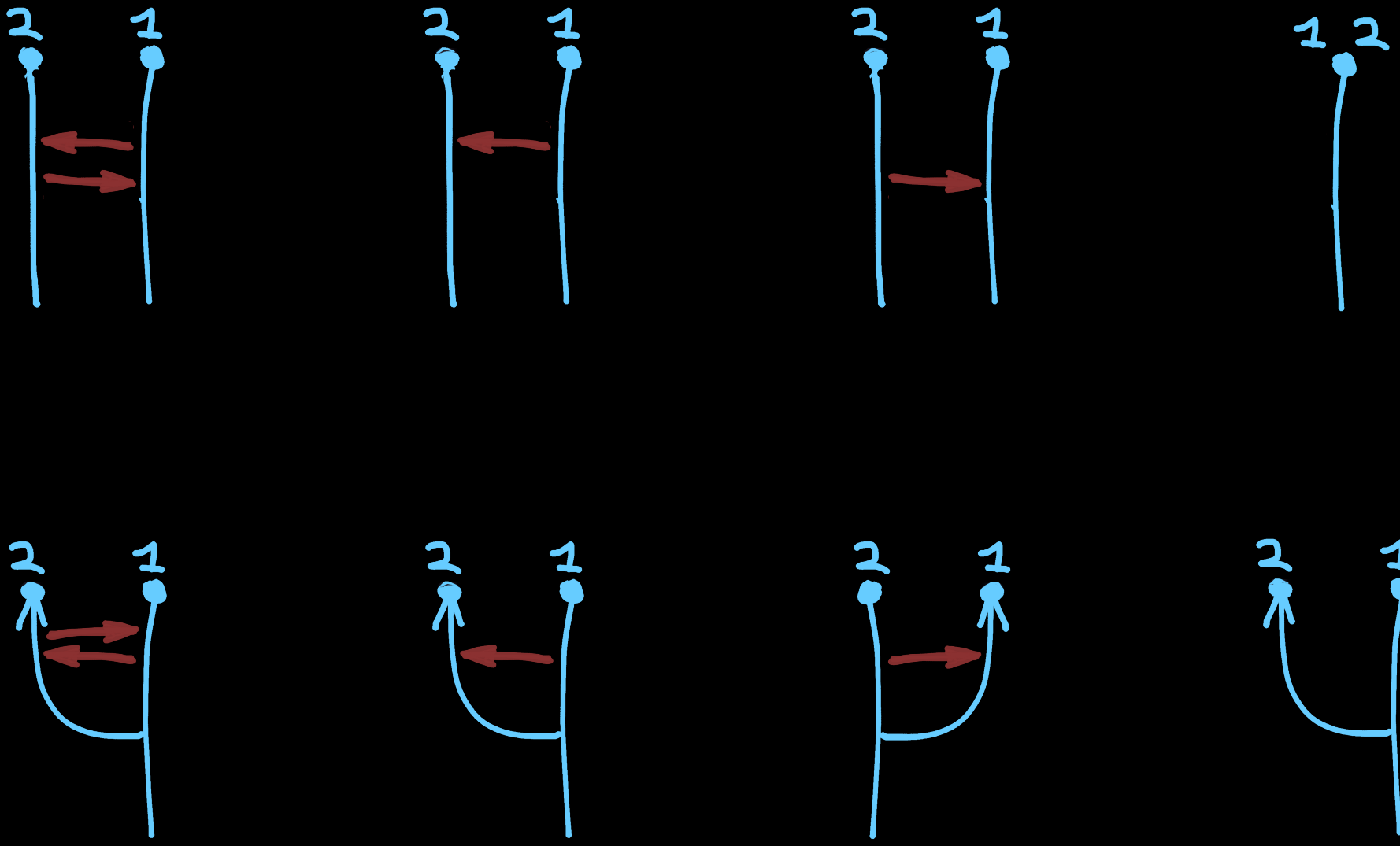
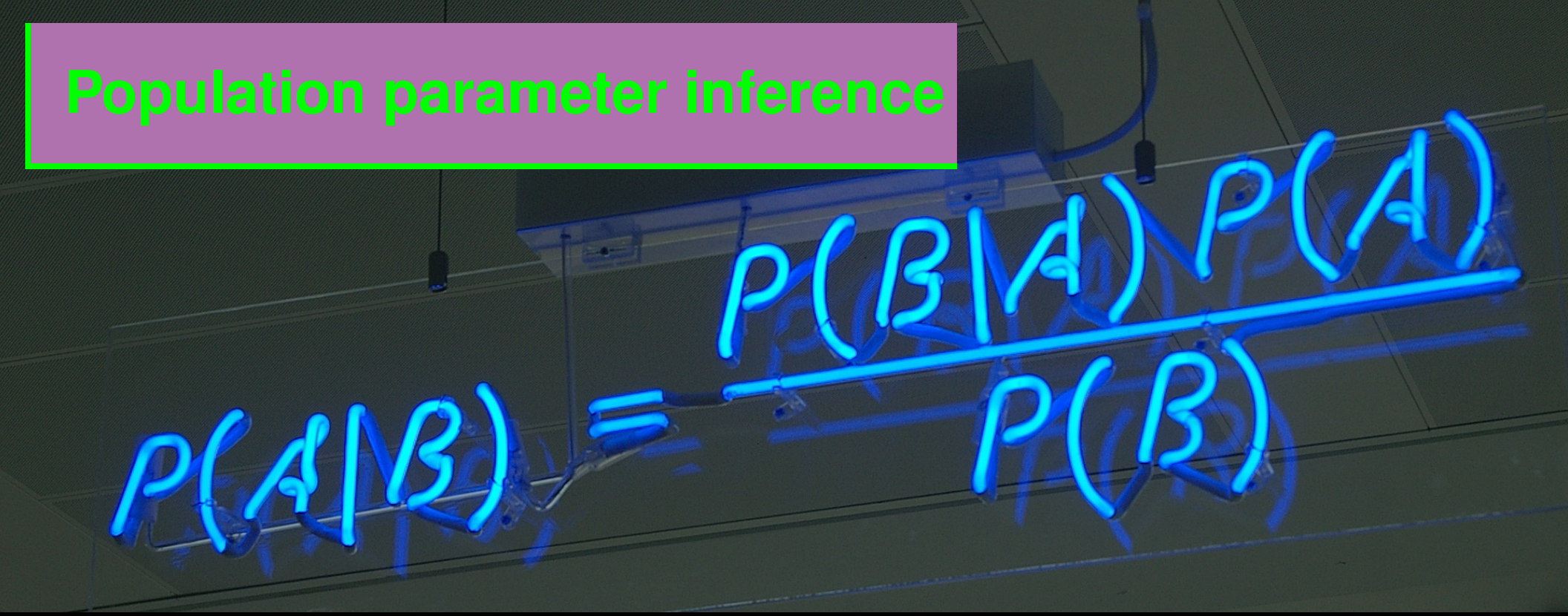
Fun with population models or Goldilocks' principle

