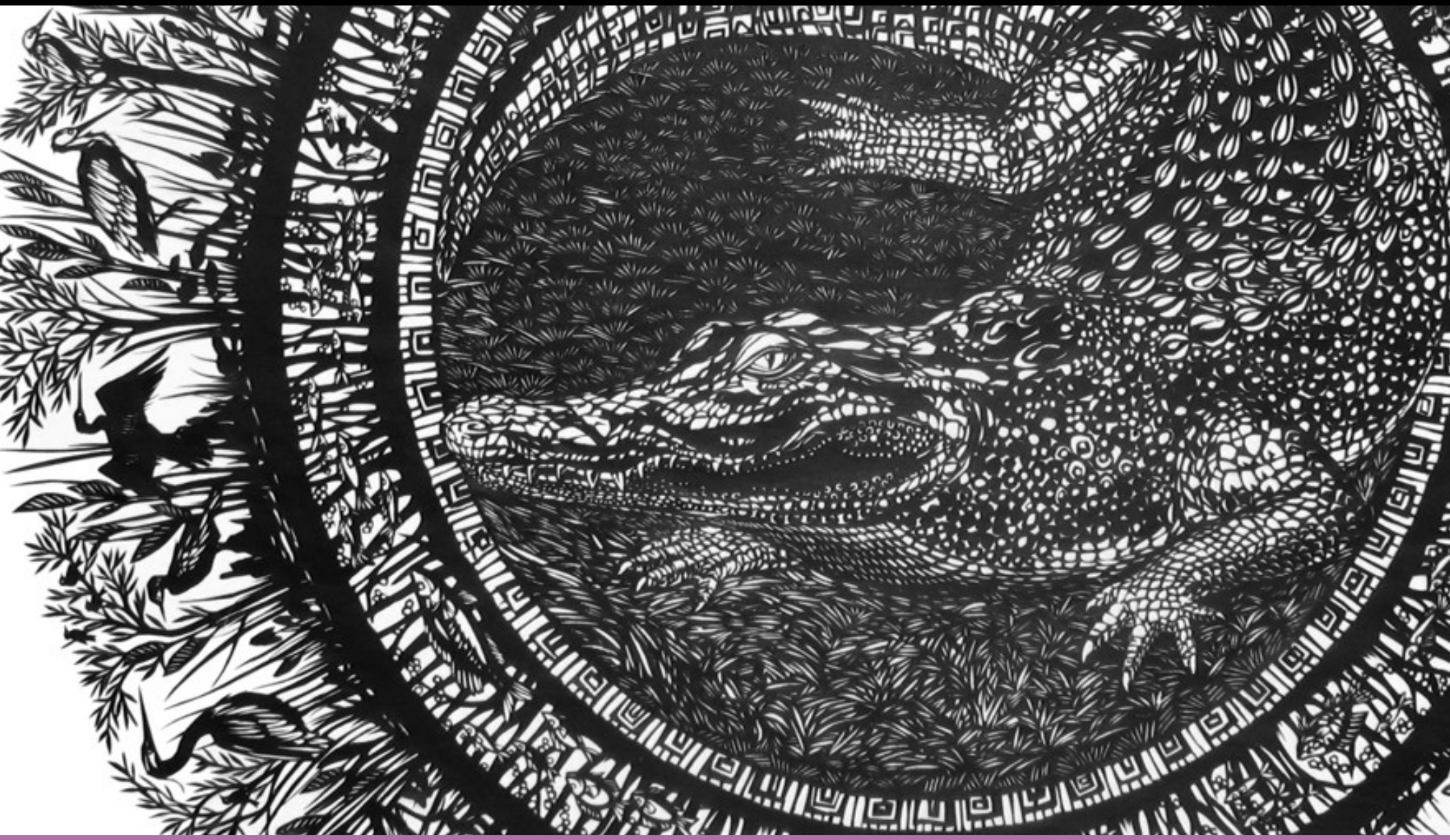


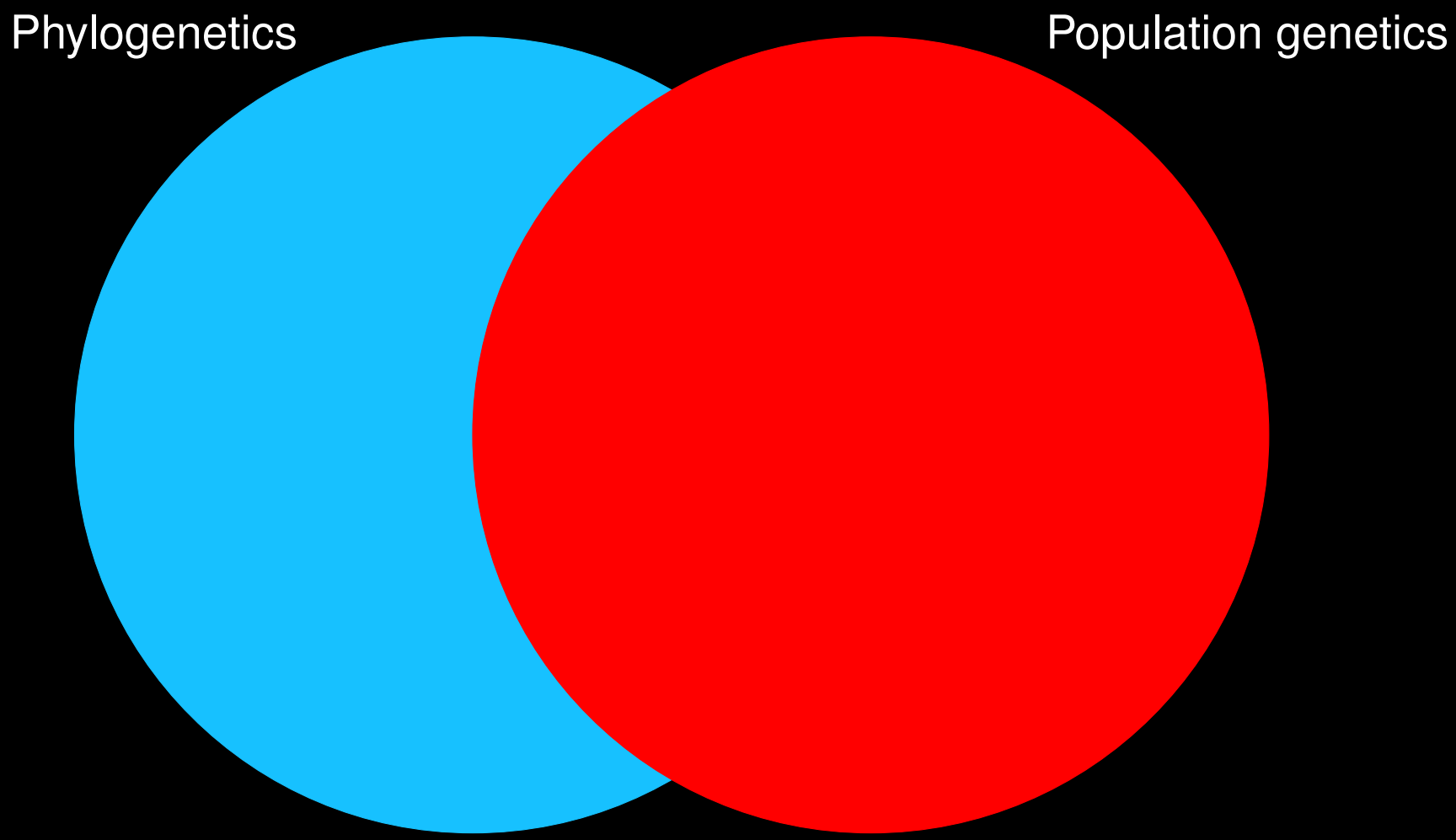
Biogeography, Models, Migrate, oh my

Peter Beerli, Scientific Computing, Florida State University

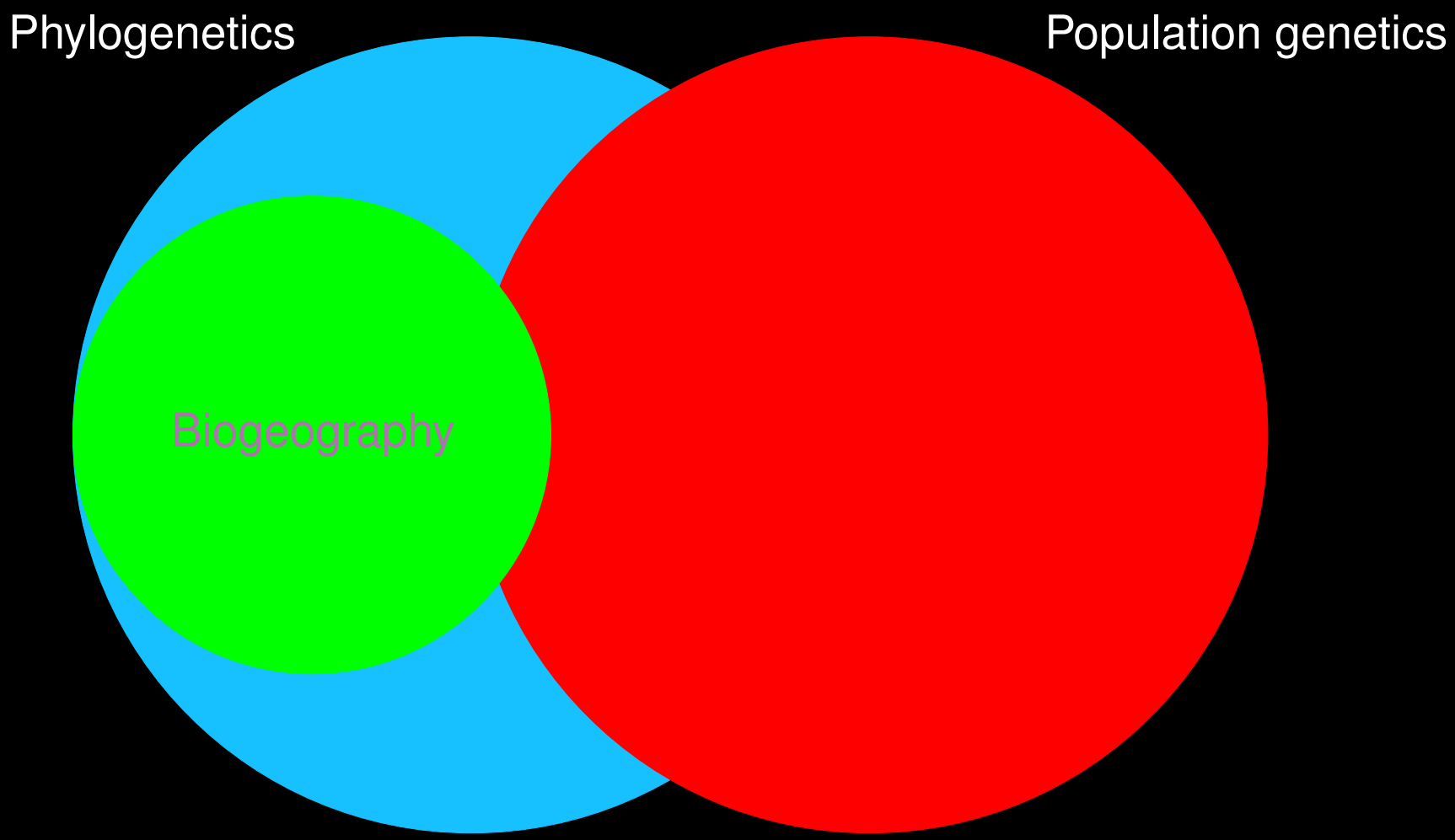
Roadmap

- Phylogenetics, Biogeography, Phylogeography, Population genetics
- Modeling a complex world
- Model comparison
  - Bayes factors, marginal likelihood
  - Parallel evaluation of many unlinked loci
- Model-based assignment of individuals

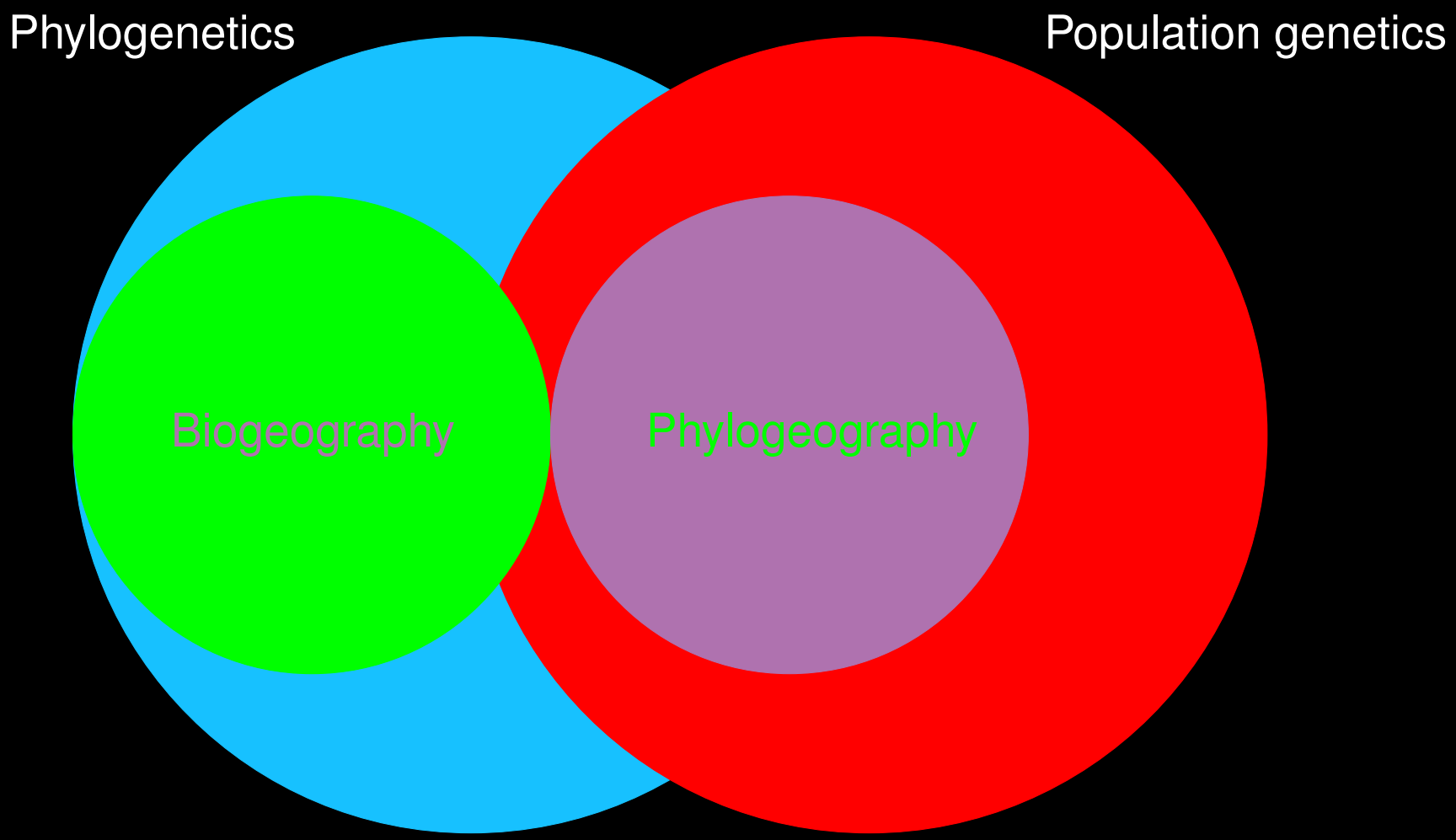
{Phylo, Population, Bio}{genetics, geography}



