Coalescent Likelihood Methods

Mary K. Kuhner
Genome Sciences
University of Washington
Seattle WA

Outline

1. **Introduction to coalescent theory**
2. Practical example
3. Genealogy samplers
4. Break
5. Survey of samplers
6. Evolutionary forces
7. Practical considerations

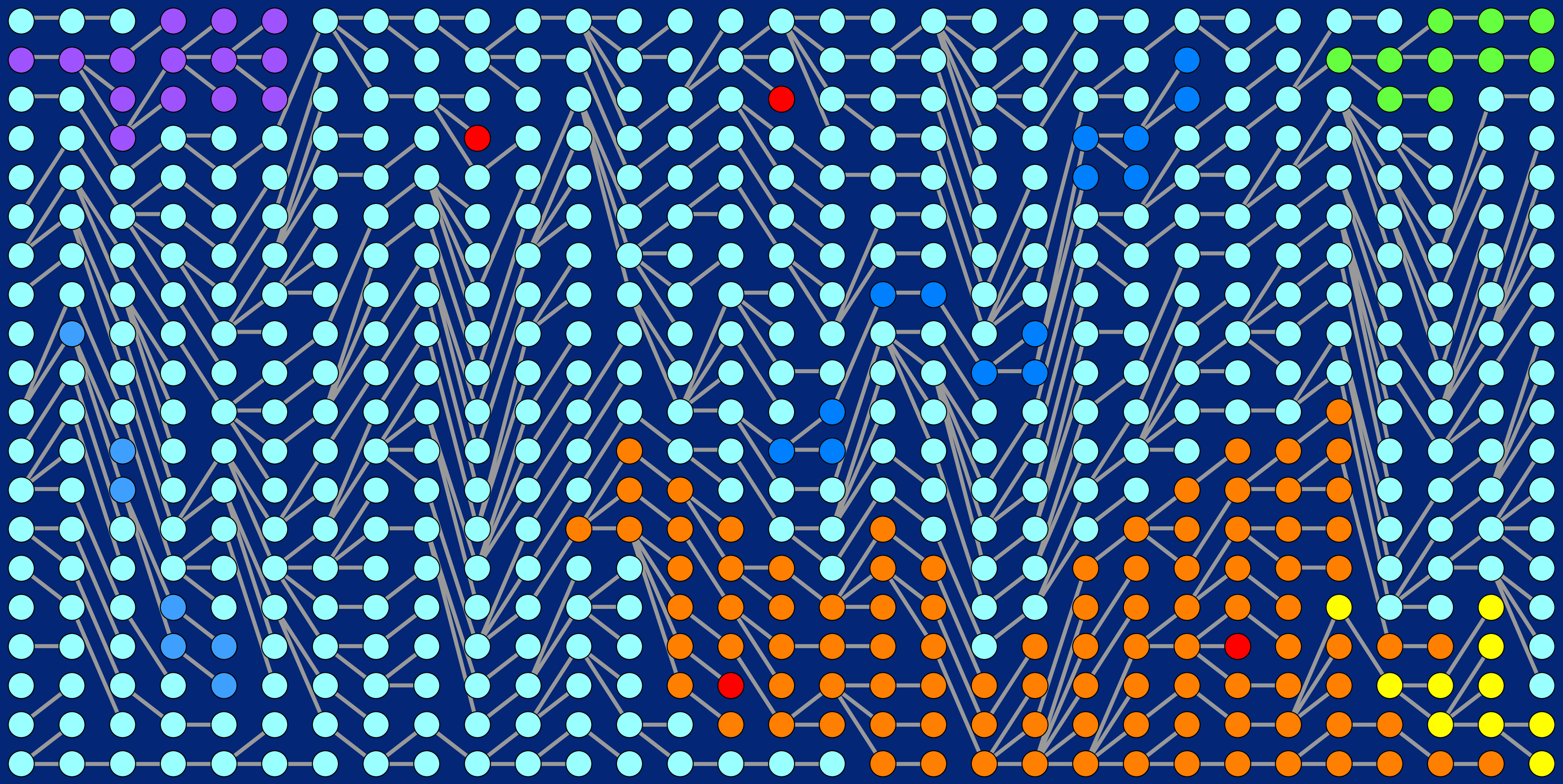
Population genetics can help us to find answers

- We are interested in questions like
  - How big is this population?
  - Are these populations isolated? How common is migration?
  - How fast have they been growing or shrinking?
  - What is the recombination rate across this region?
  - Is this locus under selection?
- All of these questions require comparison of many individuals.

Coalescent-based studies

- How many gray whales were there prior to whaling?
- When was the common ancestor of HIV lines in a Libyan hospital?
- Is the highland/lowland distinction in Andean ducks recent or ancient?
- Did humans wipe out the Beringian bison population?
- What proportion of HIV virions in a patient actually contribute to the breeding pool?
- What is the direction of gene flow between European rabbit populations?

Basics: Wright-Fisher population model



All individuals release many gametes and new individuals for the next generation are formed randomly from these.

Wright-Fisher population model

- Population size N is constant through time.
- Each individual gets replaced every generation.
- Next generation is drawn randomly from a large gamete pool.
- Only genetic drift affects the allele frequencies.

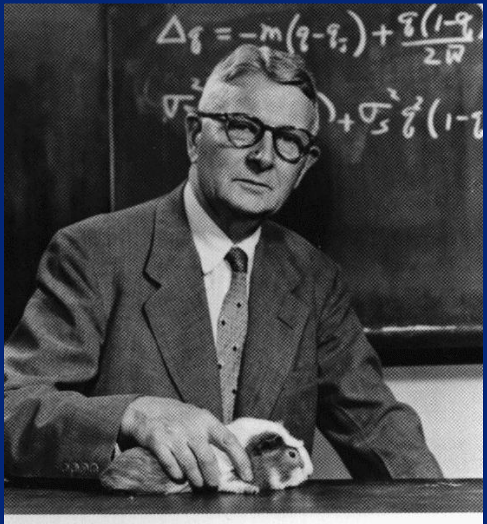
Other population models

- Other population models can often be equated to Wright-Fisher
- The N parameter becomes the effective population size N_e
- For example, cyclic populations have an N_e that is the harmonic mean of the various sizes

The big trick

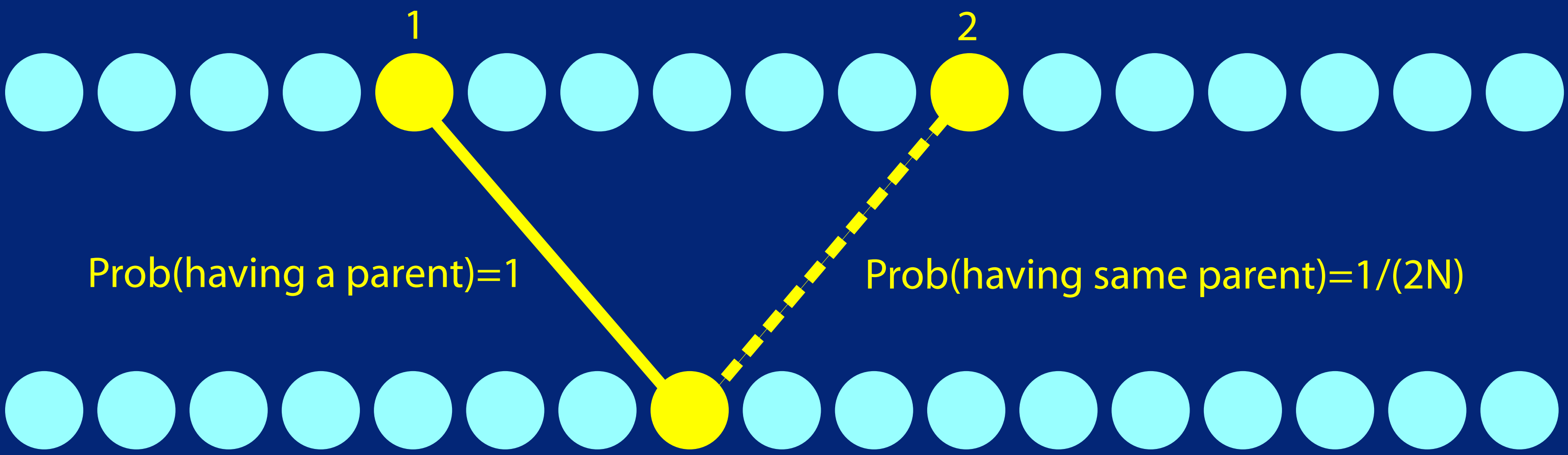
- We have a model for the progress of a population forward in time
- What we observe is the end product: genetic data today
- We want to reverse this model so that it tells us about the *past* of our sequences

The Coalescent

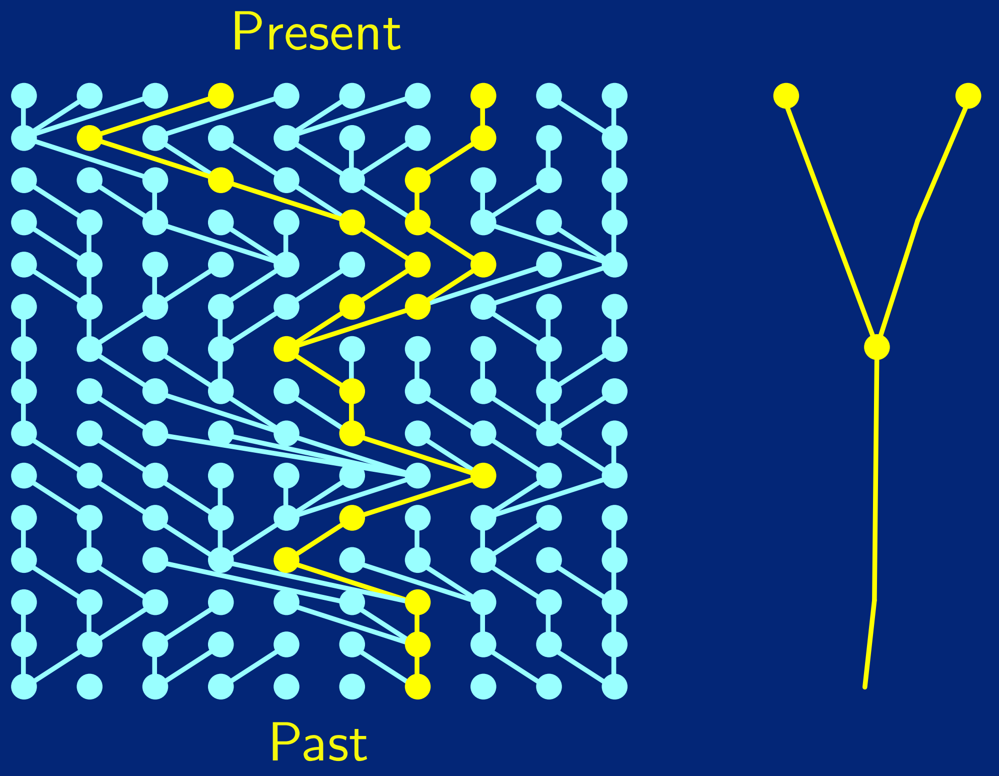


Sewall Wright showed that the probability that 2 gene copies come from the same gene copy in the preceding generation is

$$\mathrm{Prob}\ (\text{two genes share a parent}) = \frac{1}{2N}$$



The Coalescent



In every generation, there is a chance of $1/2N$ to coalesce. Following the sampled lineages through generations backwards in time we realize that it follows a geometric distribution with

