# Replicating Markov chain Monte Carlo runs

**Peter Beerli**
School of Computational Science, and
Department of Biological Sciences,
Florida State University, Tallahassee

# Inference of parameters using MCMC

Many modern methods in population genetics and phylogenetics use the Markov chain Monte Carlo (MCMC) method to approximate likelihoods or posterior probabilities

$$\mathrm{Prob}(\mathcal{P}|\mathrm{Data}) = \left(\frac{\mathrm{Prob}(\mathcal{P})}{\mathrm{Prob}(\mathrm{Data})}\right) \int_{\mathrm{objects}} f(\mathrm{object}|\mathcal{P})\mathrm{Prob}(\mathrm{Data}|\mathrm{object})d\mathrm{objects}$$

with

$$\mathrm{Prob}(\mathcal{P}|\mathrm{Data}) \simeq \left(\frac{\mathrm{Prob}(\mathcal{P})}{\mathrm{Prob}(\mathrm{Data})}\right) \frac{1}{n} \sum_{i}^{n} \frac{f(\mathrm{object_i}|\mathcal{P})}{f(\mathrm{object_i}|\mathcal{P}_0)}$$

with the driving parameters $\mathcal{P}_0$, this driving parameter is updated frequently in Bayesian analyses, but typically only few times in a maximum likelihood analysis.

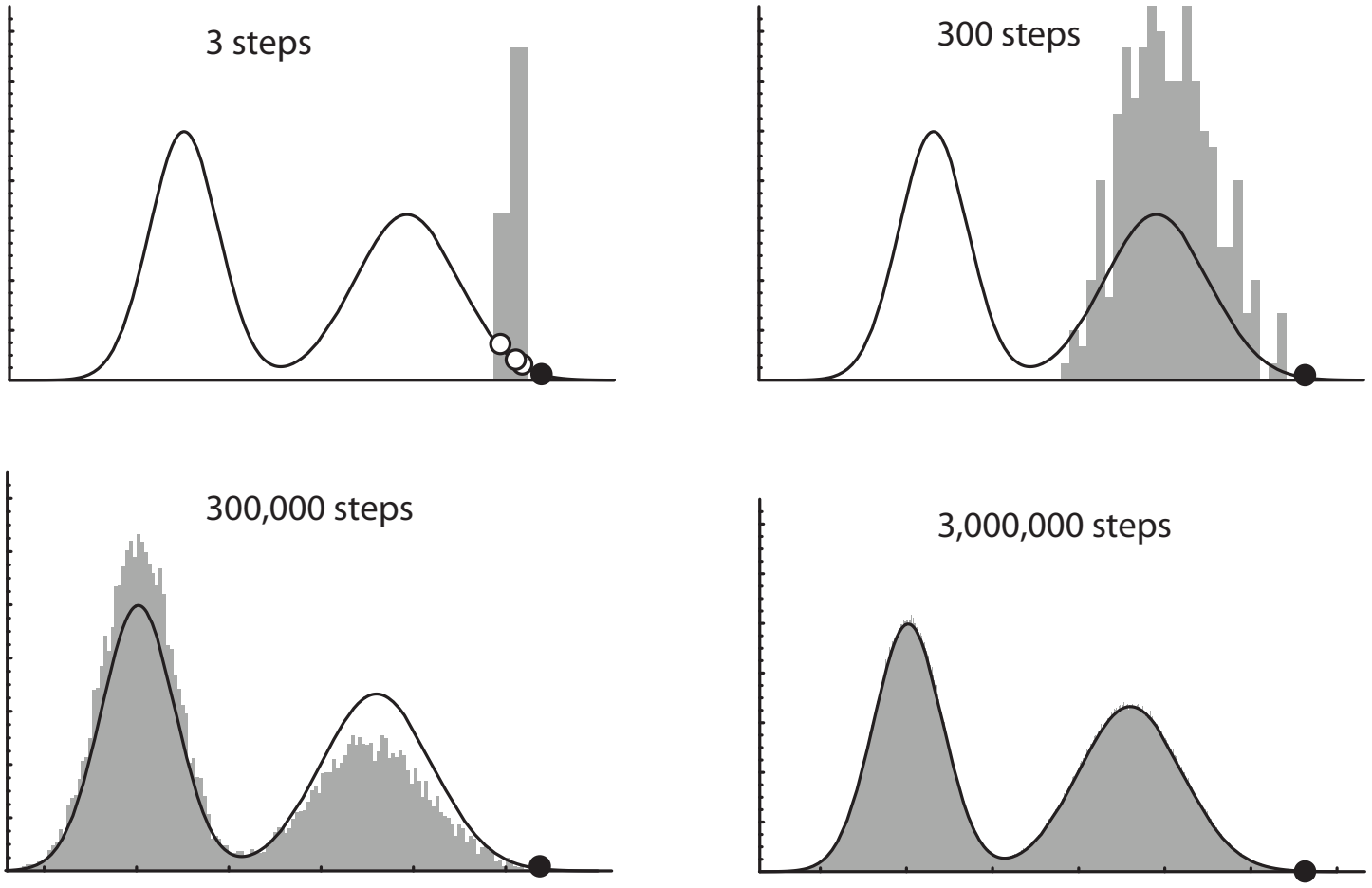
# How long to run?