Peter Beerli, Scientific Computing, FSU
Twitter: @peterbeerli

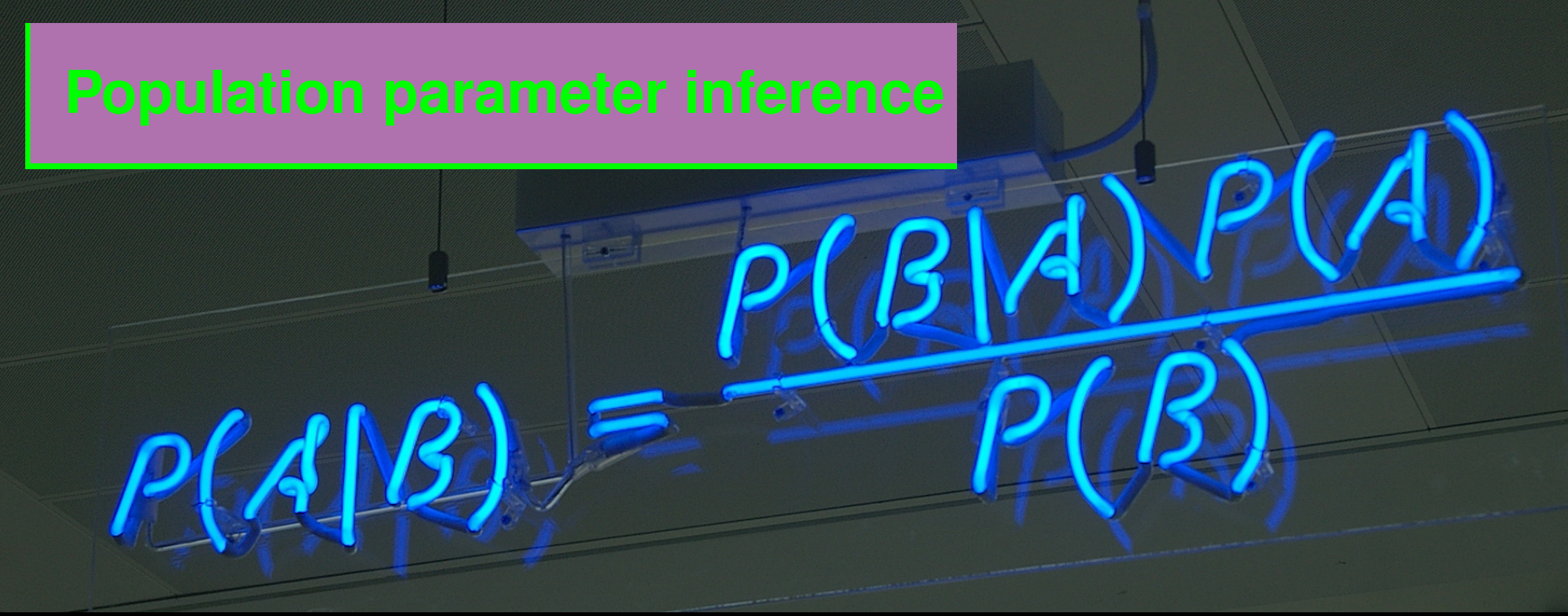
Population models

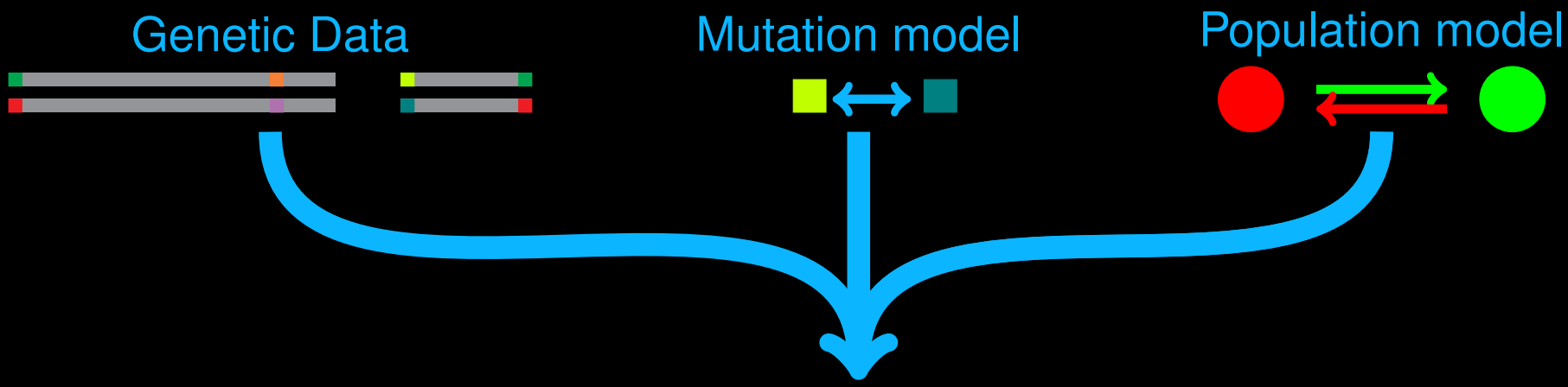


Population parameter inference



Population parameter inference





Population parameter inference

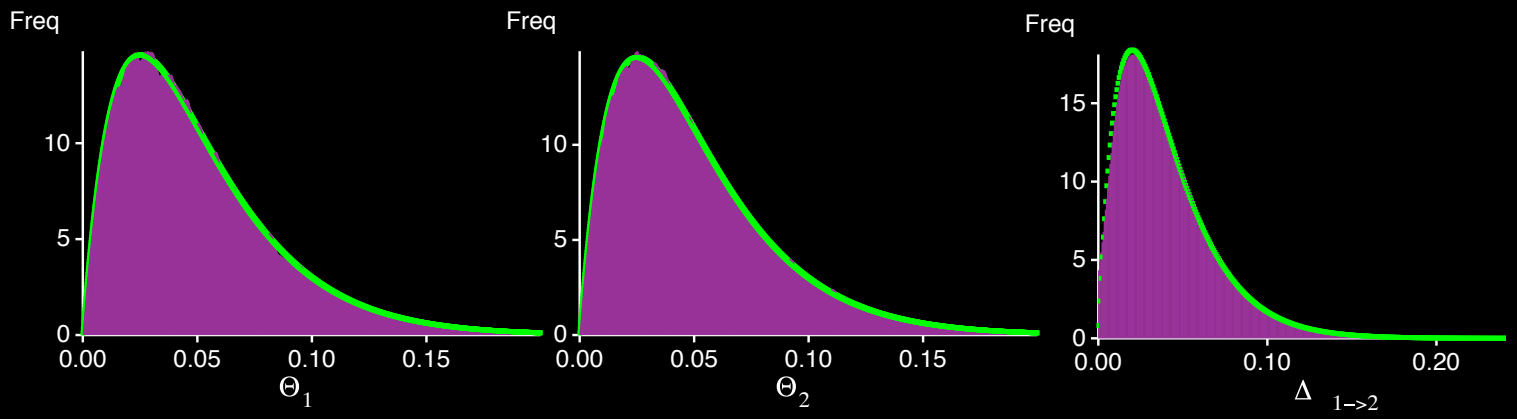
We can analyze sequence data D using a particular model M and can get answers in form of posterior distributions of the parameters of the model:

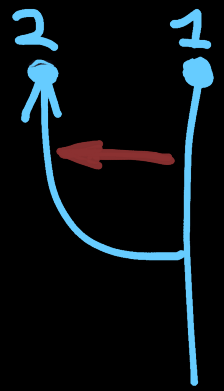
$$P(\boldsymbol{\Theta}|D,\mathbf{M}) = \frac{P(\boldsymbol{\Theta}|\mathbf{M})P(D|\boldsymbol{\Theta},\mathbf{M})}{\int_{\boldsymbol{\Theta}} P(\boldsymbol{\Theta}|\mathbf{M})P(D|\boldsymbol{\Theta},\mathbf{M})d\boldsymbol{\Theta}}$$

Population parameter inference

We can analyze sequence data D using a particular model M and can get answers in form of posterior distributions of the parameters of the model:

$$P(\mathbf{\Theta}|D,\mathbf{M}) = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{\int_{\mathbf{\Theta}} P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})d\mathbf{\Theta}}$$



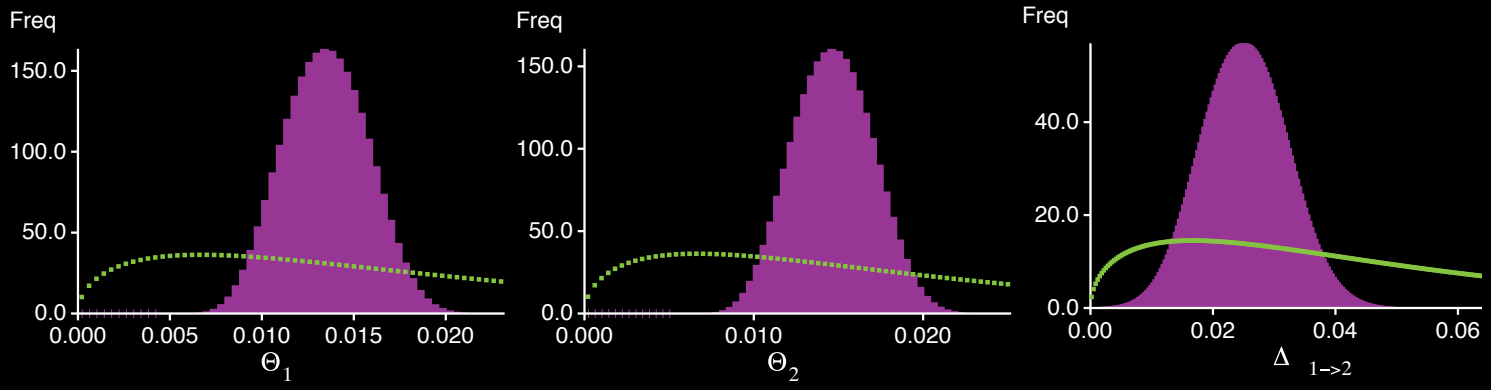


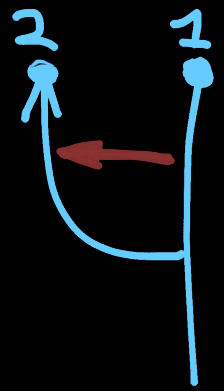
Short run without any data using Gamma distributed priors.

Population parameter inference

We can analyze sequence data D using a particular model M and can get answers in form of posterior distributions of the parameters of the model:

$$P(\mathbf{\Theta}|D,\mathbf{M}) = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{\int_{\mathbf{\Theta}} P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})d\mathbf{\Theta}}$$





Short run with data.

Comparing models

Bayes theorem:

$$P(\mathbf{\Theta}|D,\mathbf{M}) = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{\int_{\mathbf{\Theta}} P(\mathbf{\Theta}|D,\mathbf{M})d\mathbf{\Theta}} = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{P(D|M)}$$

Comparing models

We use marginal likelihoods
(in practice, this is the denominator of Bayes formula).

Bayes theorem:

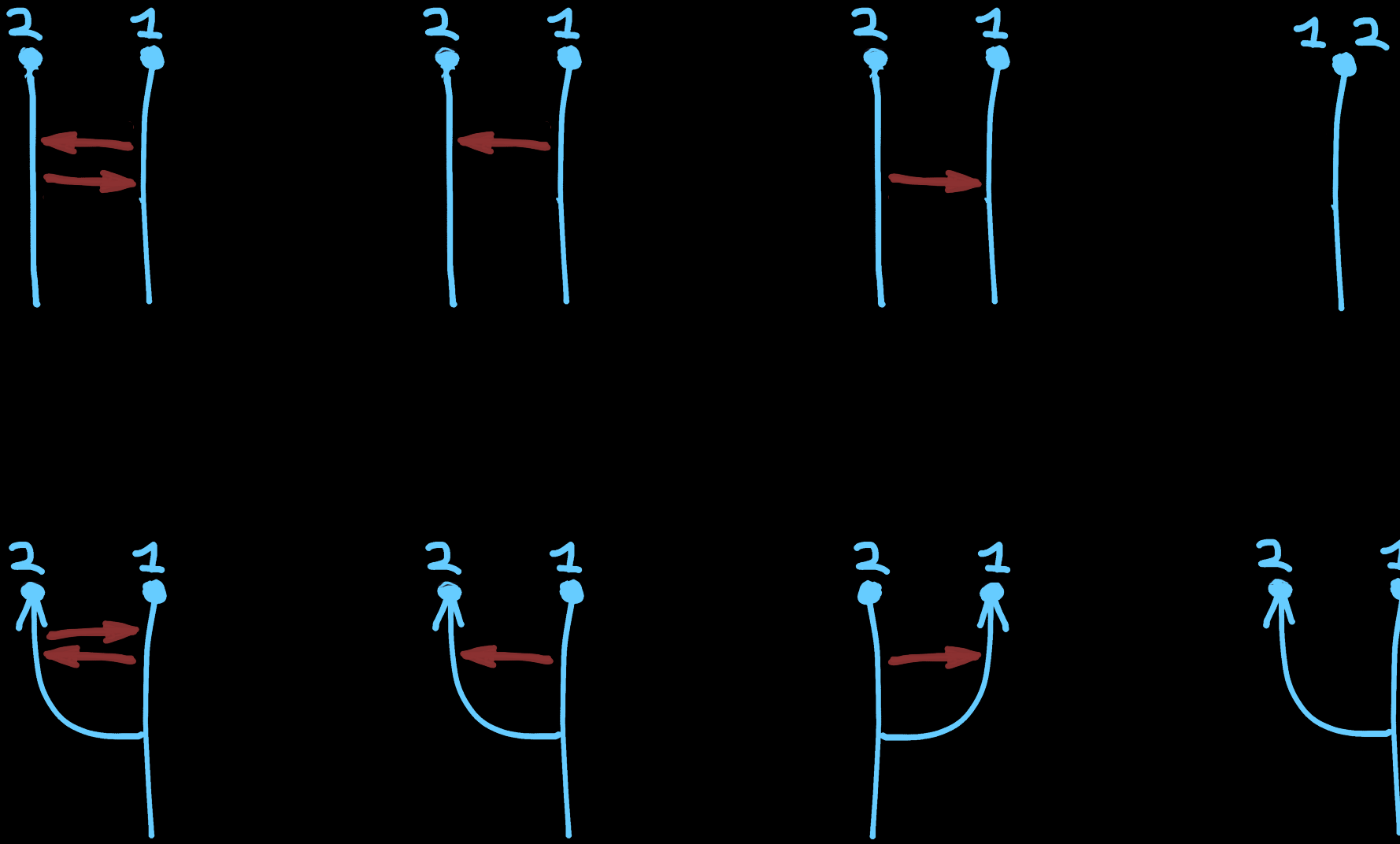
$$P(\mathbf{\Theta}|D,\mathbf{M}) = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{\int_{\mathbf{\Theta}} P(\mathbf{\Theta}|D,\mathbf{M})d\mathbf{\Theta}} = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{P(D|M)}$$

Solving for the marginal likelihood:

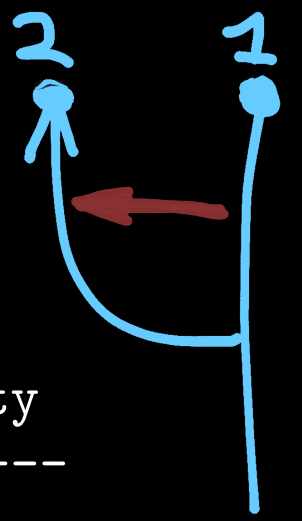
$$P(D|\mathbf{M}) = \frac{P(\mathbf{\Theta}|\mathbf{M})P(D|\mathbf{\Theta},\mathbf{M})}{P(\mathbf{\Theta}|D,\mathbf{M})}$$

In Markov chain Monte Carlo applications this is tricky, because we do not calculate the $P(D|M)$ directly, but approximate using thermodynamic integration.