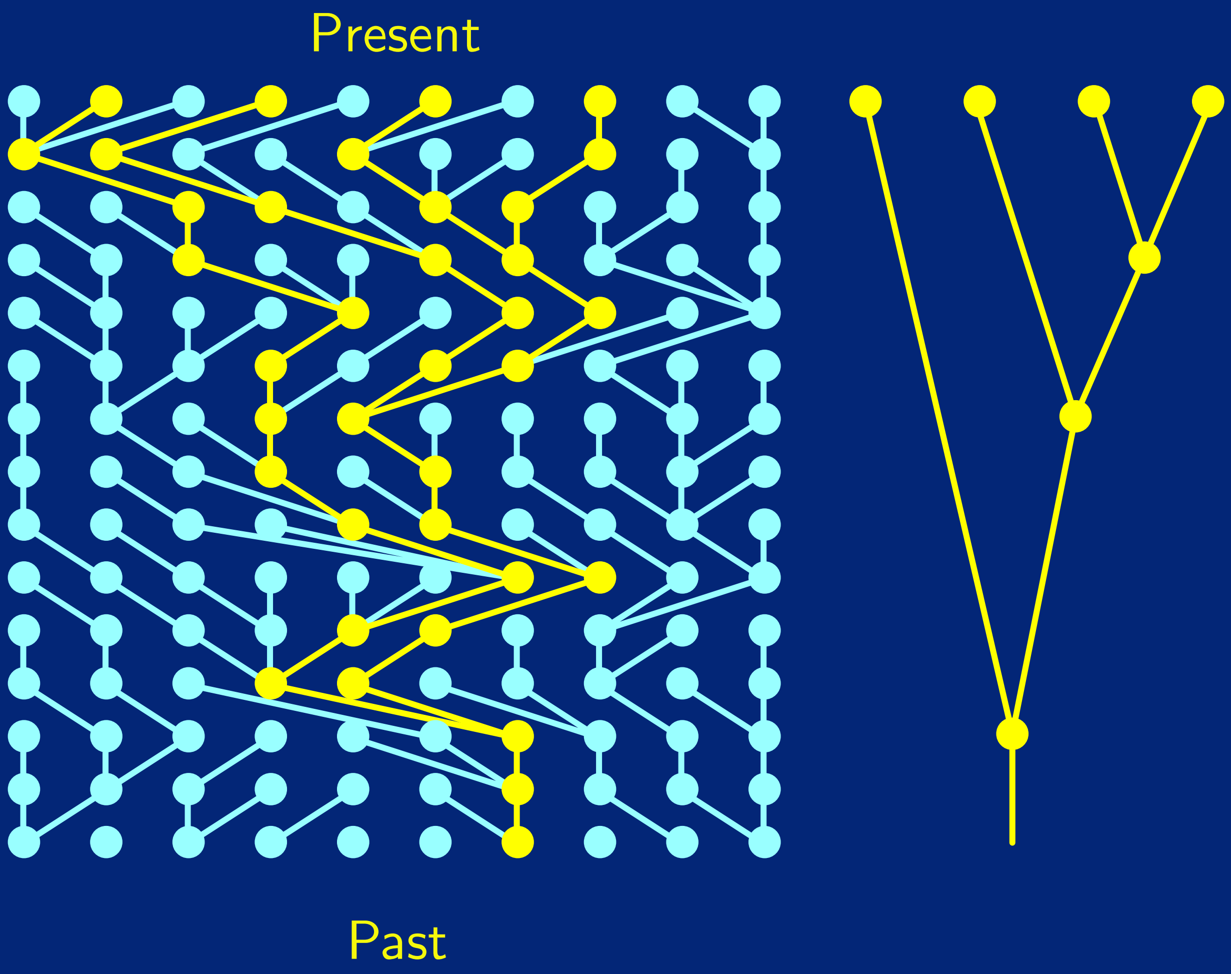
$\mathbb{E}(u) = 2N$ [the expectation of the time of coalescence u of **two** tips is $2N$]

The Coalescent

JFC Kingman generalized this for k gene copies.

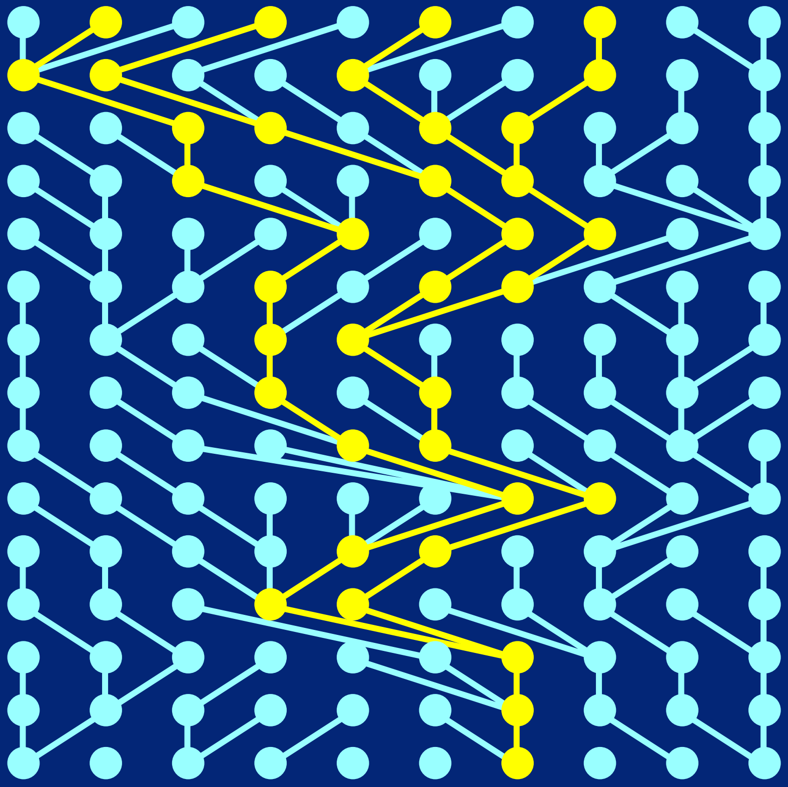
$$\text{Prob}\ (k \text{ copies are reduced to } k-1 \text{ copies}) = \frac{k(k-1)}{4N}$$

Kingman's n-coalescent



Kingman's n-coalescent

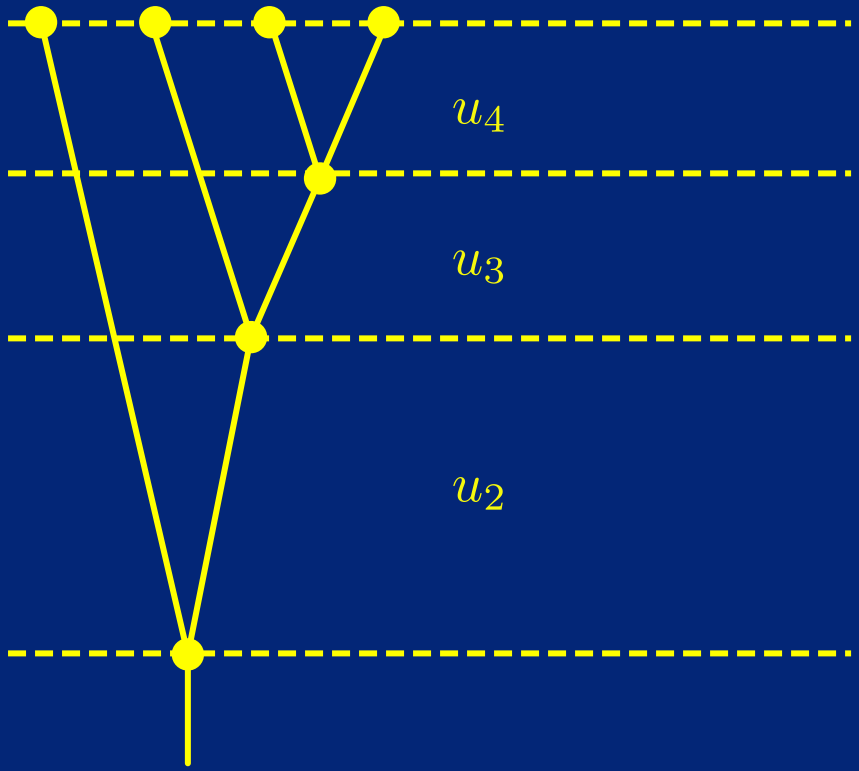
Present



Past

The expectation for the time interval u_k is

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)}$$

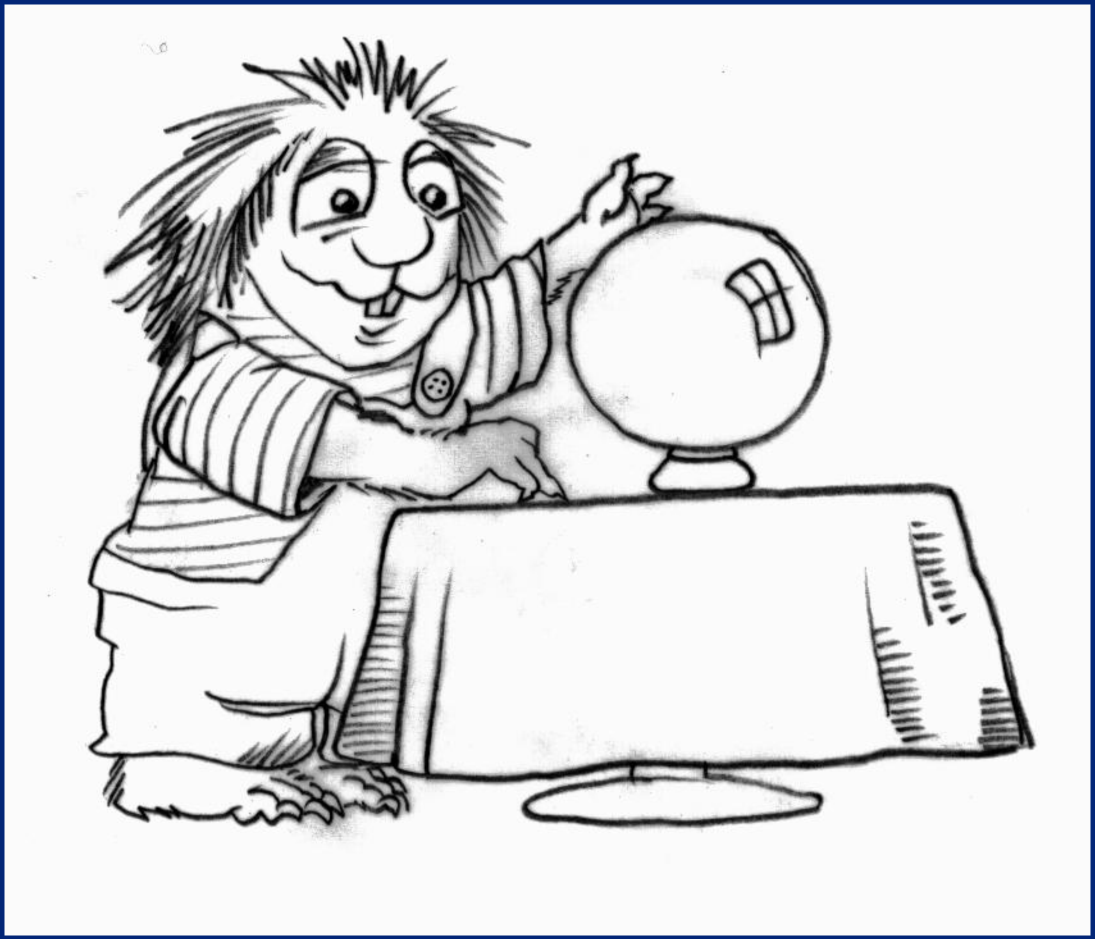


$$\mathrm{p(G|N) = \prod_i \exp(-u_i \frac{k(k-1)}{4N}) \frac{1}{2N}}$$

The Θ parameter

- The n-coalescent is defined in terms of N_e and time.
- We cannot measure time just by looking at genes, though we can measure divergence.
- We rescale the equations in terms of N_e, time, and the mutation rate μ.
- We can no longer estimate N_e but only the composite parameter Θ.
- $\Theta = 4N_e\mu$ in diploids.
- Multiple time point data can separate N_e and μ

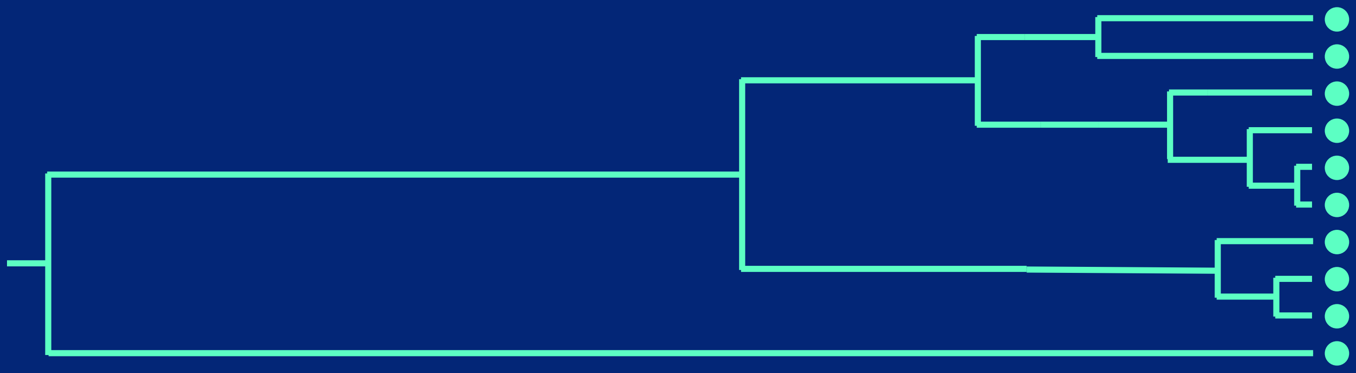
What is this coalescent thing good for?



Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

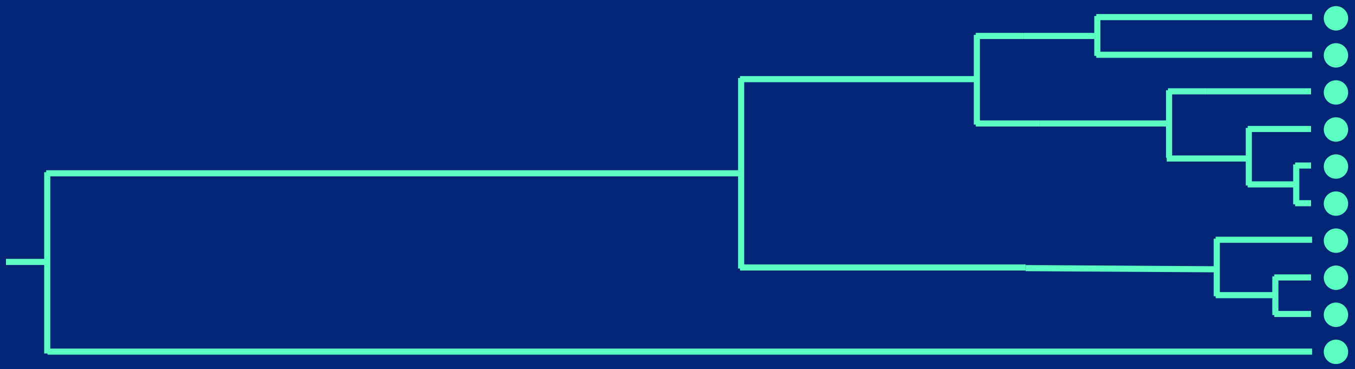
2. We know that we can calculate $\mathrm{p}(\text{Genealogy}|\mathrm{N})$



Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation

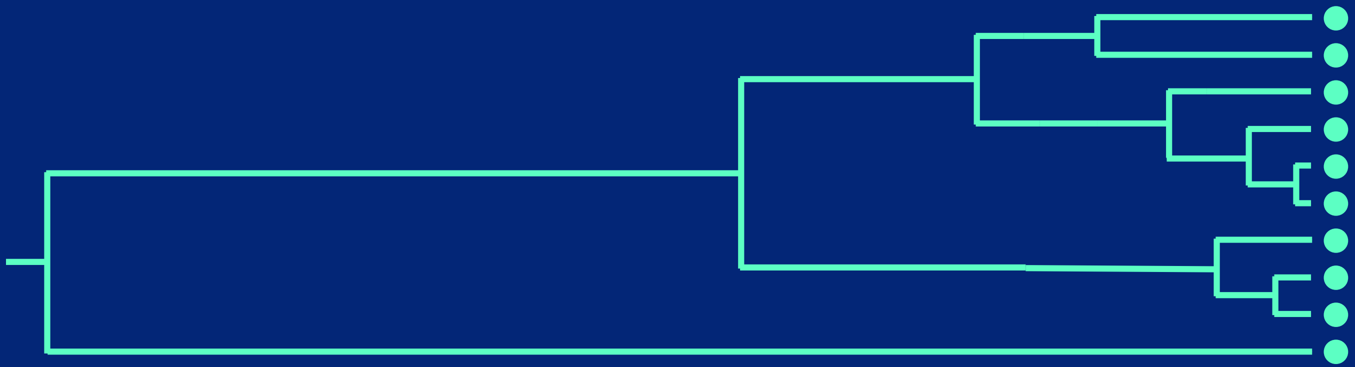


$$p(G|N) = p(u_1|N,k)\frac{1}{2N} \times p(u_2|N,k-1)\frac{1}{2N} \times .....$$

Utopian population size estimator

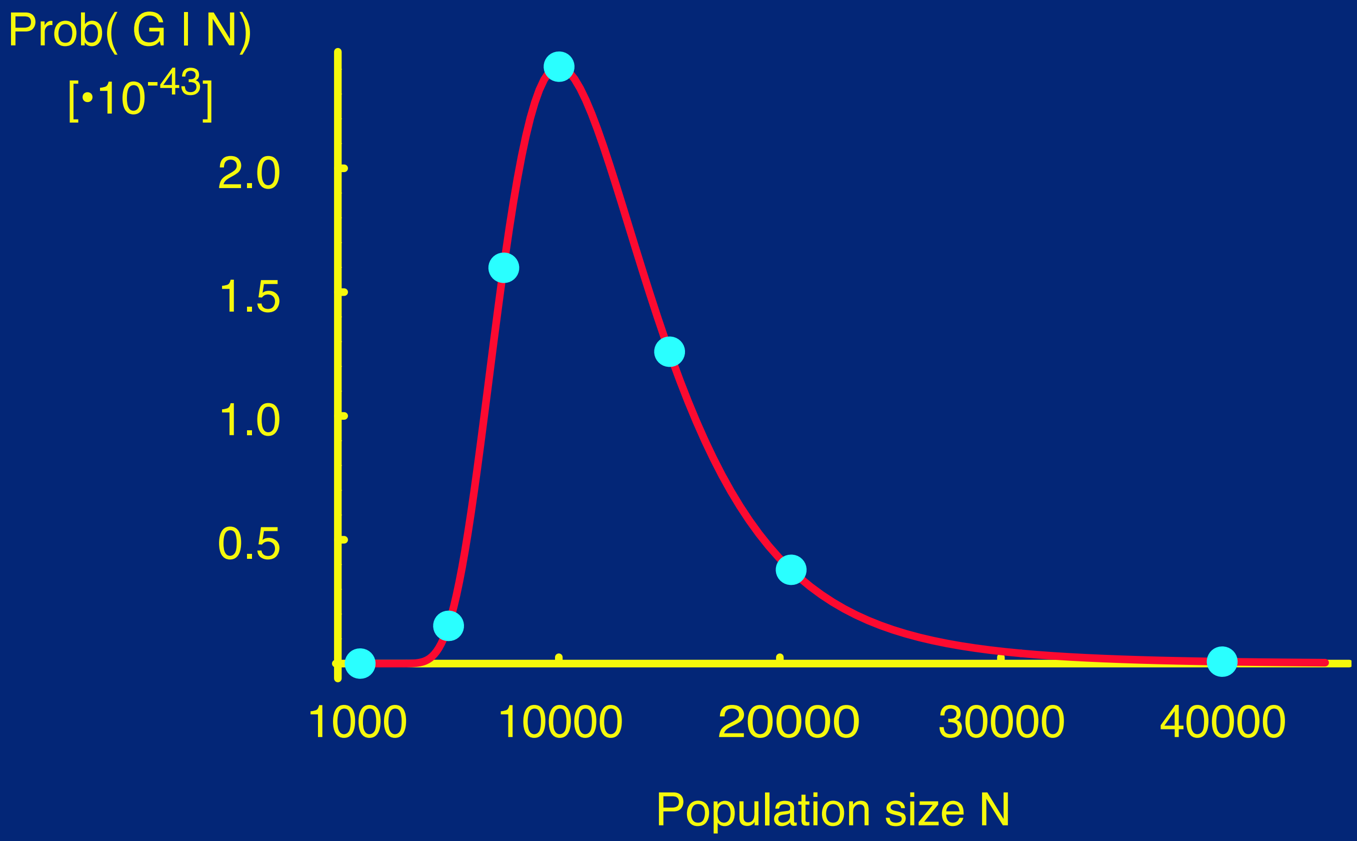
1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation

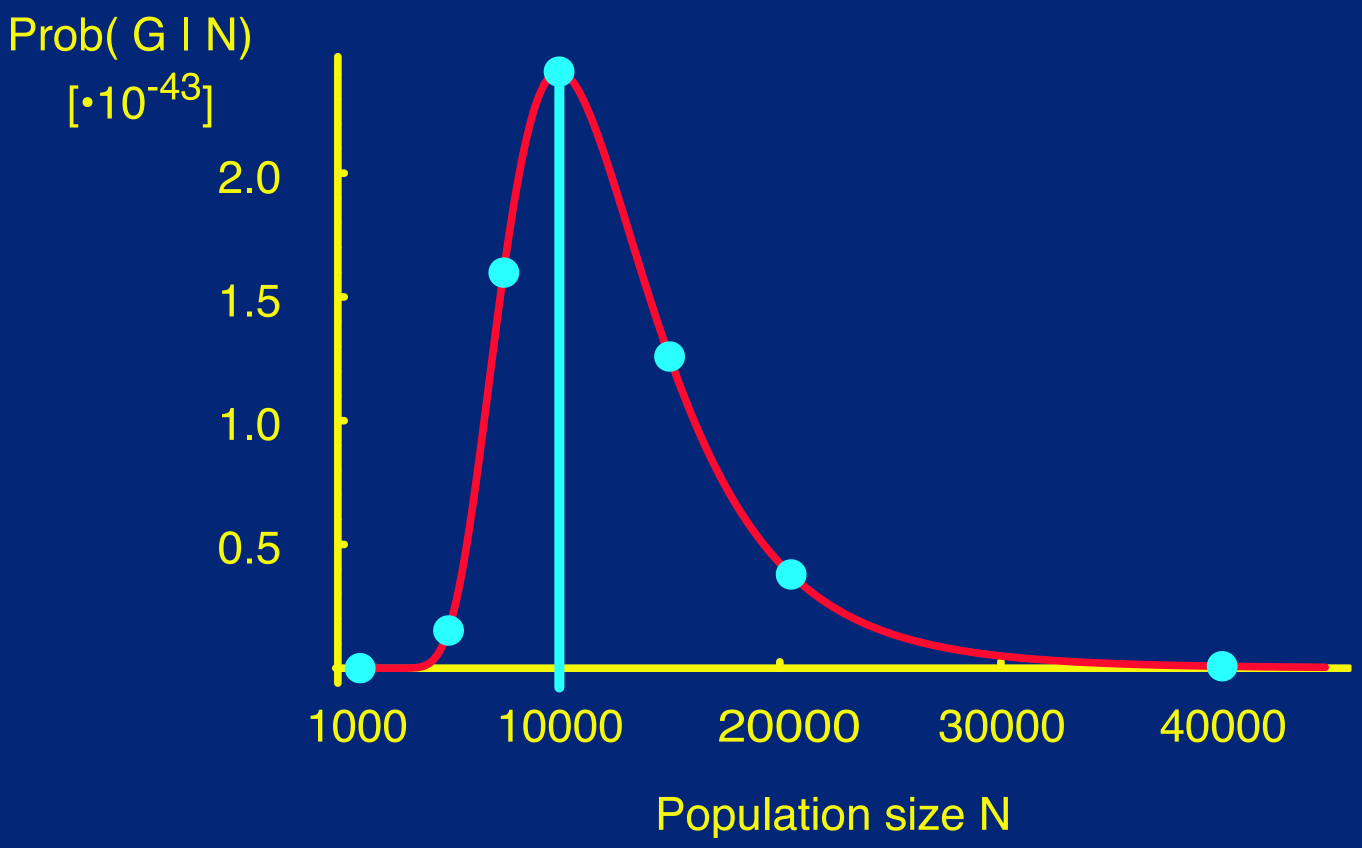


$$p(\text{Genealogy}|N) = \prod_{j}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{1}{2N}$$