$$\text{Prob}(\text{Data}|\mathcal{P}) \simeq \frac{1}{n}\sum_{i}^{n}\frac{f(\text{object}_\text{i}|\mathcal{P})}{f(\text{object}_\text{i}|\mathcal{P}_0)}$$

# How long to run?



$$\text{Prob}(\text{Data}|\mathcal{P}) \simeq \frac{1}{n}\sum_{i}^{n}\frac{f(\text{object}_\text{i}|\mathcal{P})}{f(\text{object}_\text{i}|\mathcal{P}_0)}$$

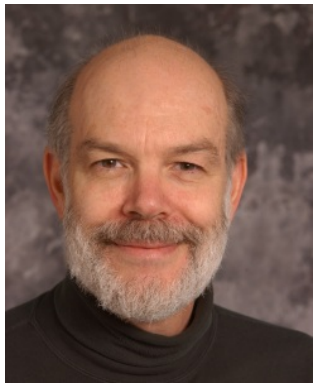


# Diagnostics

Diagnostics are often not all that useful because they only highlight the most crass errors. In my gene flow estimator MIGRATE-N, the Gelman-Rubin statistic is available. It monitors convergence by comparing expected variances within and between replicated chains that start from over-dispersed starting points. On moderately long runs convergence is not all that difficult to get for most of the parameters estimated.

Disclaimer: Do not trust any diagnostic, it will only show the worst. Getting experience with MCMC and your data is invaluable. Trace plots alone are often too optimistic about convergence.

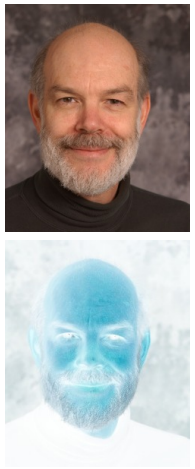
# One long run versus short runs?



Charles Geyer: *"If you can't get a good answer with one long run, then you can't get a good answer with many short runs either."* [http://www.stat.umn.edu/ charlie/mcmc/one.html; June 21 2006]

# Life is short





Charles Geyer: *"If you can't get a good answer with one long run, then you can't get a good answer with many short runs either."* [http://www.stat.umn.edu/ charlie/mcmc/one.html; June 21 2006]

¬(Charles Geyer): *"If, in principle, you can get a good answer with one long run, then perhaps you can also get a good answer with several shorter runs." [???????]*

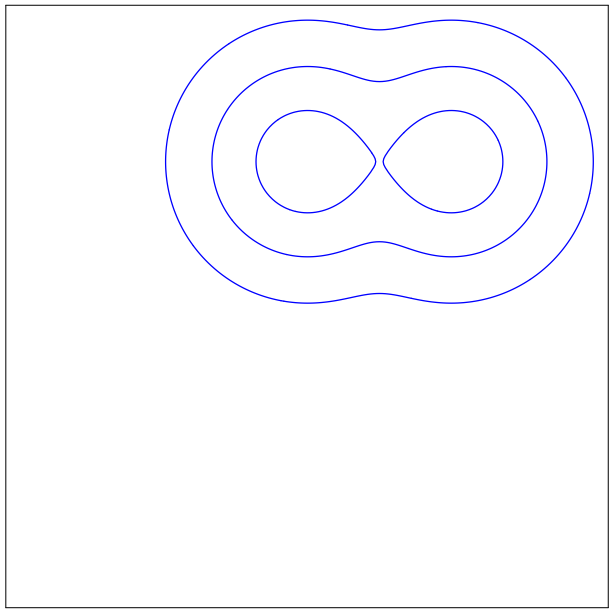
For complex problem long runs are difficult:

- Computers may fail after many hours, days, or month of operations.
- Partitioning a single long chain over multiple computers is inefficient (except for some version of heating [MC$^3$]).
- We are impatient.

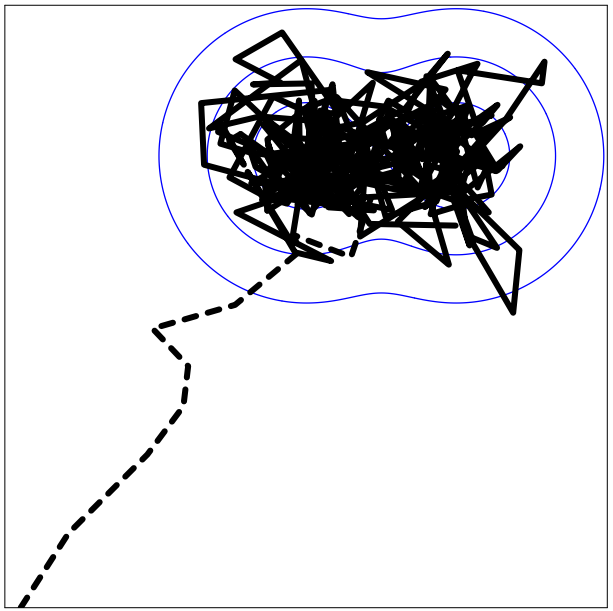
Summary: Either our life, our attention span, or most likely our computer's life is short.