Population models



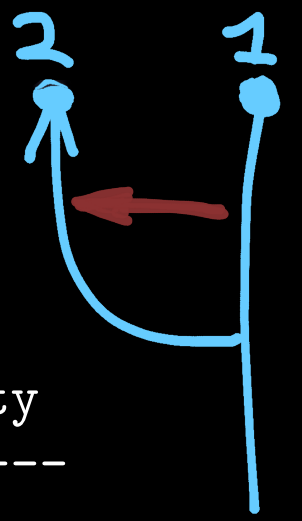
Simulated data



Two loci simulated from model `x0Dx`:

Model	Log(mL)	LBF	Model-probability
1: xxxx:	-9662.42	-23.73	0.0000
2: xDxx:	-9661.98	-23.29	0.0000
3: xxDx:	-9661.52	-22.83	0.0000
4: xd0x:	-9656.51	-17.82	0.0000
5: xD0x:	-9649.33	-10.64	0.0000
6: xx0x:	-9648.93	-10.24	0.0000
7: x0dx:	-9641.77	-3.08	0.0402
8: x0xx:	-9641.01	-2.32	0.0859
9: x0Dx:	-9638.69	0.00	0.8739

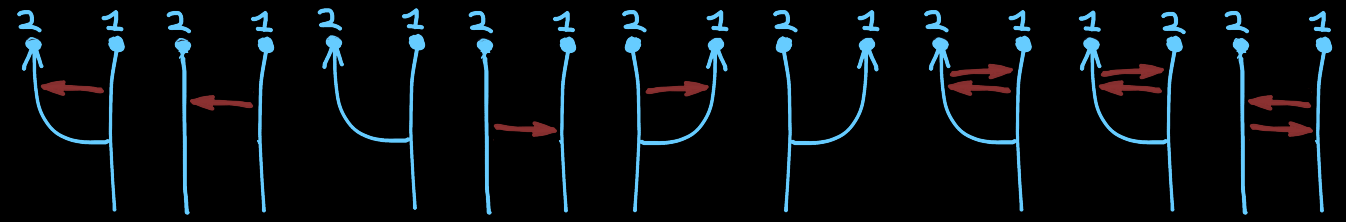
Simulated data



Two loci simulated from model `x0Dx`:

Model	Log(mL)	LBF	Model-probability
1: xxxx:	-9662.42	-23.73	0.0000
2: xDxx:	-9661.98	-23.29	0.0000
3: xxDx:	-9661.52	-22.83	0.0000
4: xd0x:	-9656.51	-17.82	0.0000
5: xD0x:	-9649.33	-10.64	0.0000
6: xx0x:	-9648.93	-10.24	0.0000
7: x0dx:	-9641.77	-3.08	0.0402
8: x0xx:	-9641.01	-2.32	0.0859
9: x0Dx:	-9638.69	0.00	0.8739

Best



Worst

A real example

1 Texas was a refugium from which populations expanded northward into other regions.
2 *P. illinoensis* is derived from *P. streckeri* that expanded through the Arkansas River Valley.
3 There is detectable genetic structure within *P. illinoensis* consistent with the disjunct range.

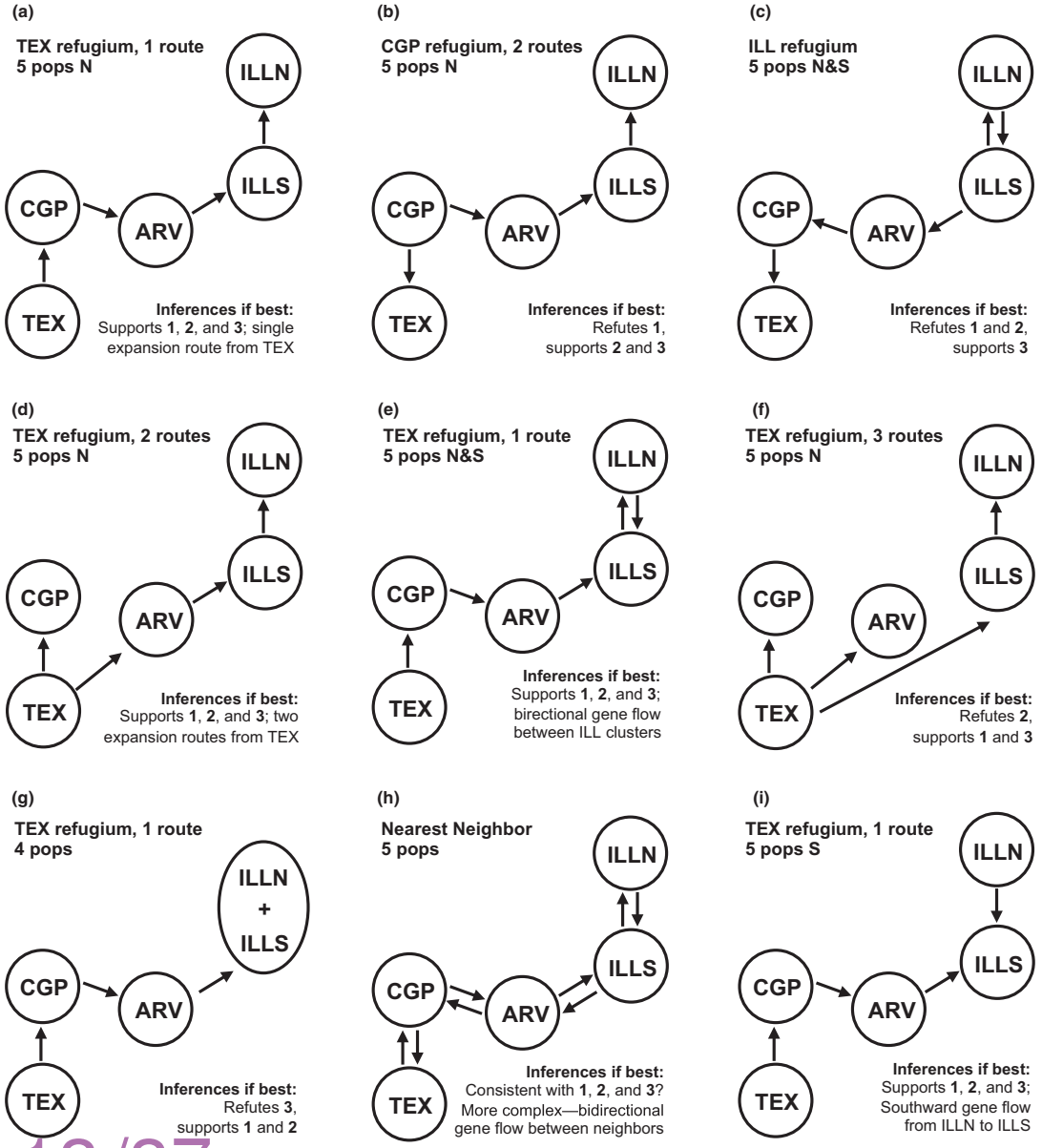
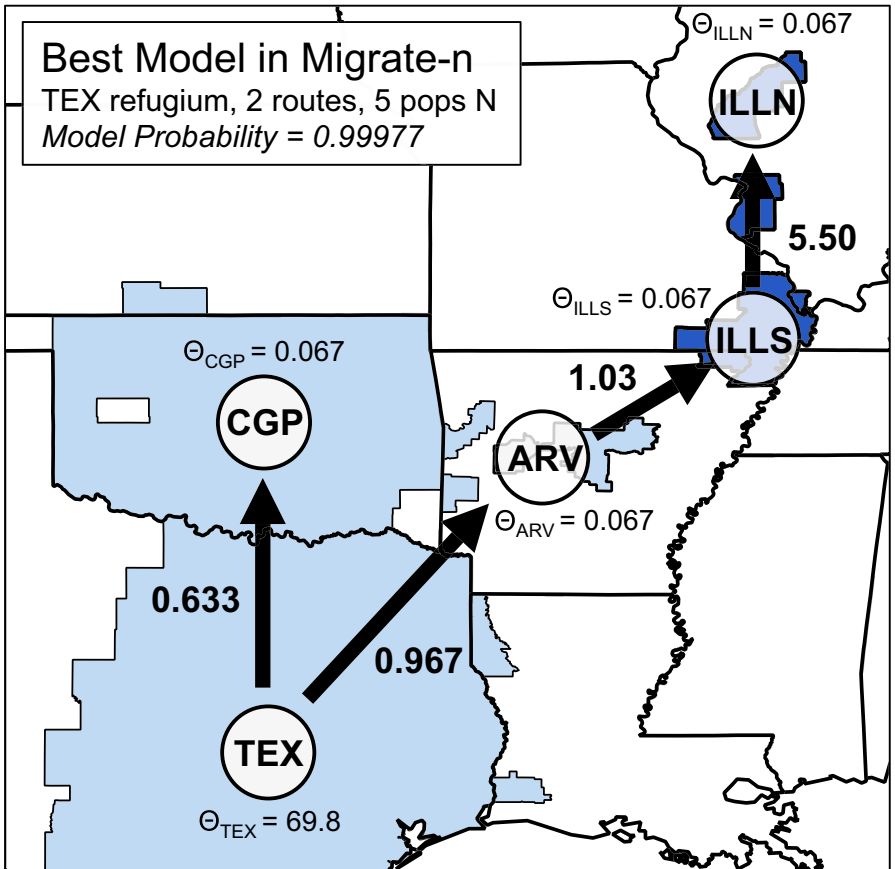


Fig. 2 Example models compared in Migrate-n to make inferences regarding the hypotheses described (1–3). Each circle represents an effective population size parameter, and each arrow represents a migration parameter. Region labels as in Table 1. Note that only a subset of the models tested is illustrated; all 24 tested models are shown in Data S1.

