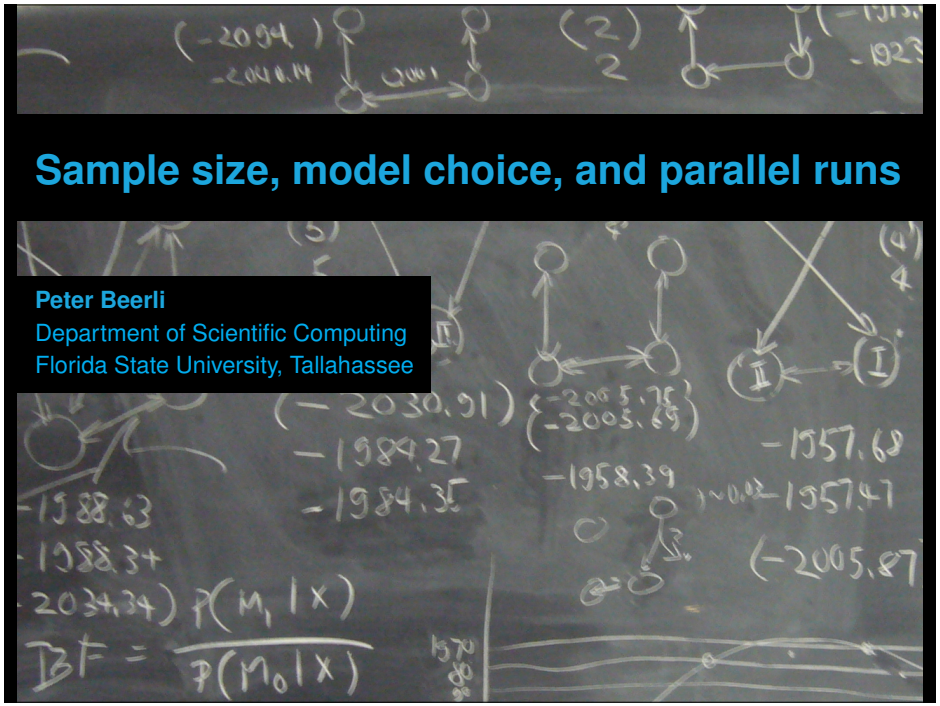


Sample size, model choice, and parallel runs

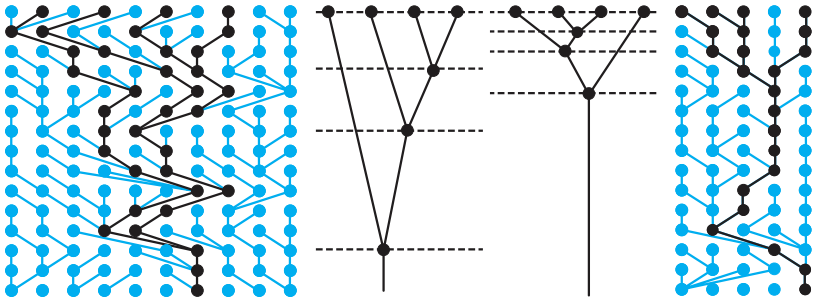
Peter Beerli
Department of Scientific Computing
Florida State University, Tallahassee

Overview

1. Sample size
2. Model parameters
3. How to reduce parameters
4. Bayes Factors (practical)
5. Replication and parallel runtime

