 contemporaneous samples, 1 Loci

detecting evolutionary bottlenecks (2)



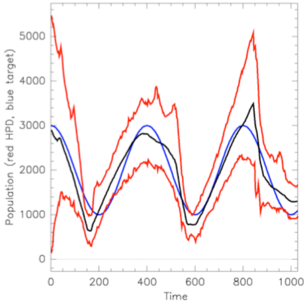
16 samples
32 Loci

Piecewise Linear kernel

Prior on number of groups:
Poisson with mean 8

Generally we don't have multiple independent loci for viruses!

detecting evolutionary bottlenecks (3)



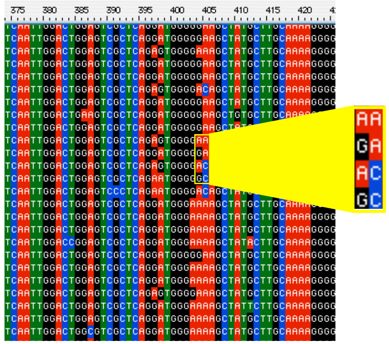
480 serial samples, uniformly distributed from zero to 1000

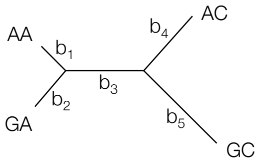
But we do often have serial samples!

Coalescent Summary

- The coalescent provides a theory of how population size is related to the distribution of coalescent events in a tree.
- Big populations have old trees
- Exponentially growing populations have star-like trees
- Given a genealogy the most likely population size (function) can be estimated.
- MCMC can be used to get a distribution of trees from which a distribution of population sizes can be estimated.

Molecular evolutionary model: Felsenstein's likelihood (1981)



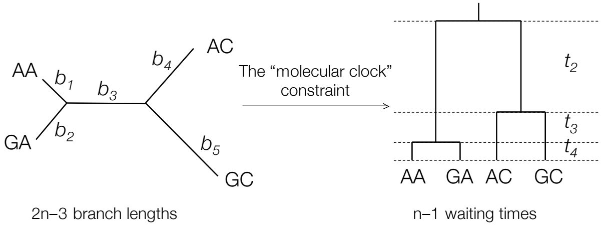


The probability of the sequence alignment,

$$\Pr\{D \mid T, Q\}$$

can be efficiently calculated given a tree and branch lengths (*T*), and a probabilistic model of mutation represented by an instantaneous rate matrix (*Q*). In phylogenetics, branch lengths are usually unconstrained.

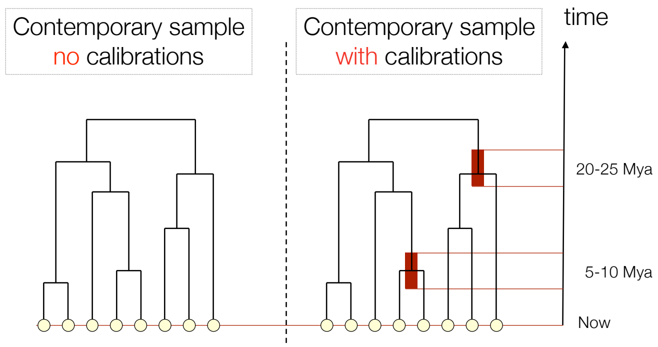
Combining the coalescent with Felsenstein's likelihood



$$p(N,g,Q \mid D) \propto \Pr\{D \mid \mu g,Q\} f_G(g \mid N) f_N(N) f_Q(Q)$$

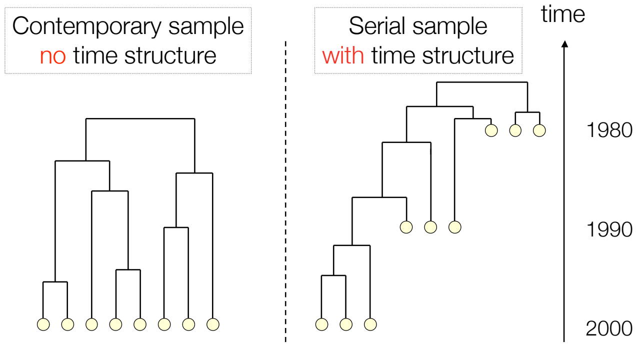
The joint posterior probability of the population history (*N*), the genealogy (*g*) and the mutation matrix (*Q*) are estimated using Markov chain Monte Carlo (Drummond et al, Genetics, 2002)

