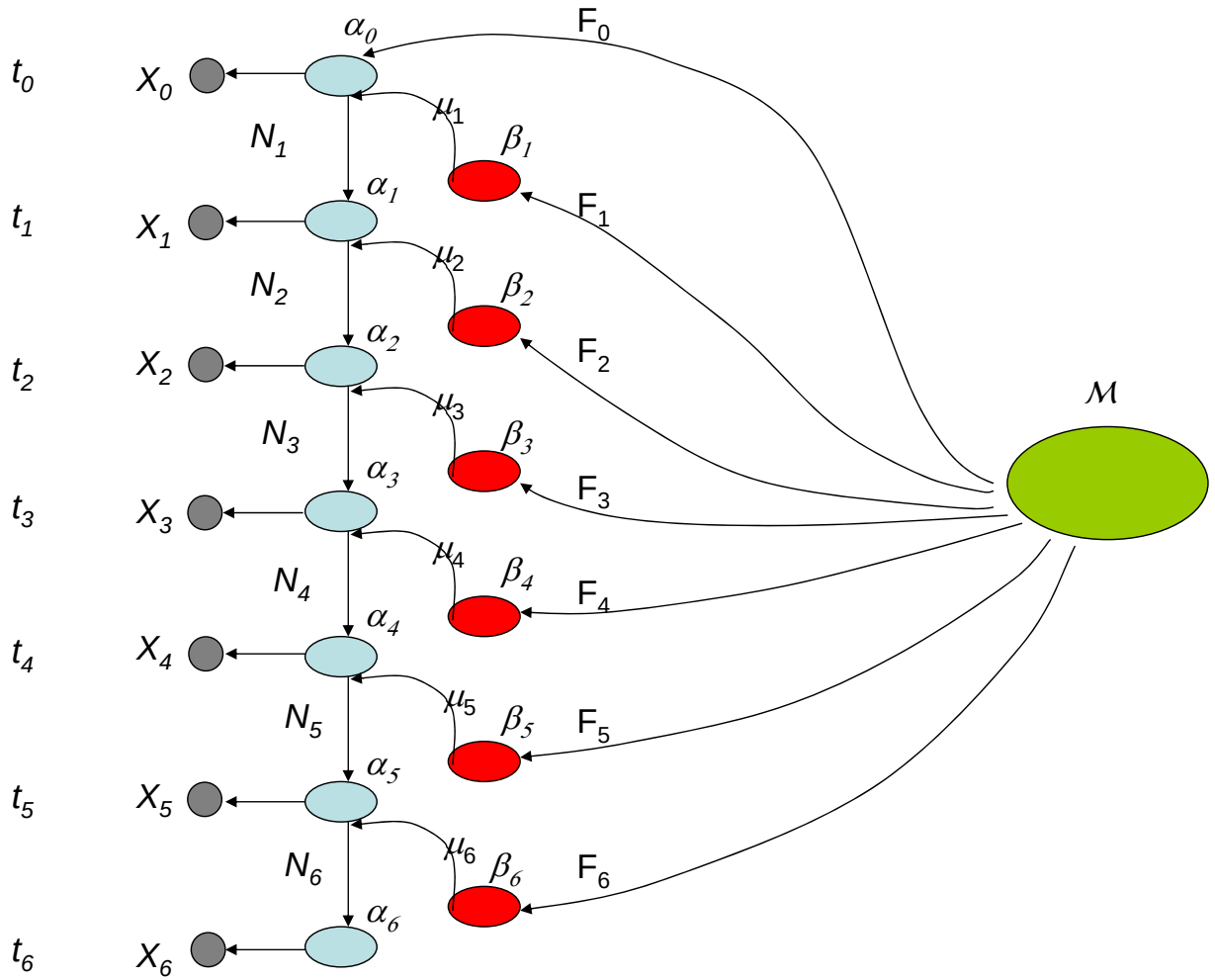
ABC for Temporally Sampled Genetic Data

Mark A. Beaumont,
Schools of Biological Sciences and Mathematics,
The University of Bristol,
Bristol, UK

05 April 2011

Temporal Change in Gene Frequency with Admixture

Aim is to infer parameters in a state space model of changes in gene frequency in the presence of admixture.



Temporal Change in Gene Frequency with Admixture

- Temporally sampled genetic data is quite commonly obtained.
- Changes are usually attributed to genetic drift (a function of the population size).
- However admixture and replacement of populations over time may be confounded with drift.
- This is a major issue for ancient DNA samples

Importance Sampling, Particles, and MCMC

- Beaumont (Genetics, 2003); GIMH algorithm; using noisy estimates of likelihood obtained from sequential importance sampling in MCMC.
- Becquet and Przeworski (Genome Research, 2007); application of GIMH idea to MCMC-ABC algorithm of Marjoram et al (PNAS, 2003).
- Andrieu and Roberts (Annal. Stat. 2009)Pseudo-marginal method: convergence proofs and generalization of GIMH.
- Andrieu, Doucet, and Holenstein (RSSB, 2010); Particle MCMC
- Peters and Cornebise (RSSB, discussion of A,D,&H, 2010); ABC and particle MCMC.

Framework for Temporal Model with Admixture (1)

S temporal samples are taken.

- t_i Time of ith sample ($i = 0, \ldots, S$).
- Δt_j Difference between time of jth and $(j-1)$th sample ($j = 1, \ldots, S$).
- N_j Effective population size for jth interval.
- μ_j Admixture proportion for jth interval.
- F_i F_{ST} of ith admixing population.