Population size mantra

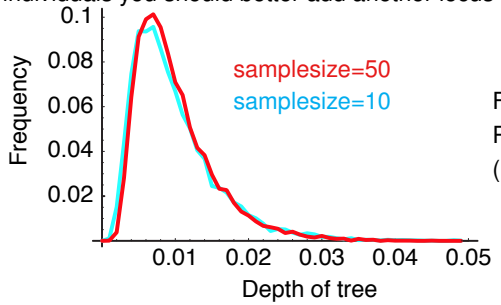
Coalescence



Required samples is small

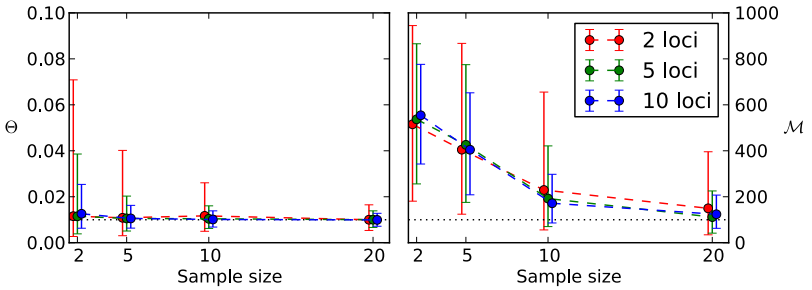
Single population

- The time to the most recent common ancestor is robust to different sample sizes.
- Simulated sequence data from a single population have shown that after 8 individuals you should better add another locus than more individuals.



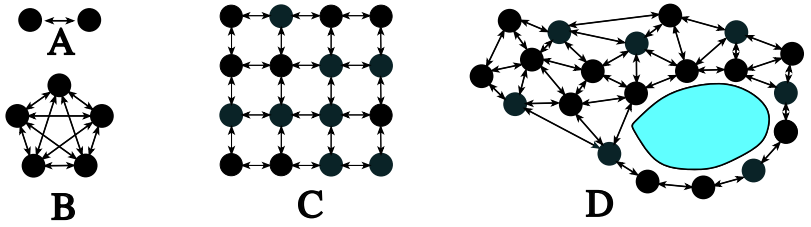
Felsenstein (2005)
Pluzhnikov and Donnelly (1996)

Required samples is small

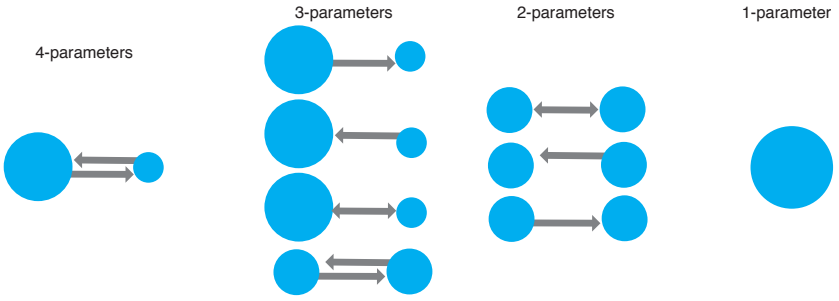


Medium variability DNA dataset: Mutation-scaled population size Θ and mutation-scaled migration rate M versus sample size for 2, 5, and 10 loci. The true $\Theta_T = 0.01$ is marked with the dotted gray line; $M = 100$

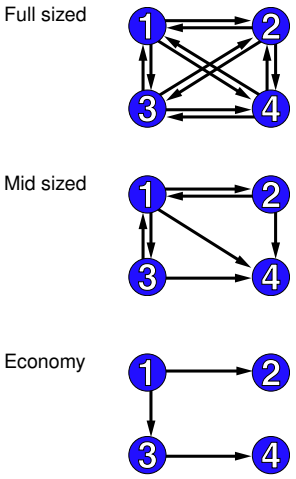
Population models



Migration model specification

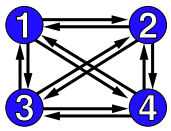


Migration model specification



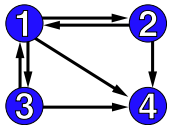
Migration model specification

Full sized



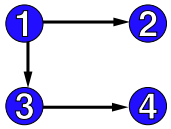
	1	2	3	4
1	-	+	+	+
2	+	-	+	+
3	+	+	-	+
4	+	+	+	-

