Summer Institute in Statistical Genetics
University of Queensland, Brisbane
Module 5: Population Genetic Data Analysis

The Coalescent Model

David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

Feb 9, 2017

Contents:

The standard coalescent model

The coalescent with population growth

Statistical inference under the coalescent

Simulation-based inference under the coalescent

Example: the Irish giant and AIP acromegaly mutation

The standard coalescent model

The coalescent with population growth

Statistical inference under the coalescent

Simulation-based inference under the coalescent

Example: the Irish giant and AIP acromegaly mutation

Going backwards in time: coalescent models

DNA sequences at the same locus from different individuals are *dependent*:

- differing amounts of common ancestry;
- so differing levels of correlation among the sequences.

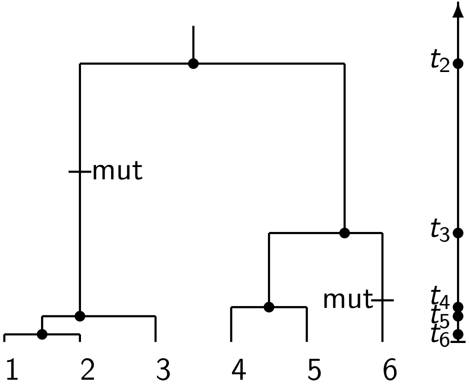
Valid inferences from DNA sequence data, e.g. about mutation rates or about the location of a gene of interest, may require modelling the relationships among the sequences.

- Incorrectly assuming independence can lead to understatement of variances of estimators – the effect can be large.
- Sometimes the relationships among the sequences are crucial e.g. for inferences about population histories.

A natural way to describe both the pattern of shared ancestry and the resulting correlations is via a genealogical tree (similar to a phylogenetic tree but for genes *within* a population, rather than from different species).

Coalescent models

Possible genealogy of a sample of 6 homologous sequences, showing two mutation events. The time arrow points backwards: e.g. t_6 denotes the most recent coalescent event, when the number of lineages decreased (going back in time) from 6 to 5.



The (standard) coalescent is a model for the genealogy underlying a sample of n genes at a neutral, non-recombining locus drawn from a large, random-mating, constant-size population.

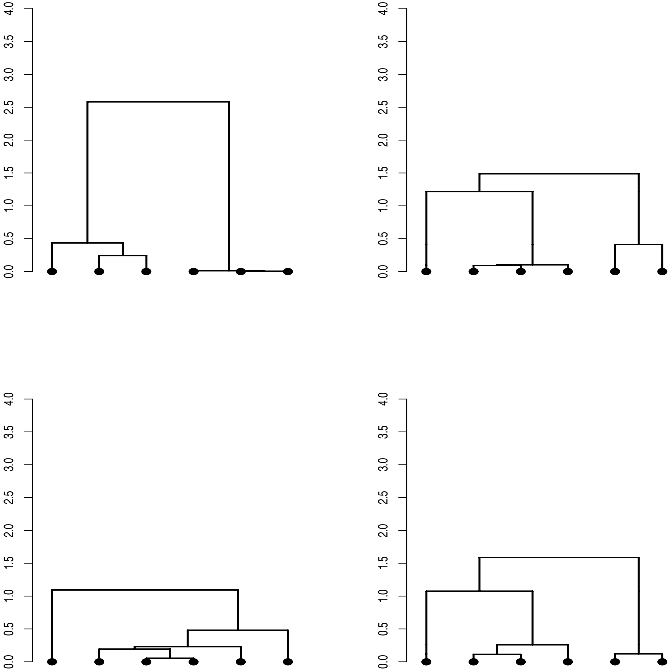
- Leaves of the tree ⇔ observed DNA sequences;
- going up the tree ⇔ tracing the ancestry of the sequences;
- branches merge, or "coalesce", when the descendant sequences first share a common ancestor;
- the root of the tree corresponds to the Most Recent Common Ancestor (MRCA) of all the sequences in the sample.

Under the coalescent model, the time during which the tree has j distinct branches has the exponential distribution with parameter $j(j-1)/2$ (we write $\text{Exp}(j(j-1)/2)$, NB mean $= 1/\text{parameter}$). The times for different j are independent. Here, one unit of "coalescent" time corresponds to NG/σ^2 years, where

N=(effective) population size,

G=generation time,

σ^2=variance in number of offspring (below assume $\sigma^2 = 1$).



Four realisations of the standard coalescent model with sample size $n = 6$. Mutations not shown.