Framework for Temporal Model with Admixture (2)

Use a Dirichlet rather than coalescent to model variance in allele frequencies:

- Laval *et al.*, (Genetics, 2003)
- Kitakado *et al.*, (Genetics, 2006)

This does not give the same allele frequency distribution as the coalescent, but for a given F_{ST}, the variance is the same (see discussant contributions to Nicholson *et al* (RSSB, 2002)).

Framework for Temporal Model with Admixture (3)

For frequency vector α_i of length K alleles, sampled at time t_i, we model the change in frequency due to drift over the interval Δt_i with effective size N_i as

$$\alpha_i \sim D(\phi_i \alpha'_{(i-1),1}, \ldots, \phi_i \alpha'_{(i-1),K})$$

where

$$\phi_i = \frac{\exp(-\Delta t_i / N_i)}{(1 - exp(-\Delta t_i / N_i))}.$$

The observed frequencies, X_i are assumed to be multinomial samples from α_i

Framework for Temporal Model with Admixture (4)

Admixture is modelled as

$$\alpha'_{i-1} = (1 - \mu_i)\alpha_{i-1} + \mu_i \beta_i.$$

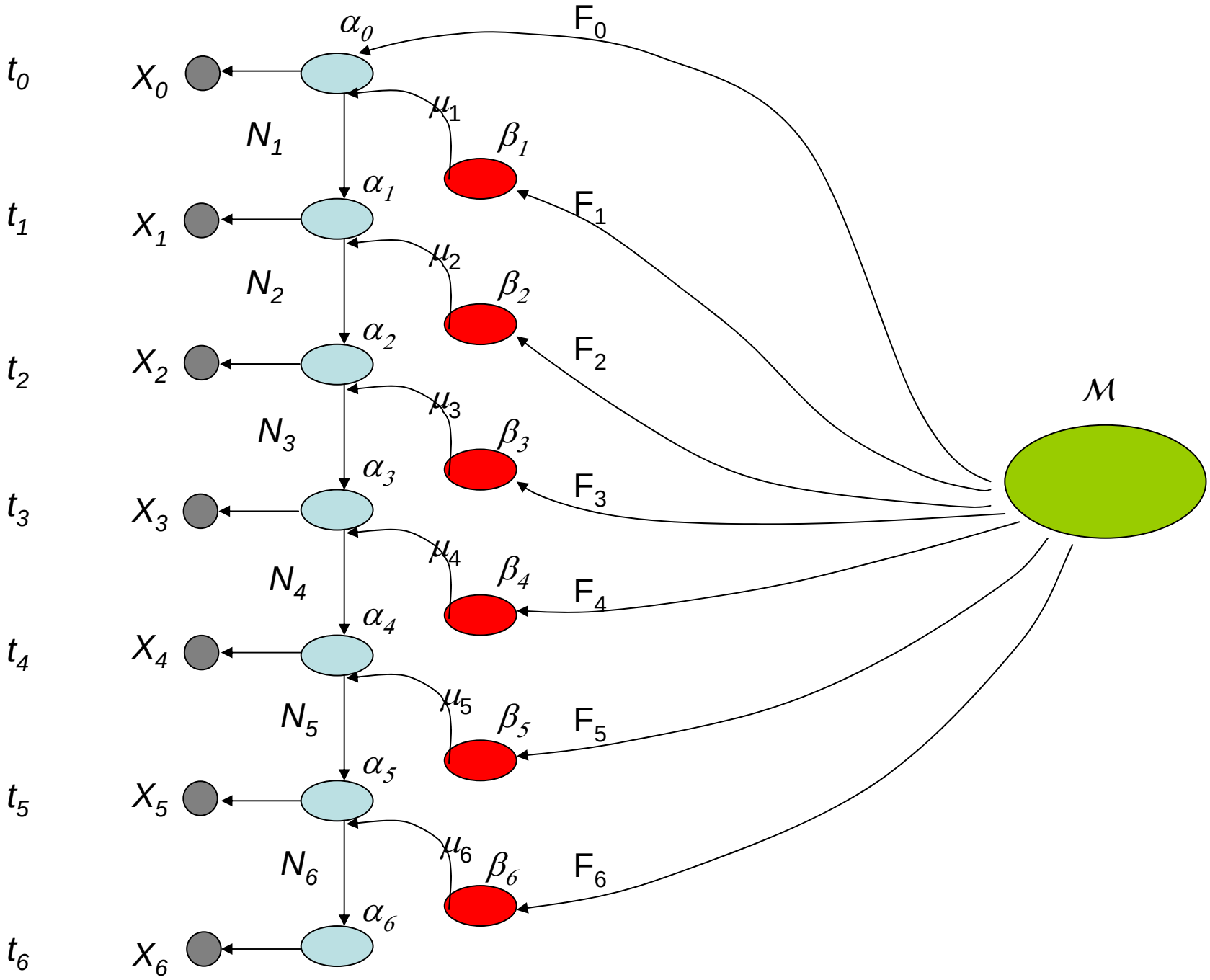
The admixing frequencies β_j, $(j = 1, \ldots, S)$, and the initial α_0, are drawn from Dirichlet distributions, parameterized by F_i $(i = 0, \ldots, S)$, and metapopulation frequency $\mathcal{M}$. *E.g*:

$$\beta_1 \sim D(\theta_1 \mathcal{M}_1, \ldots, \theta_1 \mathcal{M}_K)$$

with

$$\theta_1 = \frac{1}{F_1} - 1$$

(Sewall Wright's infinite island model)