Mid sized



	1	2	3	4
1	-	+	0	+
2	+	-	0	0
3	+	0	-	0
4	+	+	+	-

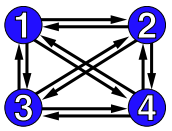
Economy



	1	2	3	4
1	-	0	0	0
2	0	-	0	0
3	+	0	-	0
4	0	+	+	-

Migration model specification

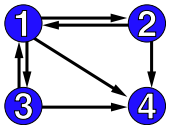
Full sized



	1	2	3	4
1	-	+	+	+
2	+	-	+	+
3	+	+	-	+
4	+	+	+	-

****
****
****
****

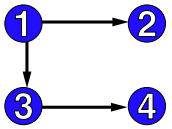
Mid sized



	1	2	3	4
1	-	+	0	+
2	+	-	0	0
3	+	0	-	0
4	+	+	+	-

**0*
**00
*0*0
****

Economy



	1	2	3	4
1	-	0	0	0
2	0	-	0	0
3	+	0	-	0
4	0	+	+	-

*000
0*00
*0*0
0***

Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.
Kass and Raftery (1995) popularized the Bayes Factor as a Bayesian alternative to the LRT.

$$p(M_1|D) = \frac{p(M_1)p(D|M_1)}{p(D)} \qquad \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(D|M_1)}{p(D|M_2)}$$

$$BF = \frac{p(D|M_1)}{p(D|M_2)} \qquad LBF = 2\ln BF = 2\ln\left(\frac{p(D|M_1)}{p(D|M_2)}\right)$$

The magnitude of BF gives us evidence against hypothesis M_2

$$LBF = 2\ln BF = z \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.
Kass and Raftery (1995) popularized the Bayes Factor as a Bayesian alternative to the LRT.

Posterior Density Prior Likelihood

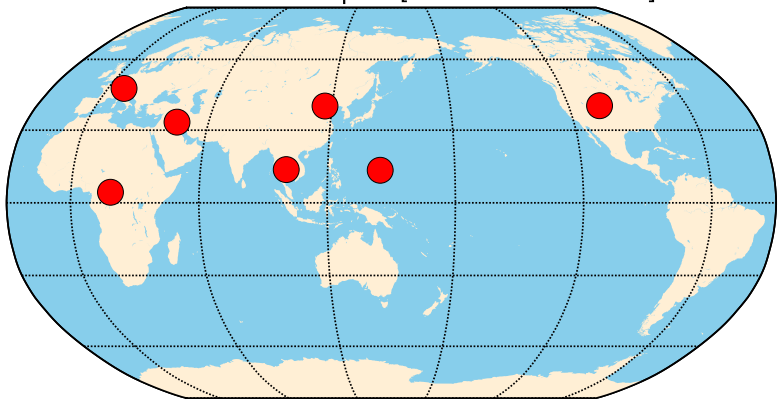
$$p(M_1|D) = \frac{p(M_1)p(D|M_1)}{p(D)} \qquad \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(D|M_1)}{p(D|M_2)}$$

$$BF = \frac{p(D|M_1)}{p(D|M_2)} \qquad LBF = 2\ln BF = 2\ln\left(\frac{p(D|M_1)}{p(D|M_2)}\right)$$

The magnitude of BF gives us evidence against hypothesis M_2

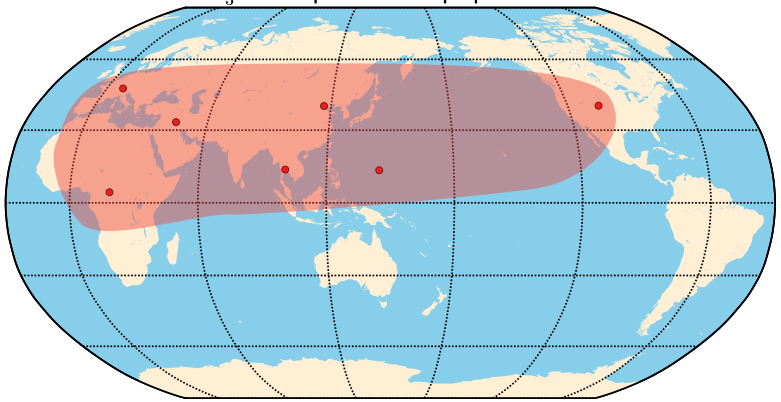
$$LBF = 2\ln BF = z \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

Locations of samples [377 microsatellites]