# Short versus long runs

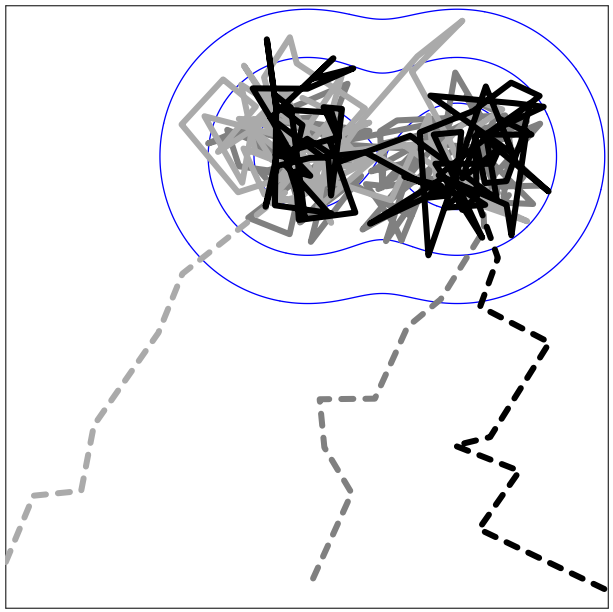
Truth



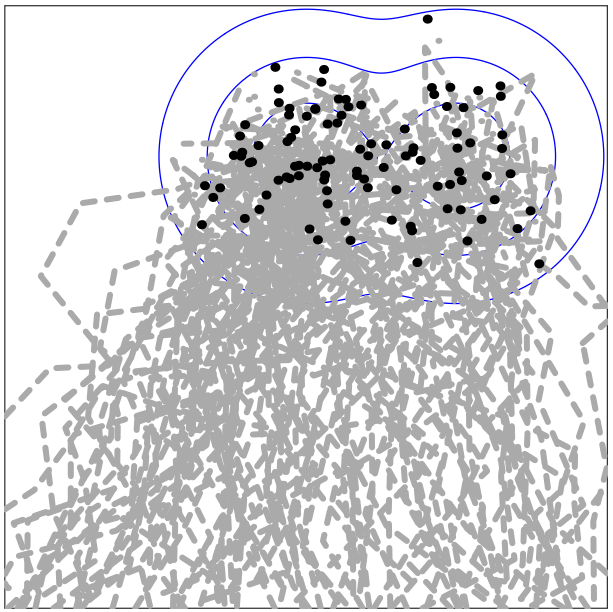
"Long" run



3 "Short" run
[start all low]