Utopian population size estimator

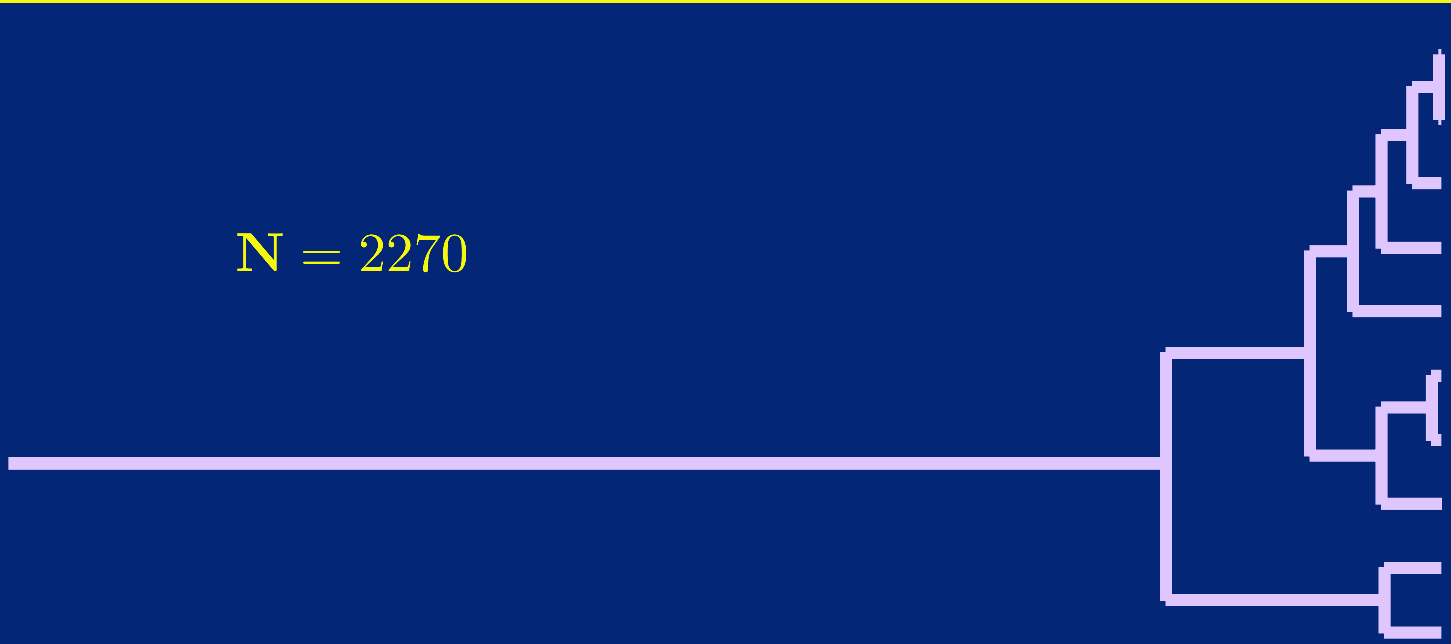


Utopian population size estimator

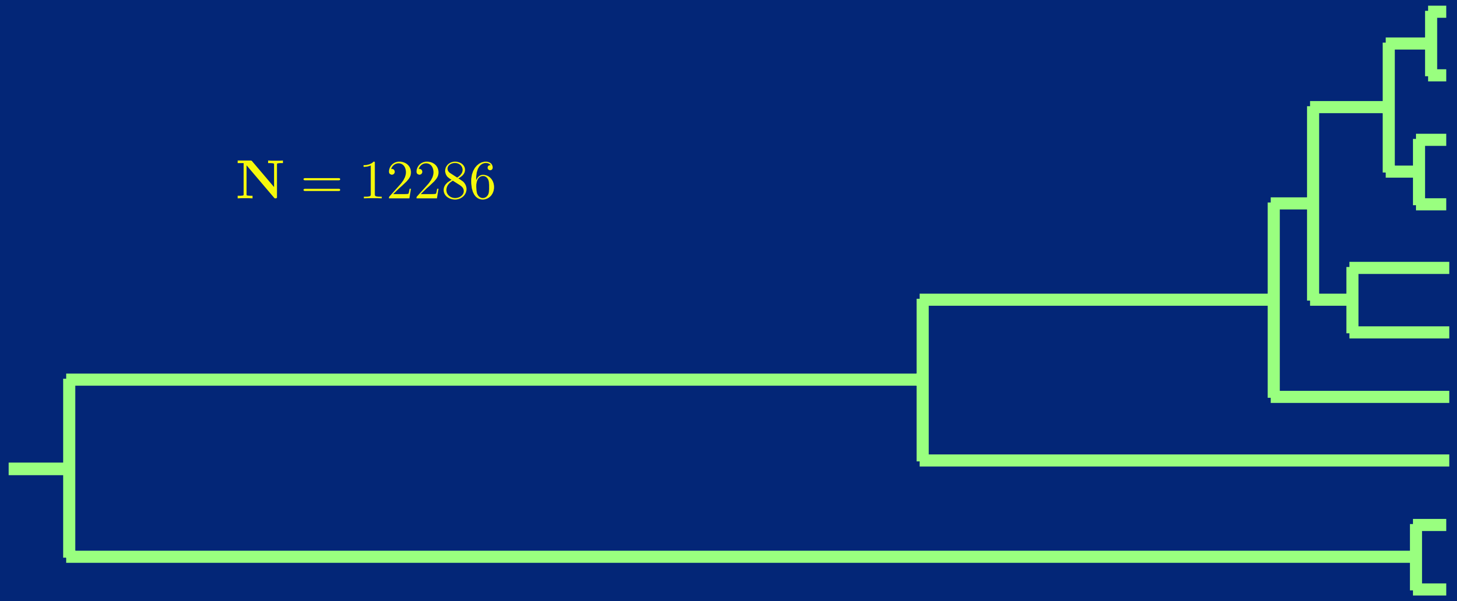


Utopian population size estimator

$\mathrm{N} = 2270$



$\mathrm{N} = 12286$



Lack of infallible oracles

- We assume we know the true genealogy including branch lengths
- We don't really know that
- We probably can't even infer it:
  - Tree inference is hard in general
  - Population data usually don't have enough information for good tree inference

Non-likelihood use of coalescent

- Summary statistics
  - Watterson's estimator of θ
  - FST (estimates θ and/or migration rate)
  - Hudson's and Wakeley's estimators of recombination rate
- Known-tree methods
  - UPBLUE (Yang)
  - Skyline plots (Strimmer, Pybus, Rambaut)

These methods are conceptually easy, but not always powerful, and they are difficult to extend to complex cases.

Genealogy samplers

- Acknowledge that there is an underlying genealogy–
  - but we don't know it
  - we can't infer it with high certainty
  - we can't sum over all possibilities
- A directed sample of plausible genealogies–
  - can capture much of the information in the unknown true genealogy
  - takes a long time but not forever
- These are **genealogy sampler** methods

Outline

1. Introduction to coalescent theory

2. **Practical example: red drum**

3. Genealogy samplers

4. Break

5. Survey of samplers

6. Evolutionary forces

7. Practical considerations

What is the effective population size of red drum?

Red drum, *Sciaenops ocellatus*, are large fish found in the Gulf of Mexico.

Turner, Wares, and Gold
Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish
Genetics 162:1329-1339 (2002)

What is the effective population size of red drum?

- Census population size: 3,400,000
- Effective population size: ?
- Data set:
  - 8 microsatellite loci
  - 7 populations
  - 20 individuals per population

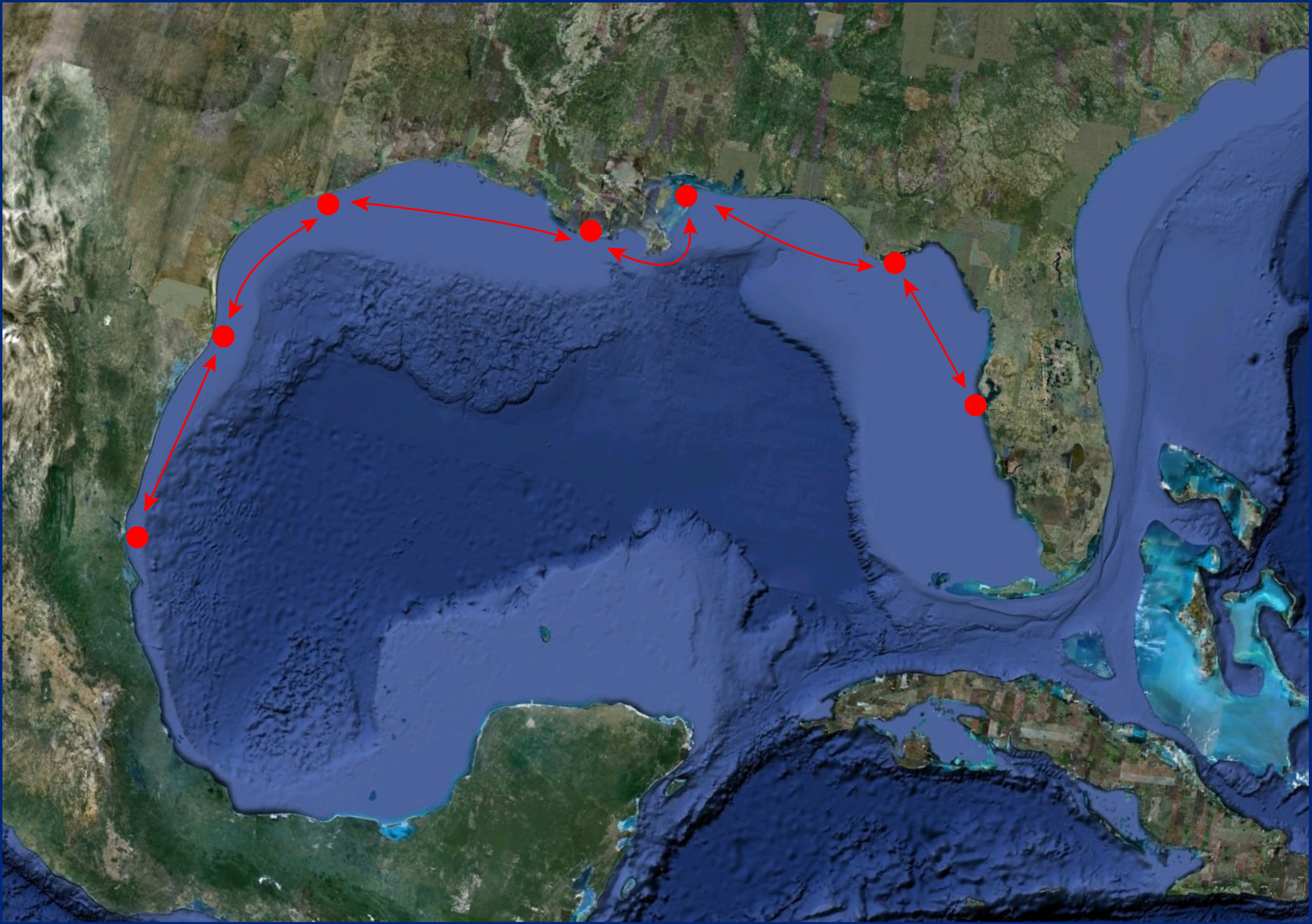
What is the effective population size of red drum?

Three approaches:

1. Allele frequency fluctuation from year to year
   - Measures current population size
   - May be sensitive to short-term fluctuations
2. Coalescent estimate from $Migrate$
   - Measures long-term harmonic mean of population size
   - May reflect past bottlenecks or other long-term effects
3. Demographic models
   - Attempt to infer genetic size from census size
   - Vulnerable to errors in demographic model
   - Not well established for long-lived species with high reproductive variability

Population model used for Migrate

- Multiple populations along Gulf coast
- Migration allowed only between adjacent populations
- Allowing for population structure should improve estimates of population size



What is the effective population size of red drum?

Estimates:

Census size (N):	3,400,000
Allele frequency method (N_e):	3,516 (1,785-18,148)
Coalescent method (N_e):	1,853 (317-7,226)

The demographic model can be made consistent with these only by assuming enormous variance in reproductive success among individuals.

What is the effective population size of red drum?

- Allele frequency estimators measure current size
- Coalescent estimators measure long-term size
- Conclusion: population size and structure have been stable

What is the effective population size of red drum?

- Effective population size at least 1000 times smaller than census
- This result was highly surprising
- Red drum has the genetic liabilities of a rare species
- Turner et al. hypothesize an “estuary lottery”
- Unless the eggs are in exactly the right place, they all die

Outline

1. Introduction to coalescent theory
2. Practical example
3. **Genealogy samplers**
4. Break
5. Survey of samplers
6. Evolutionary forces
7. Practical considerations

Coalescent estimation of population parameters

- Mutation model: Steal a likelihood model from phylogeny inference
- Population genetics model: the Coalescent

Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta)$$

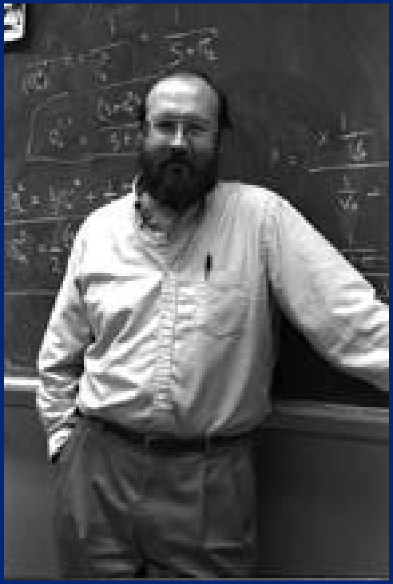
Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

$P(Data|G)$ comes from a mutational model



Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

$P(G|\Theta)$ comes from the coalescent

Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

$\sum_G$ is a problem

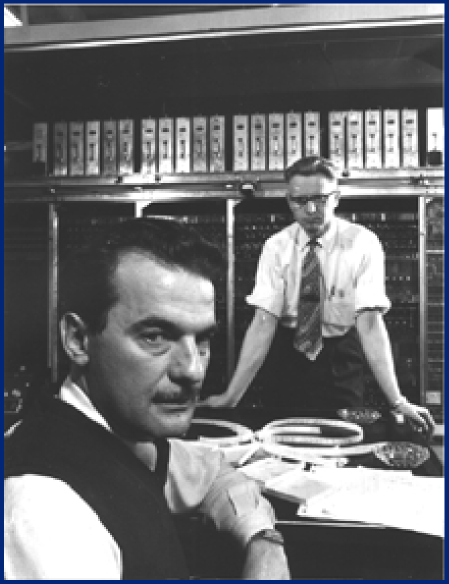
Can we calculate this sum over all genealogies?