Let T_n and L_n denote the height and the total branch length of a coalescent tree with n leaves. Then

$$E[T_n] = 2(1-1/n) \qquad E[L_n] = \sum_{j=1}^{n-1} \frac{2}{j} \approx 1+2\log(n)$$

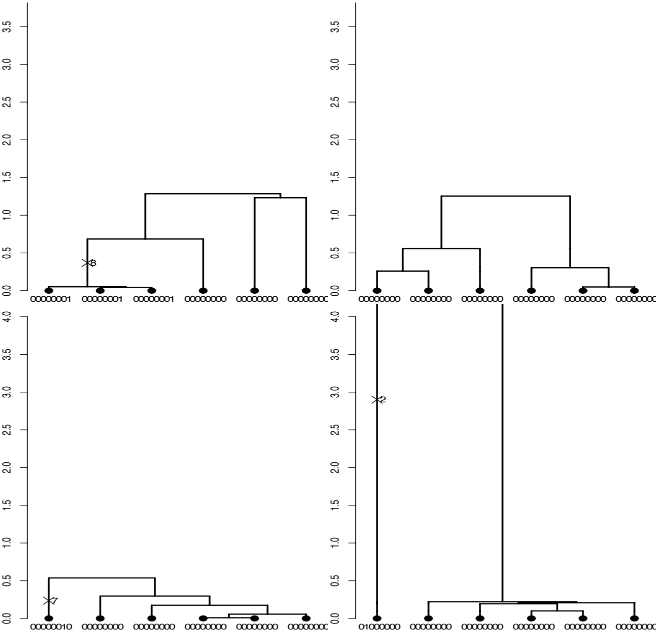
$$\text{Var}[T_n] = \sum_{j=2}^{n} \frac{8}{j^2} - 4\left(\frac{n-1}{n}\right)^2 \qquad \text{Var}[L_n] = \sum_{j=1}^{n-1} \frac{4}{j^2}$$

n	$\mathrm{E}[T_n]$	$\mathrm{V}[T_n]$	$\mathrm{E}[L_n]$	$\mathrm{V}[L_n]$
2	1	1	2	4
3	1·33	1·11	3	5
4	1·5	1·14	3·66	5·44
5	1·6	1·15	4·16	5·69
10	1·8	1·16	5·65	6·16
100	1·98	1·16	10·35	6·54
1000	2·00	1·16	14·97	6·58
10000	2·00	1·16	19·58	6·58

Features of the coalescent model:

- the mean time in which the sample has exactly two ancestors is more than half $\mathrm{E}[T_n]$, the mean total time since the MRCA (this can lead to bimodality in datasets);
- the variance of T_n is large relative to its mean; the largest contribution to $\mathrm{Var}[T_n]$ arises from the interval in which the sample has just two ancestors;
- $\mathrm{E}[L_n]$, the mean total branch length of tree (which is roughly the total amount of independent information in the data) grows only like $\log(n)$, not n as would be the case for a random sample.
- Although $\mathrm{E}[L_n]$ continues to increase with n, $\mathrm{Var}[L_n]$ does not.

These observations have big implications for patterns of DNA sequence variation along the genome.



Standard coalescent with $n = 6$, showing mutations and resulting 8-nucleotide sequences, with:

0 = ancestral,
1 = mutant nucleotide.

Coalescent with mutation

Mutations occur along the branches of a coalescent tree uniformly at random with rate $\theta/2$, where $\theta = 2N\mu$ and μ is the mutation rate per sequence per generation. Given L_n, the total branch length of the tree, the number S_n of mutations has the Poisson distribution with mean $\theta L_n/2$. The unconditional expectation is

$$E[S_n] = \frac{\theta}{2} E[L_n] = \theta \sum_{j=1}^{n-1} \frac{1}{j}.$$

If μ is small, it may be reasonable to assume the *infinite sites* model: every mutation is at a distinct site. Then S_n is just the number of variable sites and a natural estimator of θ is Watterson's estimator $\hat{\theta}_W = S_n / \sum_{j=1}^{n-1} \frac{1}{j}$, which is unbiased but the variance decreases like $1/\log(n)$, much slower than $1/n$ for estimators obtained from random samples.

- Even very big samples don't give very accurate estimates.