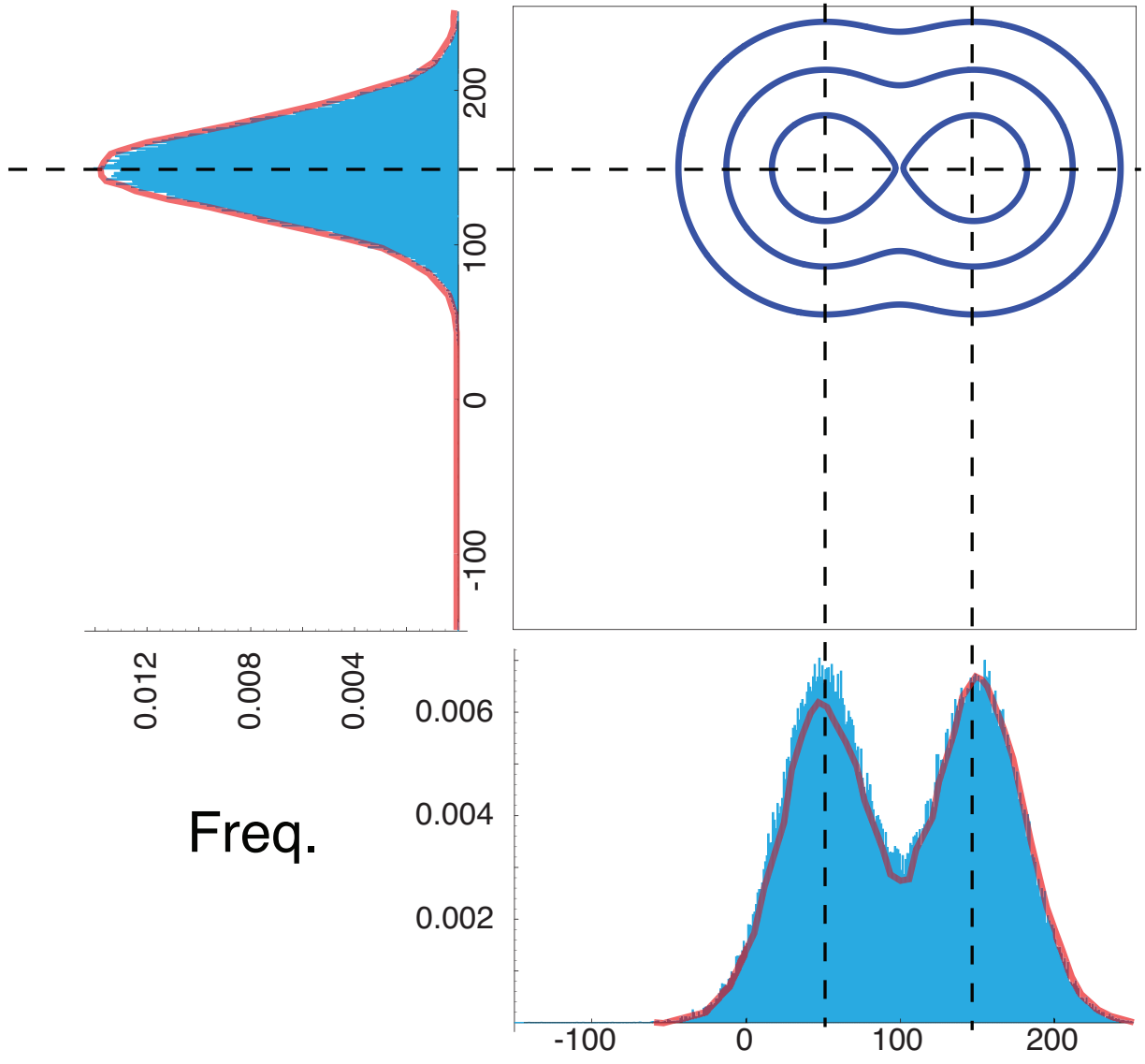


100 short runs
[last sample]

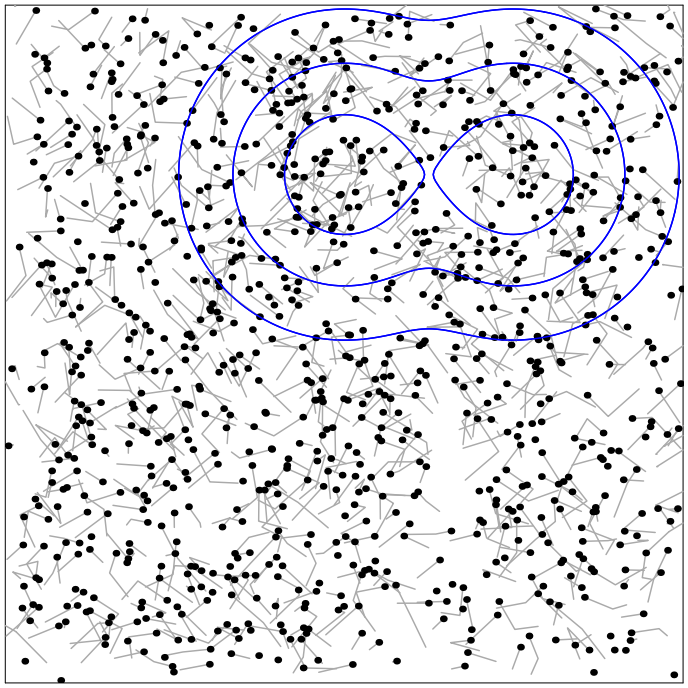


# Long run

One long chain will sample from the target distribution and deliver a decent estimate, independent on the burn-in [red line shows histogram without burn-in]