MCMC implementation of TMA

Aim is to infer parameters in this model in a Bayesian framework.
The likelihood is:

$$P(X_0|\alpha_0)P(\alpha_0|F_0,\mathcal{M})$$
$$\times \prod_{i=1}^{S}\{P(X_i|\alpha_i)P(\alpha_i|\alpha_{i-1},N_i,\Delta t_i,\mu_i,\beta_i)P(\beta_i|F_i,\mathcal{M})\}$$

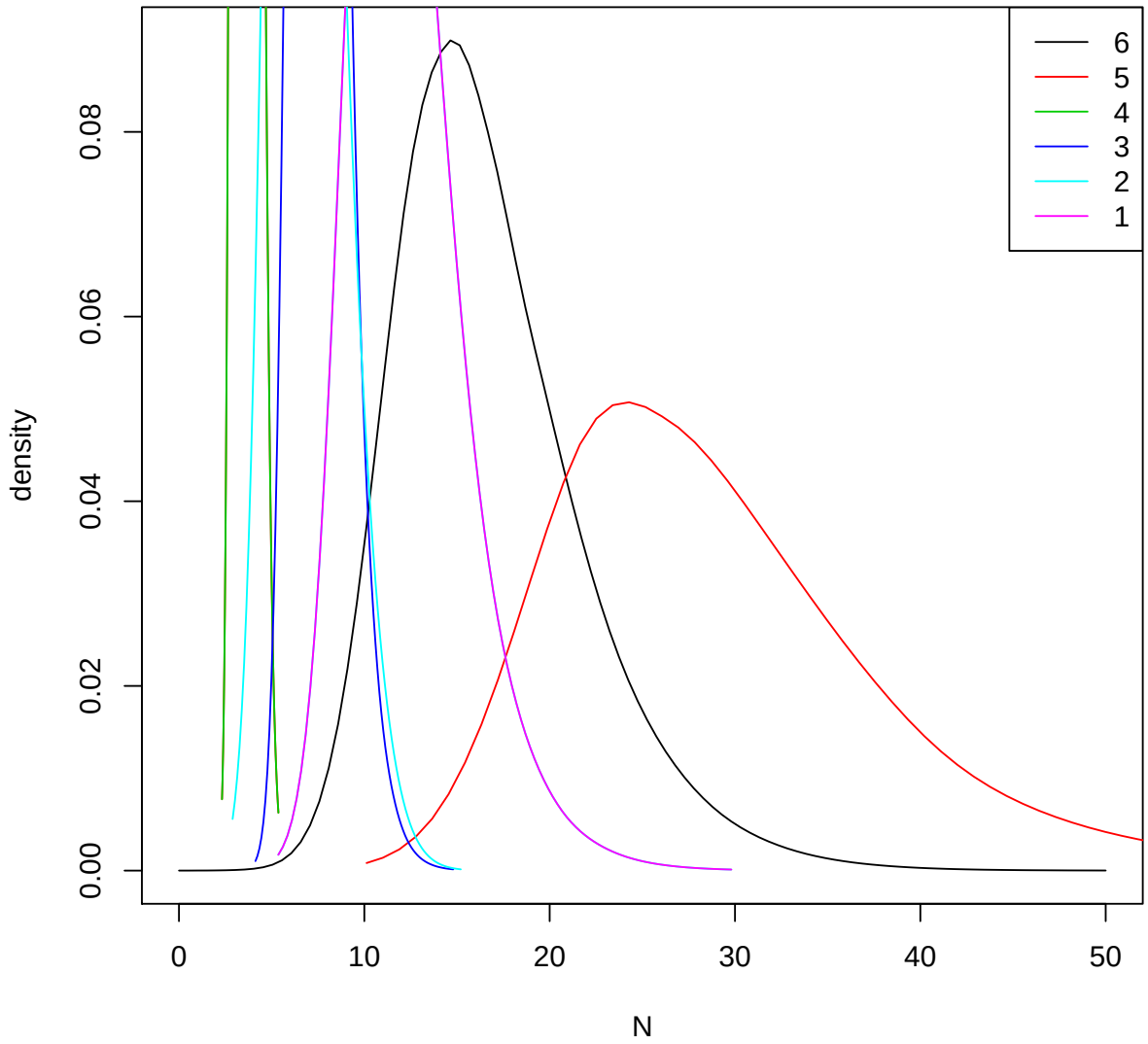
- The t_is are known.
- Assume a hierarchical prior on N_i (Gaussian on log-scale)
- Assume beta priors on μ_i and F_i
- Assume Dirichlet prior on $\mathcal{M}$

Update parameters using Metropolis-Hastings.