The standard coalescent model

The coalescent with population growth

Statistical inference under the coalescent

Simulation-based inference under the coalescent

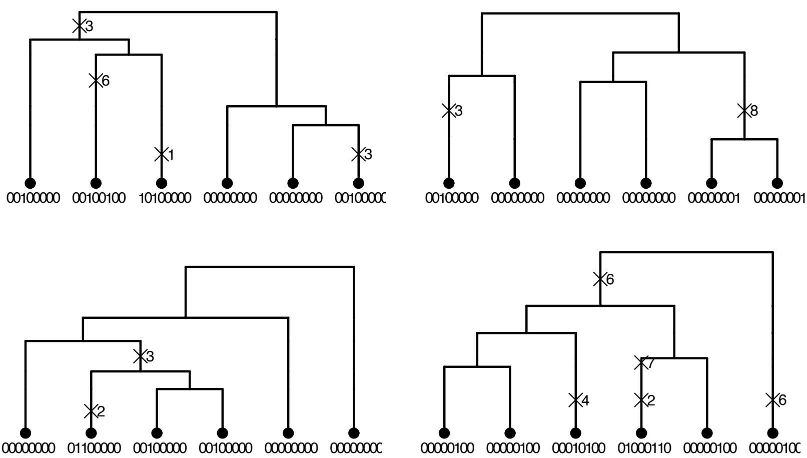
Example: the Irish giant and AIP acromegaly mutation

- Extensions of the standard model can allow for changes in population size, population subdivision, natural selection and recombination.
- The simplest extension is to incorporate **population growth**.
- Suppose that the population size Nt generations ago was $N\lambda(t)$, where N denotes the current effective population size, so that $\lambda(0) = 1$. Scaling time by N as for the standard coalescent, the waiting time for the jth coalescence event is now given by

  $$P(w_j > t) = \exp\left(-\frac{j(j-1)}{2}\Lambda(t)\right), \qquad (1)$$

  where $\Lambda(t) = \int_0^t \mathrm{d}s/\lambda(s)$.
- When the population size is large (i.e. $\lambda(t)$ is large), $\Lambda(t)$ increases only slowly with t, corresponding to the fact that coalescences rarely occur.

Four realisations of the coalescent model with mutation; sample size $n = 6$; exponential growth, R = 100.



The time scaling has been chosen so that $\mathrm{E}[L_n]$ is approximately the same as for the standard coalescent.

The standard coalescent arises in the case that $\lambda(t) \equiv 1$ and $\Lambda(t) = t$. Exponential growth/decline forward in time at rate r per generation corresponds to

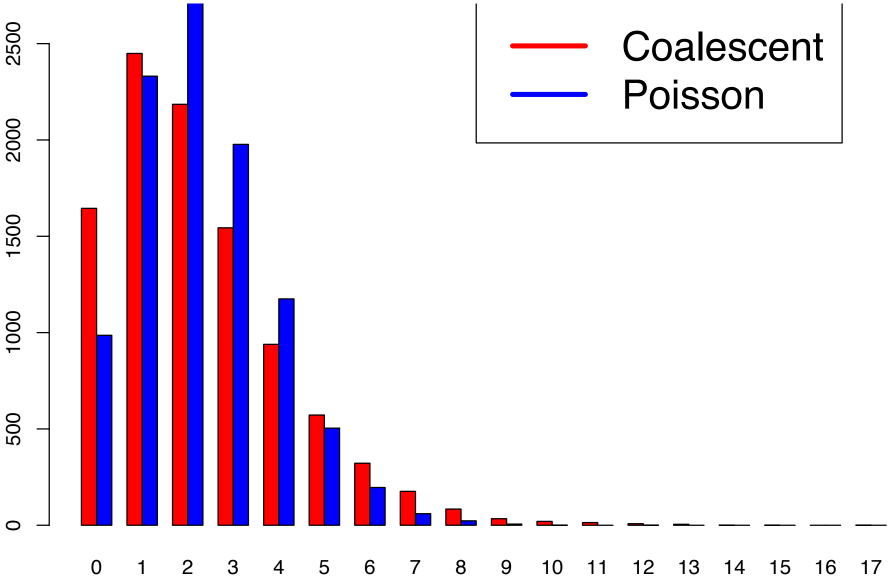
$$\lambda(t) = \exp(-Rt) \qquad \Lambda(t) = \frac{\exp(Rt) - 1}{R} \tag{2}$$

where $R = Nr$.