Lisa N. Barrow, Alyssa T. Bigelow, Christopher A. Phillips, and Emily Moriarty Lemmon (2015) Phylogeographic inference using Bayesian model comparison across a fragmented chorus frog species complex. Molecular Ecology, doi: 10.1111/mec.13343



Extending the model

Recombination

Recombination

Still, many concatenate their 'loci' or SNPs, for phylogenetic analyses:

Recombination

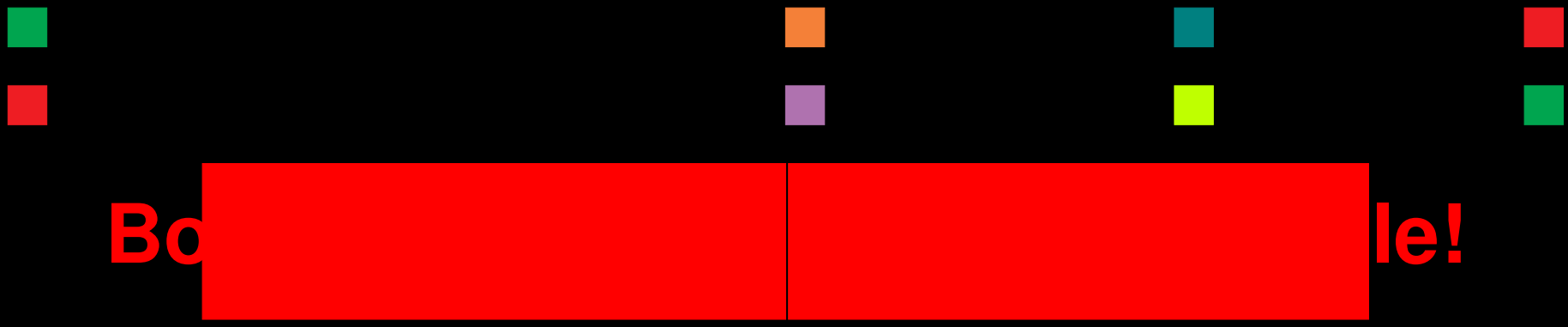
Still, many concatenate their 'loci' or SNPs, for phylogenetic analyses:

Many analyze only the SNPs independently, for population studies:

To much, too little?

Still, many concatenate their 'loci' or SNPs, for phylogenetic analyses:

Many analyze only the SNPs independently, for population studies:



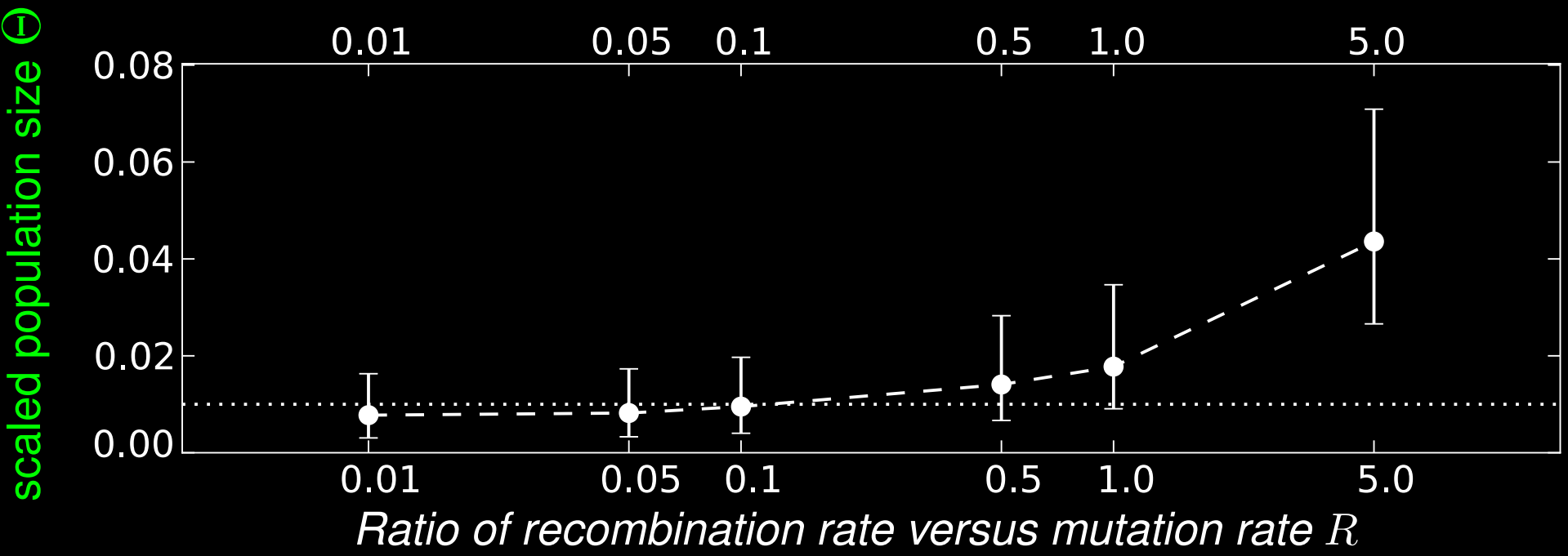
Too much, too little? Does it matter?

Too much, too little? Does it matter?

- What are the effects on our parameter estimation if we ignore recombination and analyze long stretches of contiguous sequence?
- What are the effects, if we assume recombination is rampant and we consider only small chunks of sequence?

Ignoring recombination

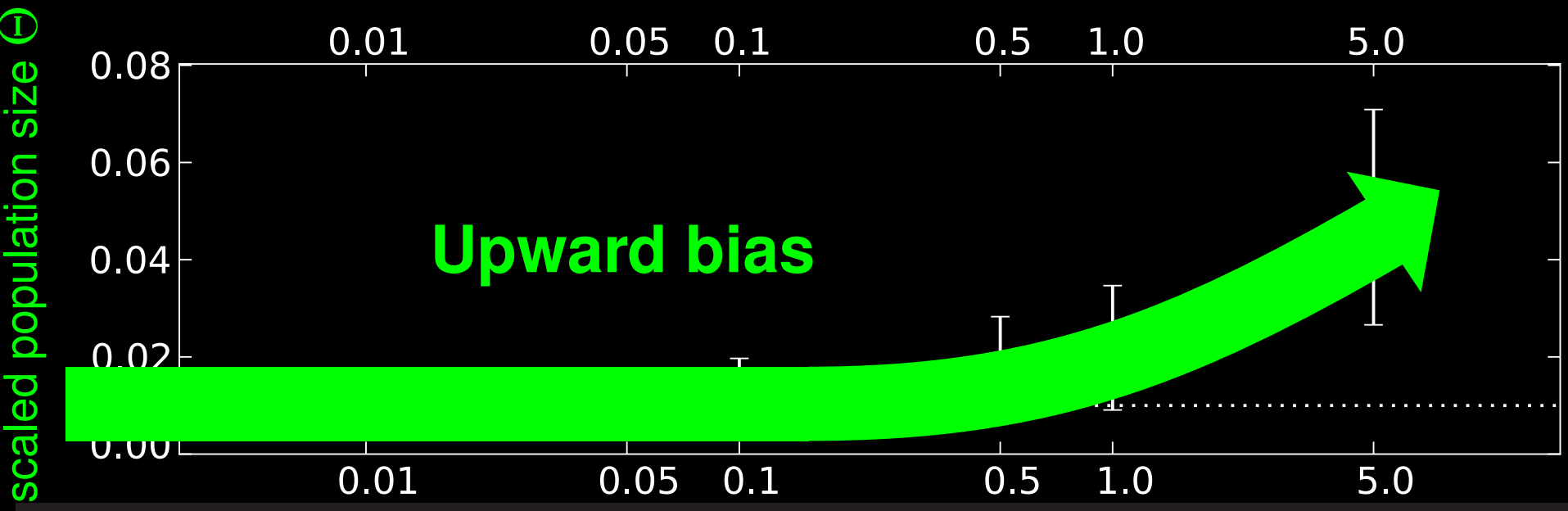
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Ignoring recombination

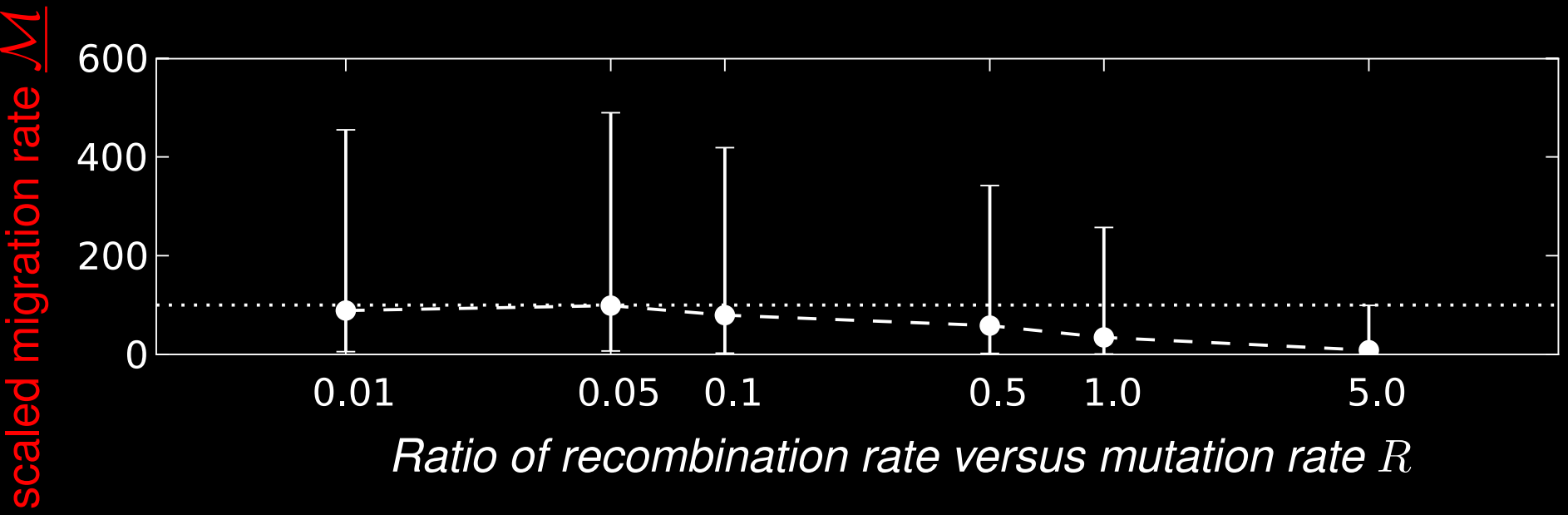
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Ignoring recombination