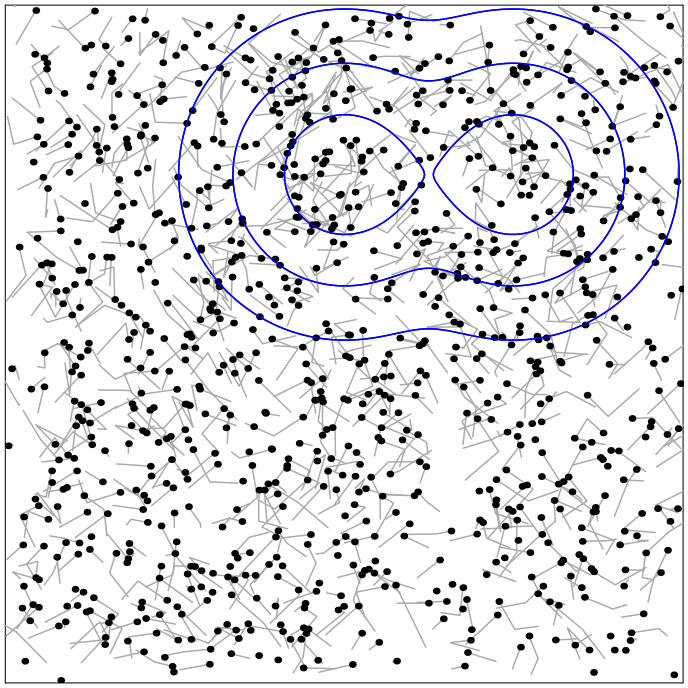


# Many short runs

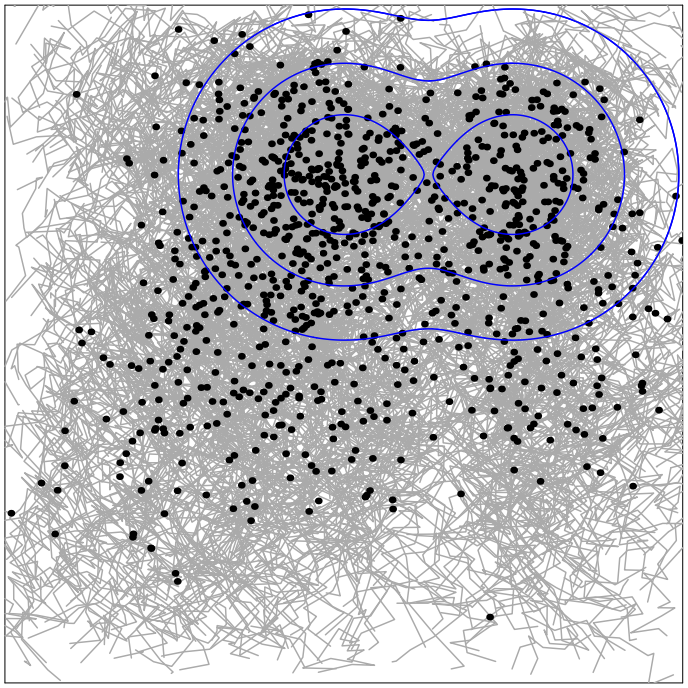


Each replicate is 3 steps

# Many short runs

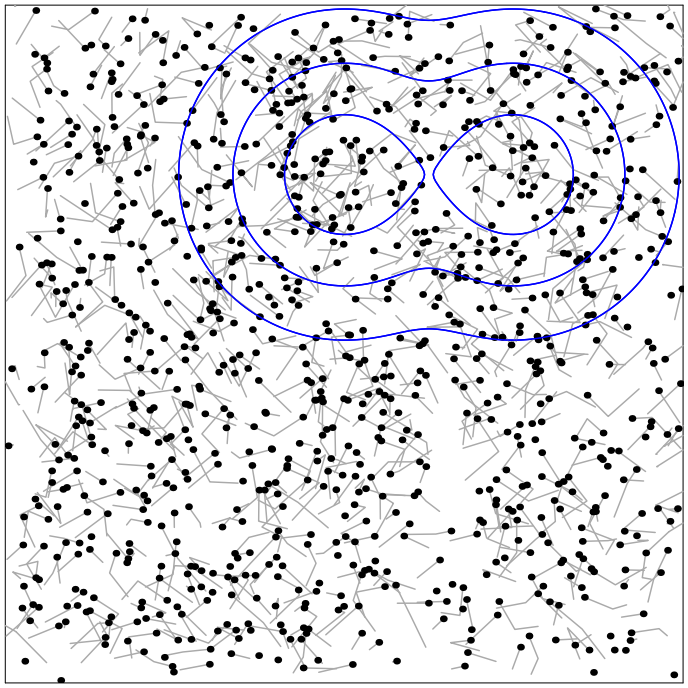


Each replicate is 3 steps

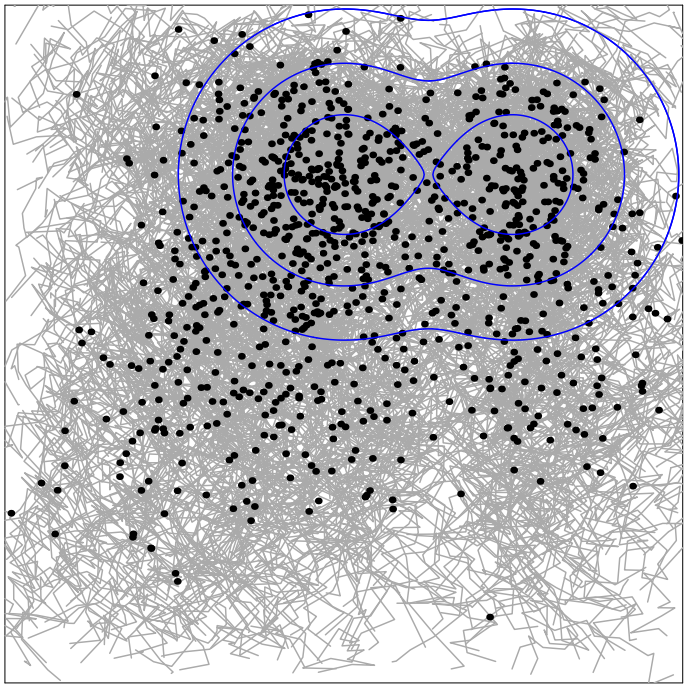


Each replicate is 56 steps

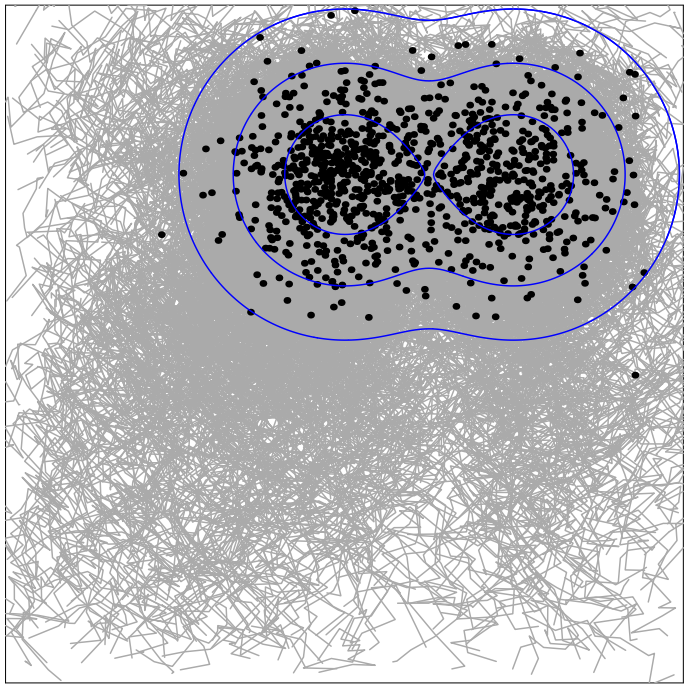
# Many short runs



Each replicate is 3 steps



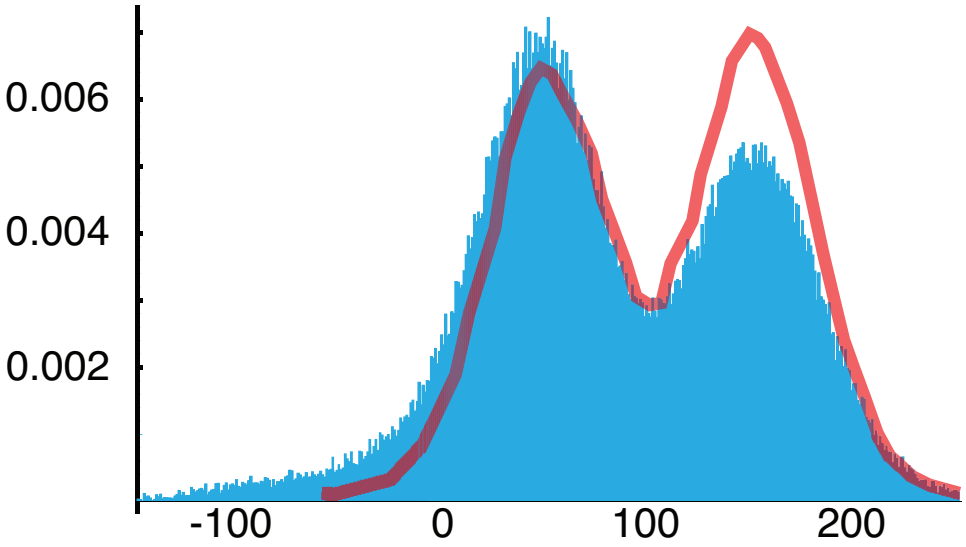
Each replicate is 56 steps



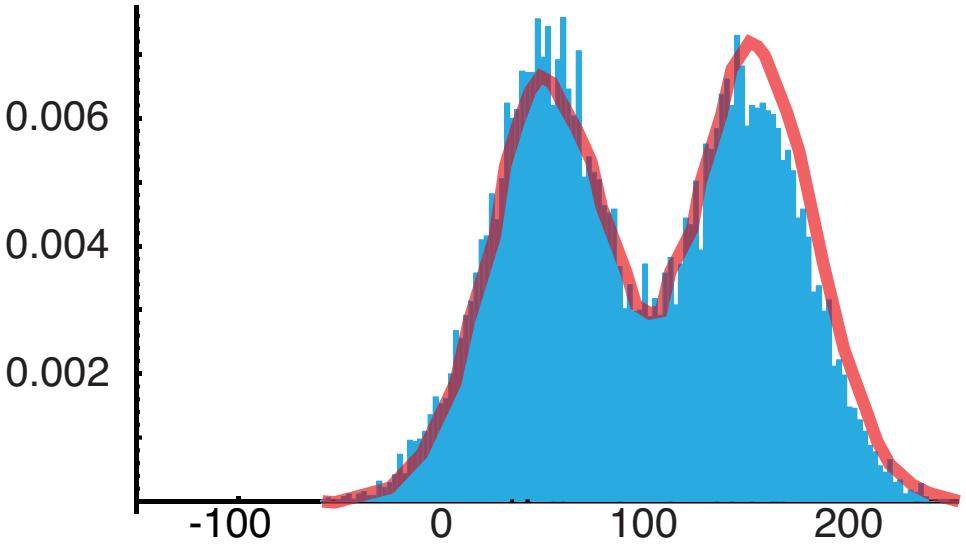
80 or more steps

# Many not so long runs

Each replicate will sample after some time from the target distribution. Once we estimated how many steps we need to reach the target distribution, an accumulation of short (but long enough) runs will do almost as good a job as a single long run if we discard some burn-in period.




No burn-in: 200 samples per replicate




Burn-in: 20 samples per replicate

Red line is marginal (from long-run slide) using one long chain without burn-in.