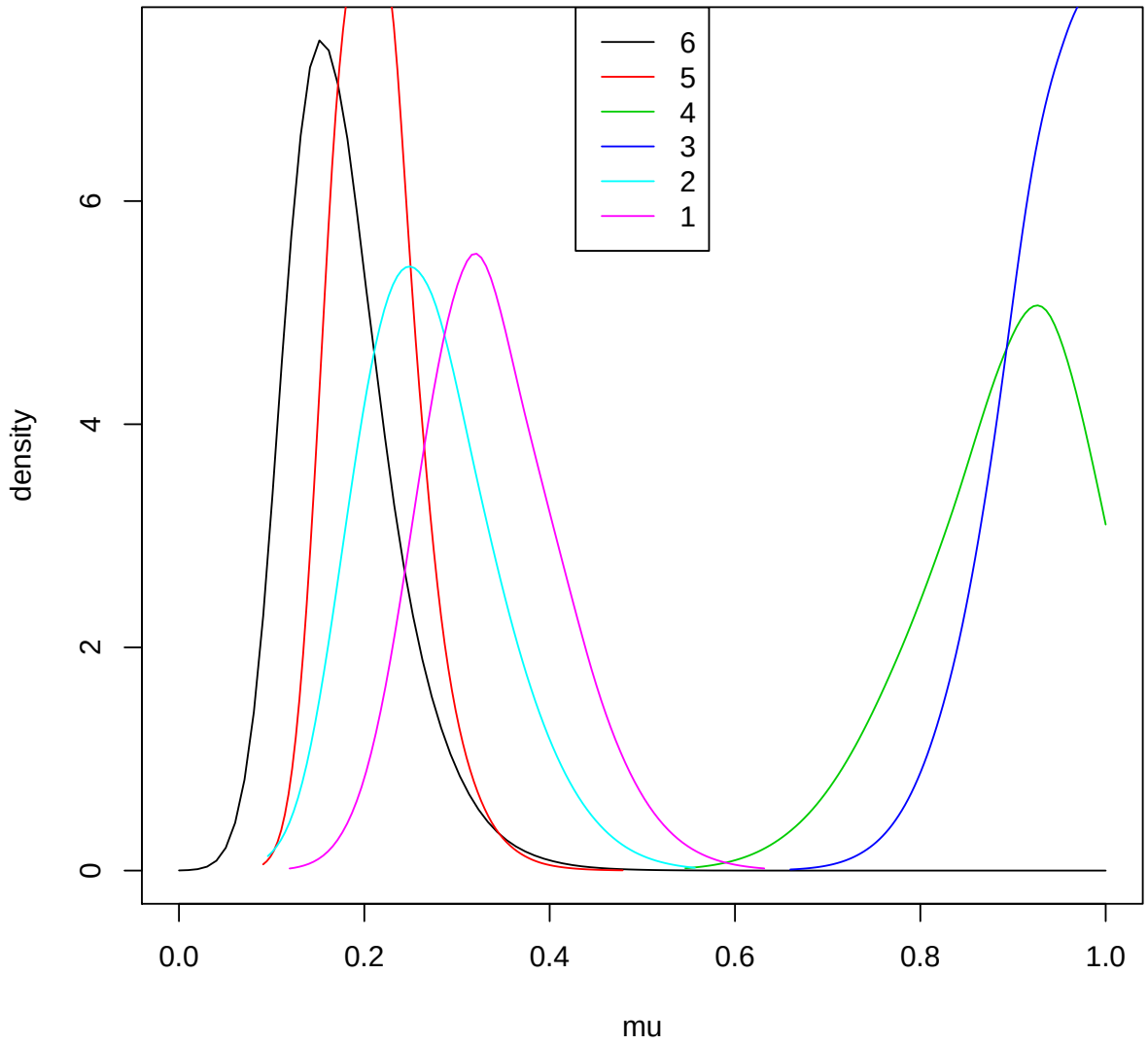
Application to Bryozoan data

1. Data from a freshwater Bryozoan, *Cristatella mucedo*, studied by Beth Okamura (NHM, London) and Sophia Ahmed (Roscoff, France).
2. 8 highly polymorphic microsatellite loci genotyped by Sophia Ahmed.
3. Sampled over 7 time periods.
4. Gene frequencies change markedly; unlikely to be due to drift.
5. Aim is to estimate effective population sizes, admixture proportions, and F_{ST} of putative admixing populations.