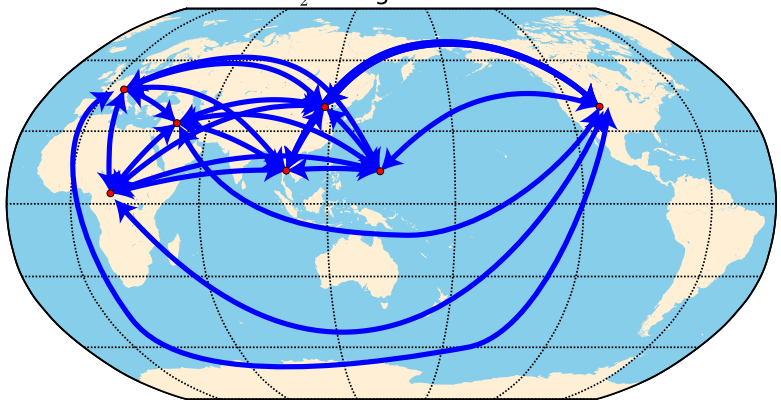


A total of 70 individuals from 7 populations analyzed for 377 microsatellite loci: Mutation model is Brownian motion approximation to the single-step mutation model

H_3 : One panmictic population

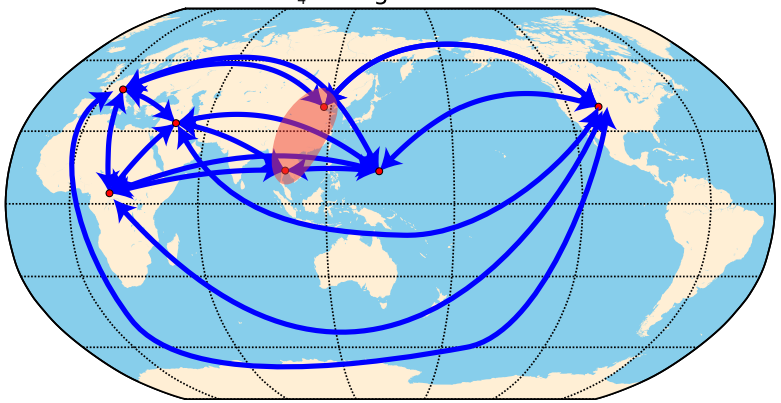


H_2 : Tangled mess

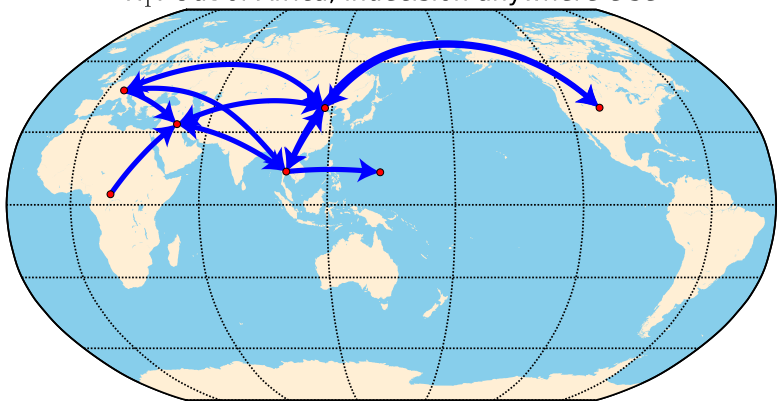


Somewhat less

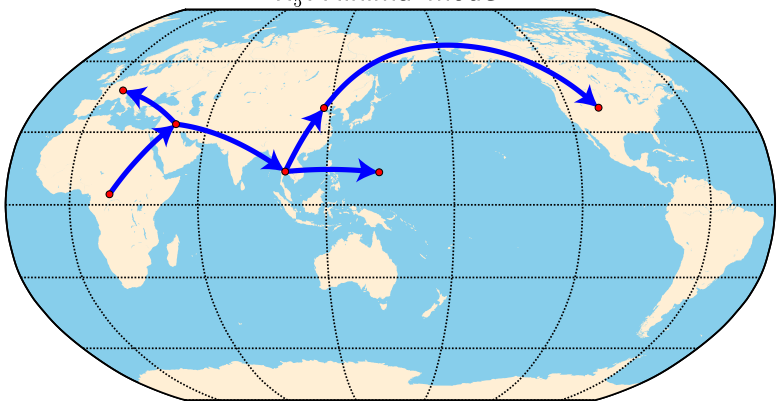
H_4 : Tangled mess



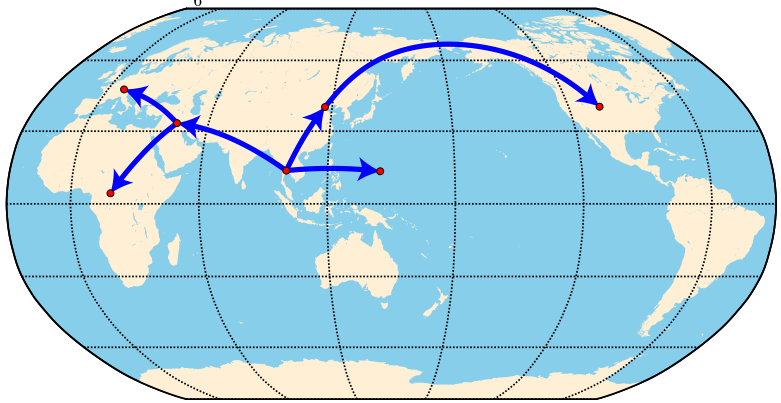
H_1: Out of Africa, indecision anywhere else



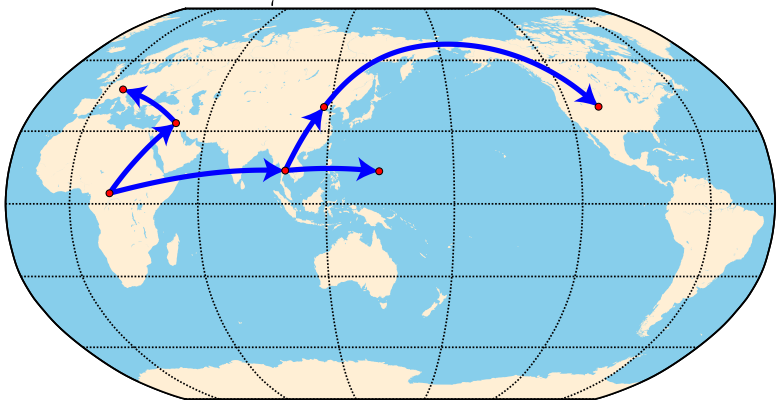
H_5: Minimal model



H_6: South-Asia is cradle of humans



H_7: Direct train to Asia



Structured populations

Model selection

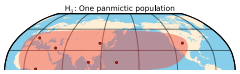
1. 