# Sooo.....

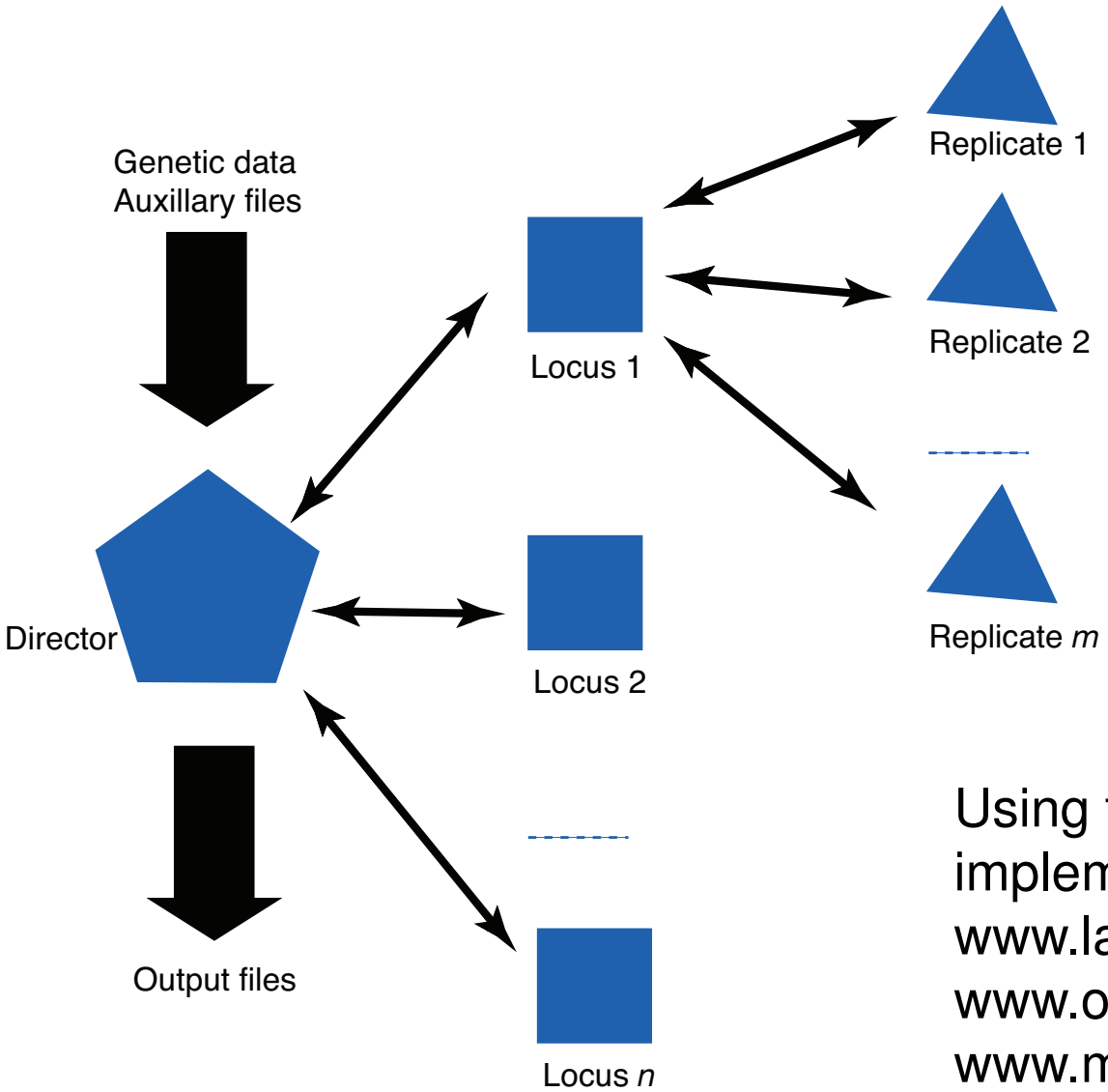
We just found out what Charles Geyer was saying a long time ago, that a single long chain is as good or better than many shorter chains, and also better than the accumulation of many shorter chains.

For toy cases there is no difficulty to run long runs, but what if we need to run a job for more than many million steps or genealogies or phylogenetic trees, ....? A single long chain will often not work.

If we think that a long chain will take weeks then we can break it up into shorter tasks and break it up (divide and conquer)

# Strategy in MIGRATE-N

Run program on a computer-cluster (dedicated or your networked lab-computers) using the standard Message Passing Interface (MPI).