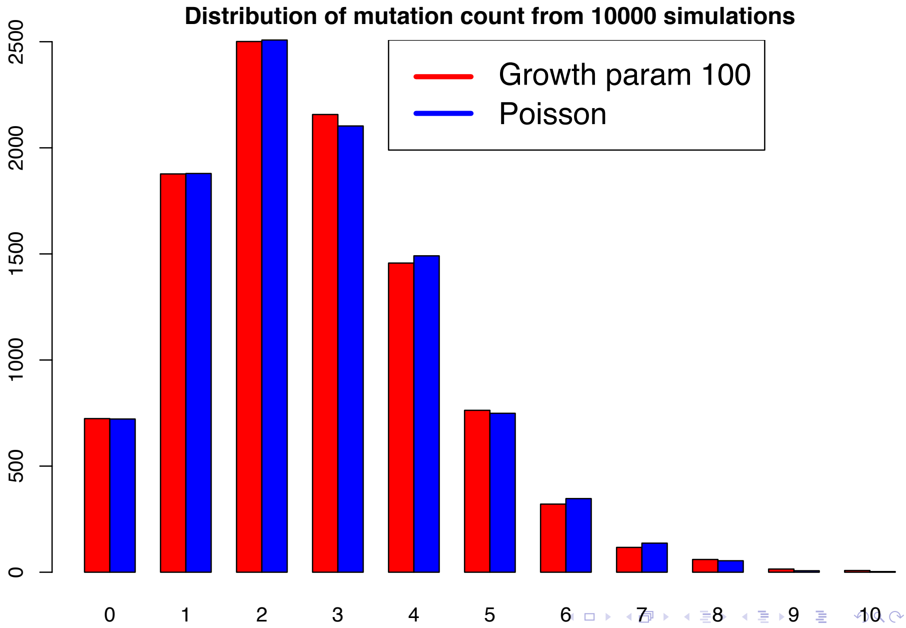
- Large R implies rapid growth forward in time $\Leftrightarrow$ rapid decline backward in time: relatively few coalescences occur in the recent past because the population size is large.
- In the limit as $R \uparrow \infty$ we obtain a "star genealogy": all coalescence events occur at about the same time and hence observed haplotypes are independent given the ancestral haplotype.

The following plots show histograms of the numbers of mutations under 10 000 realisations of (a) standard coalescent model and (b) coalescent with growth, in each case compared with expected values under the Poisson distribution with matching mean.

Distribution of mutation count: standard coalescent



Distribution of mutation count: coalescent with growth



The standard coalescent model

The coalescent with population growth

Statistical inference under the coalescent

Simulation-based inference under the coalescent

Example: the Irish giant and AIP acromegaly mutation

- Coalescent models without mutation specify a *prior* distribution for the genealogy underlying a set of DNA sequences, given the sample sizes but not the sequence data.
- Coalescent models with mutation can give predictive distributions for properties of sequence data expected under different models (with growth, structure, selection etc).
- However, often what we want to do is to infer properties of the underlying model (such as the mutation rate or time since most recent common ancestor (TMRCA)) given observed data.

One way to proceed is to seek the *posterior distribution* of parameters of interest, given the observed data, the coalescent model as prior for the genealogical tree, and assumed prior distributions for evolutionary parameters.

- To obtain the posterior from the prior, use Bayes Theorem:

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

where $\Pr(D) = \int \Pr(D|\theta)\Pr(\theta)d\theta$.

Exact Inference for TMRCA when $n = 2$

- We assume the standard coalescent model with infinite-sites mutation and suppose that θ is known;
- the unknown of interest is the coalescence time, or TMRCA of the two sequences, let's call it t_2.

Given $t_2 = t$, the number of segregating sites S has a Poisson distribution with parameter θt:

$$\Pr(S{=}s | t_2{=}t) = \frac{1}{s!}(\theta t)^s \exp(-\theta t). \tag{3}$$

By Bayes theorem we obtain the posterior pdf of t_2:

$$\mathrm{p}(t_2{=}t | S{=}s) = C(\theta t)^s \exp(-(1{+}\theta)t), \tag{4}$$

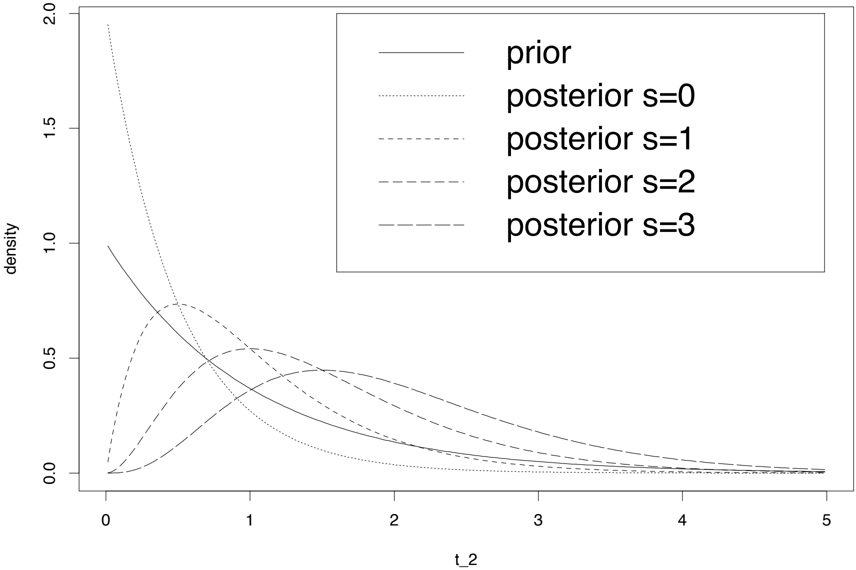
where C is a constant (does not depend on t). The RHS of (4) has the form of the Gamma$(1{+}s, 1{+}\theta)$ probability density function, and it follows that (Tajima, 1983):

$$\mathrm{E}[t_2|S{=}s] = \frac{1+s}{1+\theta} \qquad \text{and} \qquad \mathrm{Var}[t_2|S{=}s] = \frac{1+s}{(1+\theta)^2}, \tag{5}$$

which may be compared with prior moments $\mathrm{E}[t_2] = \mathrm{Var}[t_2] = 1$.

- Noting that $\mathrm{E}[S] = \theta$, we see that if $s < \mathrm{E}[S]$ then $\mathrm{E}[t_2|S{=}s] < \mathrm{E}[t_2]$, and vice-versa.
- Data usually decreases the variance: $\mathrm{Var}[t_2|S{=}s] < \mathrm{Var}[t_2]$ unless s is very large ($\geq 2\,\mathrm{E}[S] + \mathrm{E}[S]^2$).

The prior density curve and posterior curves for several values of s when $\theta = 1$ are shown on next slide.



Density curves for TMRCA (t_2, in coalescent units) when $n = 2$, $\theta = 1$. Prior: Gamma$(1, 1) \equiv$ Exp(1); posteriors: Gamma$(1+s, 2)$.

Comparison with classical estimator

- A natural estimator of t_2 within the framework of classical statistics is the method-of-moments estimator
  $$\hat{t}_2 = S/\theta,$$
  for which the mean square error (MSE) is
  $$\mathrm{MSE}(S/\theta) = \mathrm{E}_{t_2}[\mathrm{E}_{S|t_2}[(S/\theta - t_2)^2]] = 1/\theta,$$
- Uniformly larger than $1/(1+\theta)$, the MSE of $\mathrm{E}[t_2|S]$.

The use of prior distributions in statistical inference has been controversial, but here the prior is based on solid ground: the coalescent model that has been shown to provide a good approximation in many real populations.

- Additional information from prior $\Rightarrow$ more precise inferences.

An additional advantage of the Bayesian paradigm for statistical inference is that we obtain a full posterior distribution which summarises all available information about the unknown TMRCA, rather than just a point estimator and its standard error.

Exact inference for TMRCA when $S = 0$

Dorit *et al.* (1995) sequenced a 729-bp fragment in $n = 38$ human Y-chromosomes, observed $S = 0$ and reported a TMRCA estimate of $\hat{t_{38}} = 270$K years before present.

- First explicit use of coalescent theory for pop genet inference.
- A breakthrough! But unfortunately they made mistakes.

Donnelly *et al.* Science 1996, reply to Dorit:

- Assume no mutations in the underlying genealogy, then

$$\Pr(S{=}0|\theta, w_{38}, w_{37}, \ldots, w_2) = \prod_{j=2}^{38} \exp(-jw_j\theta/2), \tag{6}$$

  w_j = length of time that the coalescent has exactly j branches.

- The w_j have independent Exp($j(j{-}1)/2$) prior distributions; from (6), posteriors are independent Exp($j(j{-}1{+}\theta)/2$).

Thus posterior mean and variance of $t_{38} = \sum_{j=2}^{38} w_j$ are:

$\mathrm{E}[t_{38}|\theta, S{=}0] = \sum_{j=2}^{38} \frac{2}{j(j-1+\theta)}$, $\mathrm{Var}[t_{38}|\theta, S{=}0] = \sum_{j=2}^{38} \frac{4}{j^2(j-1+\theta)^2}$.