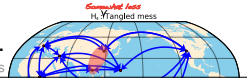
2. 

3. 

4. 

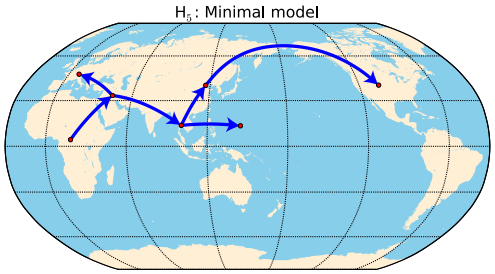
5. 

6. 

7. 

Model order and probability using Bayes factors

all other models: 0.0
Minimal model 1.0



Run time concerns

MCMC

MCMC works perfectly fine when run infinitely long. It is rather difficult to know when the (finite) run has converged and is sampling from the distribution of interest **and** is reaching all important parts. Several methods are used to improve convergence and sampling:

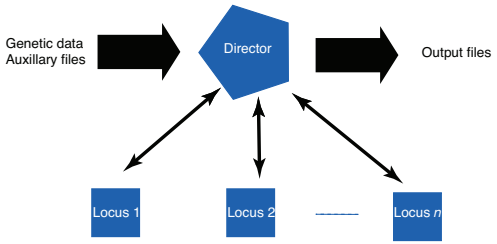
- Improve the proposal procedure
- Use Metropolis-coupled MCMC to improve finding peaks in the distribution.
- Program optimization can improve runtime considerably.
- Run several analyses in parallel