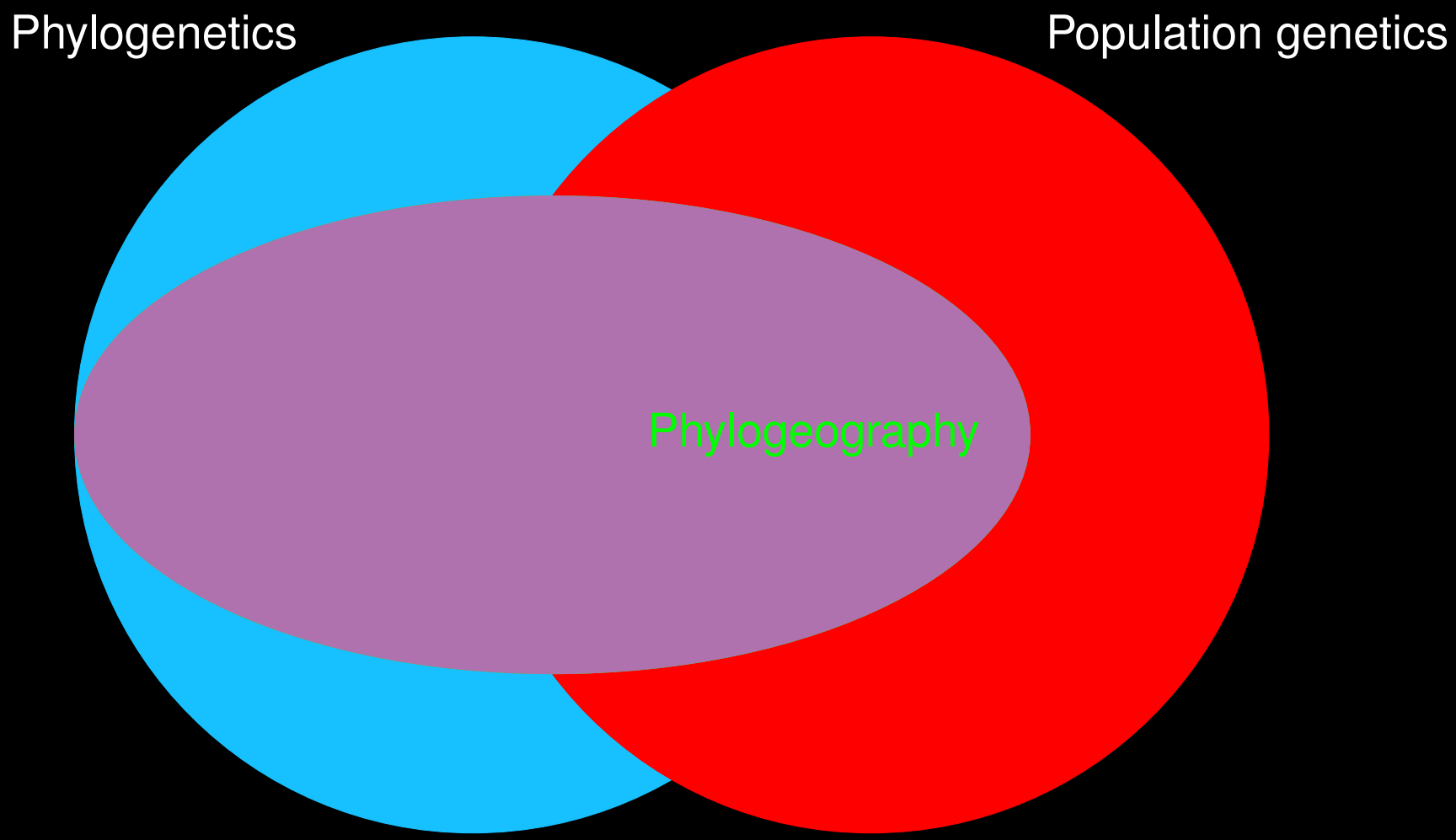
{Phylo, Population, Bio}{genetics, geography}



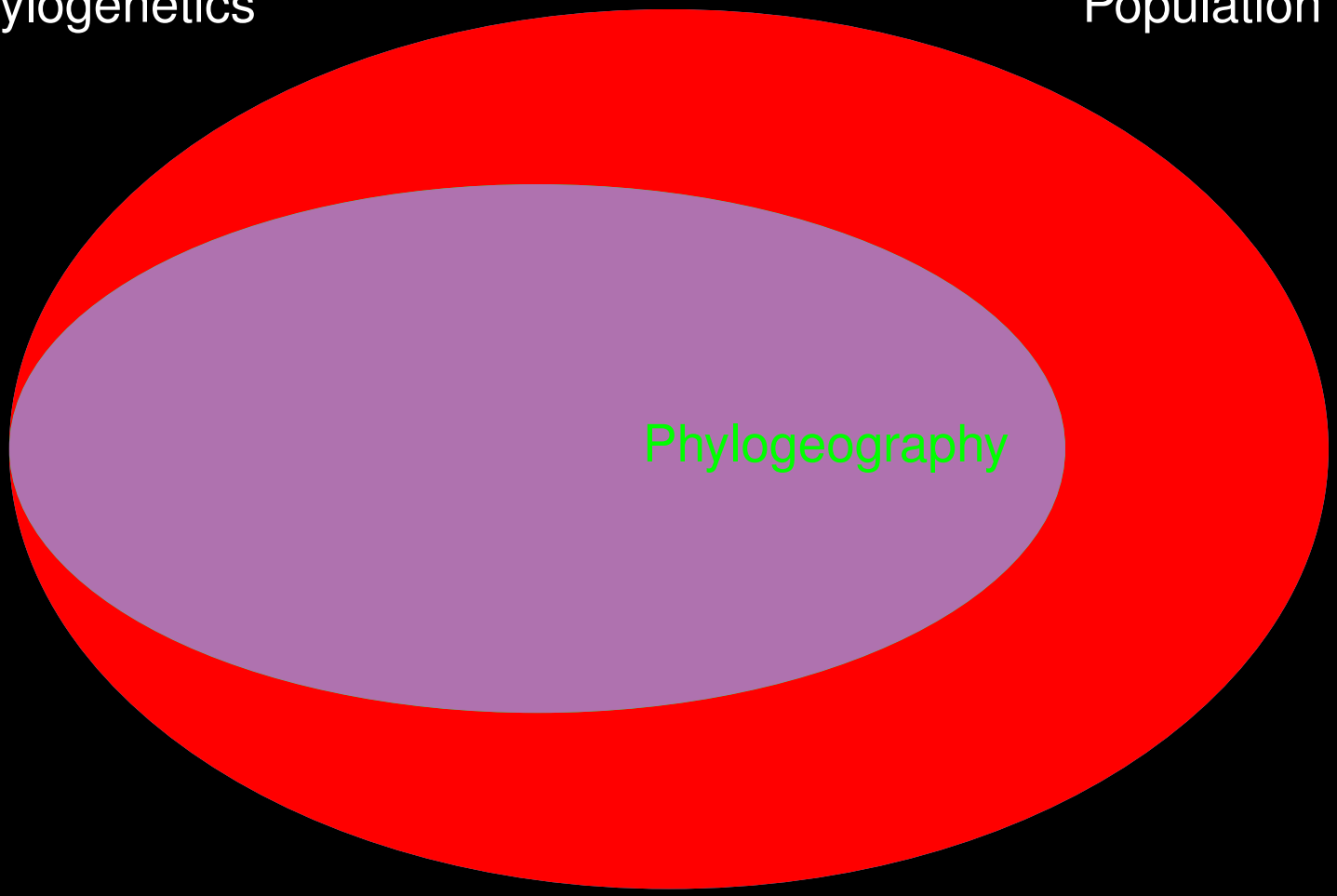
{Phylo, Population, Bio}{genetics, geography}



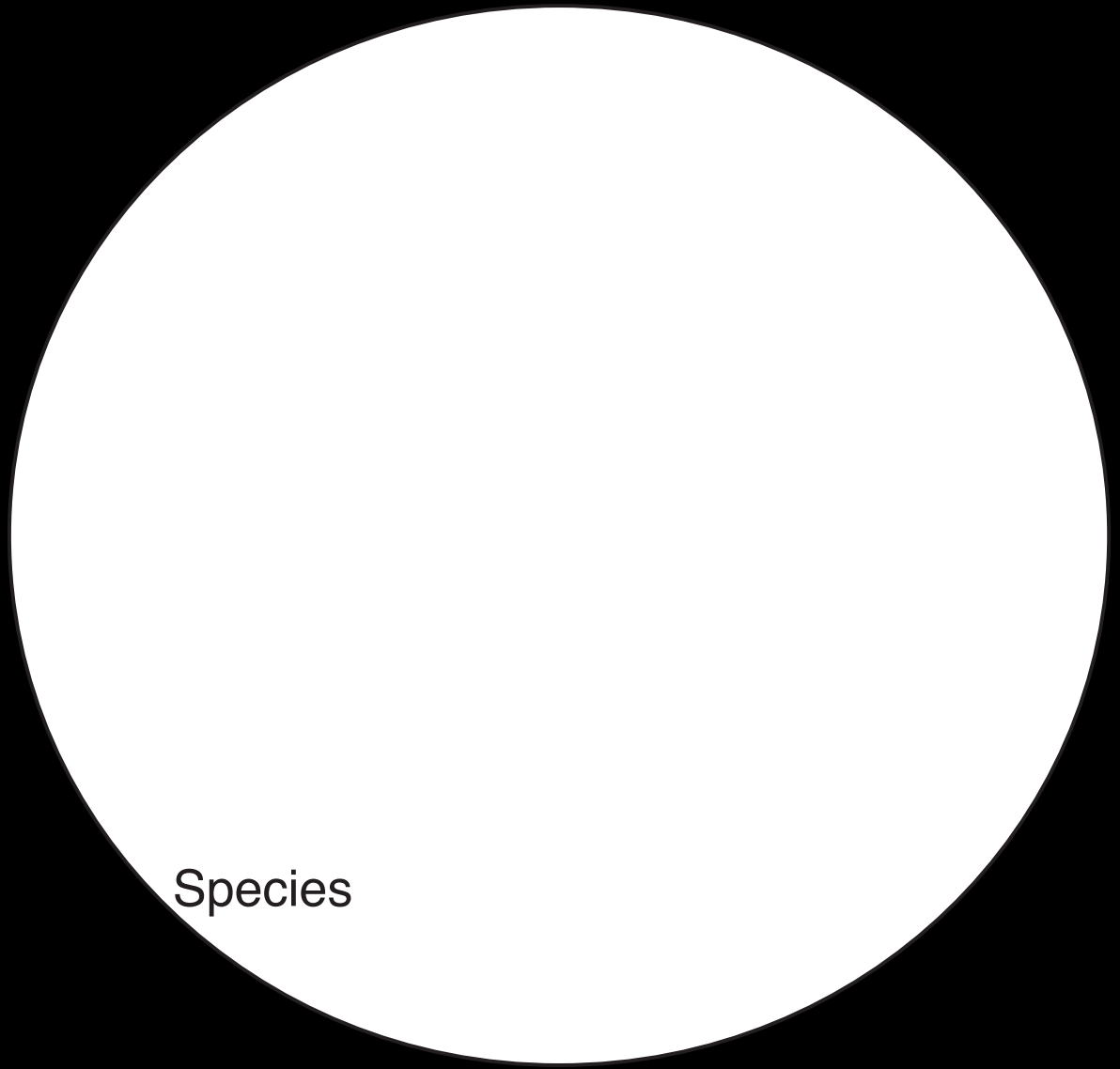
{Phylo, Population, Bio}{genetics, geography}



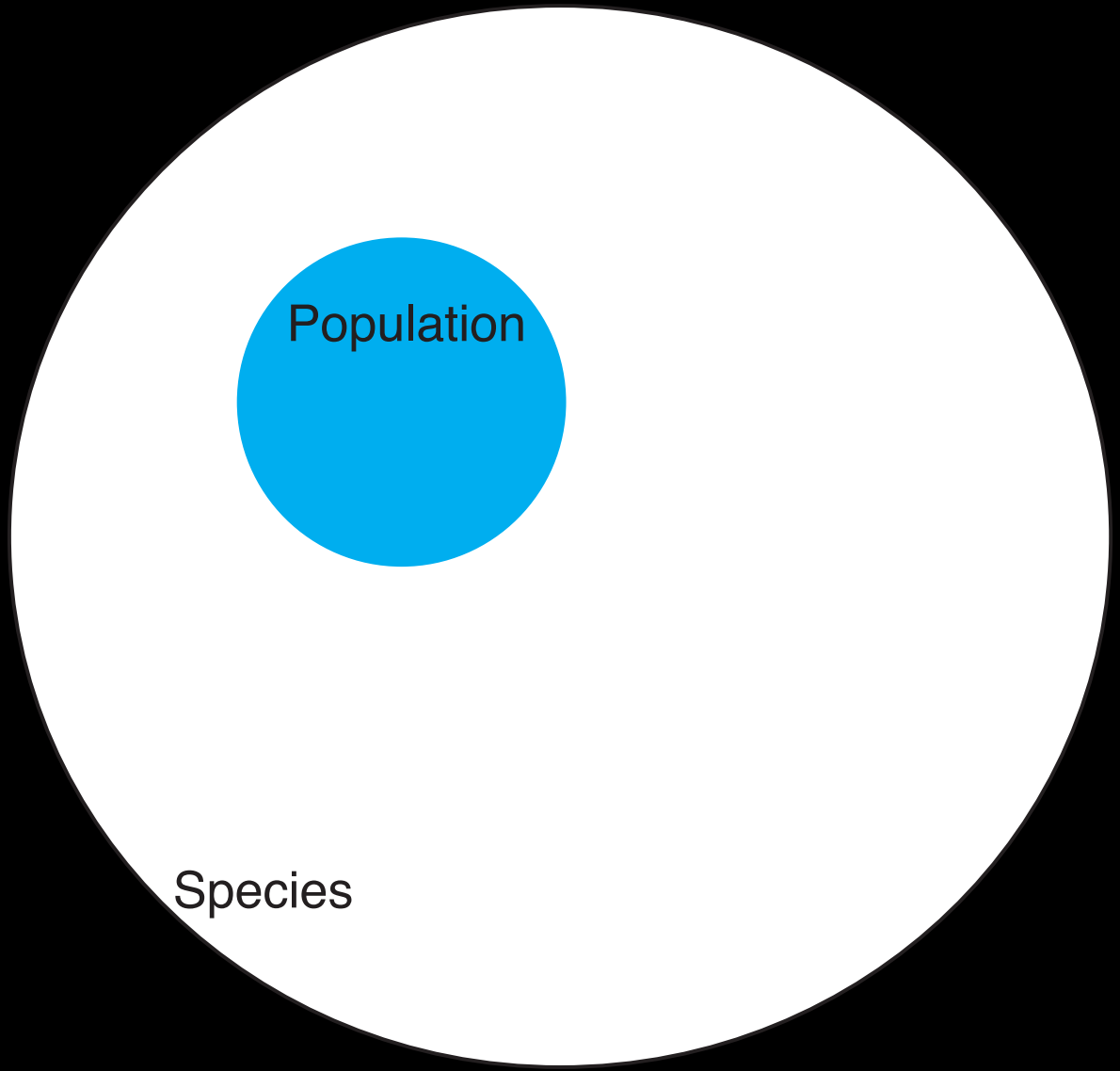
{Phylo, Population, Bio}{genetics, geography}