Time structure via calibrations


Contemporary sample
no calibrations

Contemporary sample
with calibrations

Time structure in samples themselves


Contemporary sample
no time structure

Serial sample
with time structure

Molecular evolution and population genetics of viruses

- Given sequence data that is time-structured estimate true values of:

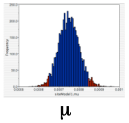
substitution parameters

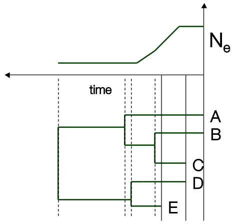
- Overall substitution rate and relative relative rates of different substitutions

population history: N(t)

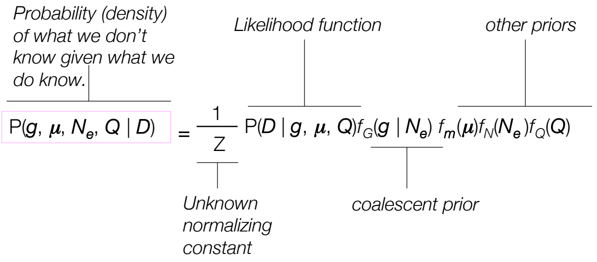
Ancestral genealogy

- Topology
- Coalescent times





Full Bayesian Model



Probability (density) of what we don't know given what we do know.

Likelihood function

other priors

$$P(g,\mu,N_e,Q\mid D)=\frac{1}{Z}P(D\mid g,\mu,Q)f_G(g\mid N_e)\,f_m(\mu)f_N(N_e)f_Q(Q)$$

Unknown normalizing constant

coalescent prior

Q = substitution parameters

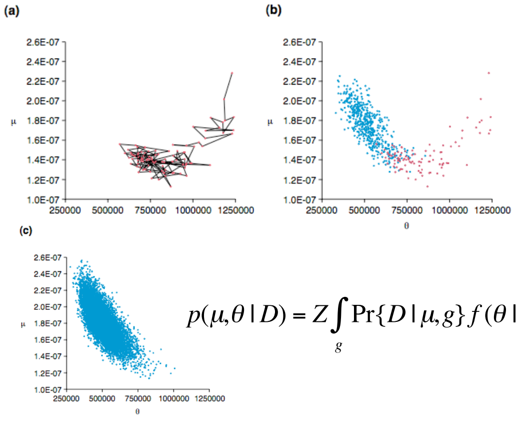
N_e = population parameters

g = tree

μ = overall substitution rate

In the software package BEAST, MCMC integration can be used to provide a chain of samples from this density.

Markov chain Monte Carlo (MCMC)



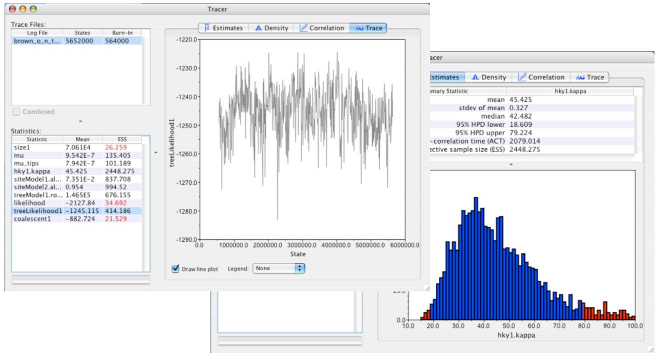
$$p(\mu,\theta\mid D)=Z\int_g \Pr\{D\mid\mu,g\}\,f(\theta\mid g)\,f(\mu,\theta)$$

TRENDS in Ecology & Evolution

BEAST

Bayesian Evolutionary Analysis Sampling Trees

http://evolve.zoo.ox.ac.uk/BEAST



Trace Files:

Log File	States	Burn-In
brown_o_n_t...	5652000	564000

□ Combined

Statistics:

Statistic	Mean	ESS
size1	7.061E4	26.259
mu	9.542E-7	135.405
mu_tips	7.942E-7	101.189
hky1.kappa	45.425	2448.275
siteModel1.al...	7.351E-2	837.708
siteModel2.al...	0.954	994.52
treeModel1.ro...	1.465E5	676.155
likelihood	-2127.84	34.692
treeLikelihood1	-1245.115	414.186
coalescent1	-882.724	21.529

☑ Draw line plot Legend: None

...mary Statistic	hky1.kappa
mean	45.425
stdev of mean	0.327
median	42.482
95% HPD lower	18.609
95% HPD upper	79.224
...-correlation time (ACT)	2079.014
...ective sample size (ESS)	2448.275

Why Bayesian?

- Probabilistic model-based inference

  Can make simple statements about the probability of alternative hypotheses given the data
- Markov chain Monte Carlo

  Convenient computational technique

  Allows for complex models: "if you can simulate you can sample"
- Incorporates prior probabilities

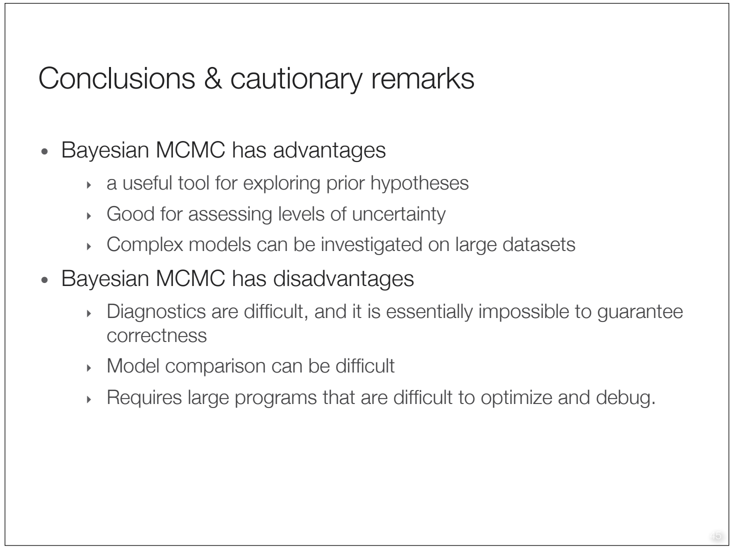
  $P(\theta|D) \propto P(D|\theta)P(\theta)$

  Convenient means of assessing alternative sets of assumptions

  Allows incorporation of independent sources of information
- Easy to include sources of uncertainty

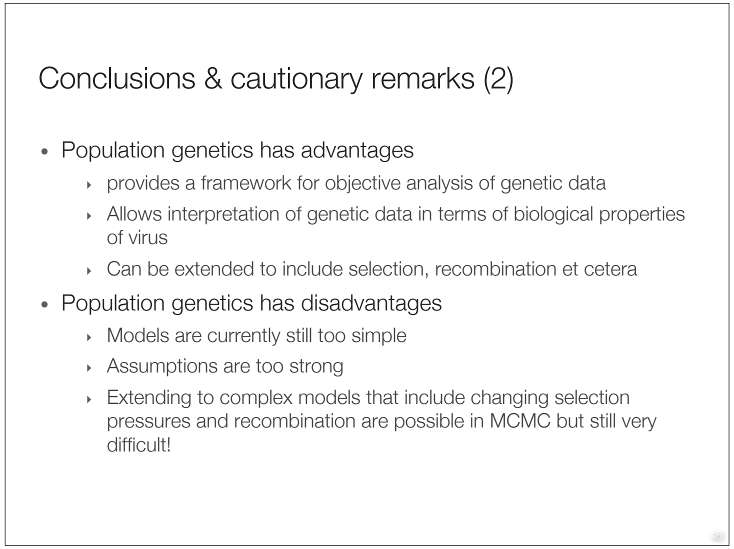
  Don't need to assume perfect knowledge of tree (for example)

  Can treat the tree and a nuisance parameter and focus on parameters of interest (strength of selection, mutation rate, growth rate, etc)