Tips	Topologies
3	3
4	18
5	180
6	2700
7	56700
8	1587600
9	57153600
10	2571912000
15	6958057668962400000
20	564480989588730591336960000000
30	4368466613103069512464680198620763891440640000000000000
40	302733382994800735654630336455145720004293943205386250170788872192000000000
50	3.28632×10^{112}
100	1.37416×10^{284}

A solution: Markov chain Monte Carlo

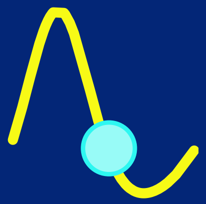
- If we can't sample all genealogies, could we try a random sample?
  - Not really.
- How about a sample which focuses on good ones?
  - What is a good genealogy?
  - How can we find them in such a big search space?

A solution: Markov chain Monte Carlo

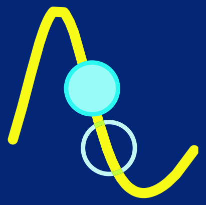


Metropolis recipe

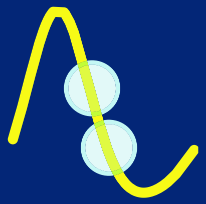
0. first state



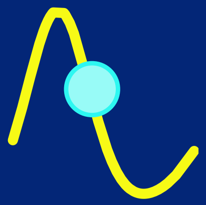
1. perturb old state and calculate probability of new state



2. test if new state is better than old state: accept if ratio of new and old is larger than a random number between 0 and 1.

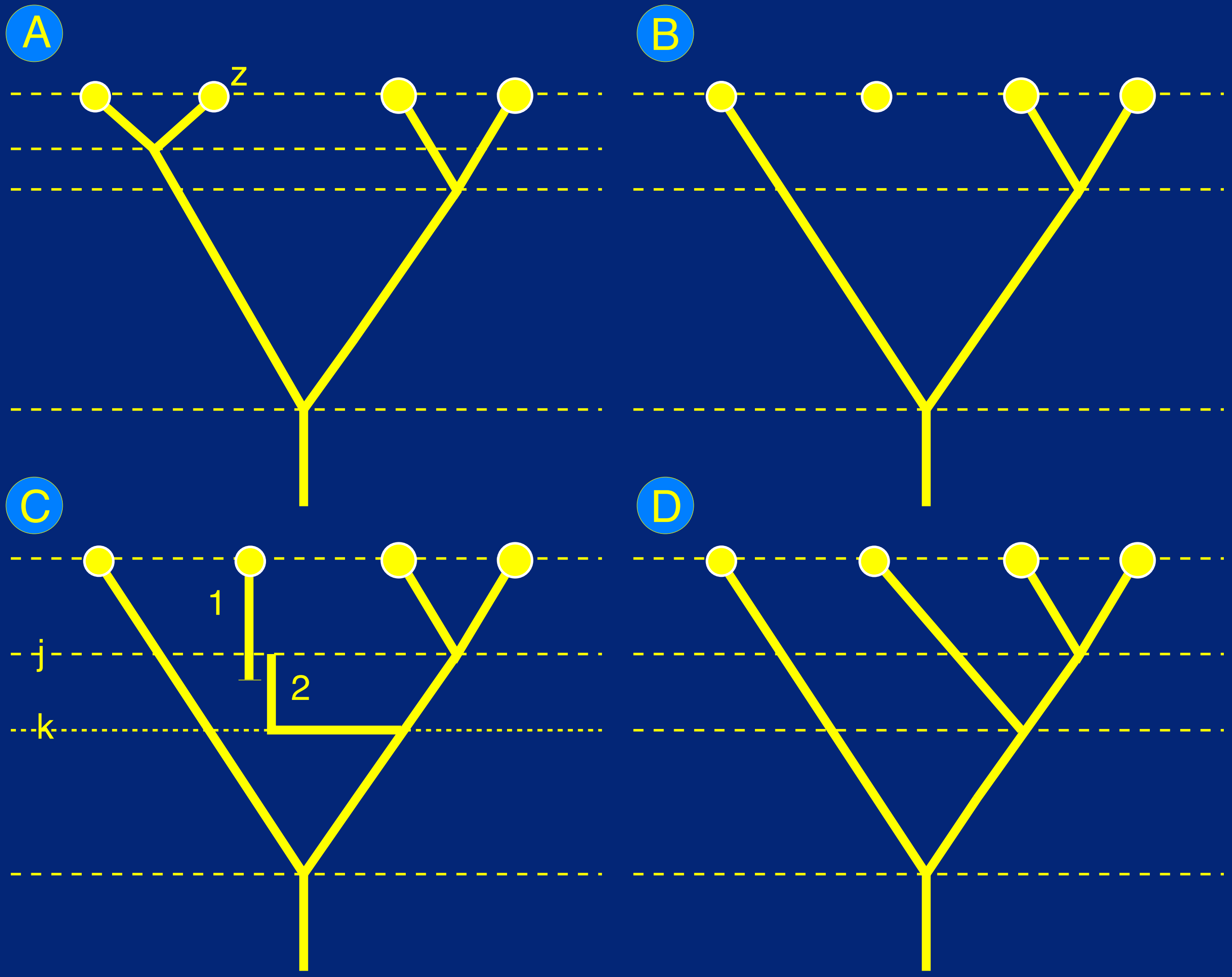


3. move to new state if accepted otherwise stay at old state

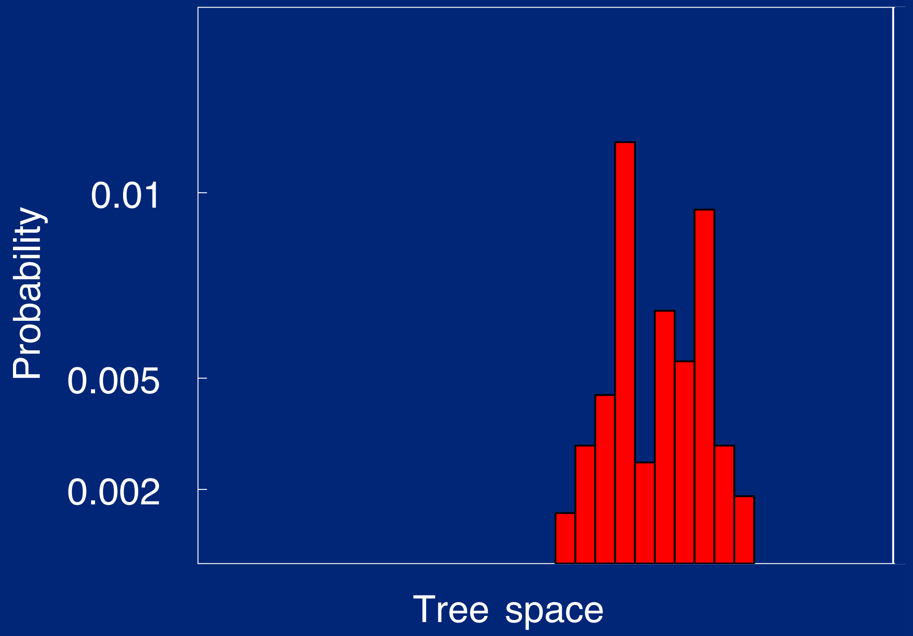


4. go to 1

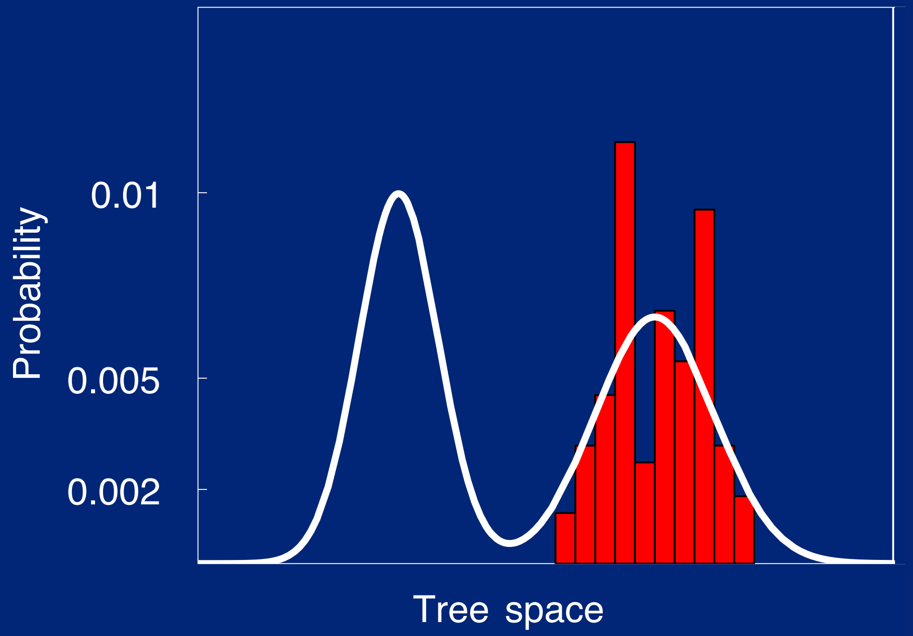
How do we change a genealogy?



MCMC walk result



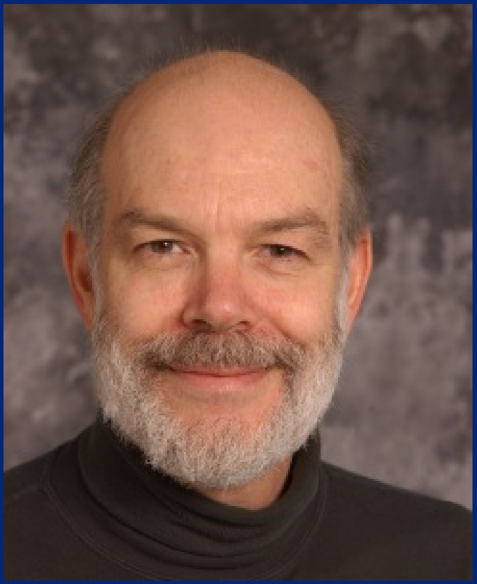
MCMC walk result–with problems



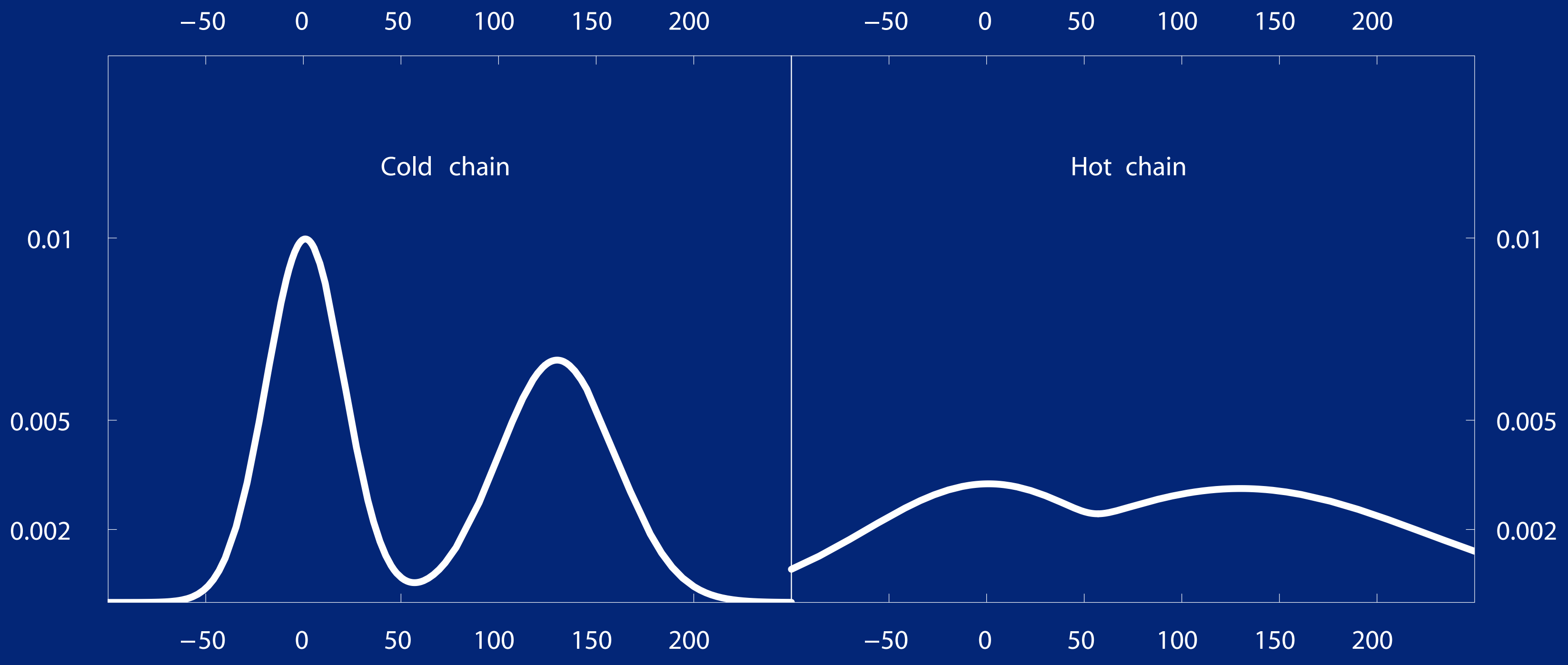
Improving our MCMC walker: Heating

Metropolis Coupled Markov chain Monte Carlo (AKA MC^3)