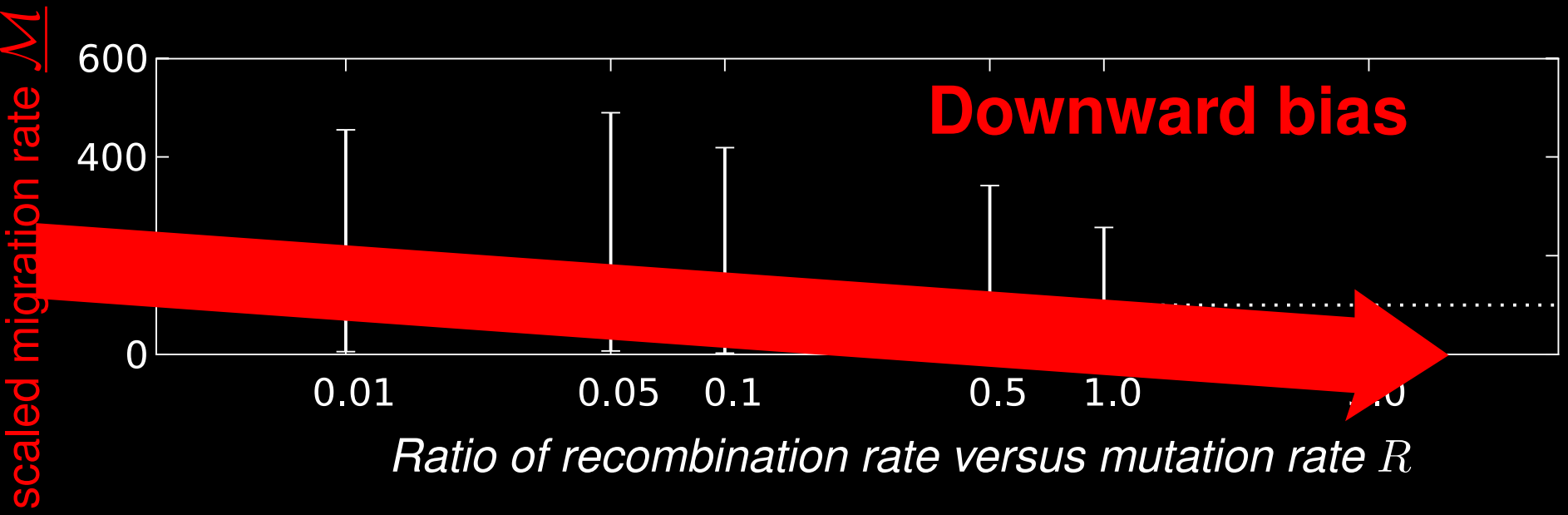
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Ignoring recombination

~500 simulated datasets

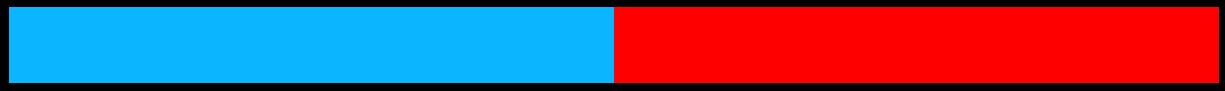


Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Breaking up long sequences

Calculate the log marginal likelihoods $\ln mL$ of models of interest and compare them. This is familiar to phylogeneticists who use mutation model partitions, but here they are analyzed independently.

H_0: 1 locus  $\ln mL = -1938$

H_1: 2 loci  $\ln mL = -1878$

H_2: 3 loci  $\ln mL = -1934$

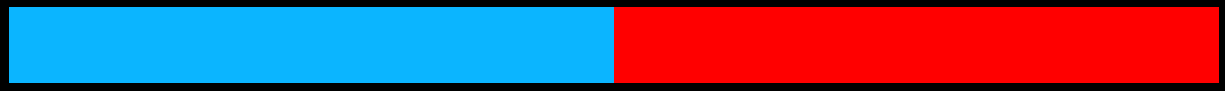
Breaking up long sequences

Calculate the log marginal likelihoods $\ln mL$ of models of interest and compare them. This is familiar to phylogeneticists who use mutation model partitions, but here they are analyzed independently.

H_0 :

$\ln mL = -1938$

H_{1a} :



$\ln mL = -1878$

H_{1b} :

$\ln mL = -1918$

H_2 :

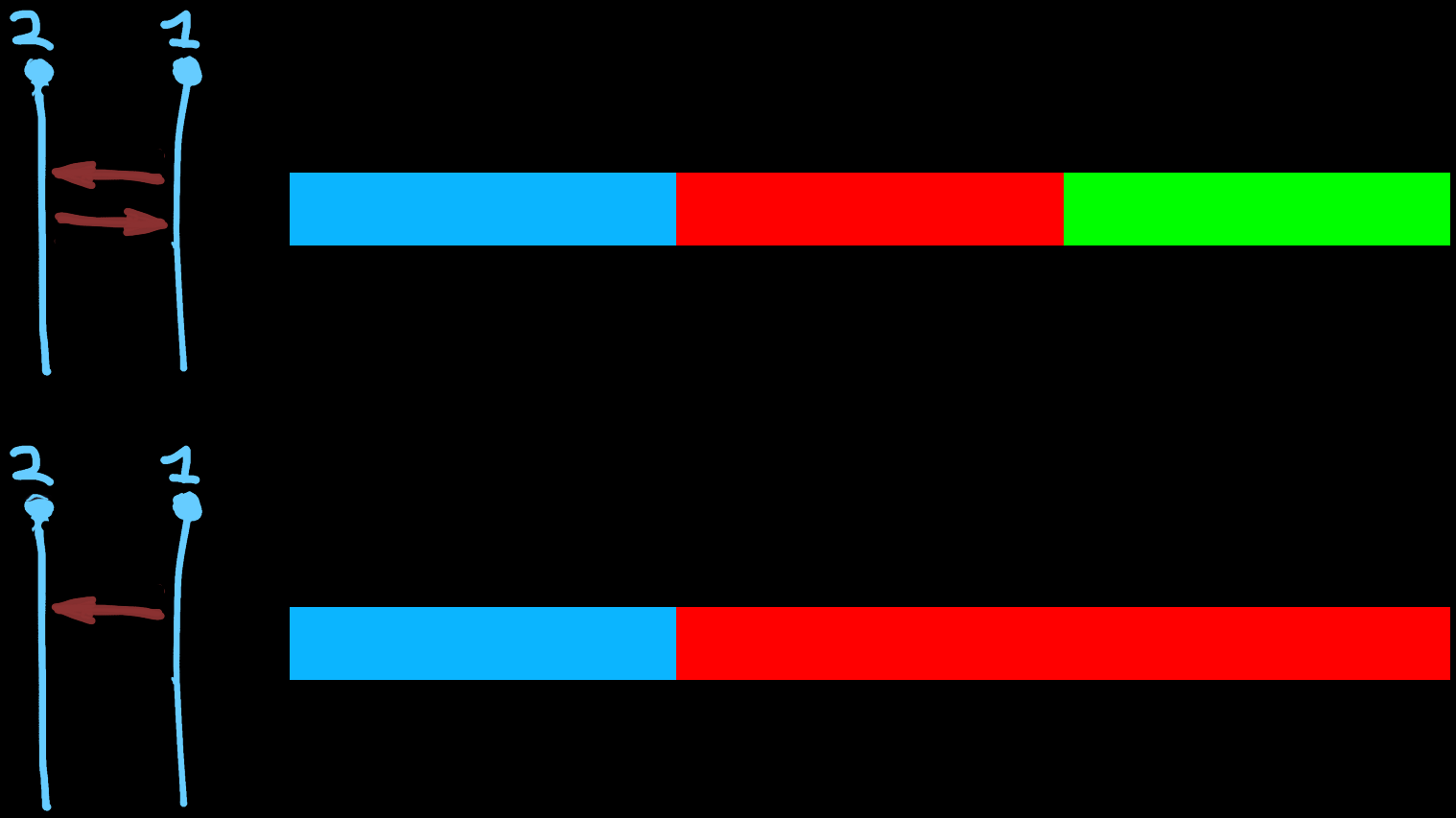
$\ln mL = -1934$

Sorting the log marginal likelihoods: $H_{1a} > H_{1b} > H_2 > H_0$

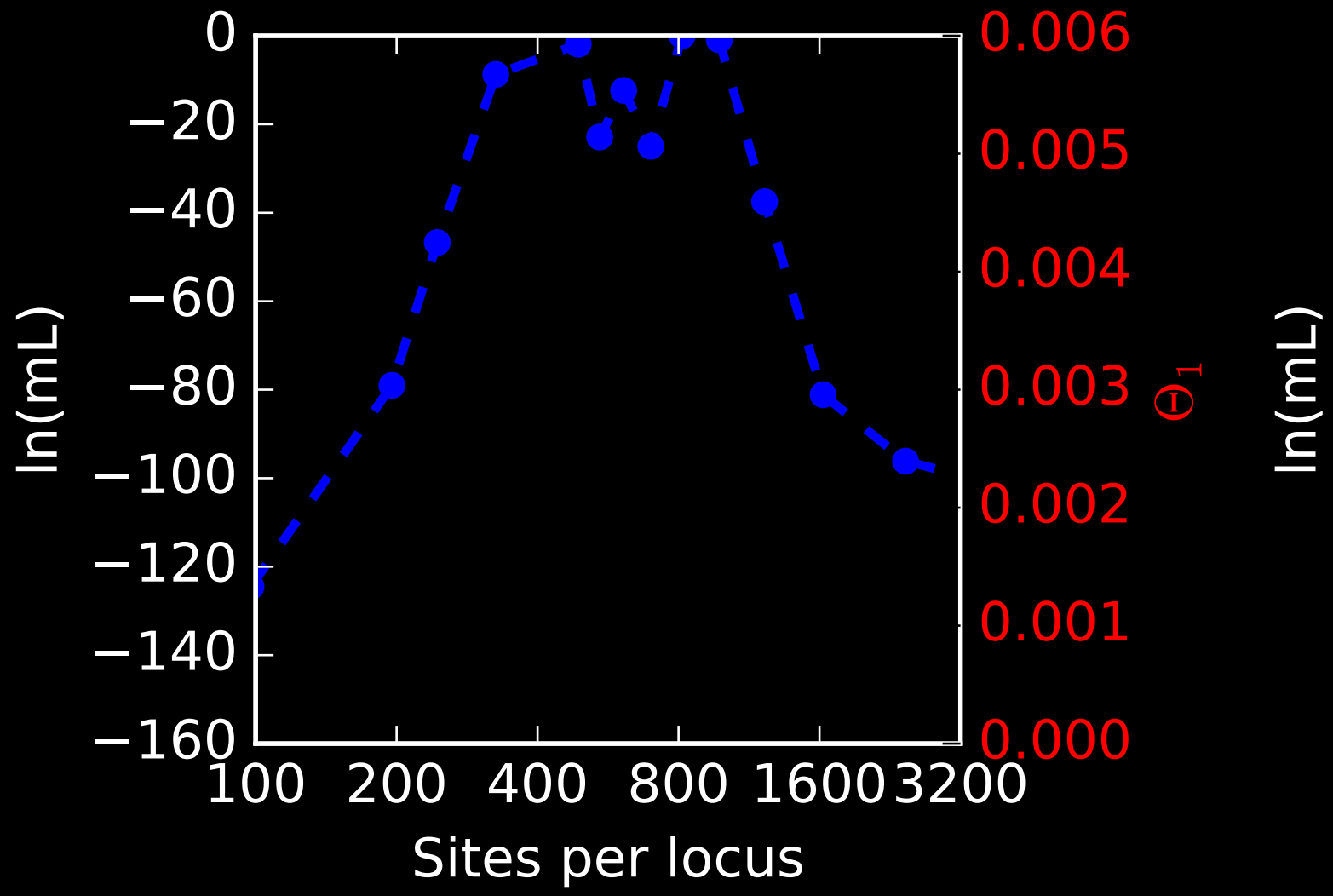
Suggests: Pick a two-locus model.