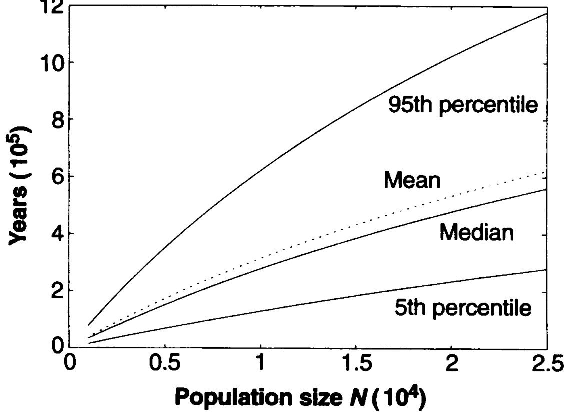
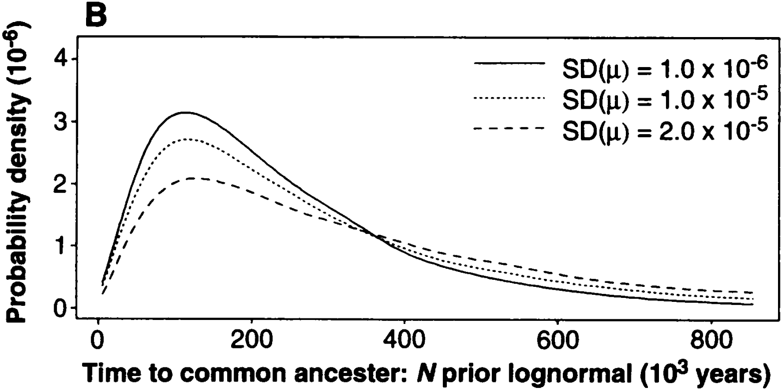


Fig. 1. Summary statistics for the conditional distribution, under the coalescent model, of the time T (in years) since the common ancestor, given a sample of 38 sequences which exhibit no variability, as a function of N, the effective population size. The generation time is assumed to be 20 years, and the mutation rate of the sequenced region per generation is taken to be 1.96×10^{-5}. Conditional distribution of T follows from equation 5.2 in (*7*).

The observation of $S = 0$ reduced the mean of t_{38} by $\approx$ 20% – 40% from prior mean for plausible N values.

Also used coalescent theory to obtain probability 7% that the TMRCA of these 38 males $\neq$ TMRCA of all human males. If so, global TMRCA is expected to be NG years further back in time.

Donnelly *et al.* Science 1996: reply to Dorit



- Here, $\mu = 1.96 \times 10^{-5}$ (from Dorit) and gen time $G = 20$ yrs.
- Modal values of t_{38} are around 120K years; variance is wide.
- More uncertainty about μ leads to more uncertainty about t_{38}: μ may be very small in which case the data are as expected, and provide little information.

Inference about θ, N and μ

- Integrating over the w_j in (6) we obtain the likelihood for θ:

$$\Pr(S{=}0|\theta) = \prod_{j=2}^{n} \frac{j{-}1}{j{-}1{+}\theta}. \tag{7}$$

  The MLE is $\hat{\theta} = 0$, which is non-sensical *a priori*.
- This defect of the MLE can be avoided by reporting a posterior 95% highest-density interval, using either an improper uniform prior for θ or a proper, informative prior.
- An additional advantage to a Bayesian approach is that it becomes possible to report inferences about N and μ separately (recall $\theta = 2N\mu$).
- The likelihood (7) only depends on N and μ through their product. So the data do not help distinguish them, but an informative prior distribution, if available, can.
- Inferences about N and μ are always sensitive to the prior assumptions, whereas in the presence of sufficient data inferences about θ will be robust to the prior.

The standard coalescent model

The coalescent with population growth

Statistical inference under the coalescent

Simulation-based inference under the coalescent

Example: the Irish giant and AIP acromegaly mutation

Rejection sampling

In most cases of interest exact inference under the coalescent is infeasible, but there are approximate methods based on simulation. A general approach to inference about θ given a sample of n DNA sequences is as follows:

1. Simulate a coalescent tree with n leaves,
2. simulate θ under an appropriate prior model,
3. simulate mutations along the branches of the tree according to a mutation model.
4. If the n resulting sequences are sufficiently close to the observed sequences, accept the simulated θ, otherwise reject.
5. The set of accepted θ values is approximately a sample from the posterior distribution of θ given the sequence data.

This is the core of the Approximate Bayesian Computation (ABC) method that has revolutionised population genetics over the past 15 years, allowing approximate inference under sophisticated models e.g. for population growth and structure.

- A key problem is to define “sufficiently close”.