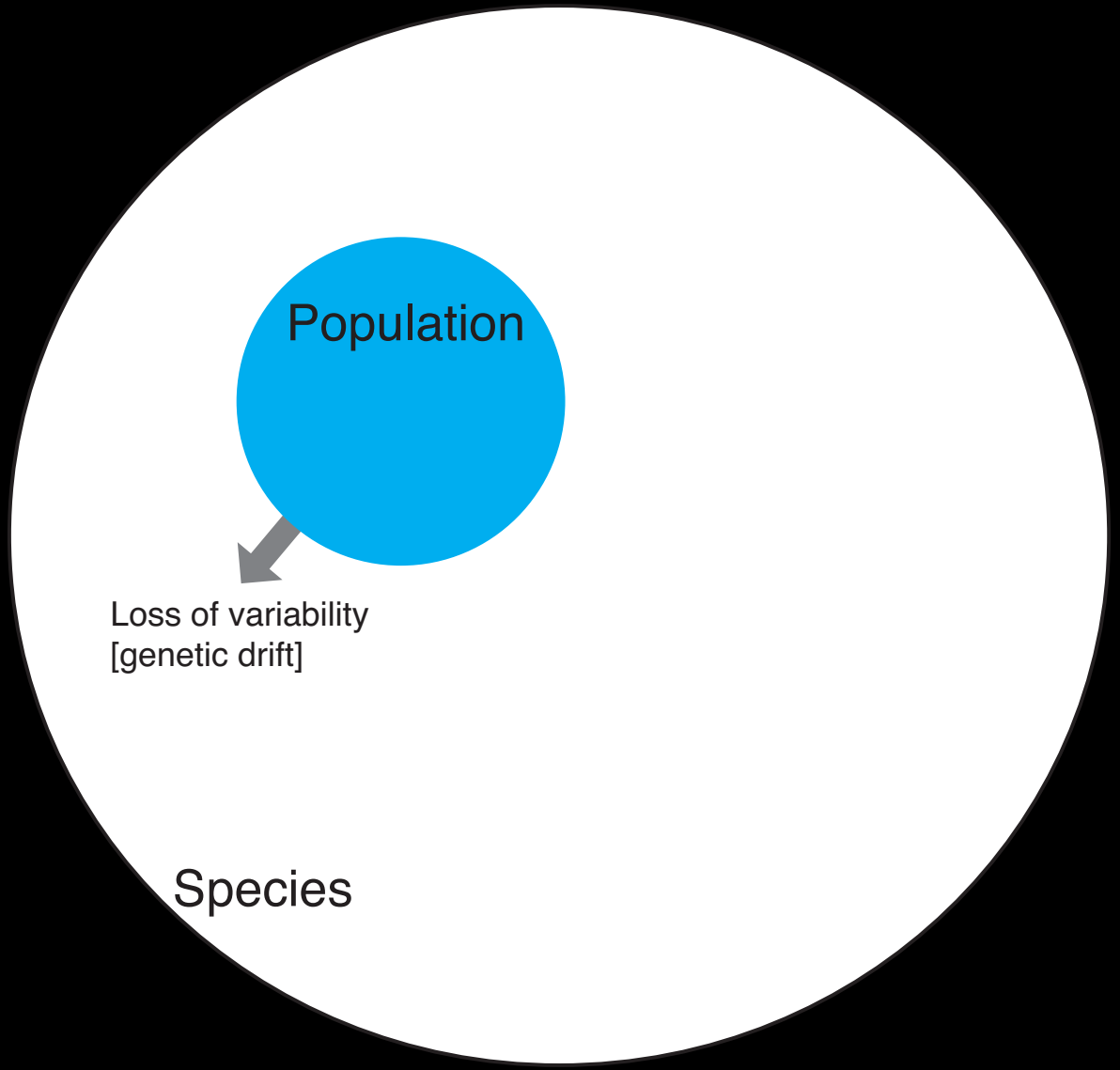
Population models



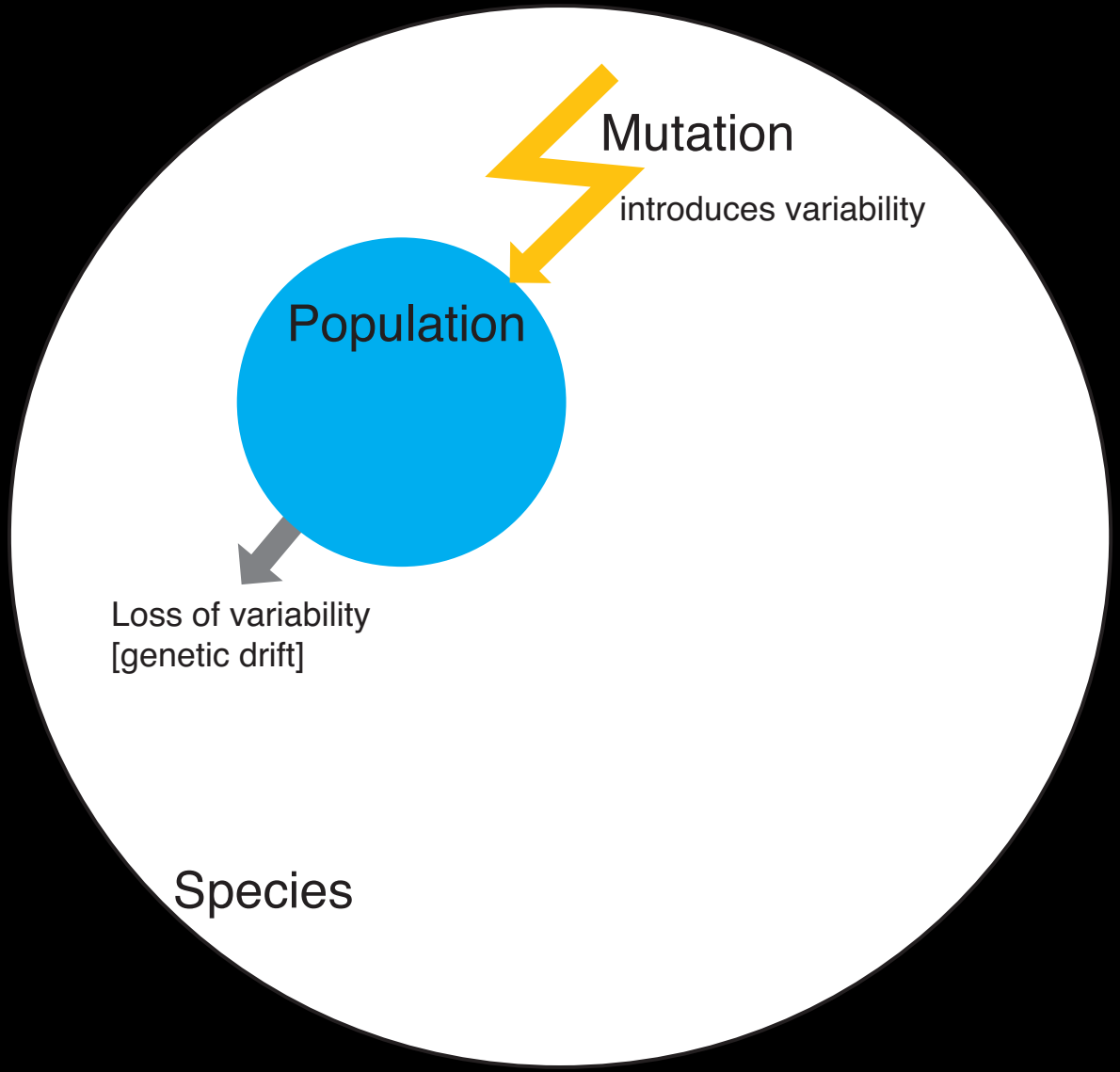
Population models



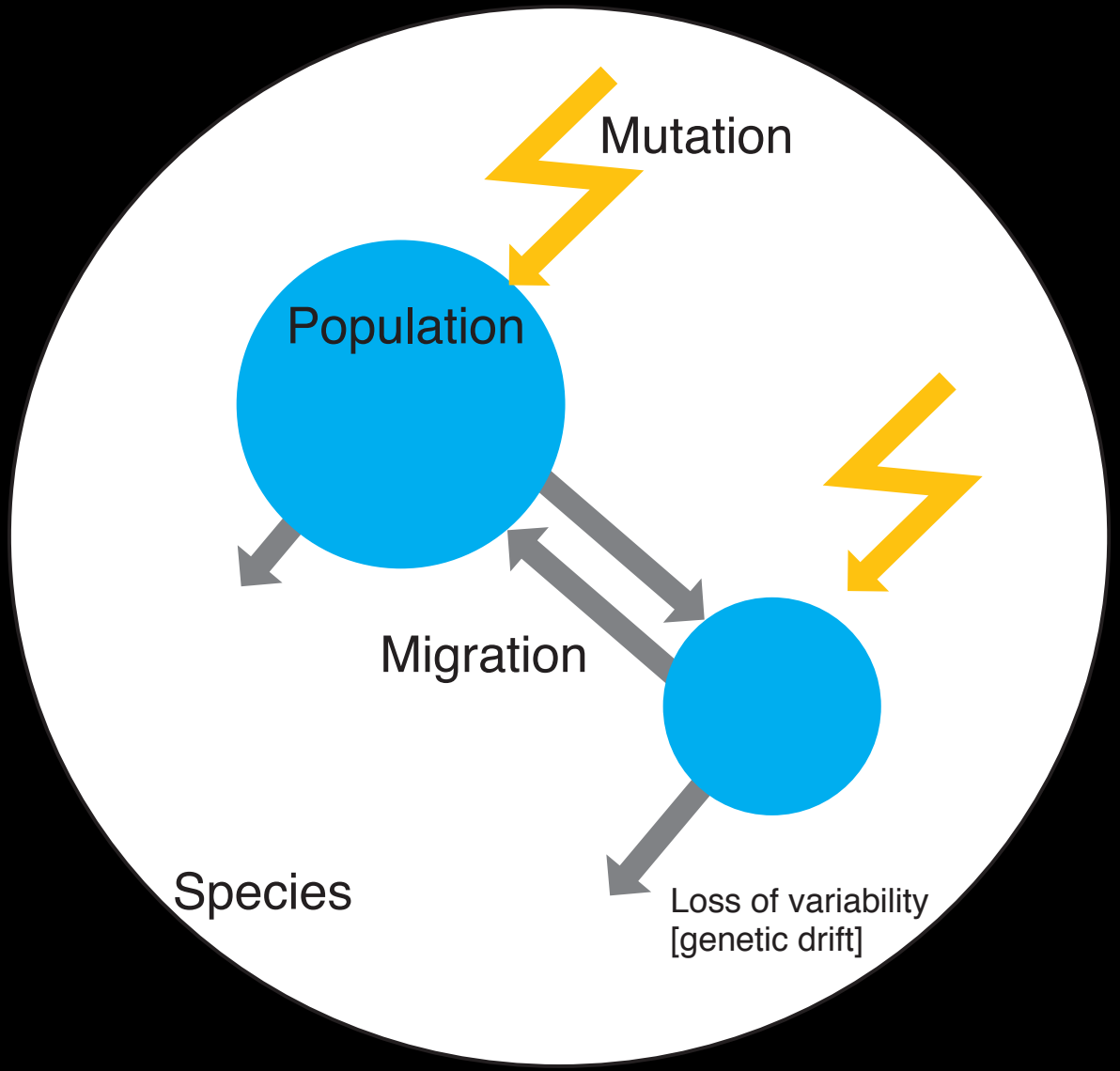
Population models



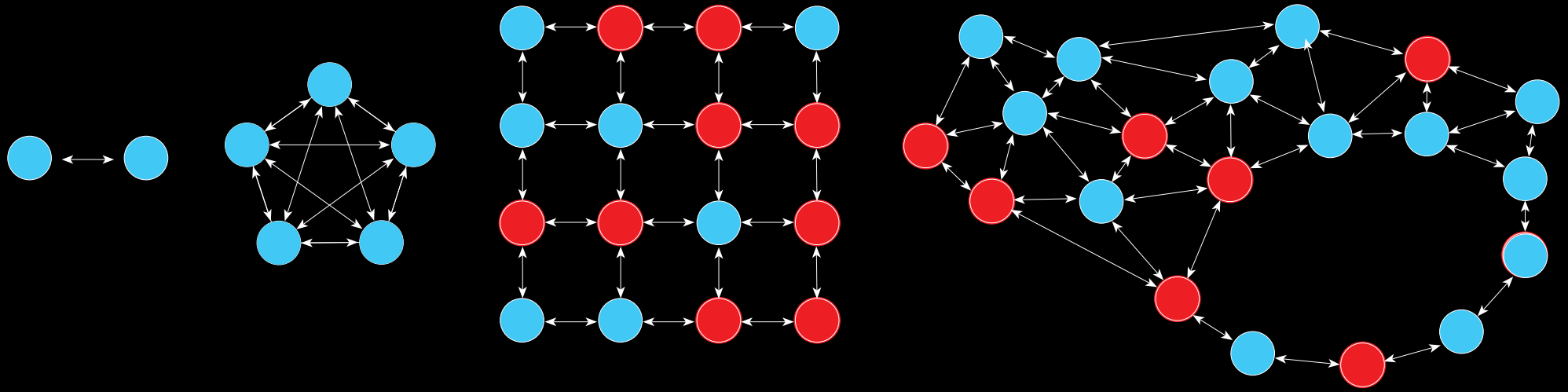
Population models



Population models

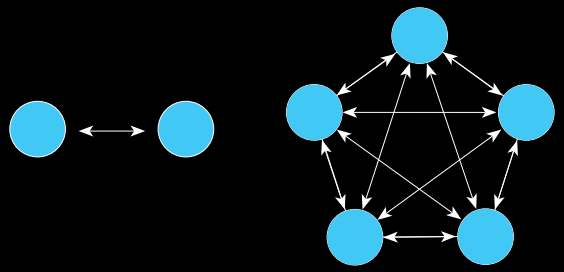


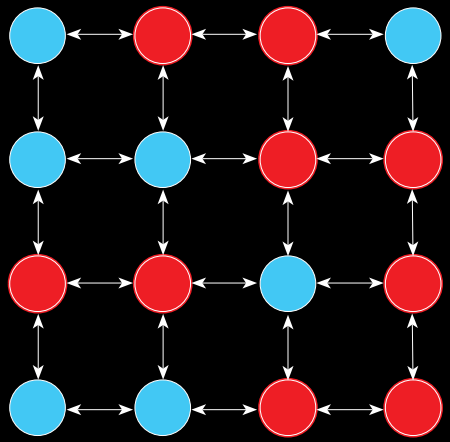
Population structure

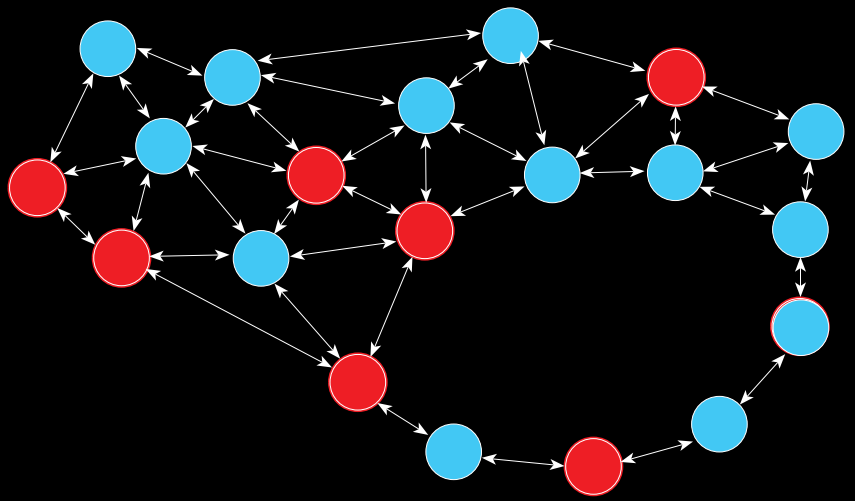


Population structure

Models available in MIGRATE







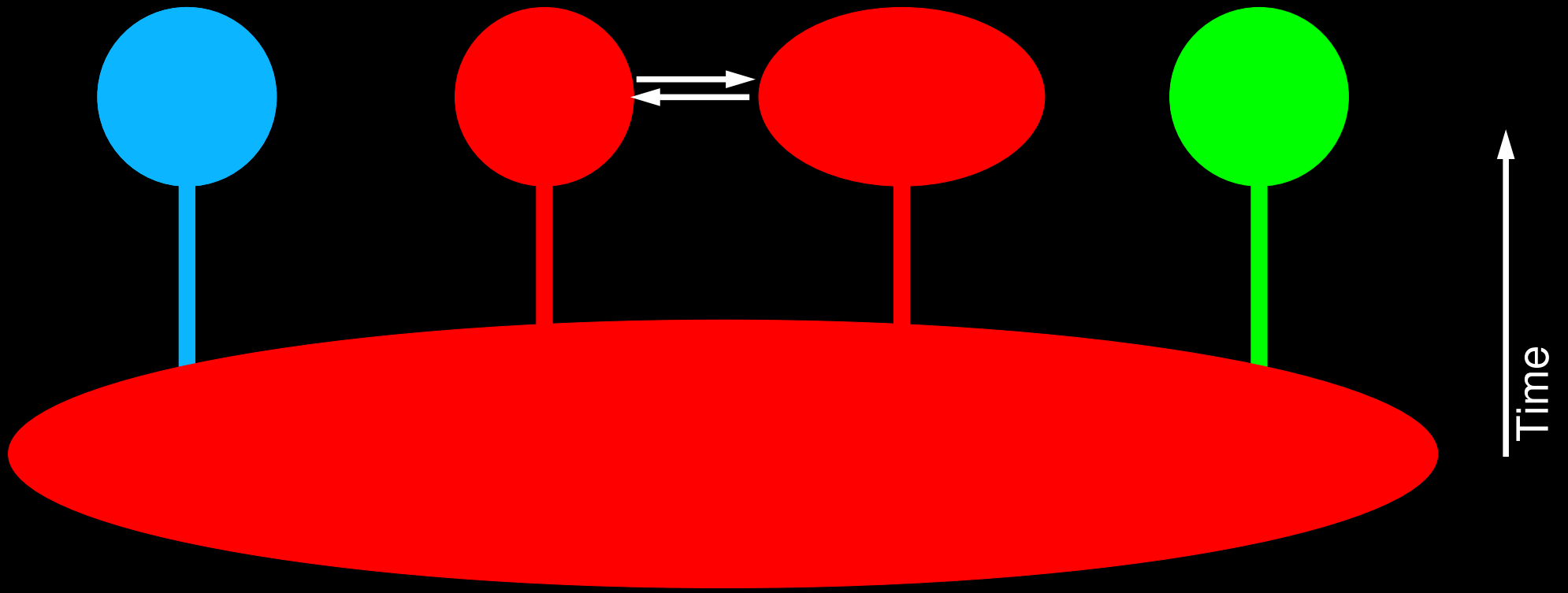
```
**
**
```

```
*mmmm
m*mmm
mm*mm
mmm*m
mmmm*
```

```
**00  *000  0000  0000
***0  0*00  0000  0000
0***  00*0  0000  0000
00**  000*  0000  0000
*000  **00  *000  0000
0*00  ***0  0*00  0000
...
```

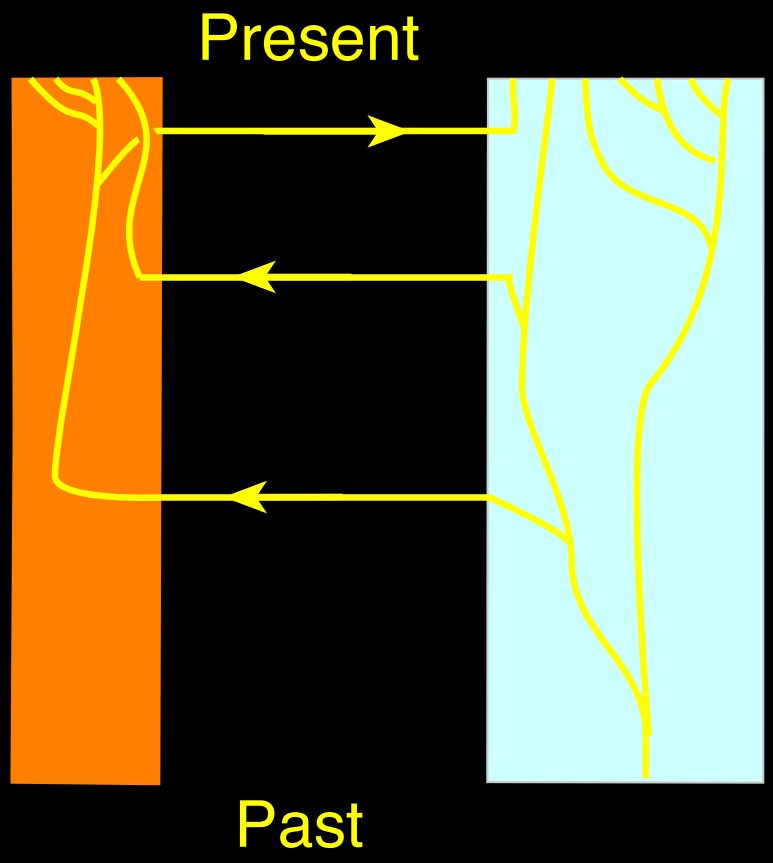
even more complex

Population through time



Populations through time

key assumption of MIGRATE



- migration rate is constant over time
- population sizes are constant over time
- [soon taking into account population splitting]

Inference of parameters

Model of prime interest:

- Geographic structure, colonization, recurrent gene flow, past population splitting, ...

Inference of parameters

Model of prime interest:

- Geographic structure, colonization, recurrent gene flow, past population splitting, ...