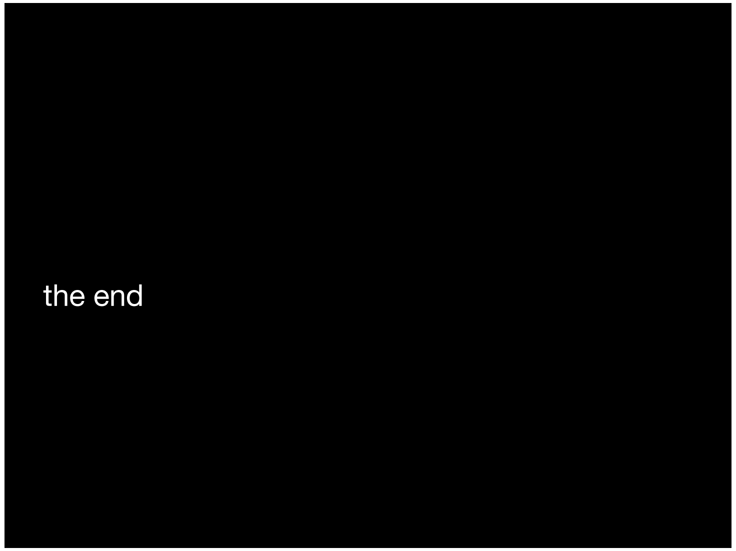
Conclusions & cautionary remarks

- Bayesian MCMC has advantages
  - a useful tool for exploring prior hypotheses
  - Good for assessing levels of uncertainty
  - Complex models can be investigated on large datasets
- Bayesian MCMC has disadvantages
  - Diagnostics are difficult, and it is essentially impossible to guarantee correctness
  - Model comparison can be difficult
  - Requires large programs that are difficult to optimize and debug.

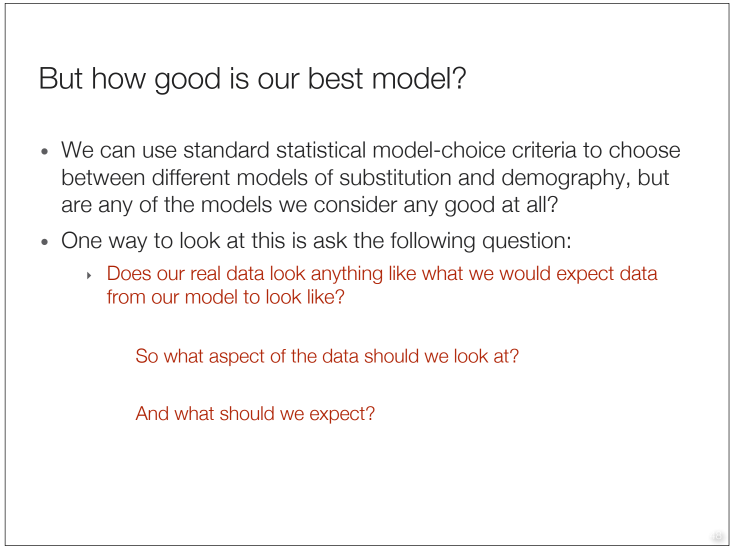


Conclusions & cautionary remarks (2)

- Population genetics has advantages
  - provides a framework for objective analysis of genetic data
  - Allows interpretation of genetic data in terms of biological properties of virus
  - Can be extended to include selection, recombination et cetera
- Population genetics has disadvantages
  - Models are currently still too simple
  - Assumptions are too strong
  - Extending to complex models that include changing selection pressures and recombination are possible in MCMC but still very difficult!