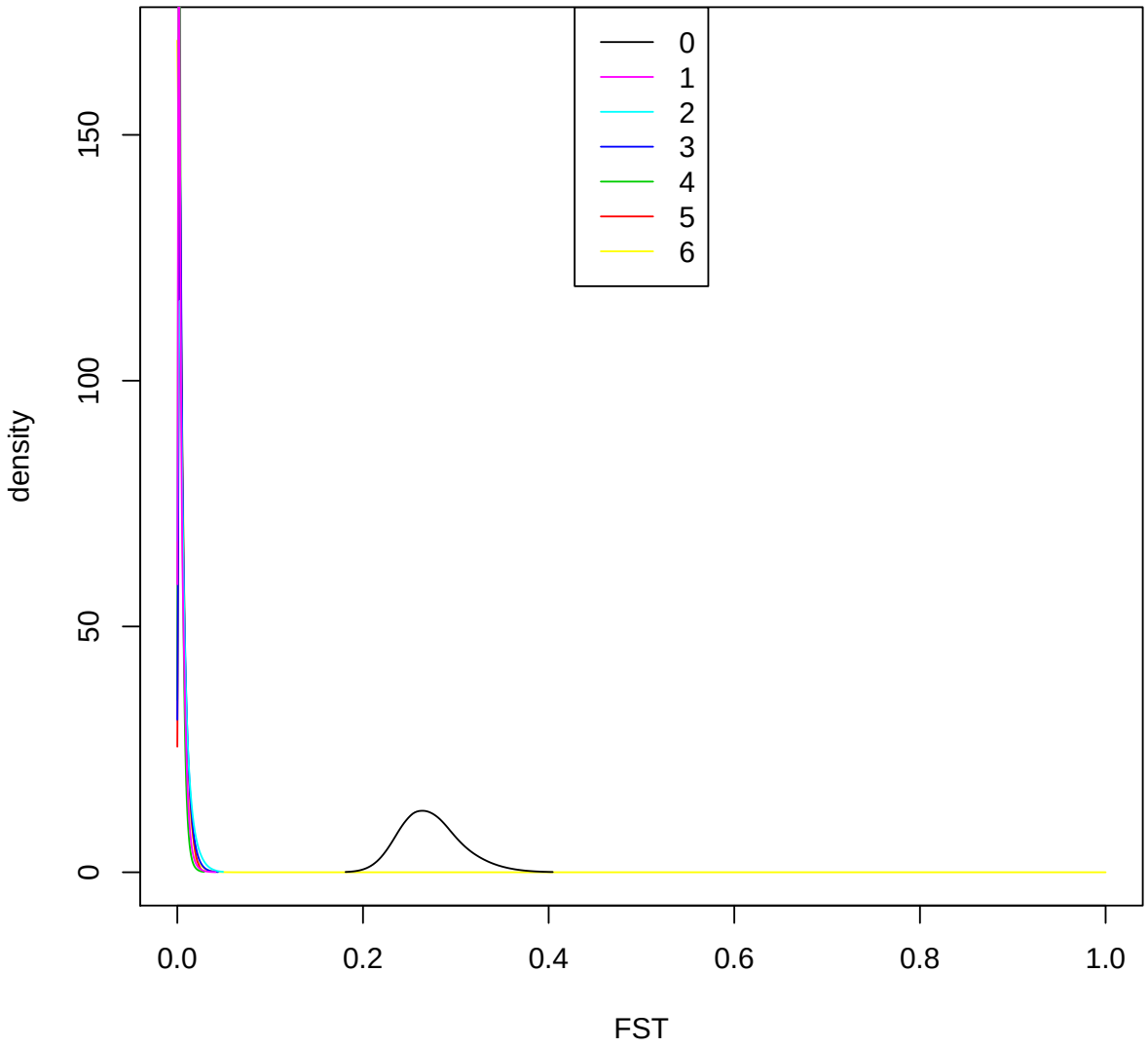
Bryozoan data: results with MCMC algorithm



Cryozoan data: results with MCMC algorithm



Bryozoan data: results with MCMC algorithm

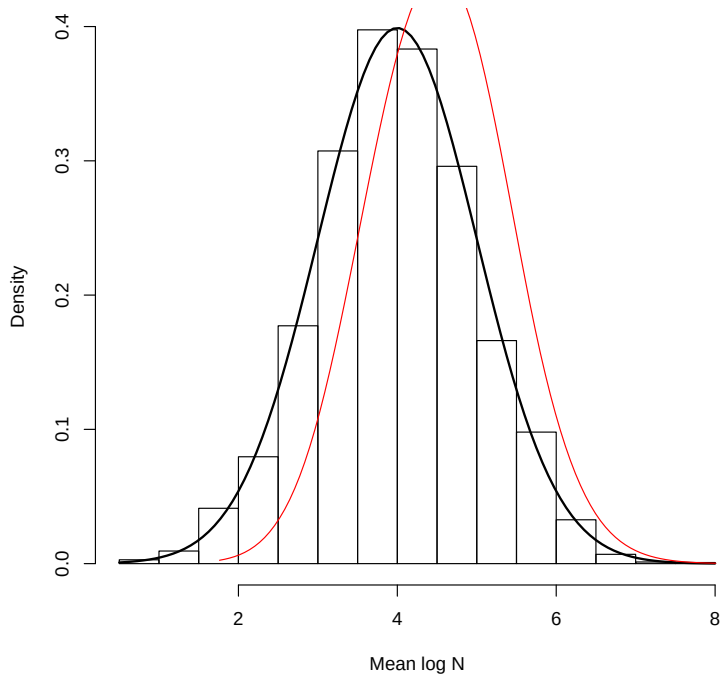


Convergence of MCMC

Comparison of runs with likelihood held constant, to check for recovery of priors.

Data sampled at 4 time points, 2 loci, 5 alleles each.

Histogram — α_i held constant Red line — α_i updated Black line — prior $\Gamma(4,1)$



Particle MCMC Implementation of TMA

Aim is to avoid MCMC updates for $\alpha_1, \ldots, \alpha_S$, but use MCMC for all other parameters (including α_0).

At each MCMC step, use importance sampling of the α_i to compute noisy likelihood estimate, conditioning on all parameter values at that stage in the MCMC.