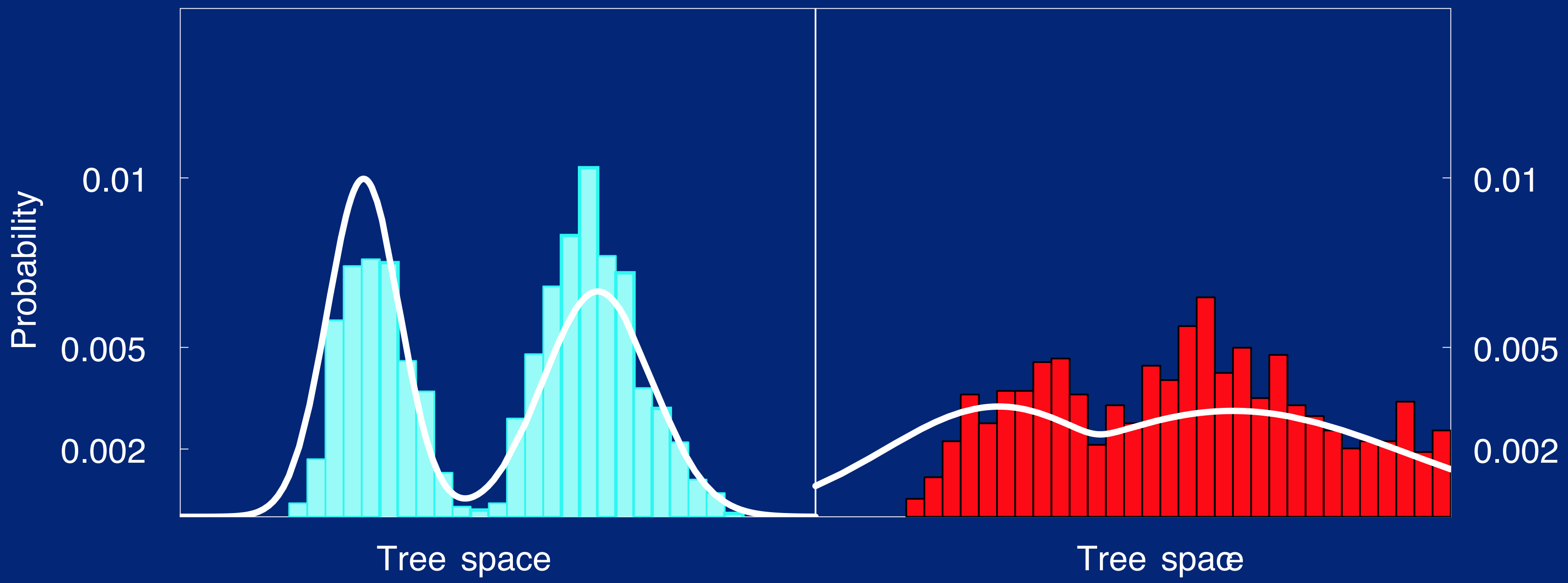
- Run several independent parallel chains: each has a different temperature
- After some sampling of genealogies, swap the genealogies of a pair of chains if the ratio between probabilities in the cold and the hot chain is larger than a random number drawn between 0 and 1.



Improving our MCMC walker: MCMCMC or MC^3



better MCMC walk result



Outline

1. Introduction to coalescent theory

2. Genealogy samplers

   (a) **Likelihood version**

   (b) **Bayesian version**

3. Practical example

4. Break

5. Survey of samplers

6. Evolutionary forces

7. Practical considerations

Likelihood and Bayesian approaches

- All genealogy samplers search among genealogies
- All of them require some type of guide value (“driving value”) to determine which genealogies will be proposed
- Two major approaches: Likelihood-based and Bayesian
- Major ideological difference, relatively small practical one

Likelihood samplers

- Use arbitrary values of the parameters to guide the search
- Sample genealogies throughout the search
- At the end of the search, evaluate $P(G|\Theta)$ for sampled genealogies
- Correct for the influence of the driving values
- Iterate to improve driving values

Bayesian samplers

- Propose new driving values throughout the run
- New driving values drawn from a prior
- Accept or reject driving values based on $P(G|\Theta)$
- Final conclusions based on histogram of driving values

Likelihood analysis

We will approximate:

$$L(\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

Likelihood analysis

We will approximate:

$$L(\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

by sampling n genealogies from $P(Data|G)P(G|\Theta_0)$:

$$L(\Theta) = \frac{1}{n}\sum_{G*} \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)}$$

Here the G^* are no longer random genealogies; they are sampled from a distribution that depends on the **driving value** Θ_0

Likelihood analysis

$$L(\Theta) = \frac{1}{n}\sum_{G} \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)}$$

Isn't this circular? We have a solution for the unknown $L(\Theta)$ in terms of the unknown $L(\Theta_0)$.

Likelihood analysis

$$L(\Theta) = \frac{1}{n} \sum_G \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)}$$

Isn't this circular? We have a solution for the unknown $L(\Theta)$ in terms of the unknown $L(\Theta_0)$.

$$\frac{L(\Theta)}{L(\Theta_0)} = \frac{1}{n} \sum_G \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)}$$

This doesn't give us the actual value of $L(\Theta)$ but it does allow us to compare various values of Θ and choose the best.

Likelihood analysis

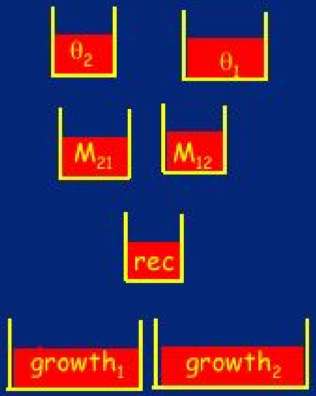
- This approach is only asymptotically correct
- For finite sample sizes, it has a bias toward its driving value
- We can greatly reduce this:
  - Start with an arbitrary Θ_0
  - Run the sampler a while and estimate the best Θ
  - It will be biased toward Θ_0, but...
  - Use it as the new Θ_0 and start over

Bayesian approach

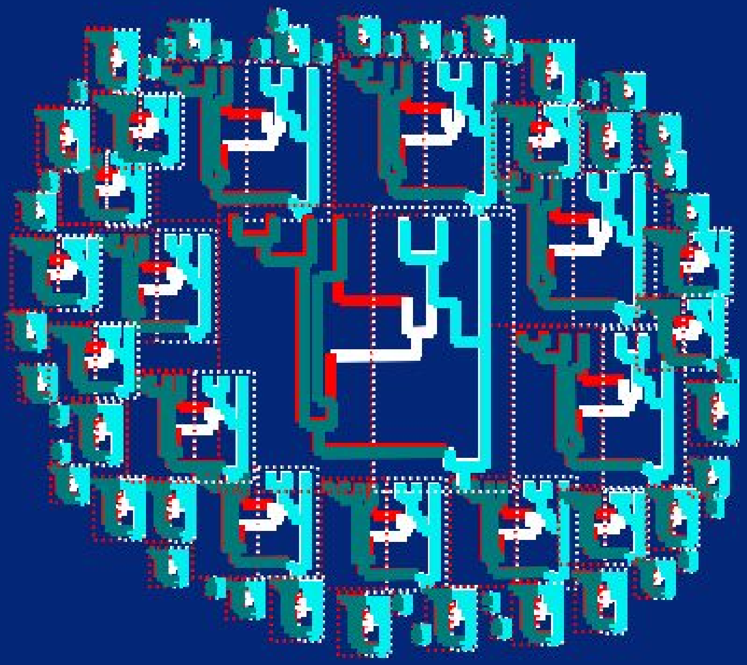
- A Bayesian analysis requires us to provide priors for all parameters
- These *could* be based on detailed knowledge of the biology
- In practice, uninformative flat priors are used

New search scheme for Bayes

Parameter space
(determined by priors)



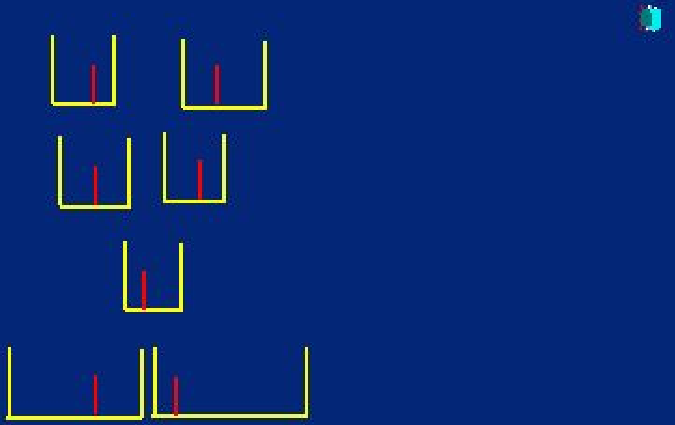
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

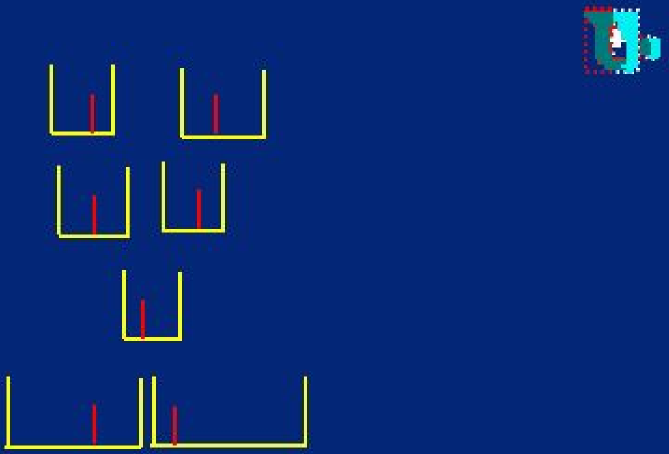
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

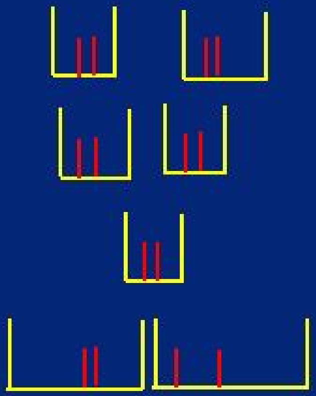
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

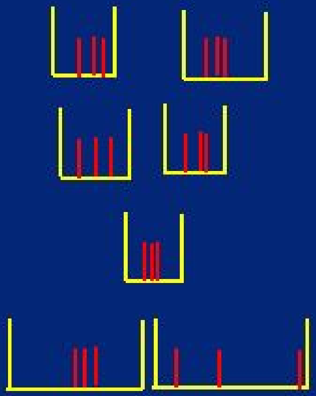
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

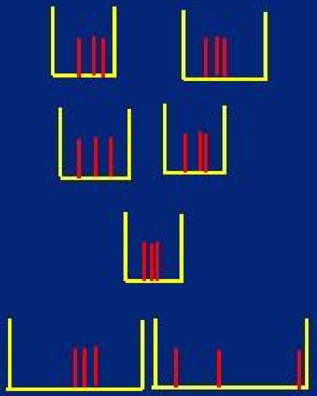
Tree space

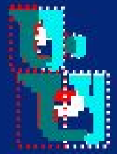


New search scheme for Bayes

Parameter space
(determined by priors)

Tree space

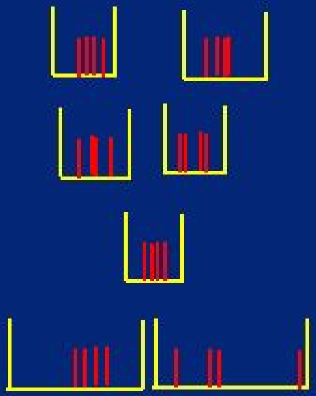


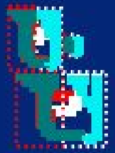


New search scheme for Bayes

Parameter space
(determined by priors)