the end

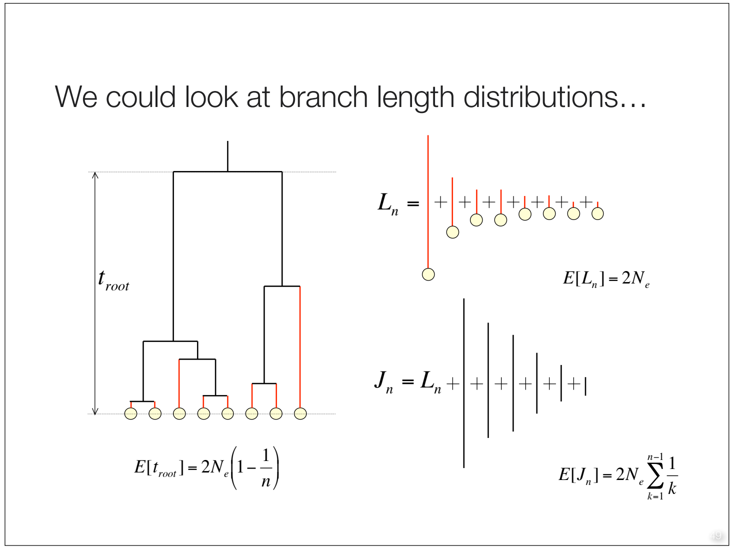


But how good is our best model?

- We can use standard statistical model-choice criteria to choose between different models of substitution and demography, but are any of the models we consider any good at all?
- One way to look at this is ask the following question:
  - Does our real data look anything like what we would expect data from our model to look like?

So what aspect of the data should we look at?

And what should we expect?

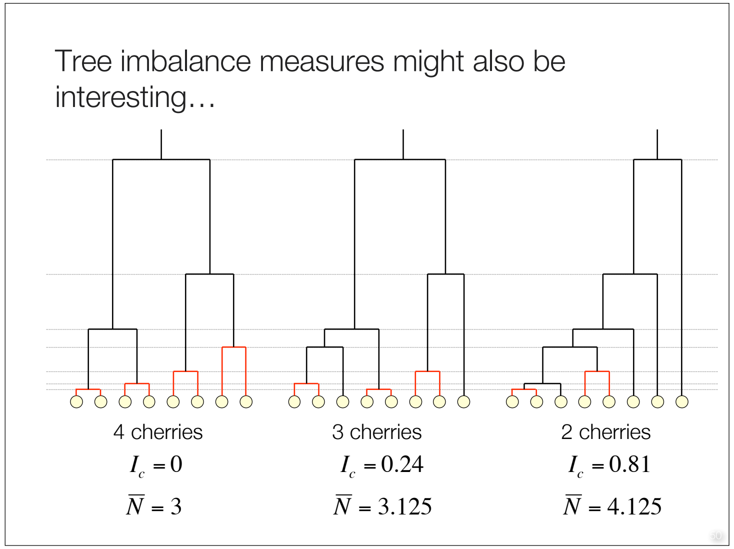


We could look at branch length distributions…

$$E[t_{root}] = 2N_e\left(1 - \frac{1}{n}\right)$$

$$E[L_n] = 2N_e$$

$$E[J_n] = 2N_e \sum_{k=1}^{n-1} \frac{1}{k}$$



Tree imbalance measures might also be interesting…

4 cherries

$I_c = 0$

$\overline{N} = 3$

3 cherries

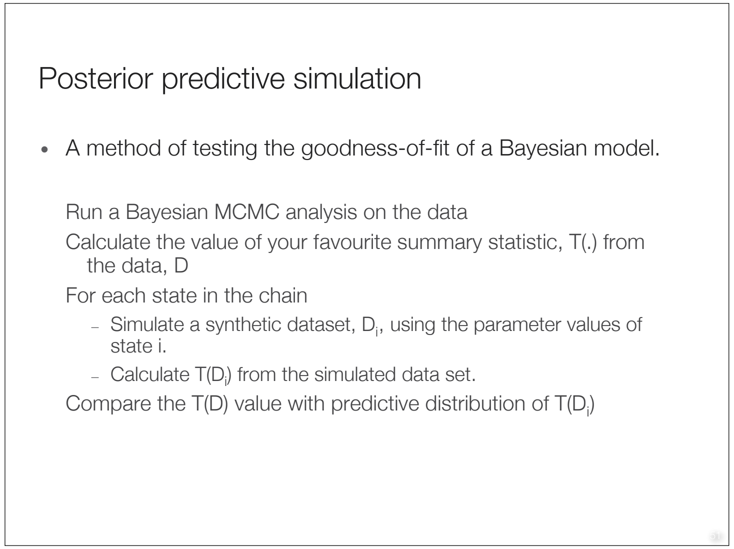
$I_c = 0.24$

$\overline{N} = 3.125$

2 cherries

$I_c = 0.81$

$\overline{N} = 4.125$



Posterior predictive simulation

- A method of testing the goodness-of-fit of a Bayesian model.