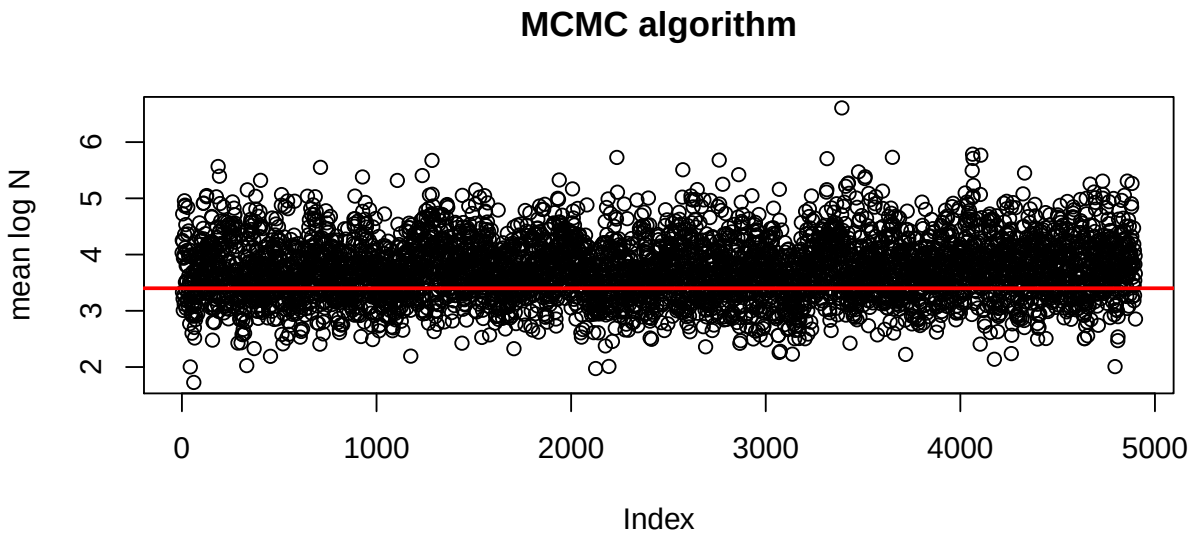
Particle MCMC Implementation of TMA

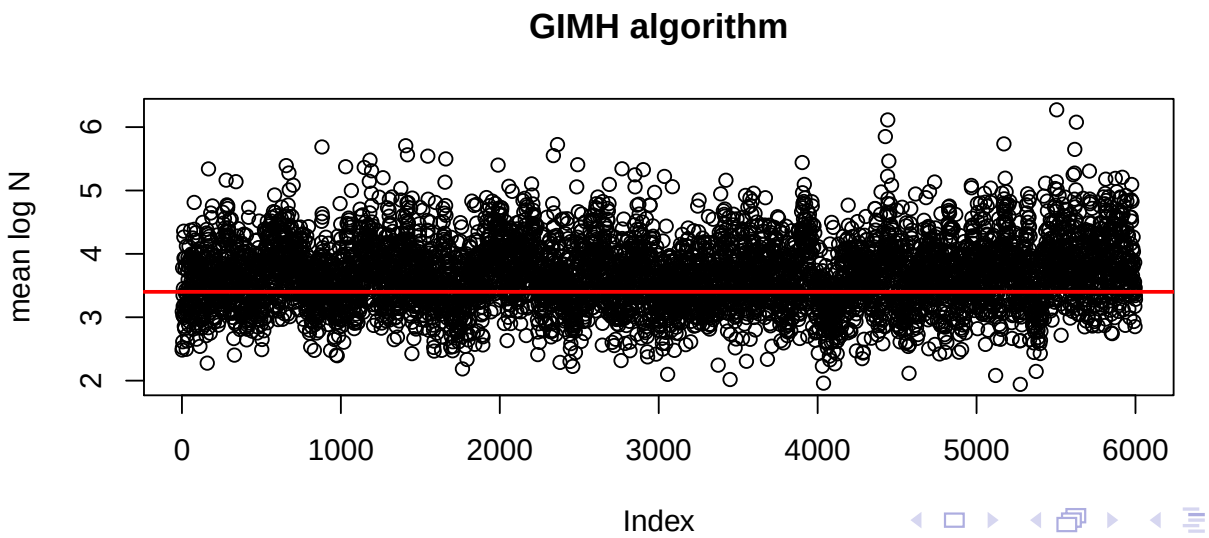
Schematic Algorithm

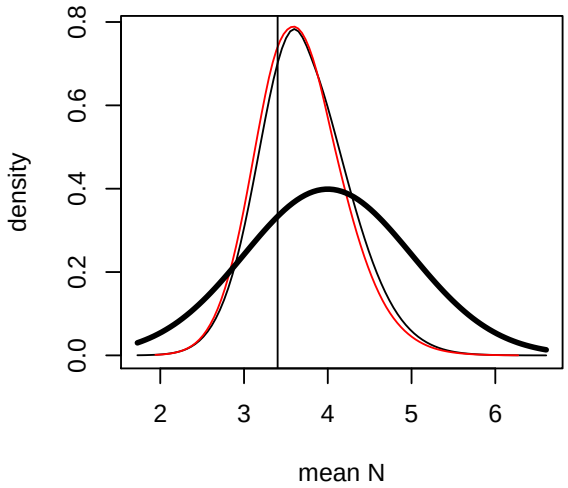
1. For sample point 1:
   1. set $\phi_1 = \frac{\exp(-\Delta t_1/N_1)}{(1-exp(-\Delta t_1/N_1))}$.
   2. Simulate M particles: $\alpha_1^{(j)} \sim q(\alpha_1^{(j)}) := D(\{\phi_1 + X_1\}\alpha_0')$.
   3. Compute importance weight $W_1^{(j)} = p(X_1|\alpha_1^{(j)})p(\alpha_1^{(j)}|\alpha_0', \phi_1)/q(\alpha_1^{(j)})$.
   4. Set $\tilde{L}_1 = 1/M \sum W_1^{(j)}$.
2. For sample points $i > 1$:
   1. Set ϕ_i.
   2. Simulate M particles: $\alpha_i^{(j)} \sim q(\alpha_i^{(j)}) := D(\{\phi_i + X_i\}\alpha_{i-1}^{\prime(j)})$, where
      $\alpha_{i-1}^{\prime(j)} = (1-\mu_i)\alpha_{i-1}^{(j)} + \mu_i\beta_i$
      where $\alpha_{i-1}^{(j)}$ is sampled from particles at step $i-1$ with weight $W_{i-1}^{(l)}$, $l = 1, \ldots, M$
   3. Compute weights *etc.* as for time step 1.
3. Set $\tilde{L} = P(X_0|\alpha_0)\prod_{i=1}^{S}\tilde{L}_i$.