Tree space

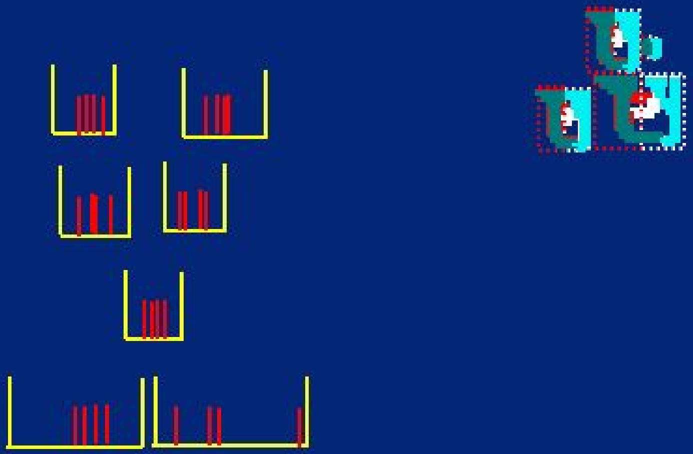




New search scheme for Bayes

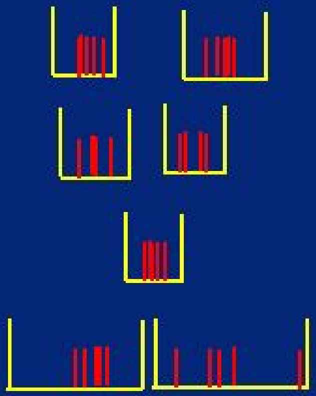
Parameter space
(determined by priors)

Tree space

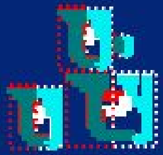


New search scheme for Bayes

Parameter space
(determined by priors)



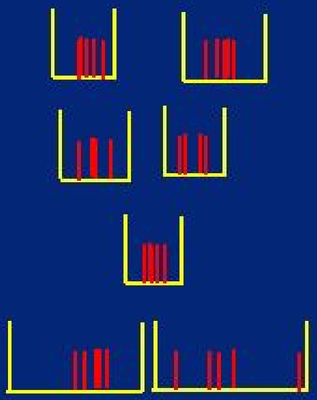
Tree space

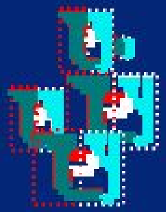


New search scheme for Bayes

Parameter space
(determined by priors)

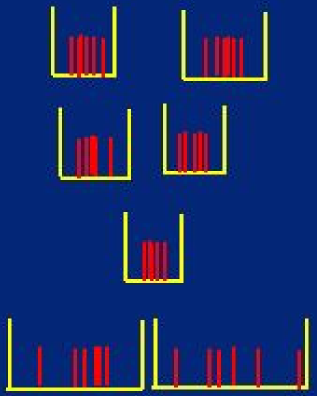
Tree space



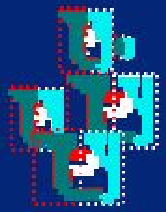


New search scheme for Bayes

Parameter space
(determined by priors)



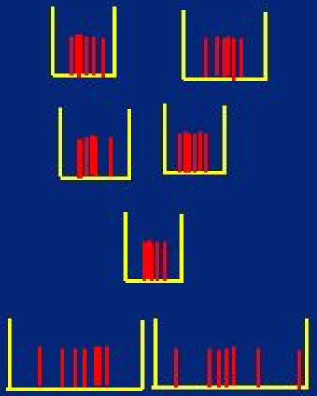
Tree space

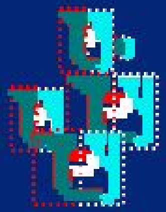


New search scheme for Bayes

Parameter space
(determined by priors)

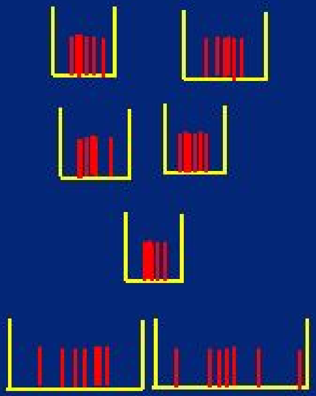
Tree space



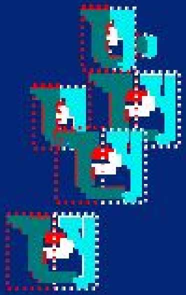


New search scheme for Bayes

Parameter space
(determined by priors)

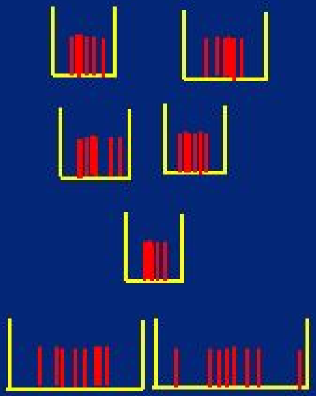


Tree space

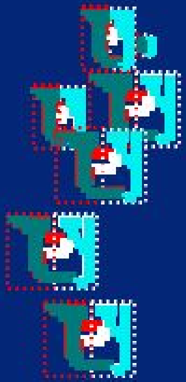


New search scheme for Bayes

Parameter space
(determined by priors)

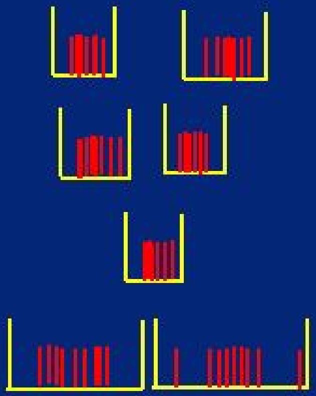


Tree space

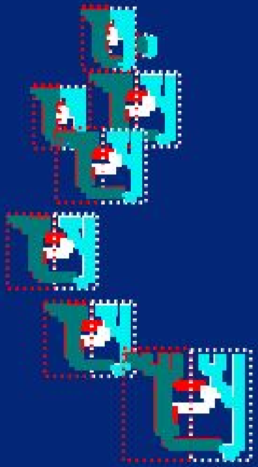


New search scheme for Bayes

Parameter space
(determined by priors)



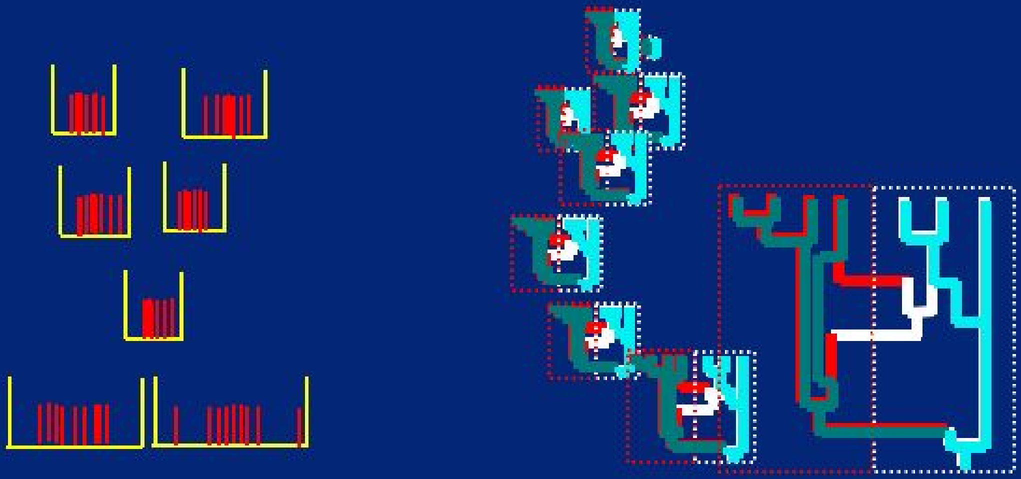
Tree space



New search scheme for Bayes

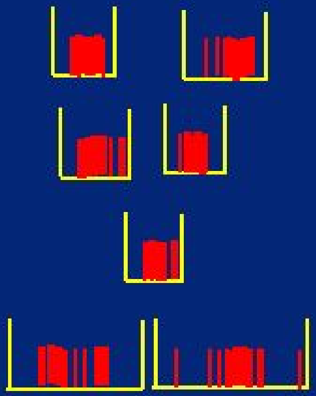
Parameter space
(determined by priors)

Tree space

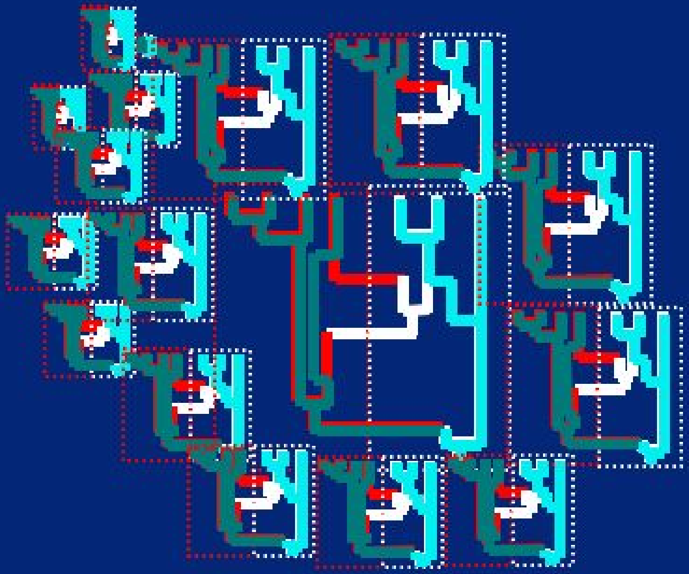


New search scheme for Bayes

Parameter space
(determined by priors)



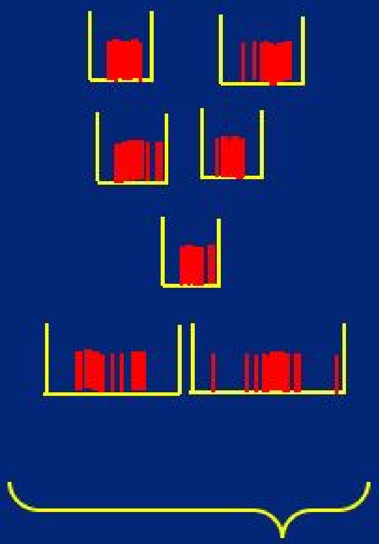
Tree space

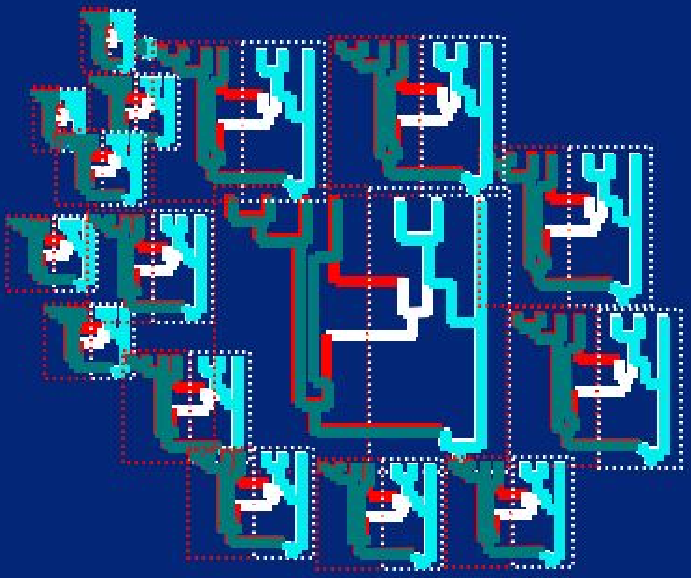


New search scheme for Bayes

Parameter space
(determined by priors)