But our data is usually not a detailed historical record, so we depend on genetic data. This is problematic because we only see differences in the sequences thus we need some more models.

Inference of parameters

Model of prime interest:

- Geographic structure, colonization, recurrent gene flow, past population splitting, ...

But our data is usually not a detailed historical record, so we depend on genetic data. This is problematic because we only see differences in the sequences thus we need some more models.

Nuisances (we are not really interested in estimating these)

- Mutation model, genealogies of individuals

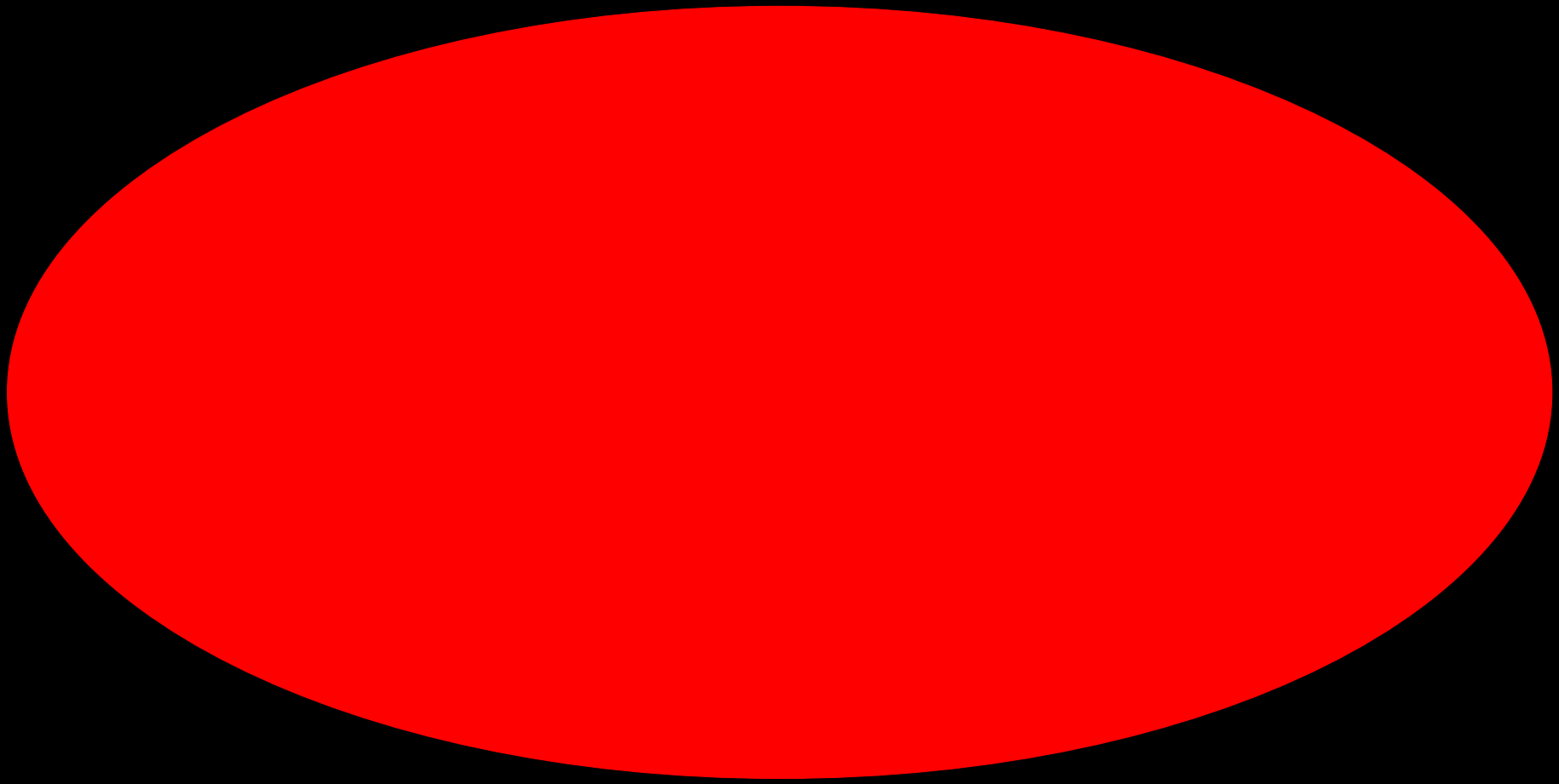
Geographic structure

The program STRUCTURE is used commonly to find the number of populations. Thus judging whether the data is from a structured population or not.

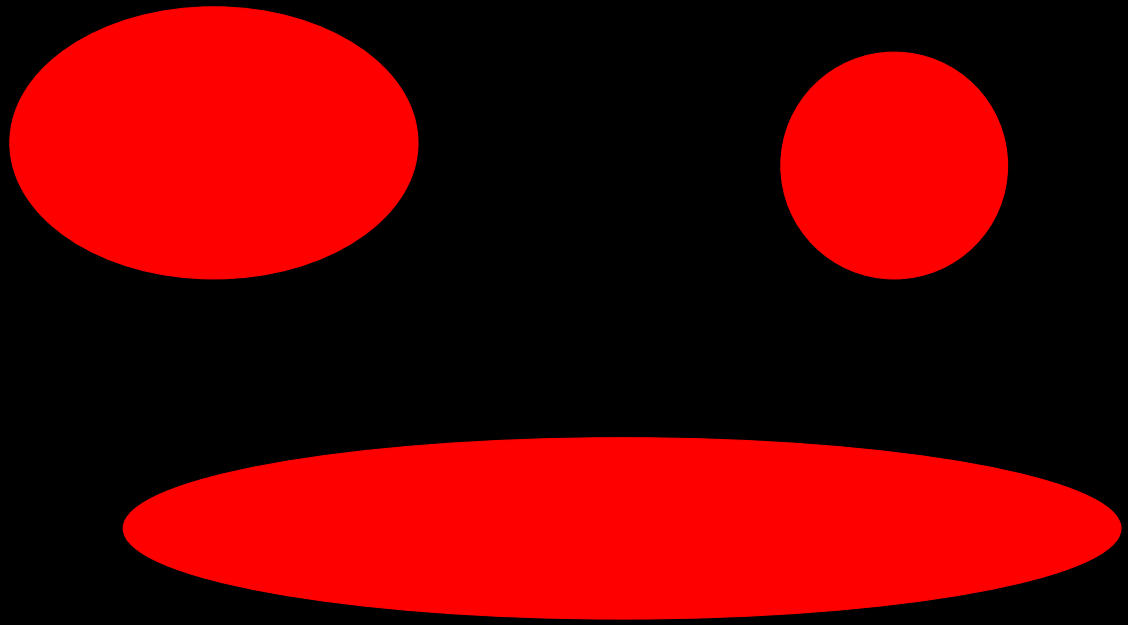
This seems an abuse of a fine program, because the main goal of the program *assigning individuals to populations* is not really used. Other programs, such as *structurama* coestimate assignment and number of populations.

The model used in STRUCTURE, does not take into account different population sizes and potential asymmetries in gene flow and thus will not really be able to give a complete picture of historical events.