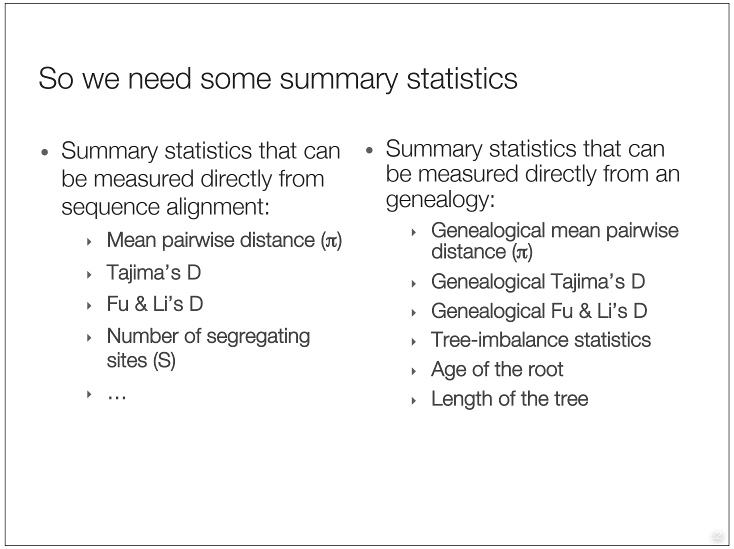
Run a Bayesian MCMC analysis on the data

Calculate the value of your favourite summary statistic, T(.) from the data, D

For each state in the chain

- Simulate a synthetic dataset, D_i, using the parameter values of state i.
- Calculate $T(D_i)$ from the simulated data set.

Compare the T(D) value with predictive distribution of $T(D_i)$



So we need some summary statistics

- Summary statistics that can be measured directly from sequence alignment:
  - Mean pairwise distance (π)
  - Tajima's D
  - Fu & Li's D
  - Number of segregating sites (S)
  - …
- Summary statistics that can be measured directly from an genealogy:
  - Genealogical mean pairwise distance (π)
  - Genealogical Tajima's D
  - Genealogical Fu & Li's D
  - Tree-imbalance statistics
  - Age of the root
  - Length of the tree

Posterior predictive simulation (2)

- Testing the goodness-of-fit of the neutral coalescent model under variable demographic functions.

Run a Bayesian MCMC analysis on the data

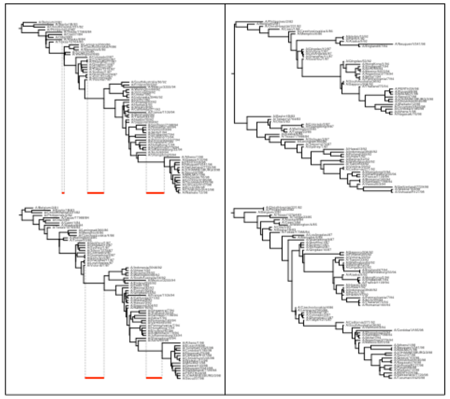
For each state in the chain

- Simulate a coalescent genealogy (G_i^S) using the population parameter values of state i.
- Calculate T(G_i^S) from the ith simulated genealogy
- Calculate T(G_i^P) from the ith posterior genealogy