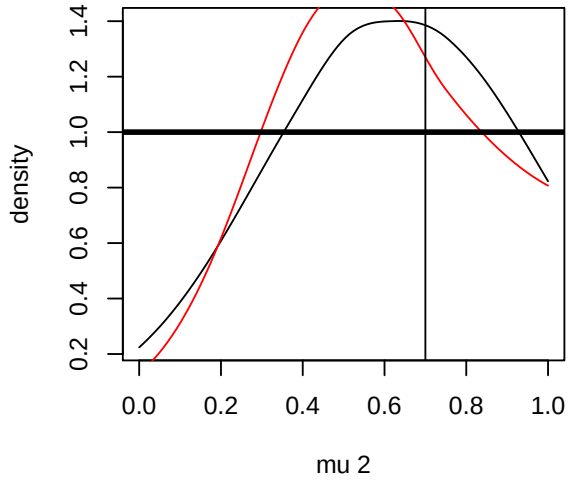
Results from Particle MCMC

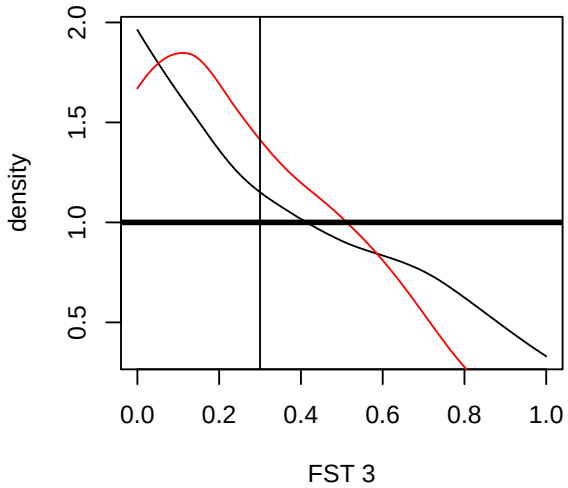
Trace of mean N











ABC and Particle MCMC

Motivation is to see whether this approach is feasible and competitive with particle MCMC.

Replace importance estimate of $\tilde{L}_i$ with proportion of simulated points that are within tolerance interval.