Embarrasingly parallel computation

MIGRATE

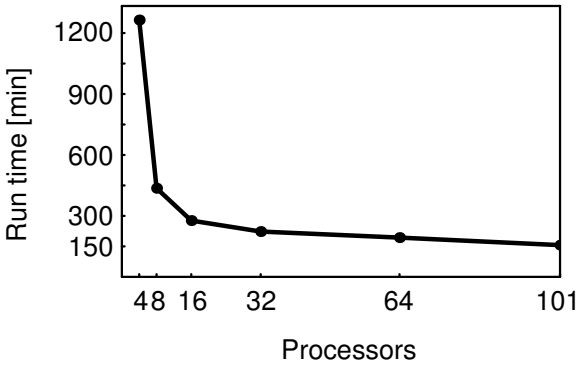
Each locus is completely independent, therefore can run on a different computer. Embarrasingly simple parallel computing can be done by splitting up data set and gathering "results" from individual nodes by "hand". This gets really tedious with 100+ loci.

MIGRATE uses a more sophisticated strategy (MPI) and can use a cluster of (loosely) connected computer nodes. With more loci than nodes a load balancing scheme is used.



Speed up

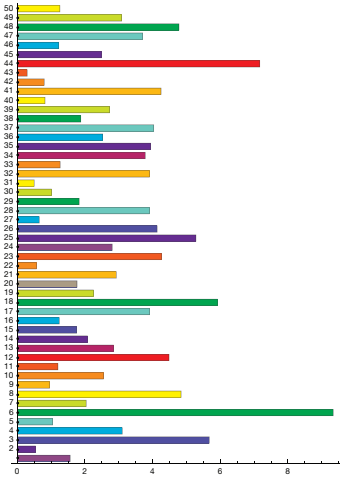
Estimation of 9 parameters in a 3 population migration model using data from a total of 100 loci, distributed over 4, 8, 16, 32, 64, 101 computer nodes.



Speed up

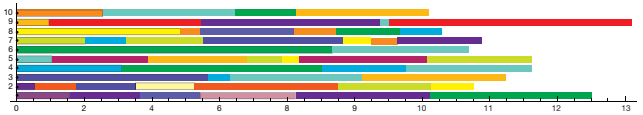
Speed of total run depends on the “slowest” locus (here out of 50)

51 nodes

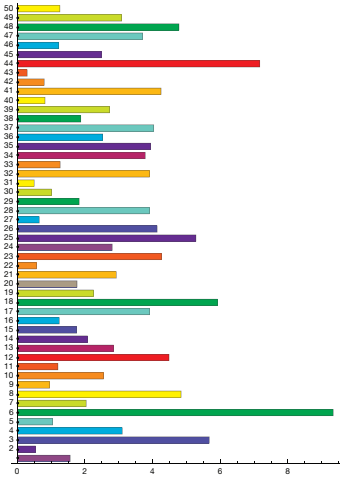


Speed up

11 nodes



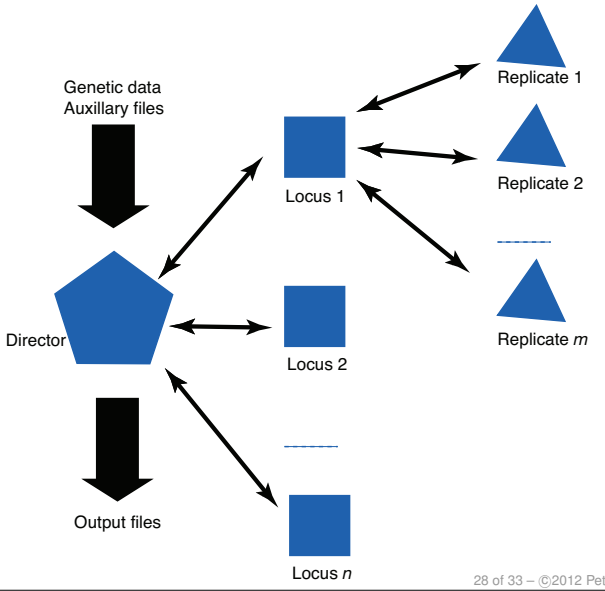
51 nodes



Load-balancing allows more effective use of all available compute nodes.