Breaking up long sequences

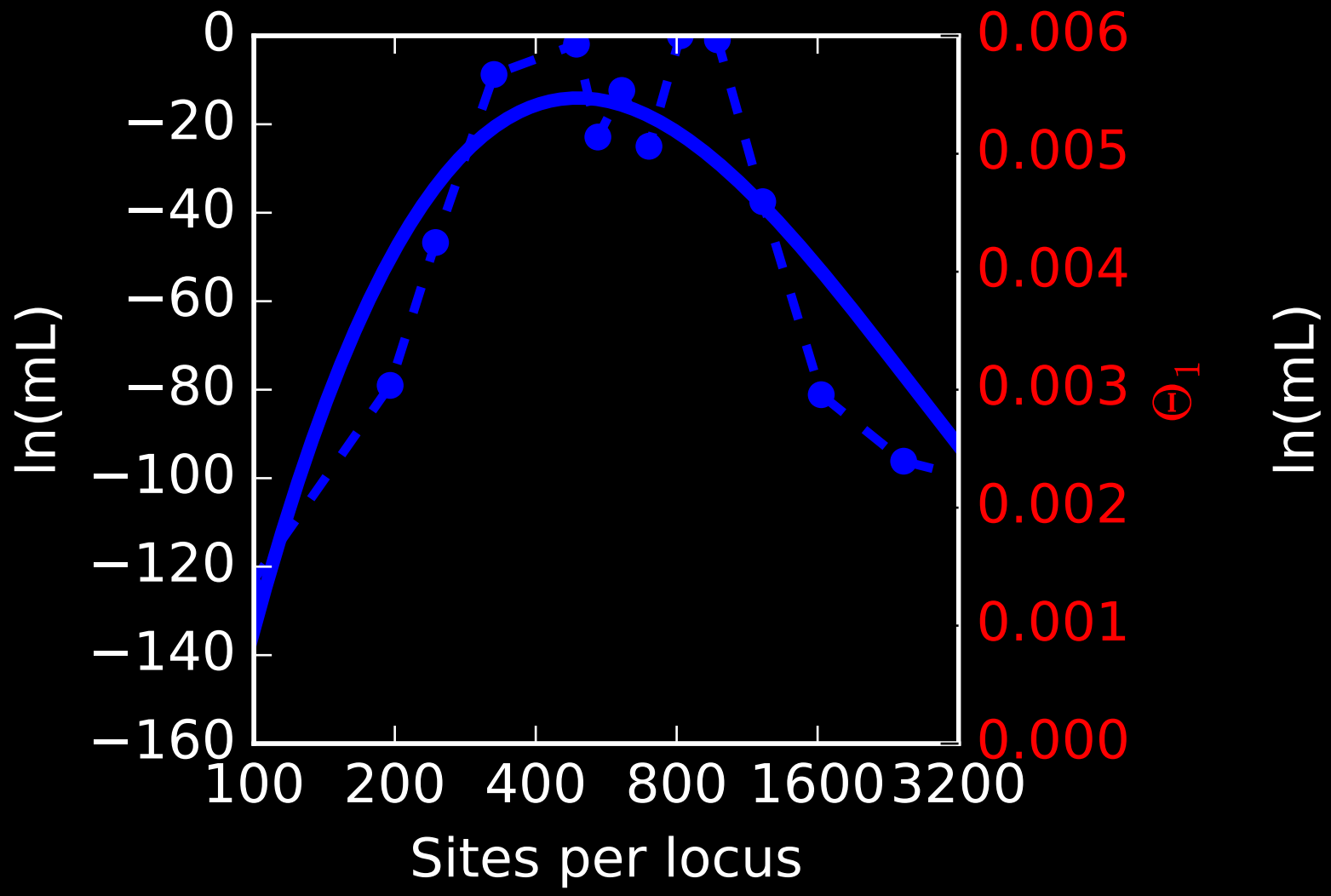
We combine now models that represent different breaks in a long sequence stretch with the population models, and this may even help to get better population parameter estimates. For example these two models:



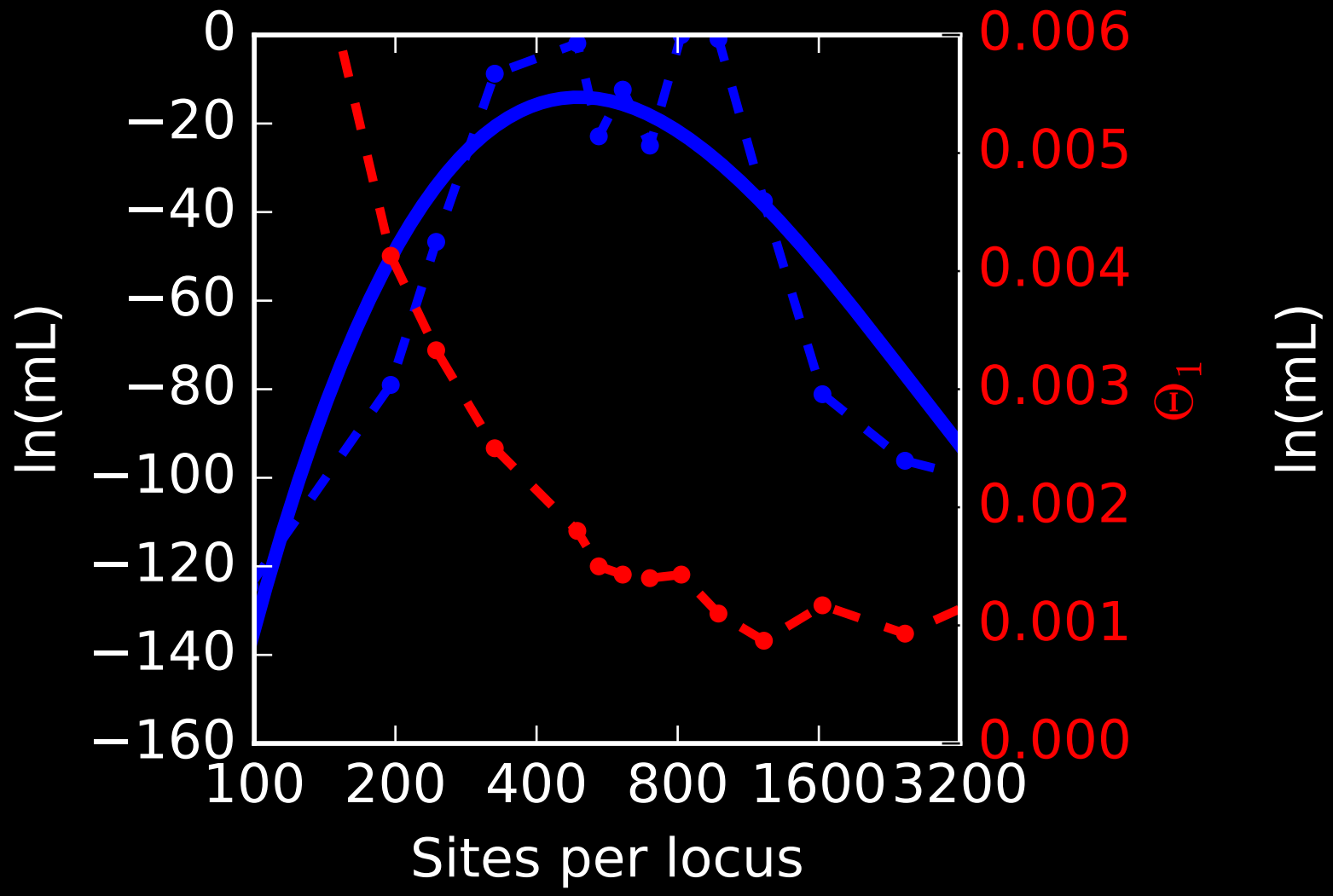
Human lipoproteine lipase: Finns



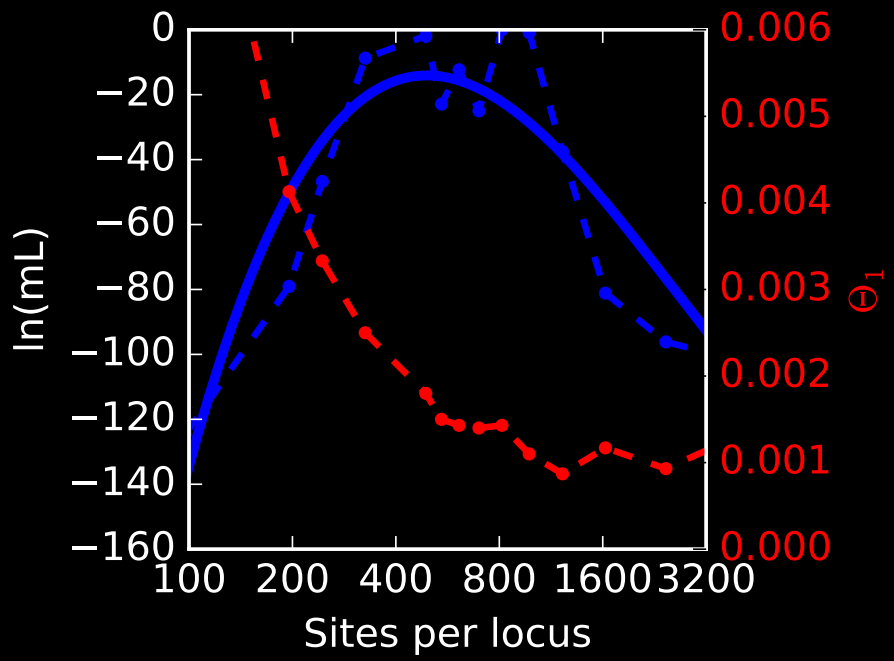
Human lipoproteine lipase: Finns

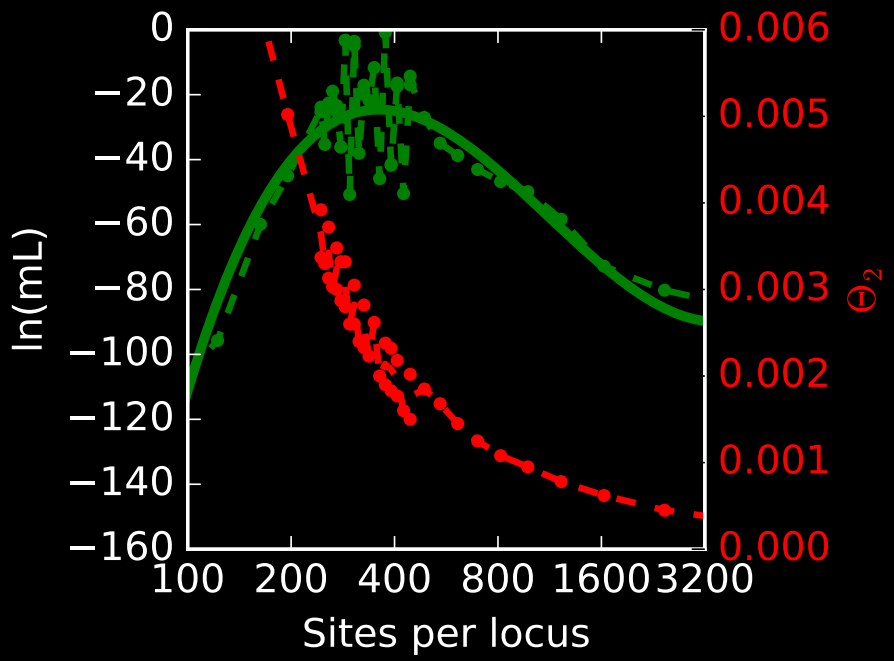


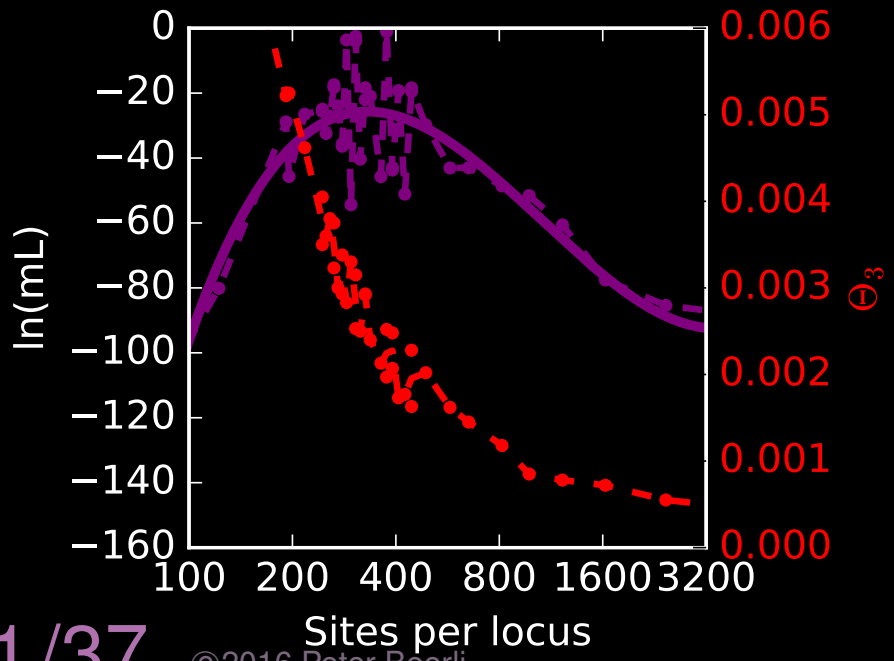
Human lipoproteine lipase: Finns

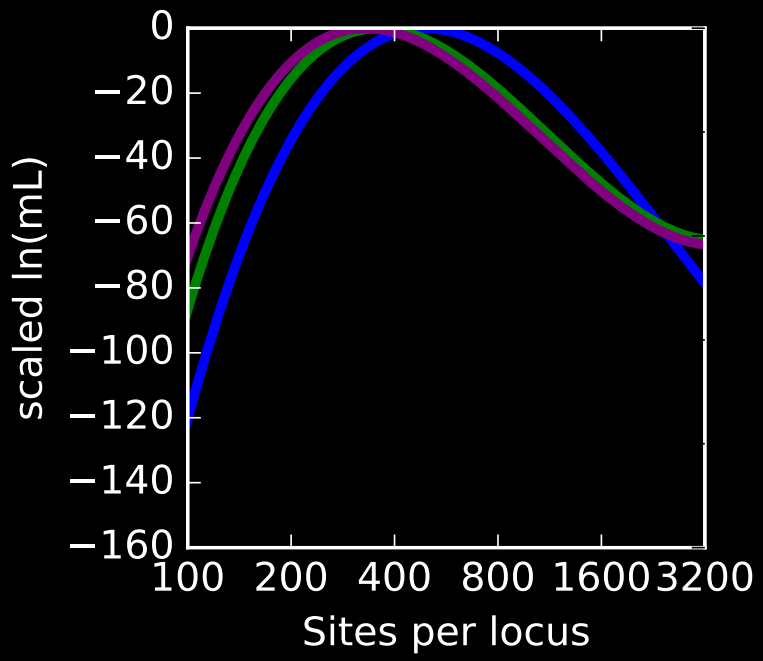


Human lipoproteine lipase: Finns, Minnesotans, Mississippians

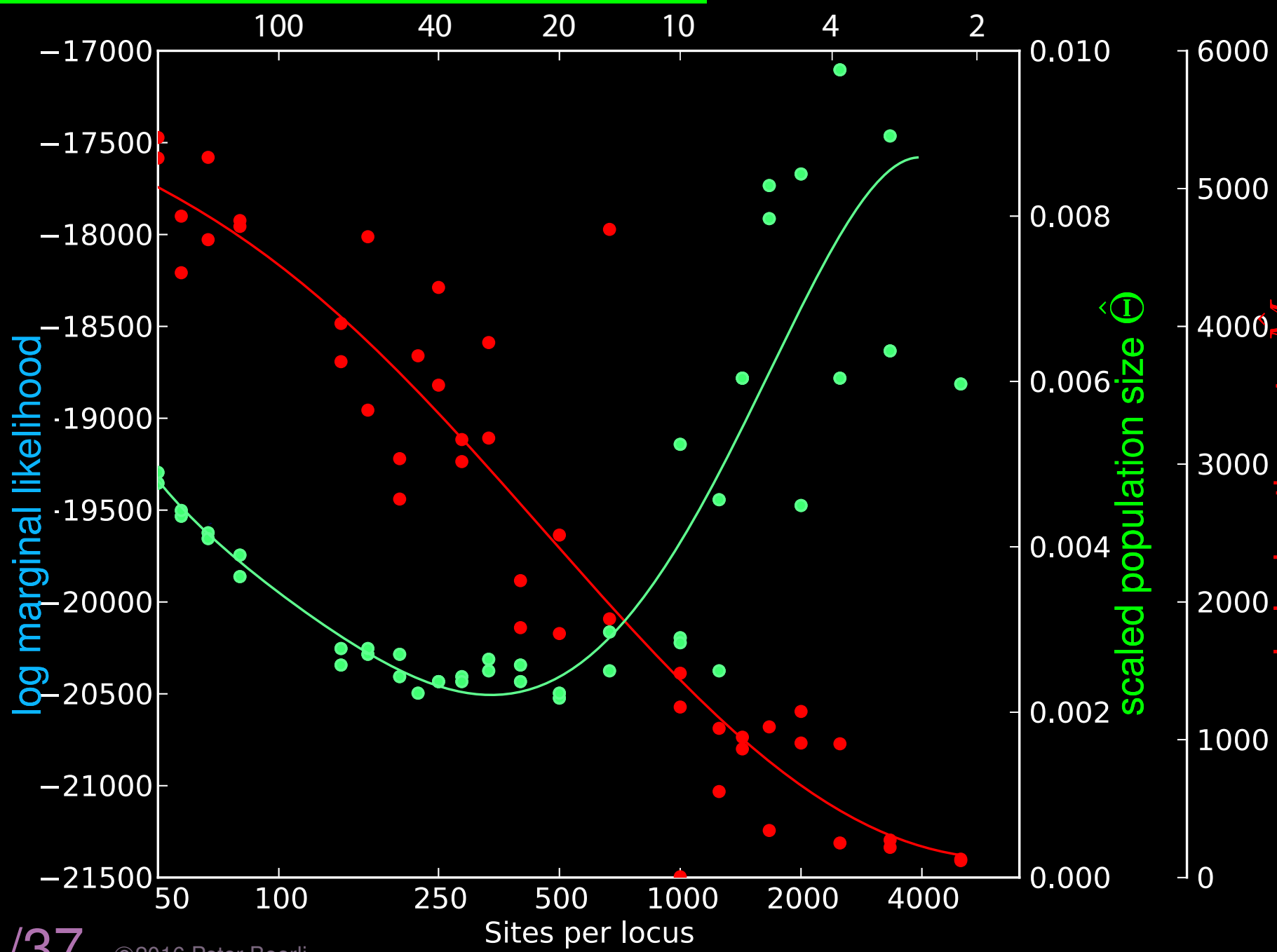








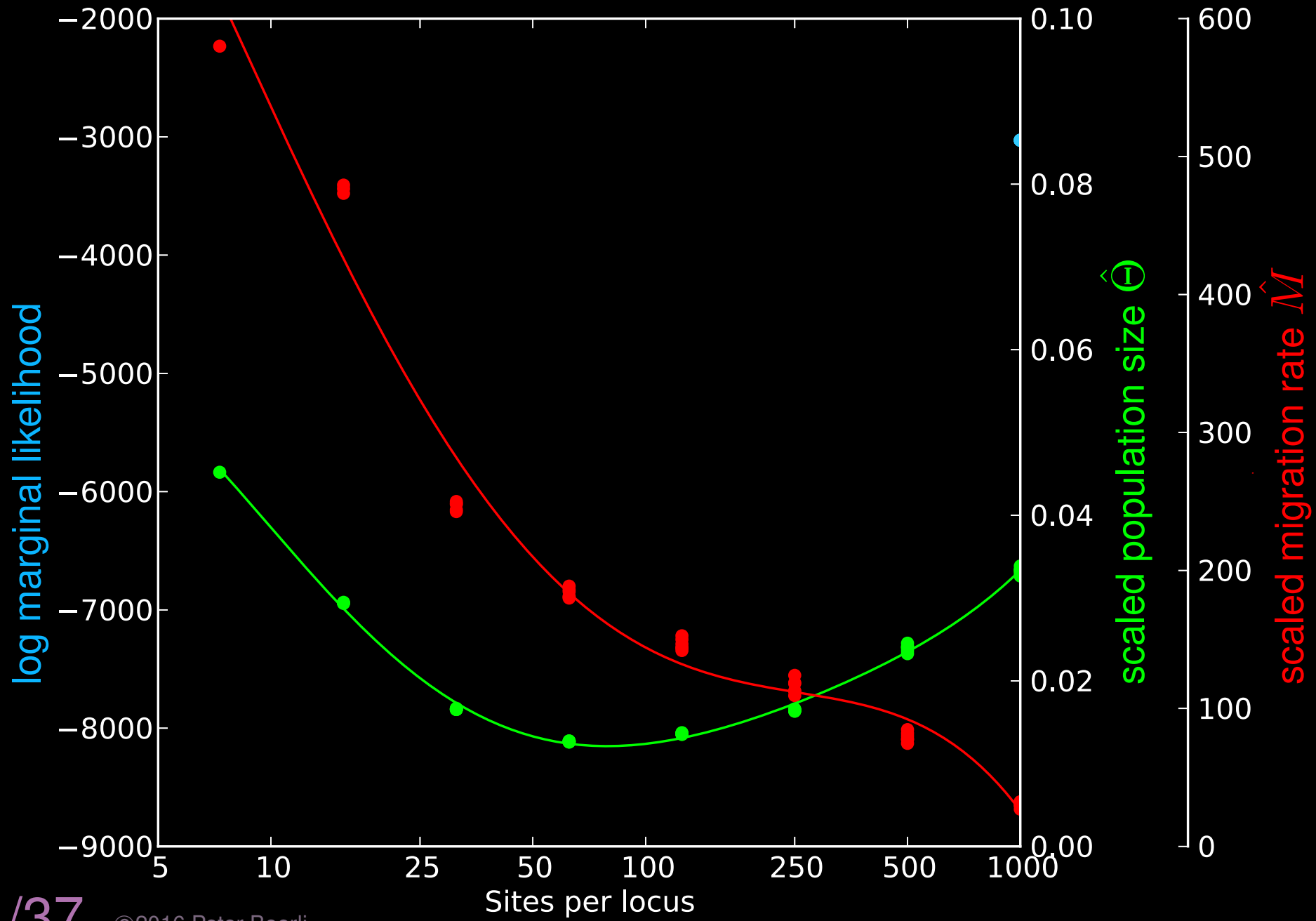