- In some settings the number of segregating sites S captures most of the information in the sequence data.
- Conditional on L, the total branch length of the coalescent tree, S has approximately a Poisson distribution with mean $\theta L/2$.
- Therefore, the accept/reject step can be performed more efficiently using Poisson probabilities, without simulating a value of S (Tavaré *et al.* 1997, Genetics).
- Coalescent tree and θ still need to be simulated unless $S = 0$.

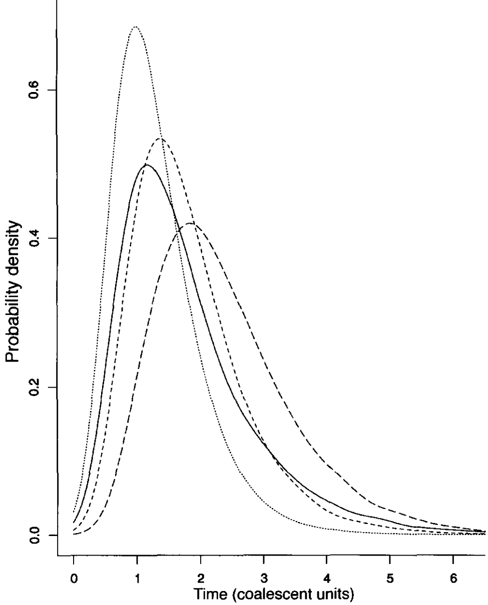


FIGURE 2.—Pre- and post-data density curves for T_{10} with $\theta = 1$. —, pre-data density; · · ·, $S_{10} = 1$; ---, $S_{10} = 3$; – – –, $S_{10} = 5$.

```
qrej = function(nacc=10000,nblk=round(nacc/2),nsamp=6,nsit=5,s=3)
# rejection inference about N, mu and TMRCA given
# s segregating sites in nsamp sequences of length nsit
{
  count = 0
  ns1 = nsamp-1
  rate = (ns1:1)*(nsamp:2)/2
  acc = matrix(0,1,3)
  while(nrow(acc)<nacc+1)
  {
    count = count + nblk
    w = matrix(rexp(ns1*nblk,rate),ns1,nblk)
    TMRCA = apply(w,2,sum)
    L = apply((nsamp:2)*w,2,sum)
    u = runif(nblk)
    N = rgamma(nblk,5,10^-3)
    mu = rgamma(nblk,2,2*10^4)
    ind = u<dbinom(s,nsit,1-exp(-L*N*mu/nsit))/dbinom(s,nsit,s/nsit)
    acc = rbind(acc,matrix(c(N[ind],mu[ind],TMRCA[ind]),,3))
  }
  list(count,acc[-1,])
}
```

Rejection-sampling inference: exercise

Use the R code `qrej` (for **q**uick **rej**ection sampling) to perform inferences under the coalescent when S segregating sites (default $= 3$) are observed in n sequences (default $= 6$). `qrej` assumes:

- gamma$(5, 10^{-3})$ prior for N (mean $= 5\,000$, SD $= 2\,236$); and
- gamma$(2, 2{\times}10^4)$ prior for μ (mean 10^{-4}, SD $= 7.07{\times}10^{-5}$).

`qrej` returns a list of length 2:

1. the number of iterations (must exceed `nacc`);
2. a matrix with 3 cols: accepted values of N, μ and TMRCA.

To obtain a density plot e.g. for N, you can do:

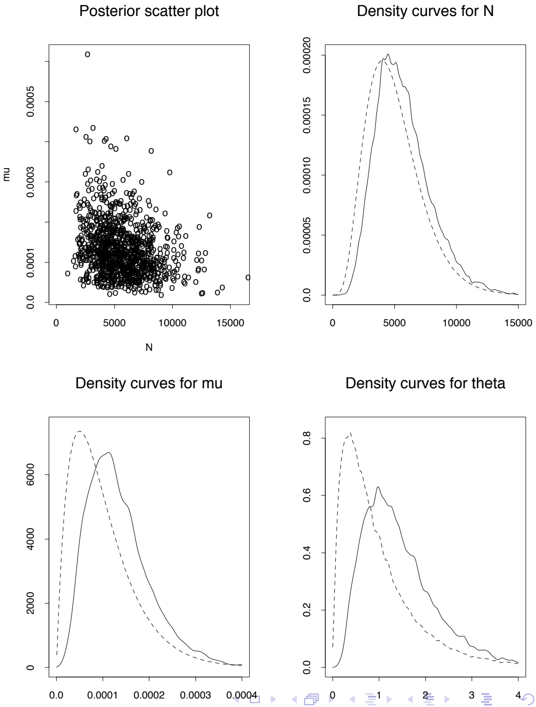
```
plot(density(res[[2]][,1]),fro=0,to=15000)
```

and you should add an `xlab` to label the x-axis. You can also add a prior density:

```
lines(x,dgamma(x,5,0.001),lty=2)
```

where `x` is a vector of grid points between 0 and 15,000.