Speed up even more?

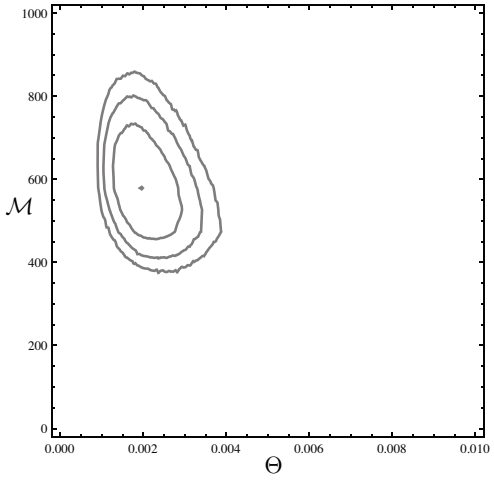
MIGRATE



Run time versus accuracy

One long run

Posterior density for a 2-parameter model (population size and gene flow) A run for 50×10^6 steps (sampling 3 quantities: 2 parameters and genealogies) took about 20 hours.



One long run:
all samples used (no burn-in)

Contour lines are at 50%, 95%, and 99% credibility level

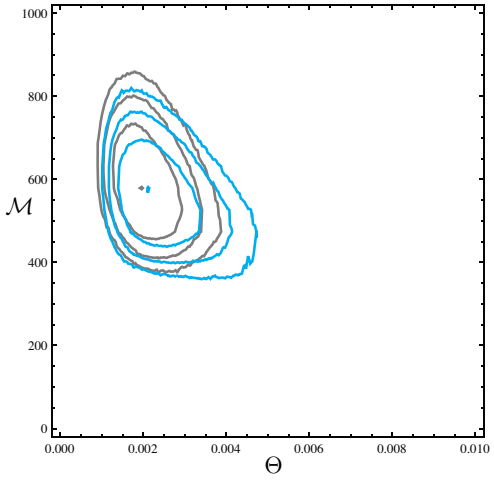
$\Theta = 4N_e\mu$ (population size scaled by mutation rate)

$\mathcal{M} = \frac{m}{\mu}$ (immigration rate scaled by mutation rate)

Run time versus accuracy

0 replicated runs

Posterior density for a 2-parameter model (population size and gene flow) 10 runs each for 5×10^6 steps took about 2 hours.



One long run: 20 hours
all samples used (no burn-in)
10 replicates: 2 hours
all samples used (no burn-in)

Contour lines are at 50%, 95%, and 99% credibility level

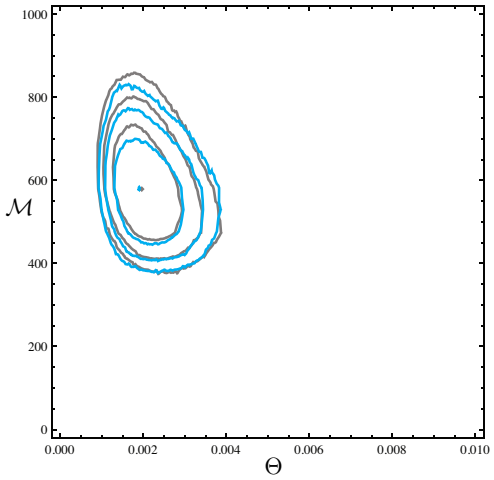
$\Theta = 4N_e\mu$ (population size scaled by mutation rate)

$\mathcal{M} = \frac{m}{\mu}$ (immigration rate scaled by mutation rate)

Run time versus accuracy

0 replicated runs*

Posterior density for a 2-parameter model (population size and gene flow) 10 runs each for 5×10^6 steps took about 2 hours.



One long run: 20 hours
all samples used (burn-in)
10 replicates: 2 hours
first 50% of samples discarded

Contour lines are at 50%, 95%, and 99% credibility level

$\Theta = 4N_e\mu$ (population size scaled by mutation rate)

$\mathcal{M} = \frac{m}{\mu}$ (immigration rate scaled by mutation rate)


Learn a computer scripting language today to be ready for tomorrow, the parallel genome sequencing revolution has begun.