D. melanogaster Chr 2L



D. melanogaster Chr 2L

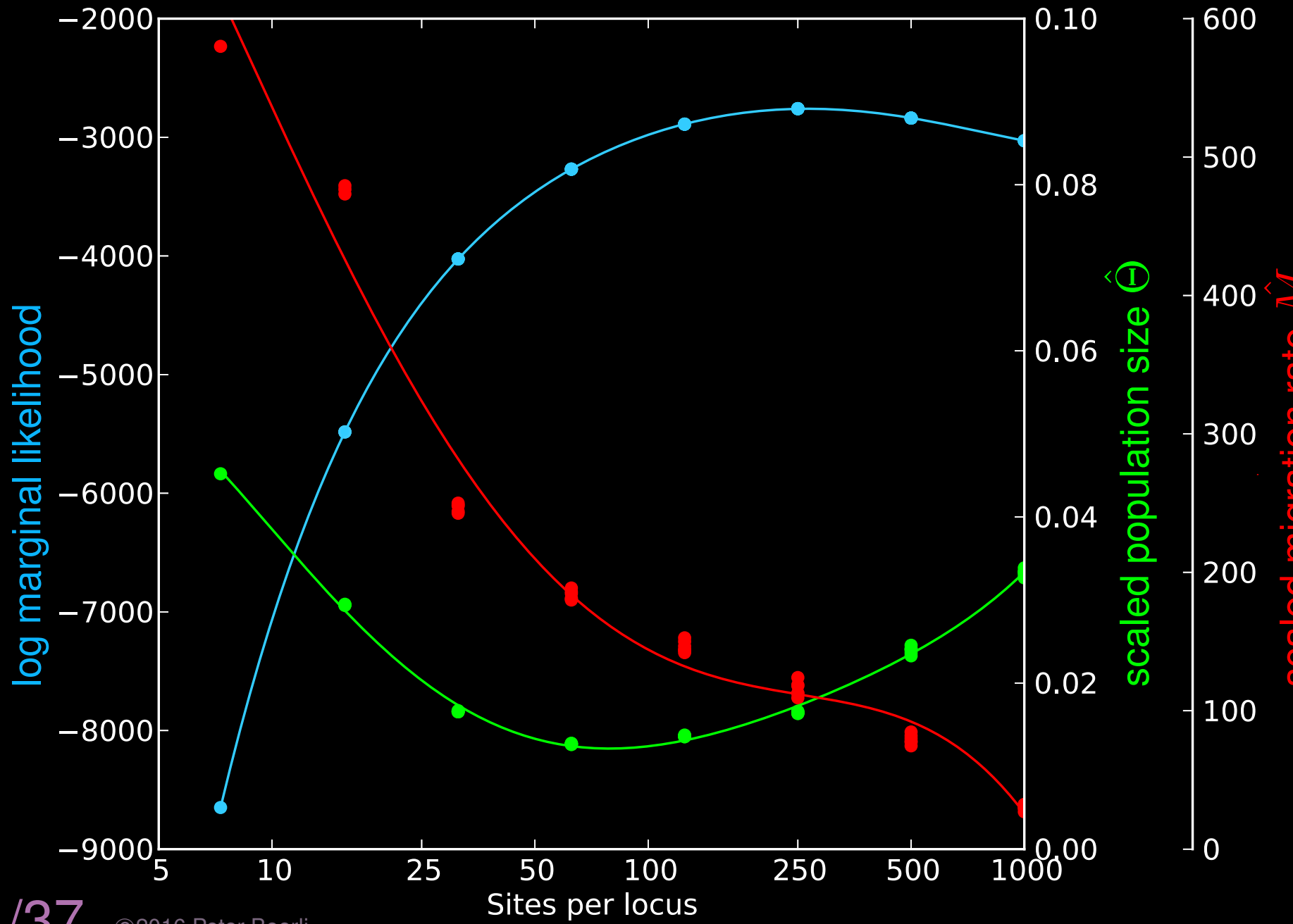
Chopping a simulated data set

1 simulated dataset (1000bp)
with high recombination rate $\frac{C}{\mu} = 1$



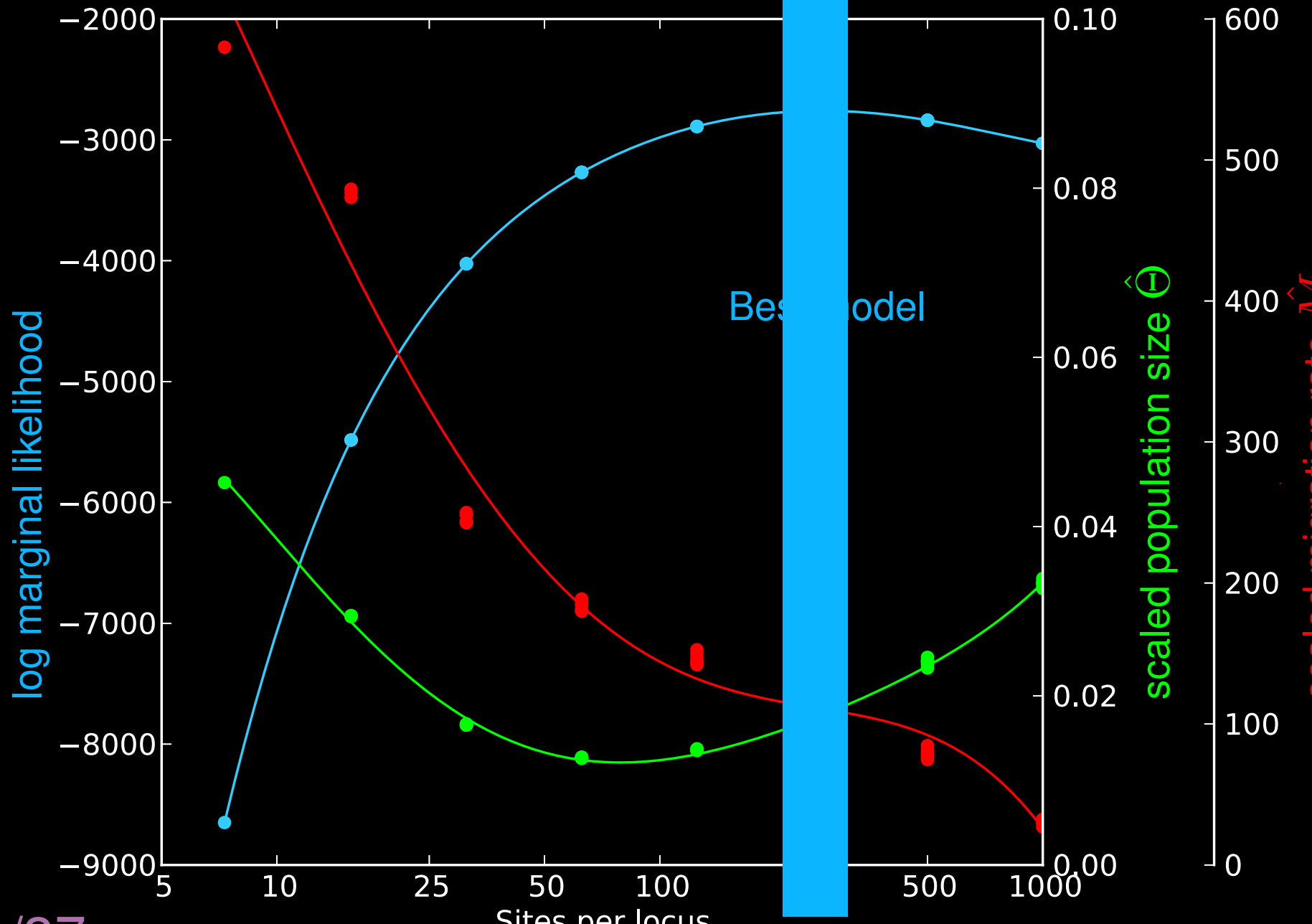
Chopping a simulated dataset

1 simulated dataset (1000bp)
with high recombination rate $\frac{C}{\mu} = 1$



Chopping a simulated dataset

1 simulated dataset (1000bp)
with high recombination rate $\frac{C}{\mu} = 1$



Summary