ABC and Particle MCMC: application to TMA

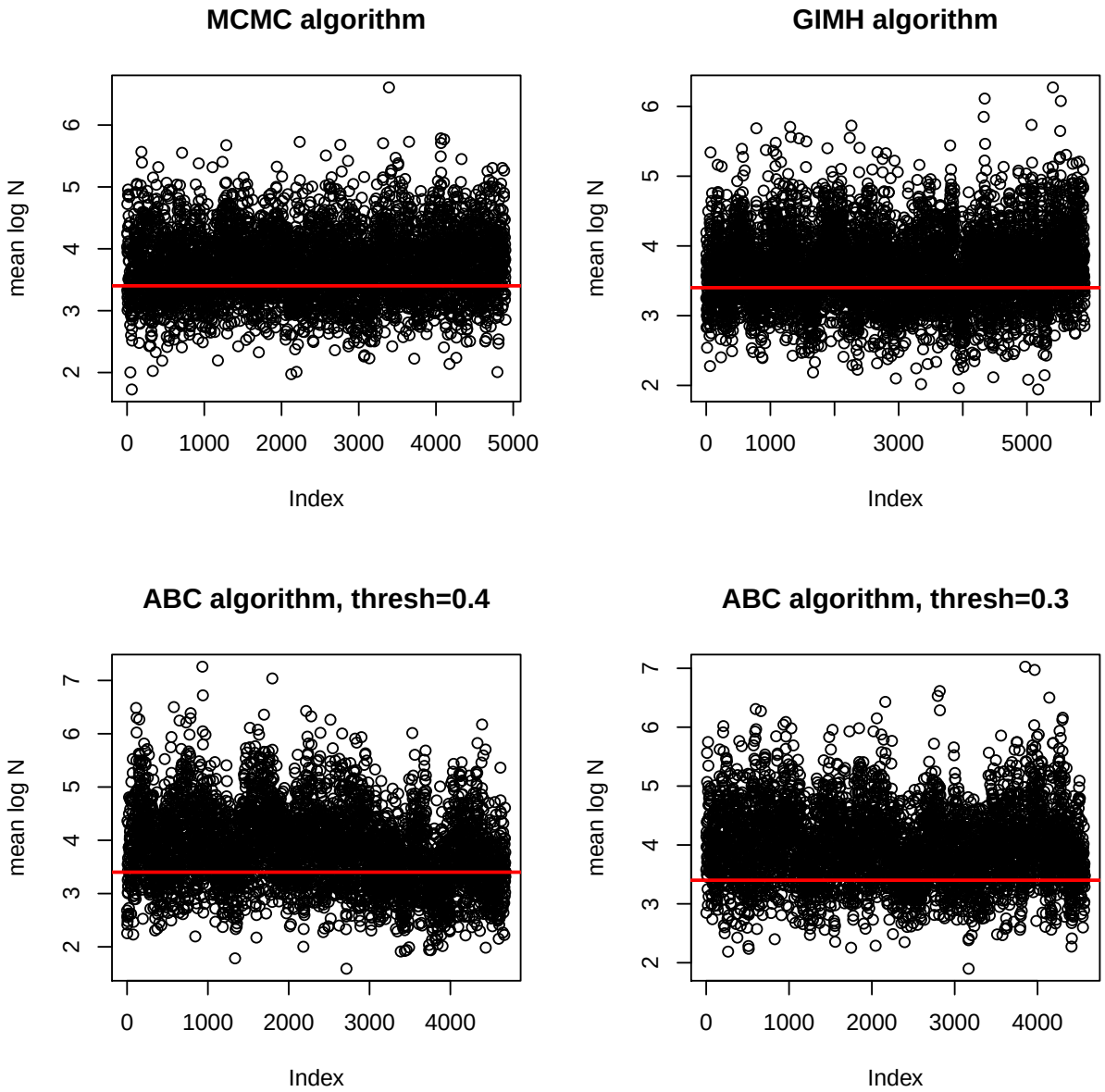
Implementation

1. For sample point 1:
   1. set $\phi_1 = \frac{\exp(-\Delta t_1/N_1)}{(1-exp(-\Delta t_1/N_1))}$.
   2. Simulate M particles: $\alpha_1^{(j)} \sim D(\phi_1 \alpha_0')$, $X_1' \sim \text{Multinom}(\alpha_1^{(j)})$.
   3. Compute (0,1) weight $W_1^{(j)} = I(|X_1' - X_1| < \delta)$.
   4. Set $\tilde{L}_1 = 1/M \sum W_1^{(j)}$.
2. For sample points $i > 1$:
   1. Set ϕ_i.
   2. Simulate M particles: $\alpha_i^{(j)} \sim D(\phi_i \alpha_{i-1}'^{(j)})$, where $\alpha_{i-1}'^{(j)} = (1-\mu_i)\alpha_{i-1}^{(j)} + \mu_i \beta_i$ where $\alpha_{i-1}^{(j)}$ is sampled from particles at step $i-1$ with weight $W_{i-1}^{(l)}$, $l = 1, \ldots, M$.
   3. Compute weights *etc.* as for time step 1.
3. Set $\tilde{L} = P(X_0|\alpha_0) \prod_{i=1}^{S} \tilde{L}_i$.

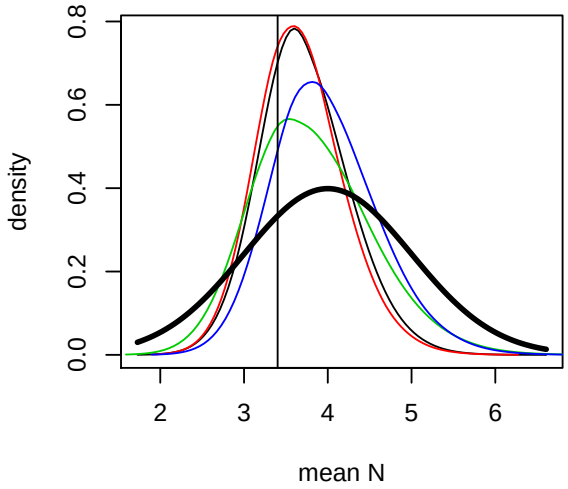
Summary Statistics

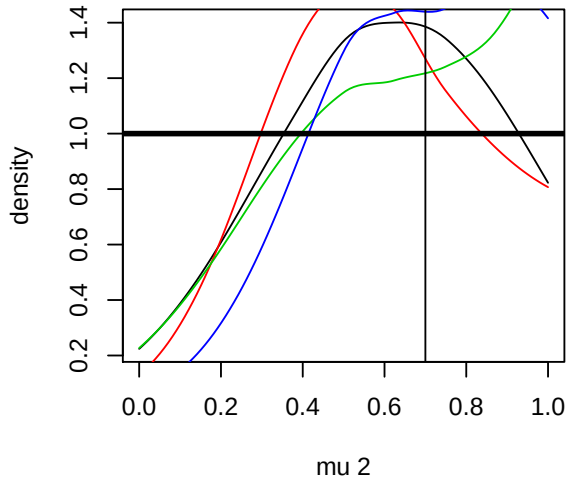
1. Allele frequency is used.
2. Compute $Q = 1/(K-1)\sum(X_i' - X_i)/(X_i + g)$ for alleles $i = 1, \ldots, K$.
3. For threshold R, accept if $Q < R$.
4. In examples, $R = 0{\cdot}3$ or $0{\cdot}4$ and $g = 1$.

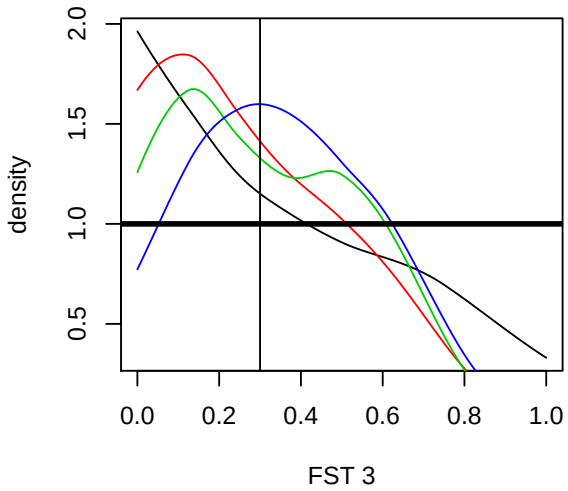
Results from Particle MCMC with ABC



Results from Particle MCMC with ABC







Acknowledgments

Beth Okamura Sophia Ahmed