Using free MPI implementations, for example: www.lam-mpi.org, www.openmpi.org, or www.mcs.anl.gov/mpi/mpich2

# Timings

Green crabs:

4 populations (total 310 samples), 1 locus, estimating 2 parameters

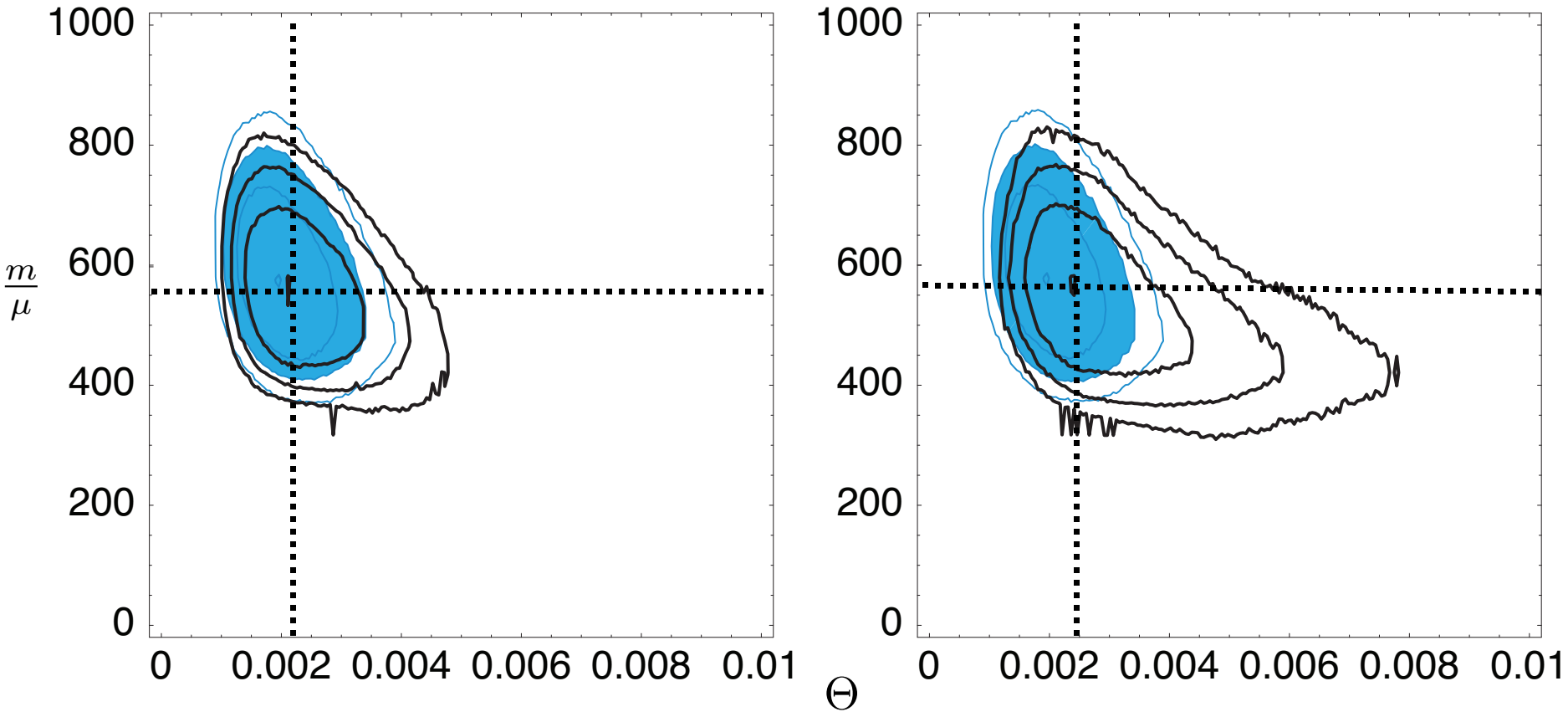
| Visited $[10^6]$ | Replicates | Recorded $[10^6]$ | CPUs [+director] | Wall-clock time | Speed gain |
|---|---|---|---|---|---|
| 50 | 1 | 1 | 1 | 20 : 26 : 34 | 1 |
| 50 | 10 | 1 | 11 | 2 : 15 : 03 | 9.1 |
| 50 | 50 | 1 | 51 | $0 : 51 : 06^{(1)}$ | 24.0 |
| 50 | 100 | 1 | 51 | 0 : 39 : 21 | 31.2 |
| 50 | 1000 | 1 | 51 | 0 : 46 : 20 | 26.5 |
| 200 | 10 | 10 | 11 | 10 : 55 : 45 | $\sim 8^{(2)}$ |

$^{(1)}$some other processes were also running on the cluster

$^{(2)}$16 parameters estimated instead of 2.

# Accuracy

Posterior densities for a 2-parameter model (size and gene flow) for two different runs (one with 10 replicates and one with 50, visiting the same number of trees and parameters. For comparison:Blue credibility area and contours are from results of one single long run ($50 \times 10^6$ steps).




10 replicates = $50 \times 10^6$ steps

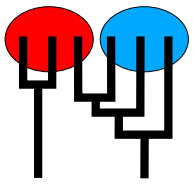
50 replicates = $50 \times 10^6$ steps

# Conclusions

- Charles Geyer is correct.
- Accumulation of short (too short) runs returns worse samples than a long run.
- Many runs that have a large burn-in period that needs to be discarded are almost as good as a single long run.
- If you have more than one computer use them and run multiple long replicates in parallel.

# Thanks + Program

- Charles Geyer for clear and unambiguous statements on MCMC
- Dave Swofford for critical questions
- Joe Roman gave me his green crab data set
- Research is supported by the joint NSF/NIGMS Mathematical Biology program with NIH grant R01 GM 078985.



download MIGRATE-N from

`http://popgen.scs.fsu.edu`