Calculate the predictive probability by comparing the posterior distribution of T(.) with predictive distribution of T(.):

$$P_T^* = \frac{1}{n}\sum_{i=1}^{n} I(T(G_i^S) \geq T(G_i^P))$$

Human influenza A (HA gene) trees



State 5m

$N_e\tau = 9.12$

$t_2 = 11.03$ years

State 10m

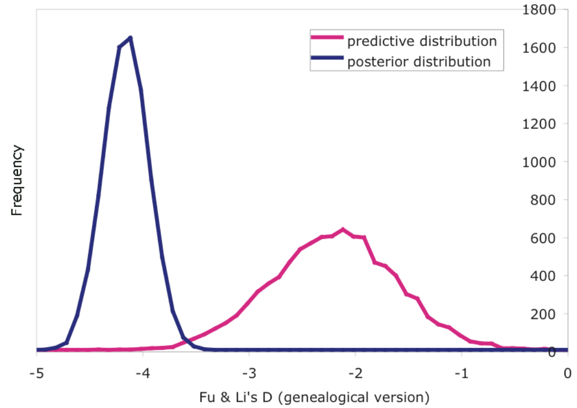
$N_e\tau = 5.00$

$t_2 = 15.29$ years

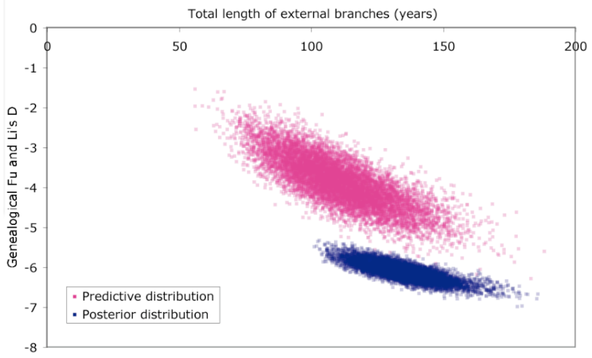
Posterior genealogy

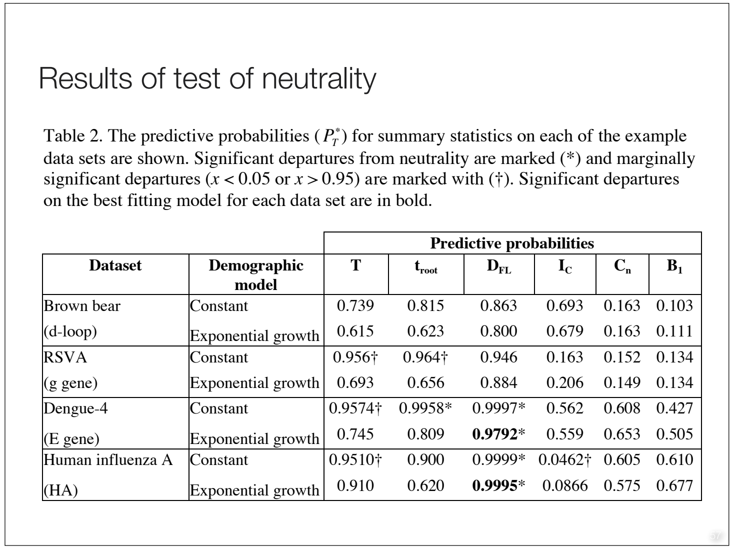
Predictive simulations

Human influenza A trees: Genealogical Fu & Li's D statistic



Puerto Rican Dengue-4 gene trees: multivariate summary statistics





Results of test of neutrality

Table 2. The predictive probabilities (P_T^*) for summary statistics on each of the example data sets are shown. Significant departures from neutrality are marked (*) and marginally significant departures ($x < 0.05$ or $x > 0.95$) are marked with (†). Significant departures on the best fitting model for each data set are in bold.

		Predictive probabilities					
Dataset	**Demographic model**	**T**	**t_{root}**	**D_{FL}**	**I_C**	**C_n**	**B_1**
Brown bear (d-loop)	Constant	0.739	0.815	0.863	0.693	0.163	0.103
	Exponential growth	0.615	0.623	0.800	0.679	0.163	0.111
RSVA (g gene)	Constant	0.956†	0.964†	0.946	0.163	0.152	0.134
	Exponential growth	0.693	0.656	0.884	0.206	0.149	0.134
Dengue-4 (E gene)	Constant	0.9574†	0.9958*	0.9997*	0.562	0.608	0.427
	Exponential growth	0.745	0.809	**0.9792***	0.559	0.653	0.505
Human influenza A (HA)	Constant	0.9510†	0.900	0.9999*	0.0462†	0.605	0.610
	Exponential growth	0.910	0.620	**0.9995***	0.0866	0.575	0.677