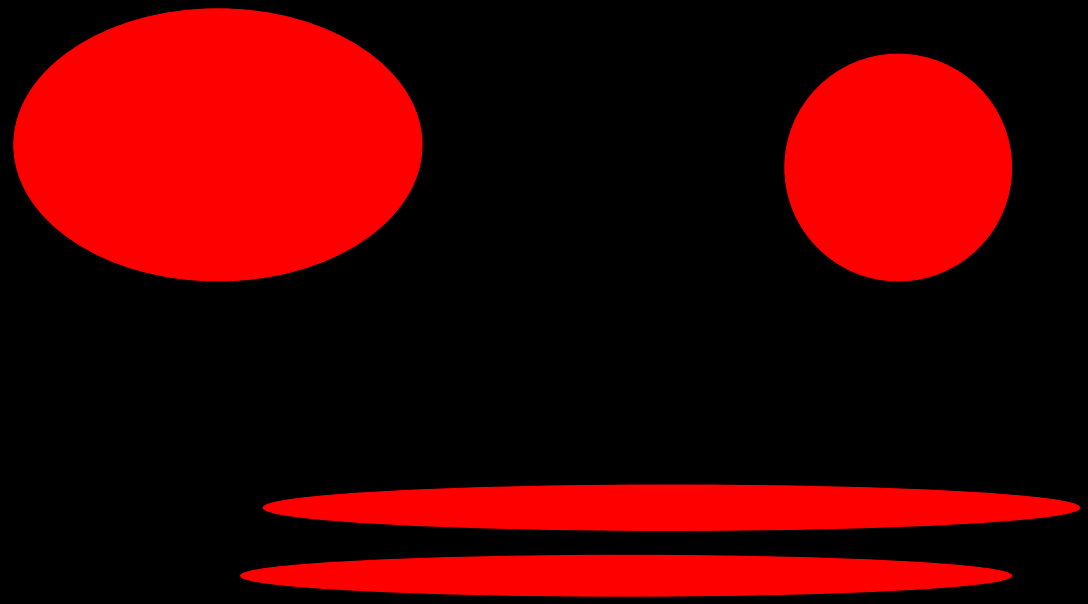
Geographic structure



Geographic structure

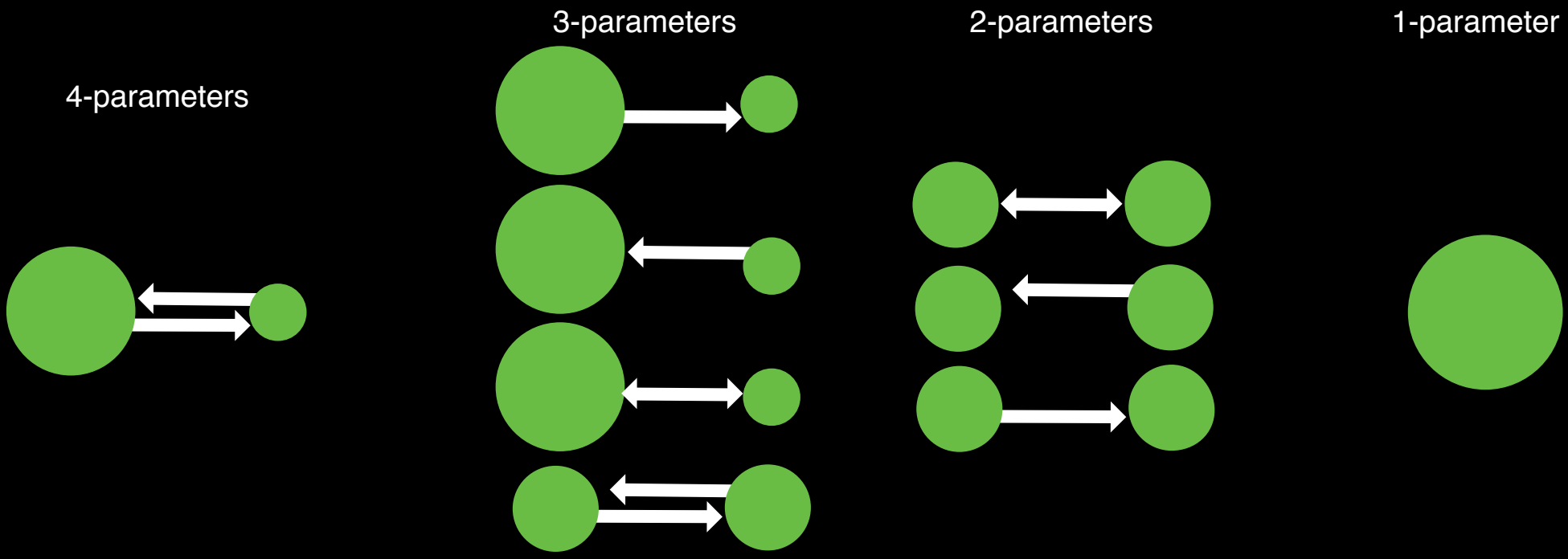


Geographic structure



Model comparison

All simple “two-population” population models that can be use in my software MIGRATE to estimate population parameters using Bayesian inference.



[The size difference of the disks marks independent estimation of population sizes.]

The nitty gritty detail

$P(G, E, S) = P(G)\,P(E|G)\,P(S|E,G)$

$P(G) = \exp\left(-u\left(\sum_i \frac{k_i(k_i-1)}{\theta_i} + \sum_j \sum_i k_i M_{ji}\right)\right)\left(\sum_i \frac{k_i(k_i-1)}{\theta_i} + \sum_j \sum_i k_i M_{ji}\right)$

$P(E|G) = \frac{\eta}{\lambda}\sum_i \partial_i + \frac{\upsilon}{\lambda}\left(1 - \sum_j \partial_j\right)$

$P(S|G,E) = \frac{k_j(k_j-1)}{\theta_j}\,\partial_j + \sum k_i M_{ji}\,(1-\partial_j)$

so we can assemble all parts to get the final result

$\exp\left(-u\left(\sum \frac{k_i(k_i-1)}{\theta_i} + \sum\sum k_i M_{ji}\right)\left(\sum \frac{k_i(k_i-1)}{\theta_i} + \sum\sum k_i M_{ji}\right) *\right.$

$\left.\left(\frac{k_j(k_j-1)}{\theta_j}\,\partial_j + \sum k_i M_{ji}\,(1-\partial_j)\right)\right)$

M_{ji} = migration rate scaled for

$\theta_i = P$

∂_j η 0 if there is a migrat
1 if there is a codes

there are
coalescent ev
induced and
migration
per lineage

1 if coalescence, ∂_j if coalescence, otherwise

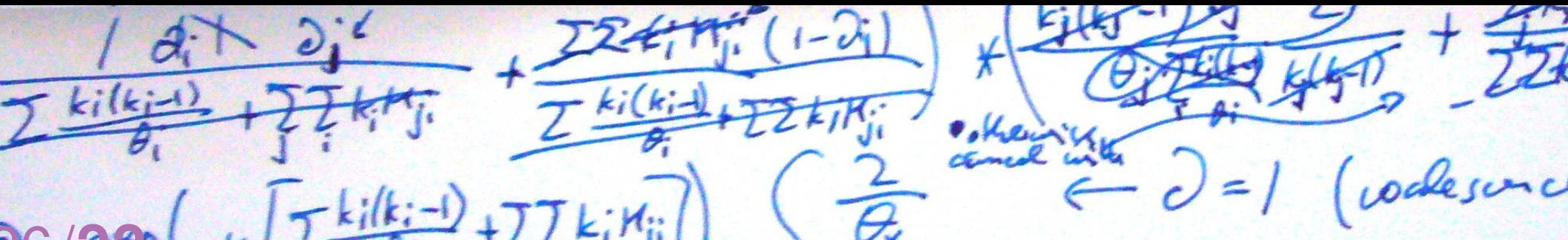
$= \lambda$ ex

$\leftarrow \partial = 1$ (coalesce

The nitty gritty detail

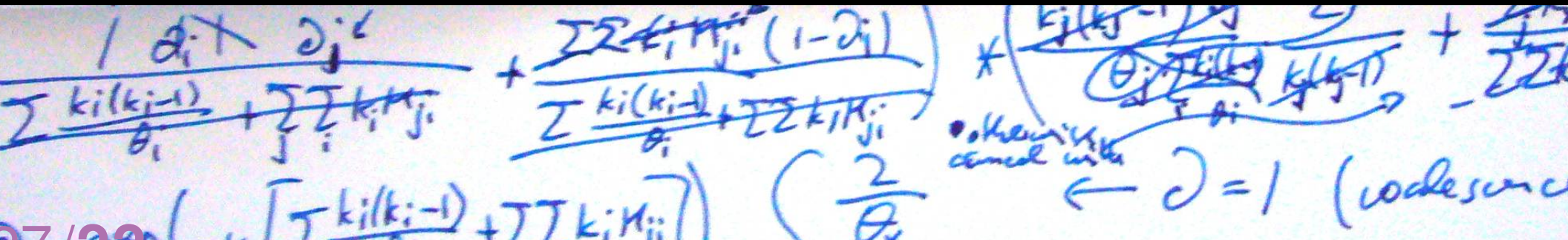
- infer the posterior probability of parameters of a population model

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta)\int_G P(G|\theta)P(D|G,\mu)dG}{\int_P(\theta)\int_G P(G|\theta)P(D|G,\mu)dGd\theta}$$



The nitty gritty detail

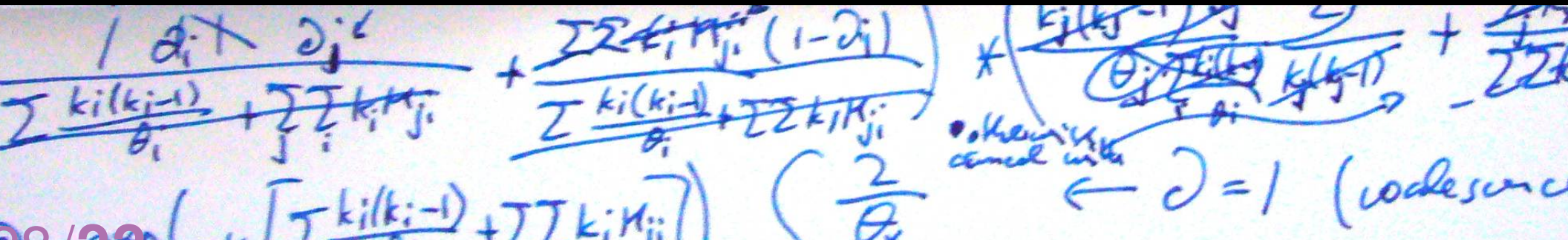
- infer the posterior probability of parameters of a population model, usually using Markov Chain Monte Carlo
- report the posteriors and highlight some differences of the parameter, done!?



The nitty gritty detail

- infer the posterior probability of parameters of a population model, usually using Markov Chain Monte Carlo
- report the posteriors and highlight some differences of the parameter, done!

We can do better than that and stastistically compare different models.



Bayesian Odds Ratios

Using Bayes' theorem:

$$p(M_1|X) = \frac{p(M_1)p(X|M_1)}{p(X)}$$

we can express support of one model over another as a ratio:

$$\frac{p(M_1|X)}{p(M_2|X)} = \frac{\frac{p(M_1)p(X|M_1)}{p(X)}}{\frac{p(M_1)p(X|M_1)}{p(X)}}$$

$$\underbrace{\frac{p(M_1|X)}{p(M_2|X)}}_{\text{Posterior Odds}} = \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{Prior Odds}} \times \underbrace{\frac{p(X|M_1)}{p(X|M_2)}}_{\text{Bayes Factor}}$$

Bayes factor

We can use the posterior odds ratio or equivalently the Bayes factors for model comparison:

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

The magnitude of BF gives us evidence how different the models are

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

Marginal likelihood calculation

In MCMC application it is often complicated to calculate marginal likelihoods. Several approaches were put forward, of which the easiest, the harmonic mean estimator, has turned out to be unreliable and sometimes wrong.

Several other methods give accurate marginal likelihoods:

- Thermodynamic integration [MIGRATE uses this]
- Stepping-stone integration
- Inflated Density Ratio

A simple example

We want to establish a direction of geneflow between n populations.

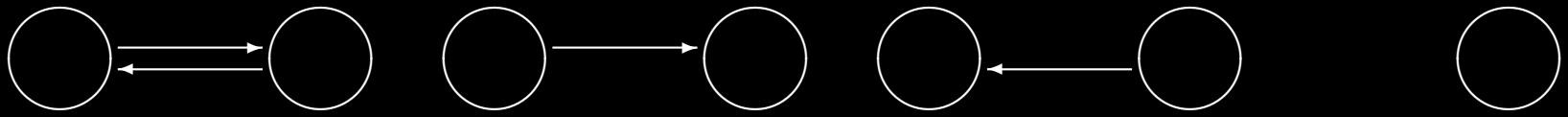
A simple example

We want to establish a direction of geneflow between 2 populations.

A simple example

We want to establish a direction of geneflow between 2 populations.

We generate 4 hypotheses



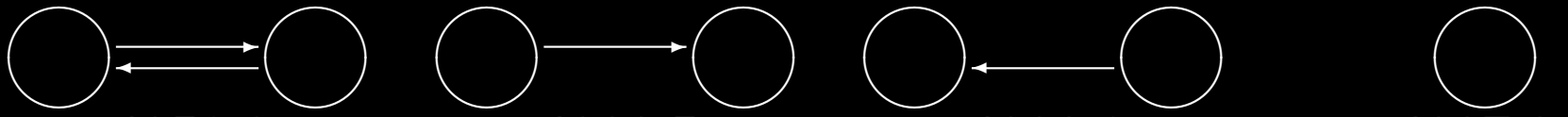
We collect data from individuals in the two populations

Analyze the data in MIGRATE

A simple example

Recipe: starting with the finished dish

Log Marginal likelihoods [lmL] of the 4 hypotheses:



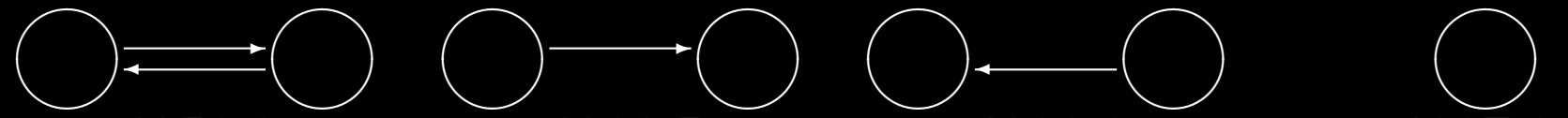
lmL	-4856.2	-4822.5	-4832.6	-4837.8

Data was simulated using the second model (2) from the left.

A simple example

Recipe: starting with the finished dish

Log Marginal likelihoods [lmL] of the 4 hypotheses:



lmL -4856.2 -4822.5 -4832.6 -4837.8

The best model (highest lmL) is the model second from left (model 2).
We can calculate the log Bayes factor for two leftmost models as

$$LBF_{12} = 2(lmL_1 - lmL_2) = 2(-4856.2 - -4822.5) = -67.4$$

The value suggests that we should strongly prefer model 2 over model 1.

Data was simulated using the second model from the left (model 2).

A simple example

Recipe:

1. Pick the hypothesis with largest number of parameters
2. Set priors and run parameters (use heated chains) so that you are comfortable with the result (converged, etc)
3. Record the log marginal likelihood from the output.
4. Pick next hypothesis, adjust migration model, and run and record the log marginal likelihood.
5. Repeat (4) until all log marginal likelihoods are calculated
6. Compare the log marginal likelihoods, for example order the hypothesis accordingly, or calculate the model probability

A simple example

Ordered models

lmL	-4822.5	-4832.6	-4837.8	-4856.2
P(model)	0.99	0.01	0.0	0.0

Model probability (Burnham and Anderson 2002) calculation:

$$\mathrm{P}(M_i) = \frac{\exp(lmL_i)}{\sum_j \exp(lmL_j)} = \frac{mL_i}{\sum_j mL_j}$$

Marginal likelihood for lots of data

Running complex models with many genetic loci using MCMC are often very time consuming (hours, days, weeks computing time). In genetics we have often independent genes (loci) that allow us to treat the analysis as if we evaluate multiple independent replicates.

$$P(D|M_1) = P(D_1, ..., D_n|M_1)$$

Unfortunately live is an inter-dependent mess and we cannot do

$$P(D_1, ..., D_n|M_1) \neq \prod_i^n P(D_i|M_1)$$

First we thought that we are doomed to run all the independent data blocks in sync to calculate the combine marginal likelihood.