- The plot shows some results from inference using `qrej` with default settings.
  - Dashed curves: prior,
  - Solid curves: posterior.
- You should try to replicate these plots and obtain a similar plot for TMRCA.
- Explore the effects of choosing different values for `nsamp` or `s`.
- Also try altering the prior distributions (you will need to edit the `rgamma` commands in the `R` code).



The standard coalescent model

The coalescent with population growth

Statistical inference under the coalescent

Simulation-based inference under the coalescent

Example: the Irish giant and AIP acromegaly mutation

Radian *et al. Human Mutation* (2016). Coalescent-based estimate of number of carriers of AIP risk allele in Ireland.

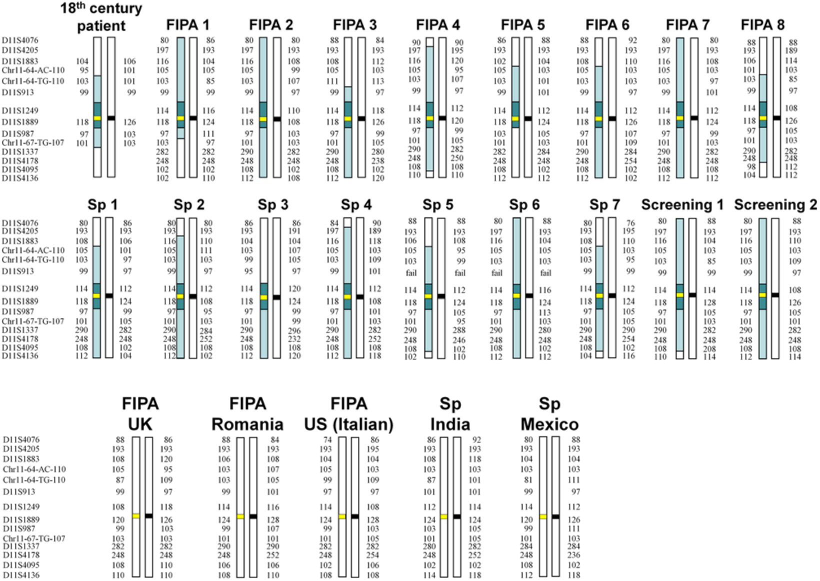
AIP mutations cause autosomal dominant familial isolated pituitary adenomas (FIPA), most commonly manifesting as acromegaly or gigantism. Due to incomplete penetrance, the disease can also manifest as apparently sporadic pituitary adenoma (PA).

- Chahal *et al.* NEJM (2011): 5 carriers identified in mid-Ulster, including a proband case: a C18 "Irish giant".
- Coalescent simulation-based analysis predicted a large number of carriers concentrated in mid-Ulster

Population screening in mid-Ulster for AIP mutations:

- 81 carriers (30 affected, 18 pedigrees) identified in mid-Ulster.
- Low prevalence in Belfast ($n = 1\,000$), no carriers found in Republic of Ireland ($n = 2\,000$).
- Haplotype conservation suggested a recent TMRCA.

Now we seek to update predictions of the TMRCA and consequently the number of carriers not yet identified.



Haplotypes on chromosome 11q12.2–13.3 of individuals carrying AIP R304*. Dark shading: haploblock shared by all Irish pedigrees; light shading: additional shared haploblocks. AIP alleles (black = wild-type, yellow = R304*).

Radian *et al.* (2016): Methods

- Haplotypes inferred from genotypes using PHASE.
- We performed exact coalescent inference for the fully conserved haplotype, the result of which became a prior for ABC inference of the varyingly-shared haplotypes.
- Since we were concerned only with conserved haplotypes, recombination and mutation have the same effect (of destroying conservation) and so we treated recombination events like mutations and combined the two rates.
- The statistic used to compare simulated and observed datasets was the number of haplotypes sharing each genome segment (defined by consecutive short tandem repeat markers).

Results:

- TMRCA estimated at 2 550 (1 275 – 5 000) years.
- Forward simulations using TMRCA distribution predicted 432 (90 – 5 175) current carriers, including 86 affected assuming 20% penetrance.