Marginal likelihood for lots of data

Theorem: The combined marginal likelihoods over all independent data blocks can be calculated as a product of independently calculated marginal likelihoods for each data block and a constant. (Proof in Beerli and Palczewski 2010)

$$P(D_1, ..., D_n|M_1) = K \prod_i^n P(D_i|M_1)$$

$$K = \int_\theta \prod_i^n P(\theta|D_i, M_1) P(\theta|M_1)^{1-n} d\theta.$$

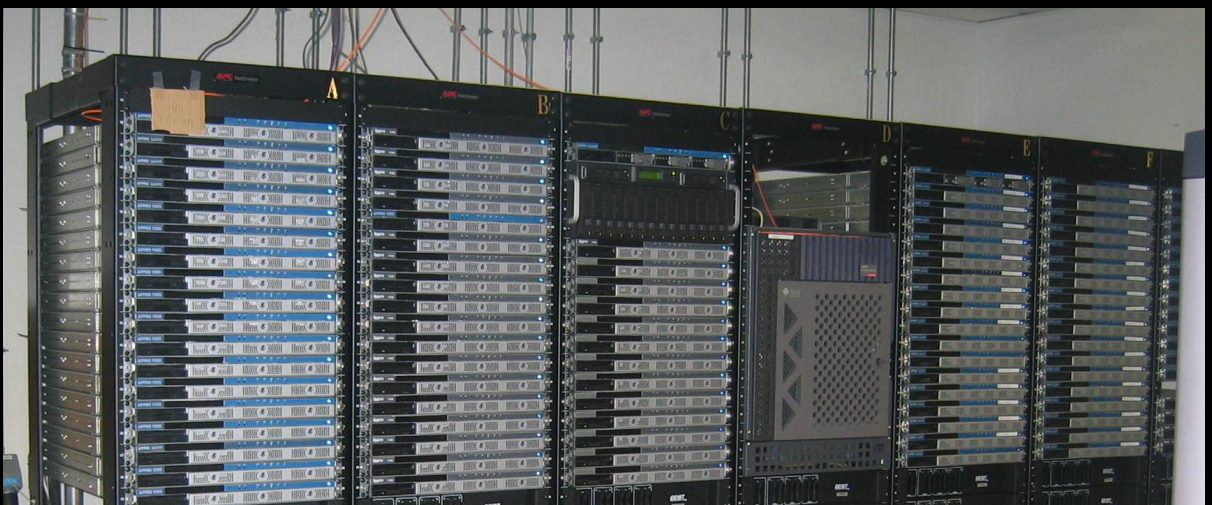
Marginal likelihood for lots of data

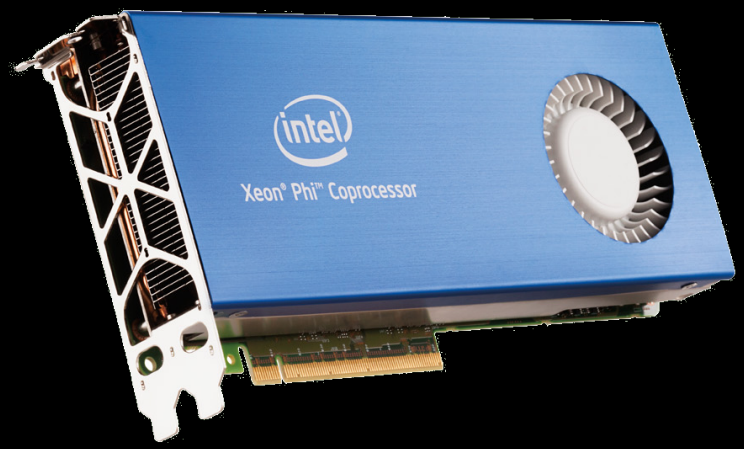
Theorem: The combined marginal likelihoods over all independent data blocks can be calculated as a product of independently calculated marginal likelihoods for each data block and a constant. (Proof in Beerli and Palczewski 2010)

$$P(D_1, ..., D_n|M_1) = K \prod_i^n P(D_i|M_1)$$

$$K = \int_\theta \prod_i^n P(\theta|D_i, M_1) P(\theta|M_1)^{1-n} d\theta.$$

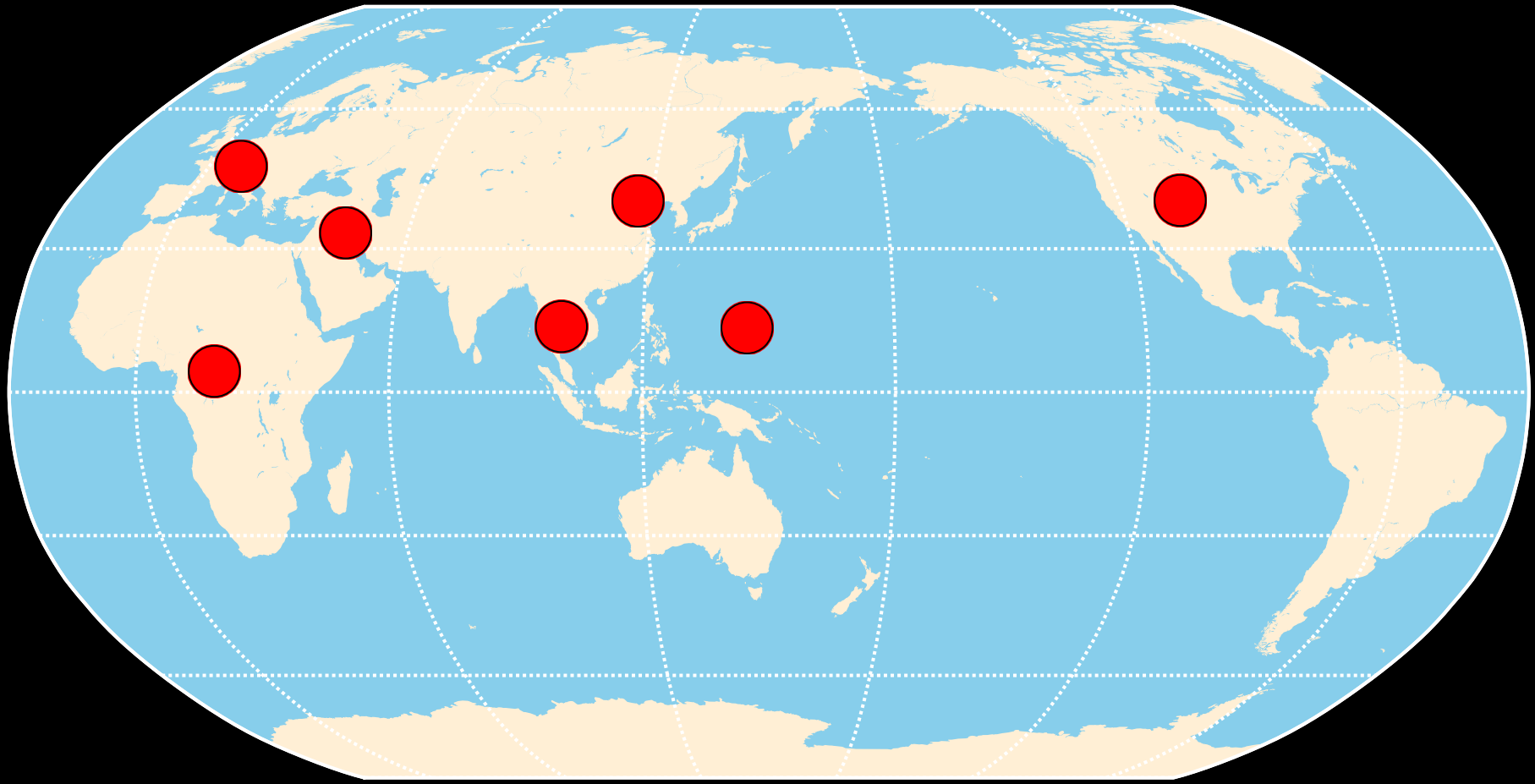
This allows to run independent data blocks in parallel on a computer cluster!





Marginal likelihood for lots of data

Locations of samples [377 microsatellites]

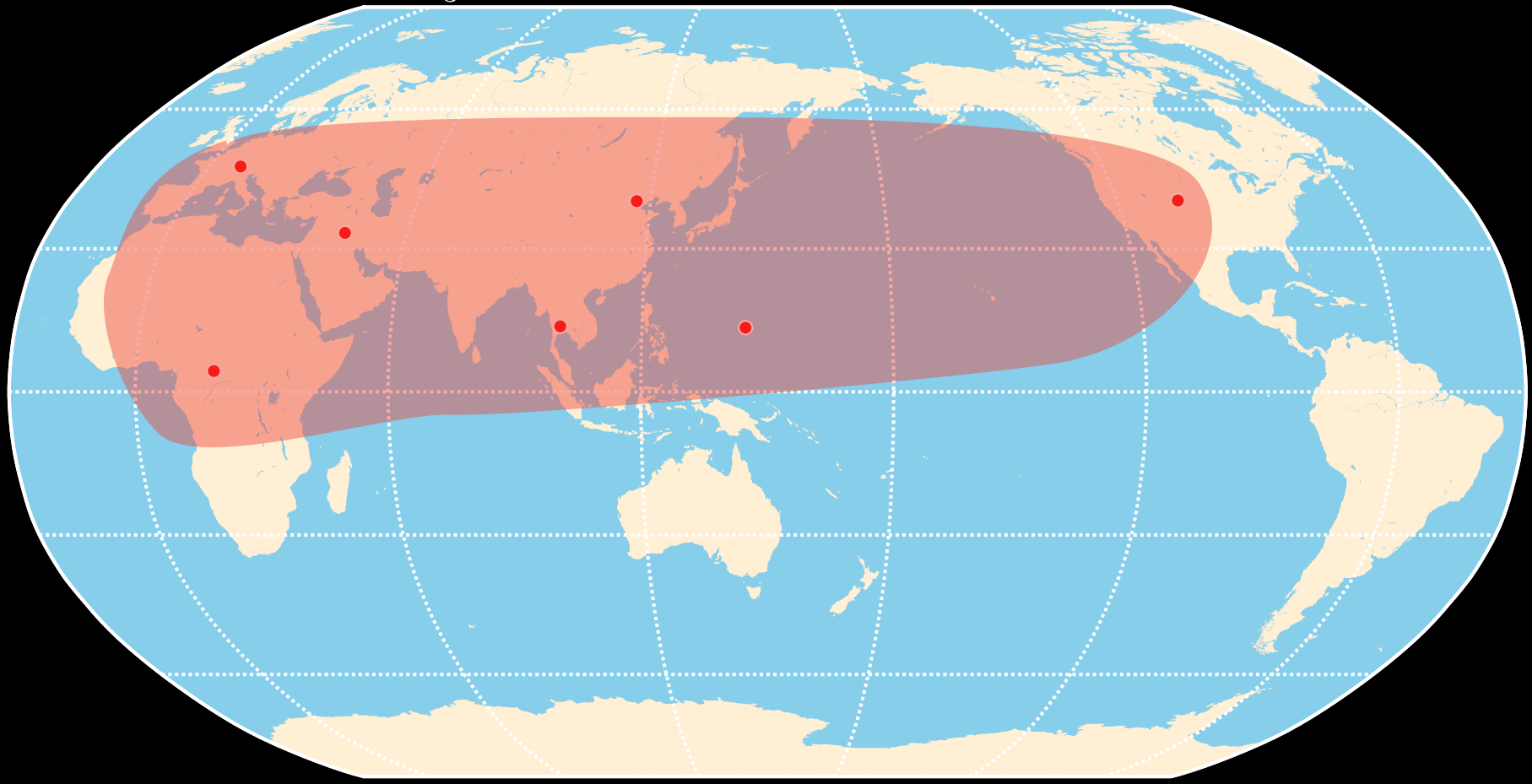


70 individuals from 7 populations analyzed for 377 microsatellite loci:
Brownian motion approximation to the single-step mutation model

Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

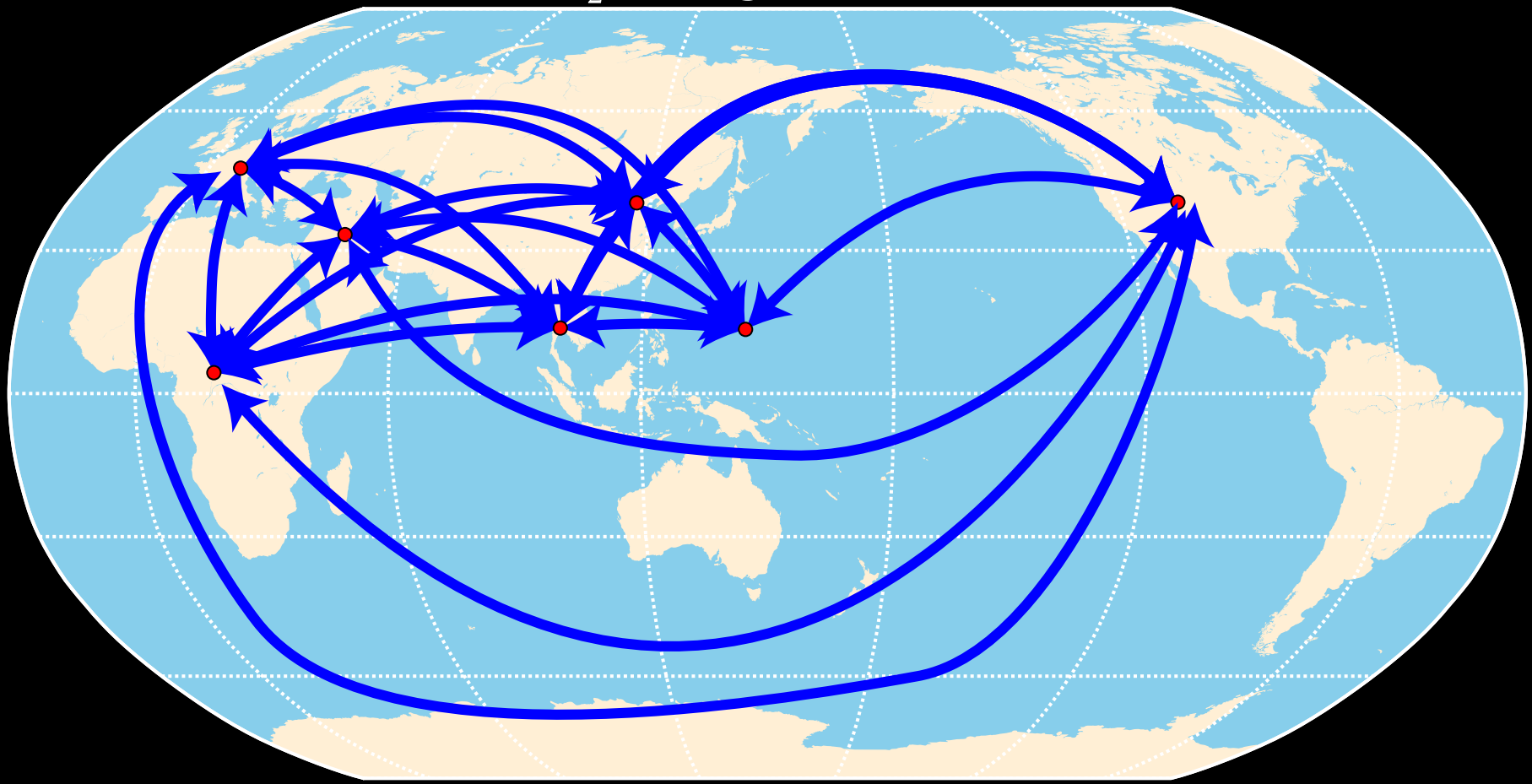
H_3 : One panmictic population



Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

H_2 : Tangled mess

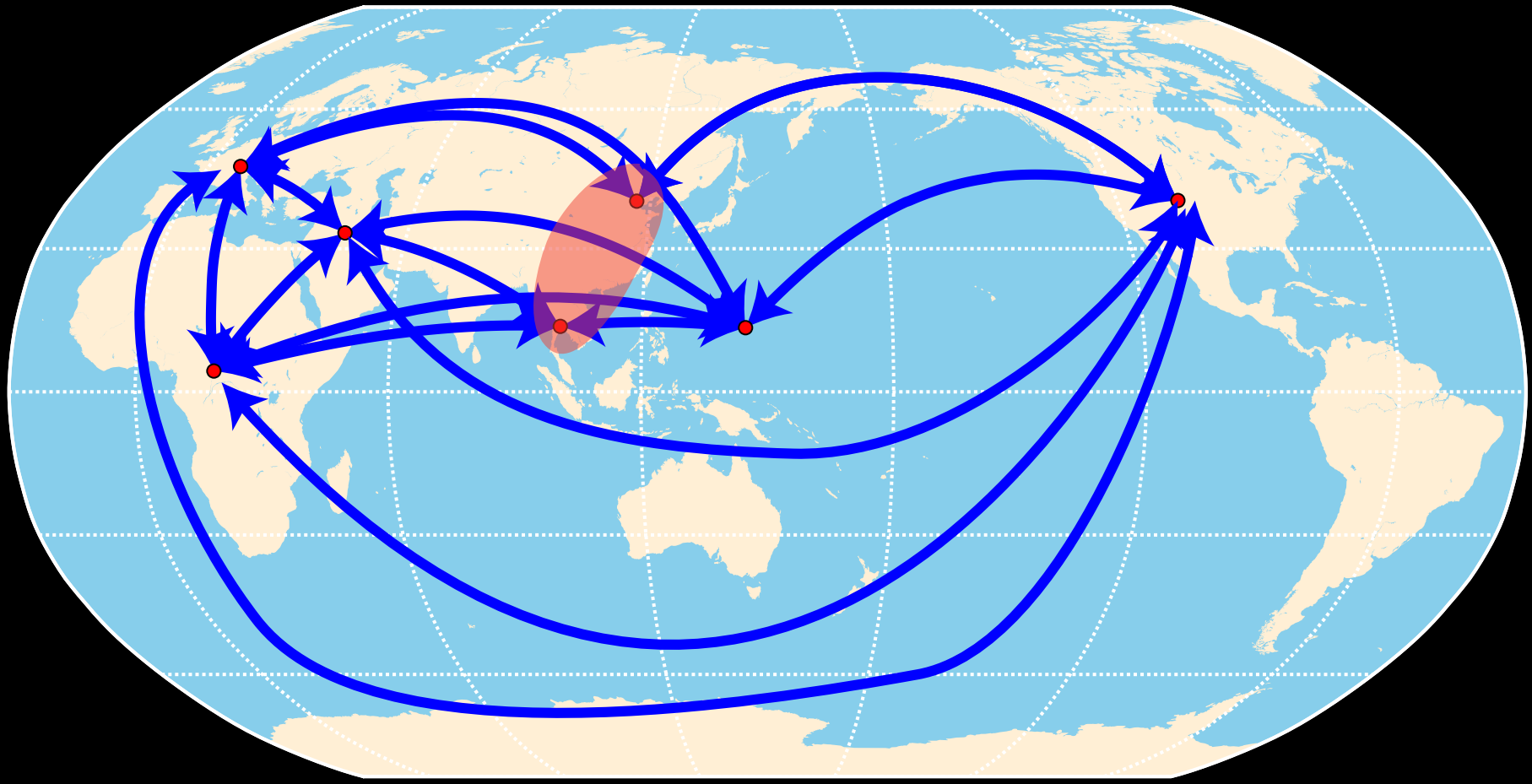


Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

Somewhat less

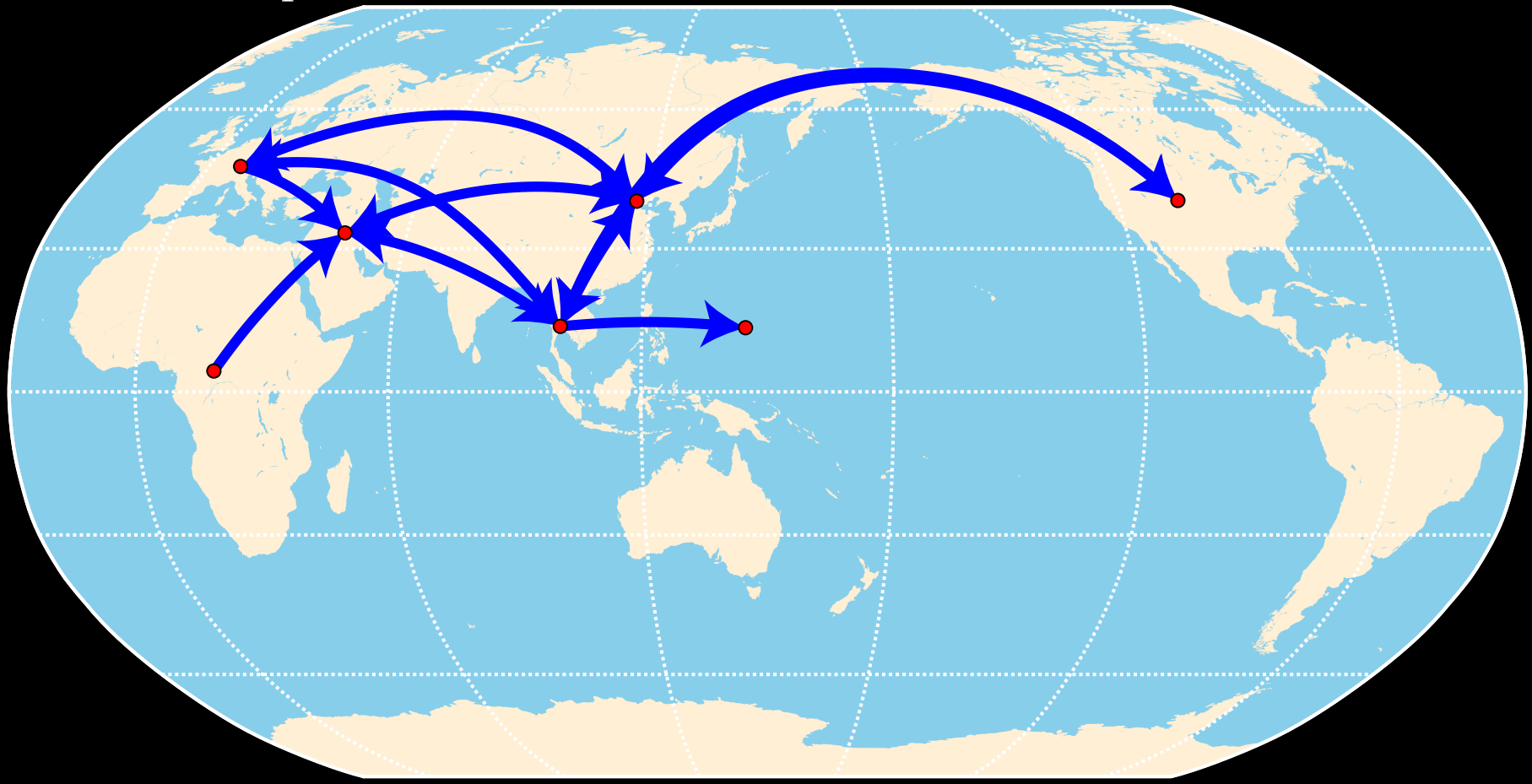
H_4 : Tangled mess



Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

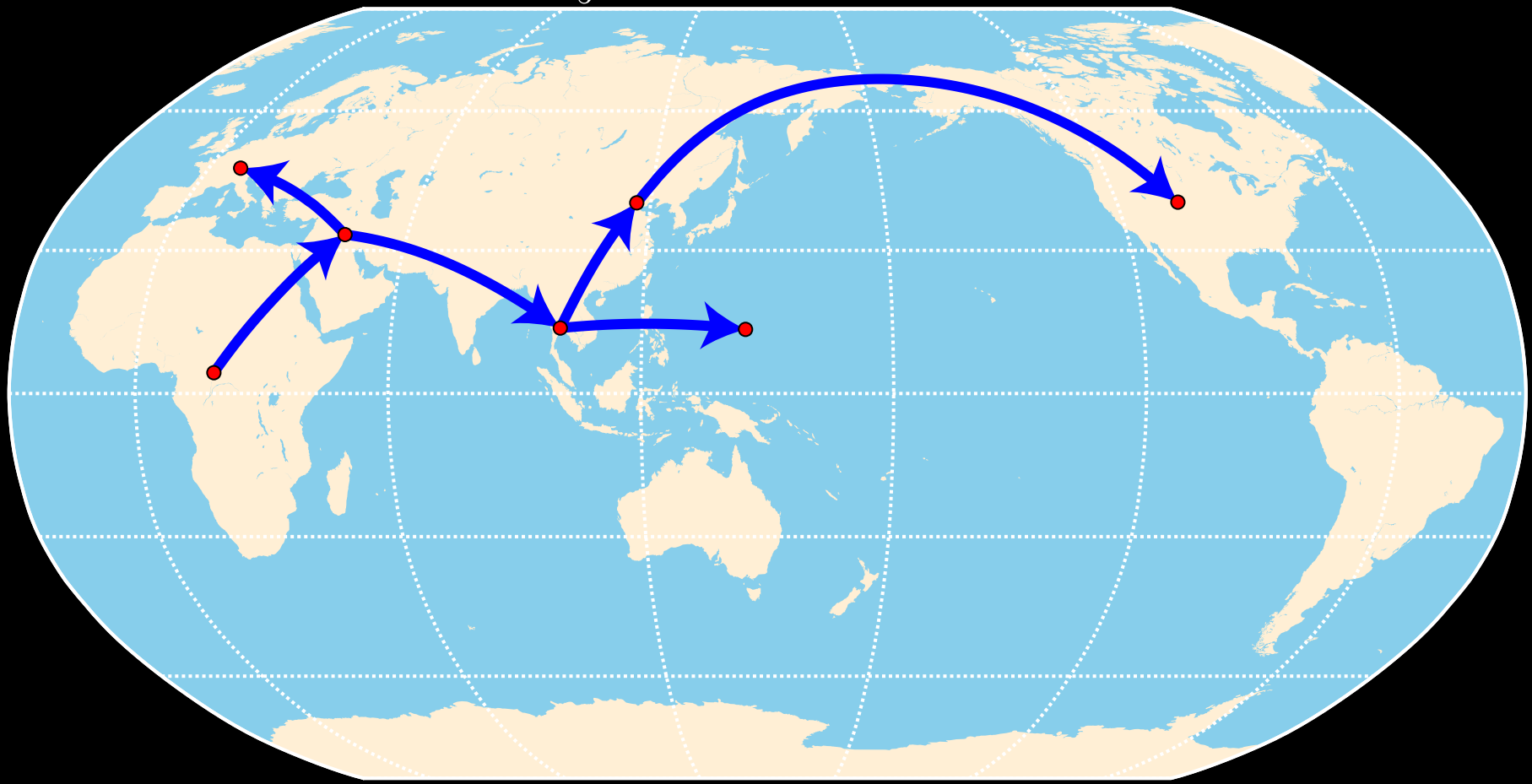
H_1 : Out of Africa, indecision anywhere else



Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

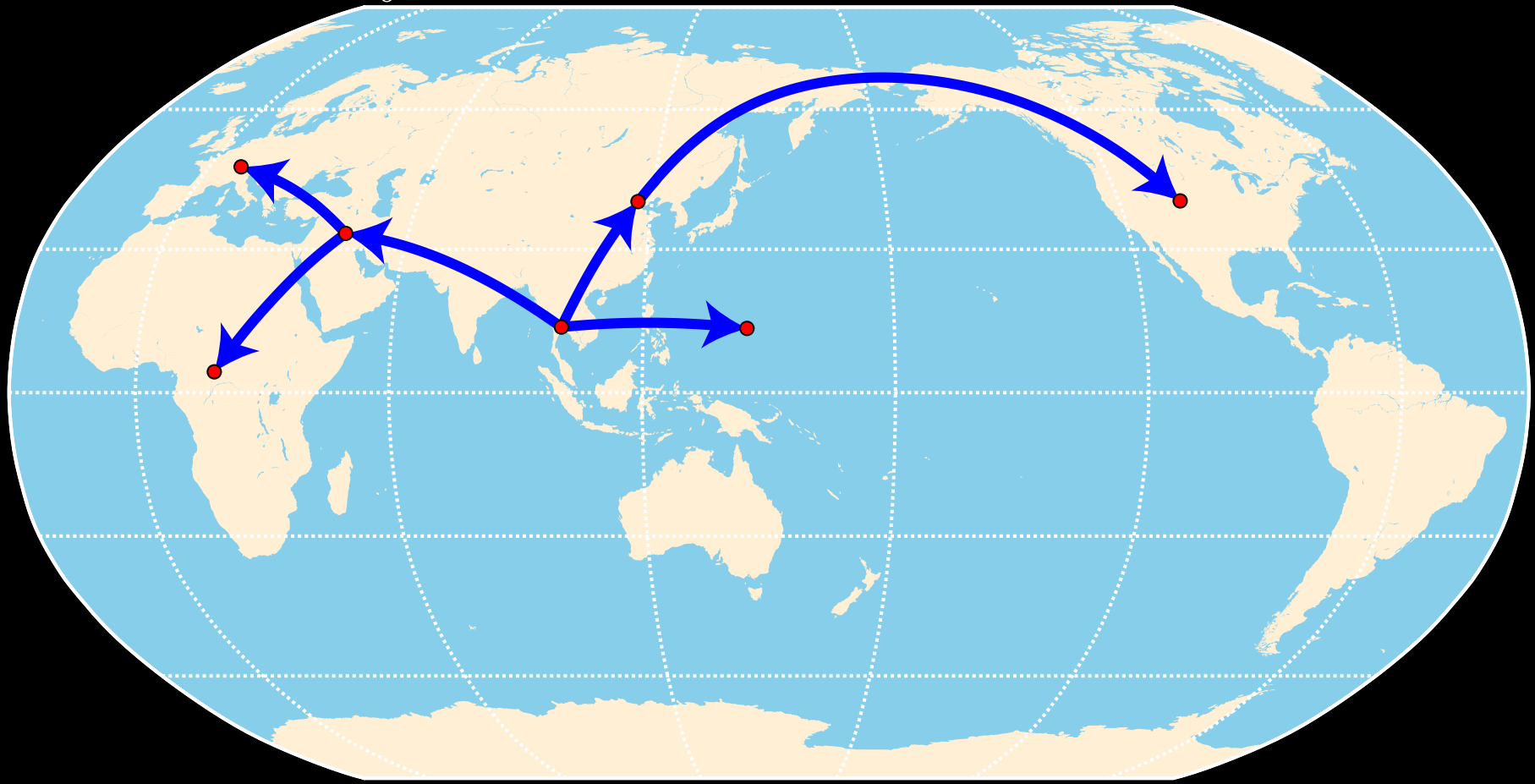
H_5 : Minimal model



Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

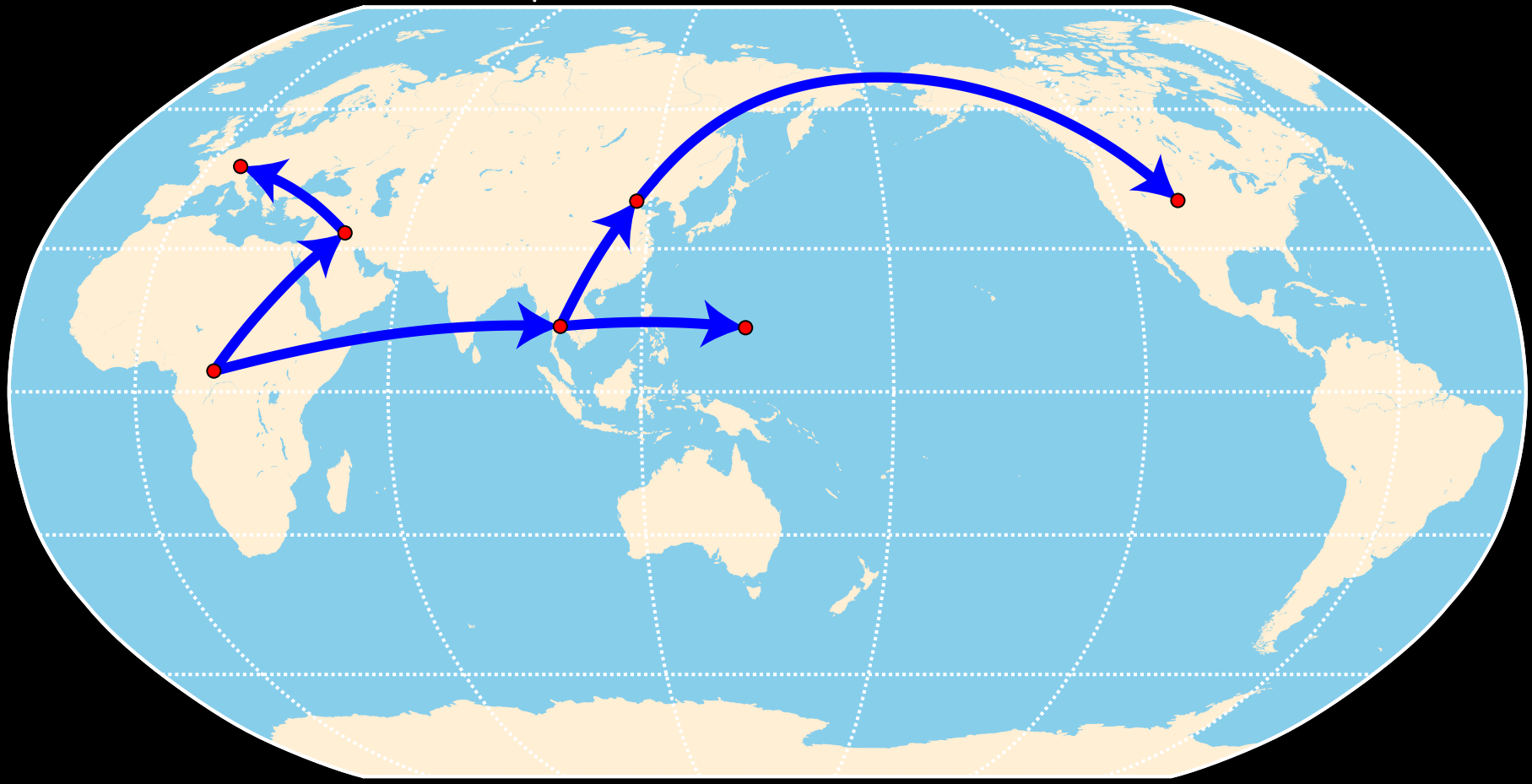
H_6: South-Asia is cradle of humans



Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

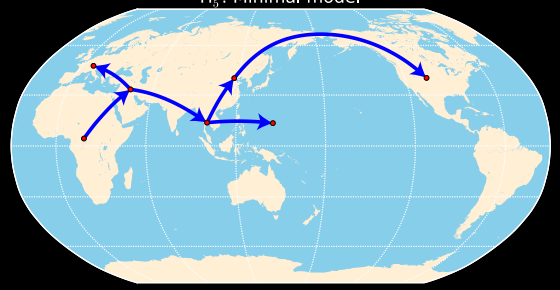
H_7 : Direct train to Asia



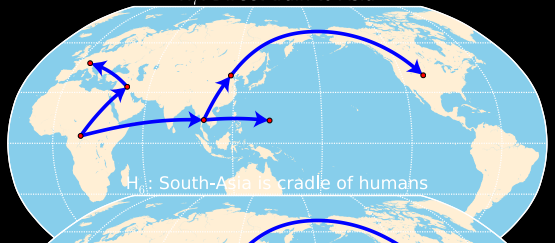
Reanalysis of data from Rosenberg et al. Science 2001

Pick a model

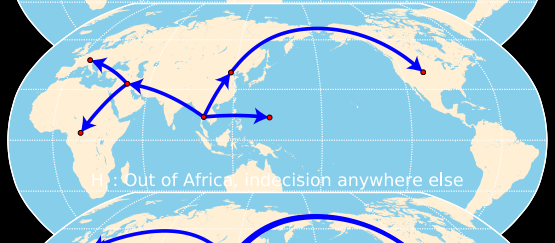
1. H_5: Minimal model



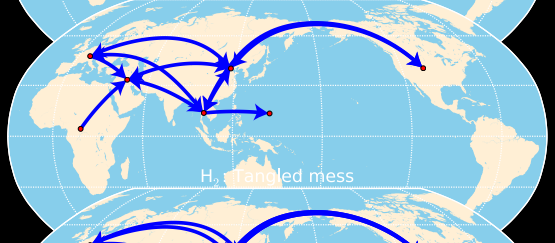
2. H_7: Direct train to Asia



3. H_6: South-Asia is cradle of humans

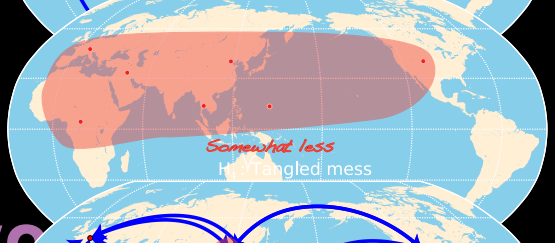


4. H_1: Out of Africa, indecision anywhere else



5. H_3: Tangled mess

6. H_4: One panmictic population

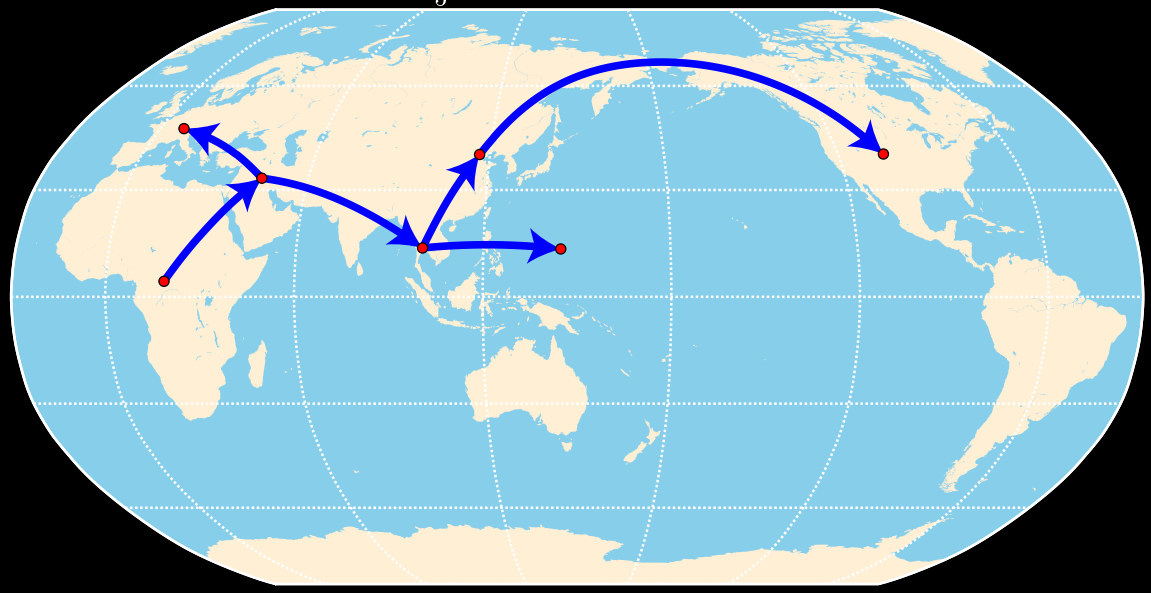


7. H_2: Tangled mess

Model order and probability using Bayes factors

all other models: 0.0
Minimal model 1.0

H_5: Minimal model



Reanalysis of data from Rosenberg et al. Science 2001

Summary

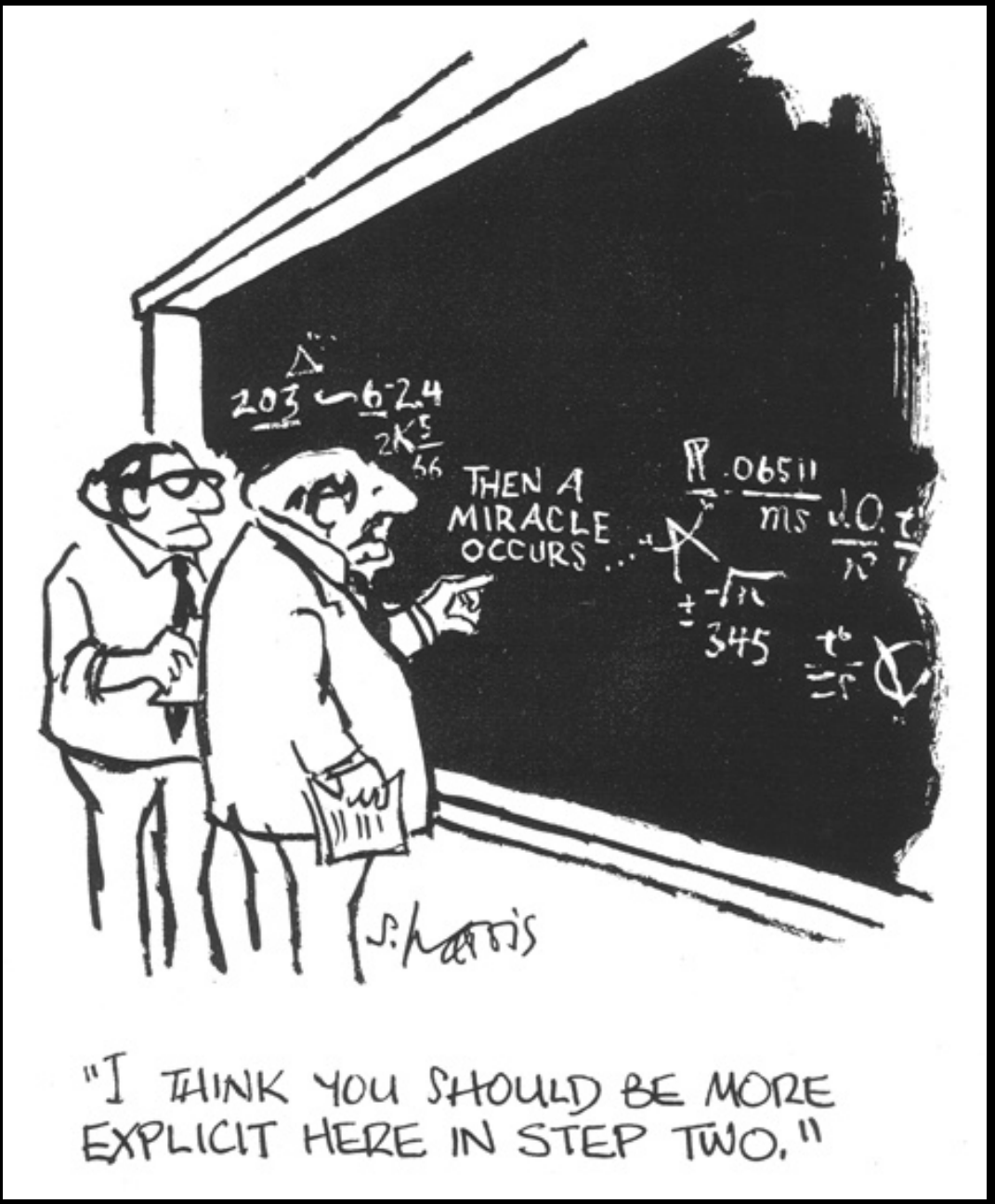
Caveats:

- MIGRATE supports a large list of models but that may not be sufficient for your hypothesis.
- MIGRATE assumes a simple population model that may not fit your data.

Plus:

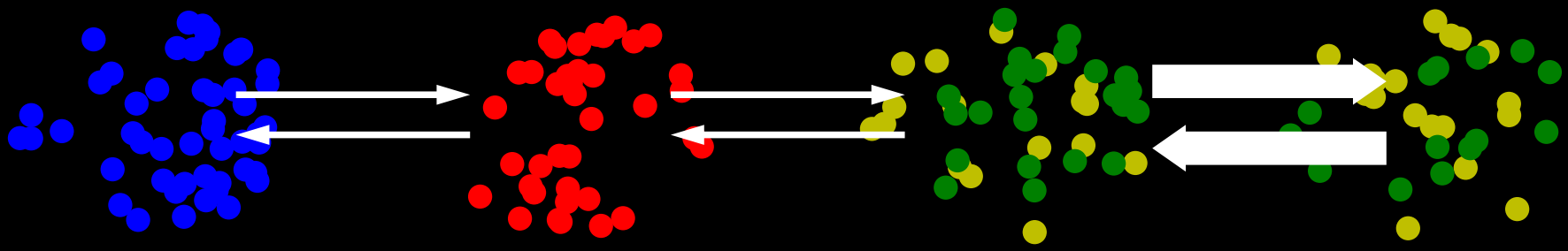
- MIGRATE supports a large list of models.
- MIGRATE can run in parallel allowing to analyze large numbers of loci in decent time.
- Bayesian model selection allows comparison of non-nested models.

Questions?



MIGRATE website:
http://popgen.sc.fsu.edu

Assignment of individuals



Migrate (this summer) will be able to assign individuals to populations, for example we collect individuals from 3 locations and wonder whether these individuals at the center location are local or recognizable immigrants.

Migration rate M	Average assignment probability			
	1	2	3	4
-	3	2	2	2
low migration	1.0	1.0	0.86	0.81
medium	0.98	0.99	0.64	0.53
high	0.33	0.20	0.68	0.36