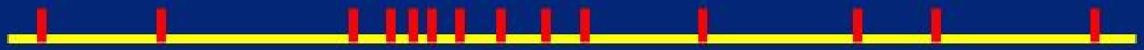
Tree space



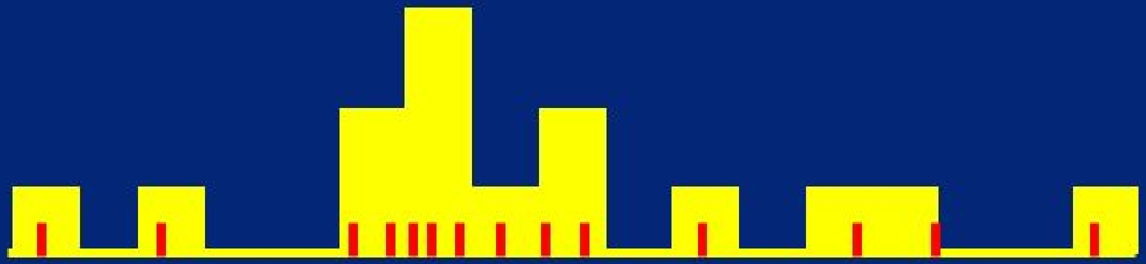


Keep a list of all accepted parameters

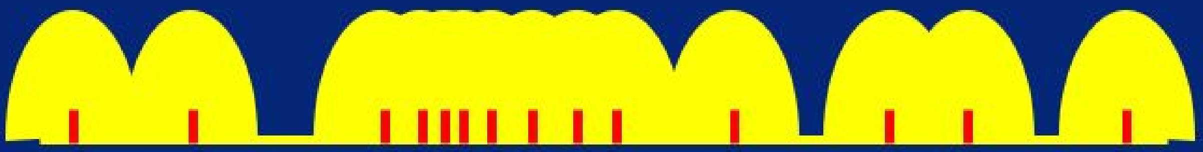
Data collection and curve smoothing



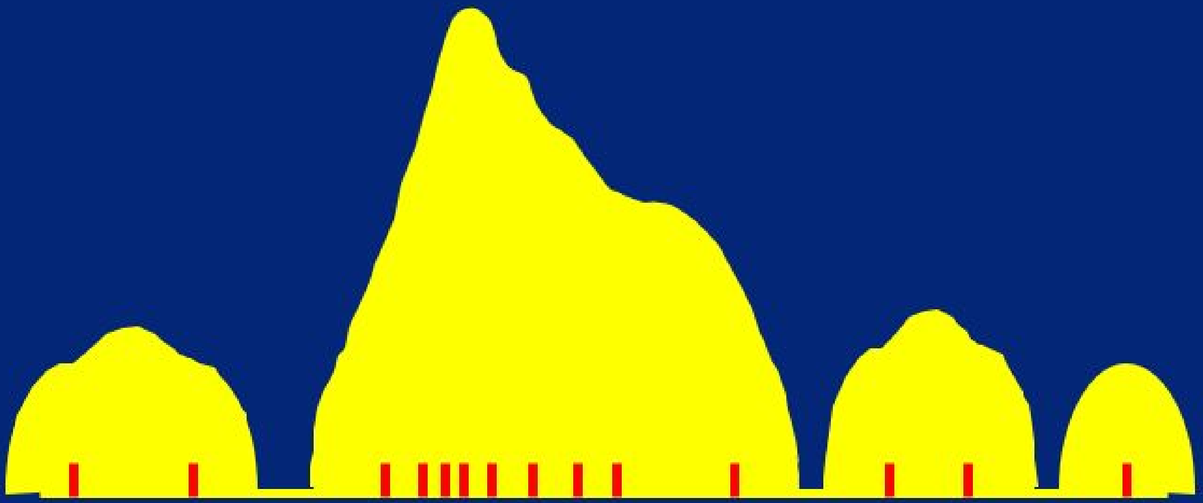
Data collection and curve smoothing



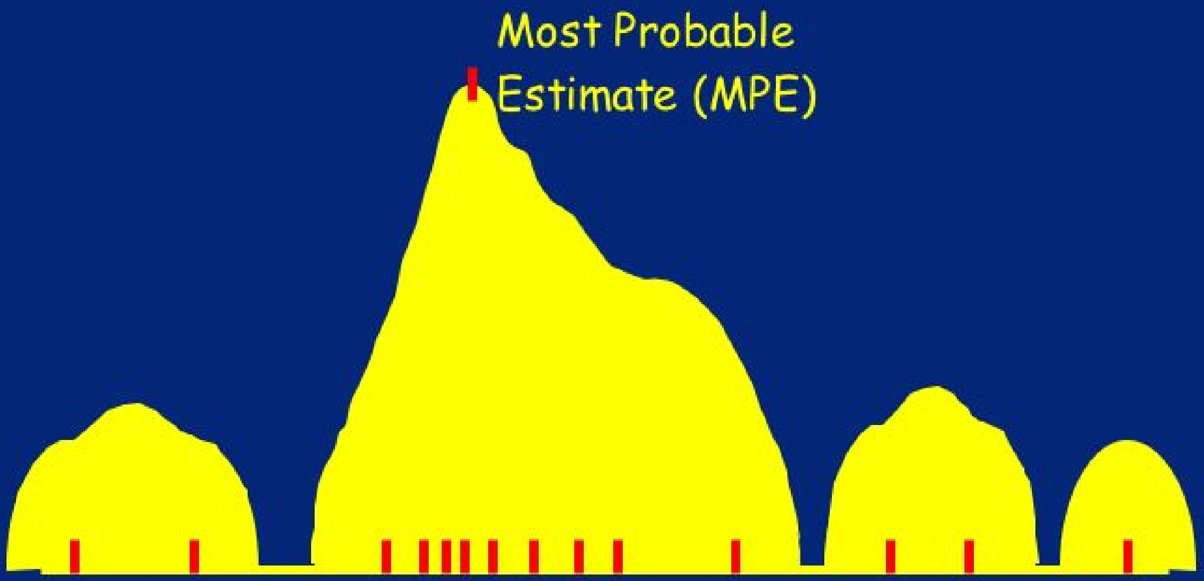
Data collection and curve smoothing



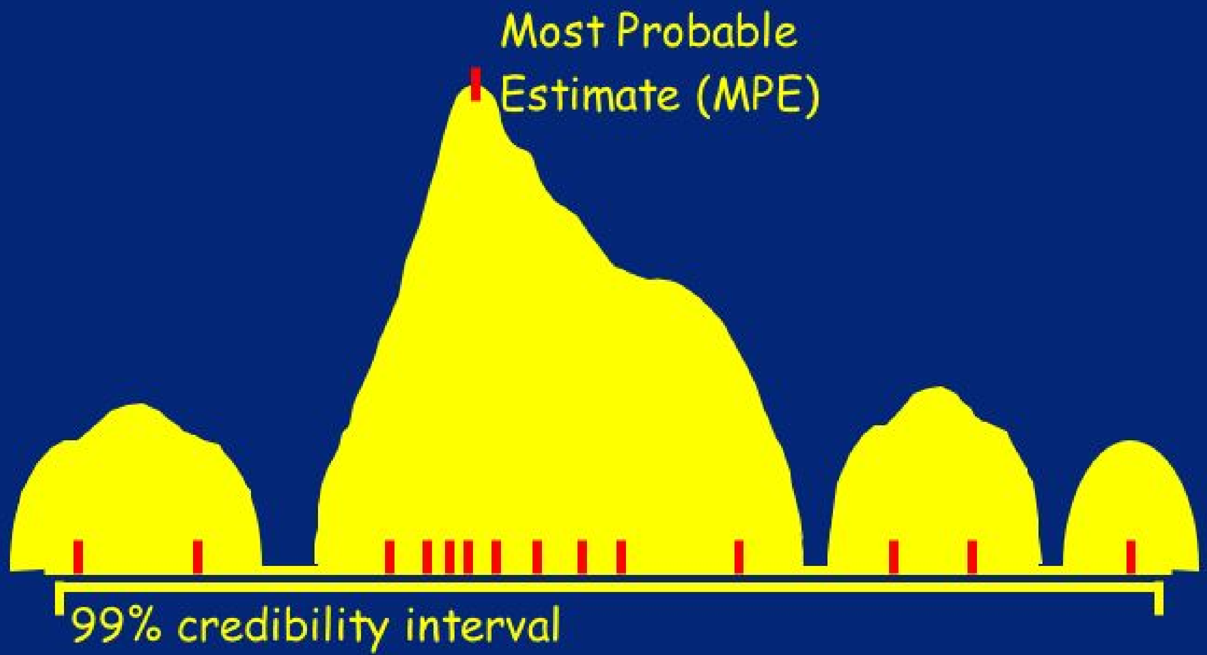
Data collection and curve smoothing



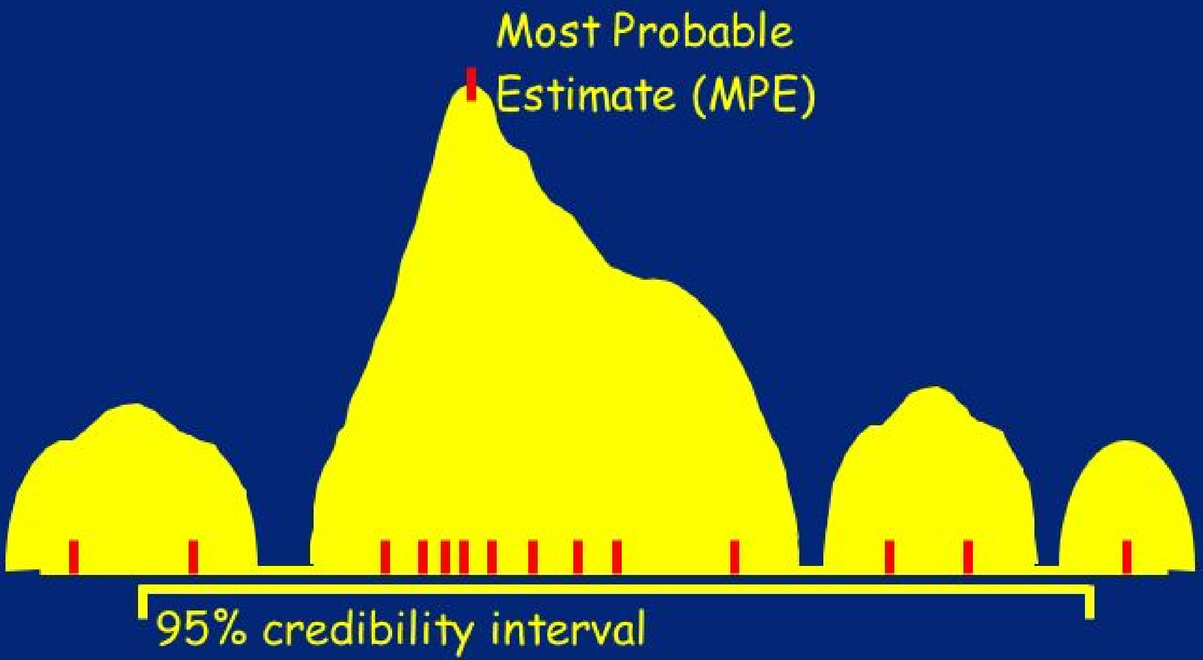
Data collection and curve smoothing



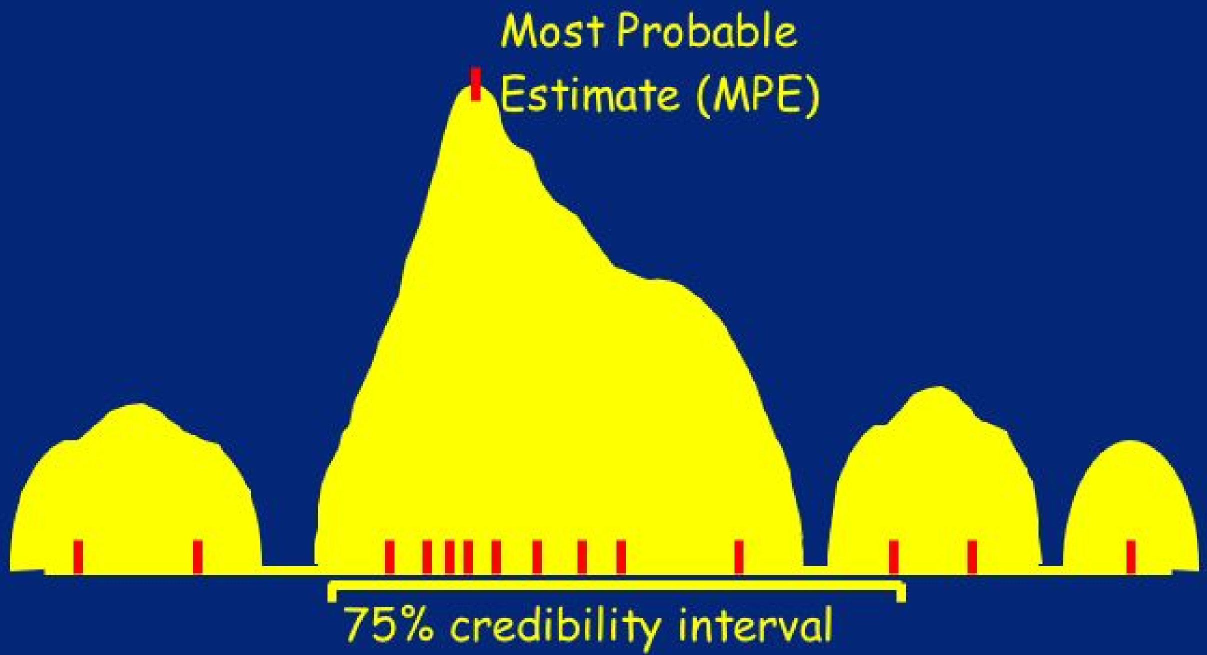
Data collection and curve smoothing



Data collection and curve smoothing



Data collection and curve smoothing



Advantages of Bayesian analysis

- Easier to interpret probabilities than likelihoods
- Smoothing a histogram is quicker than finding maxima of a likelihood curve
- Not dependent on starting driving values
- Parameter values near zero estimated more accurately
- Prior information can be incorporated (in theory)
- Trendy!

Disdvantages of Bayesian analysis

- No information currently available on correlation of parameters
- Dependent on good priors; results can be severely distorted by bad priors

Bottom line

- Kuhner 2006: Bayes and likelihood almost identical
- Beerli 2006: Bayes has edge with sparse data
- My recommendations:
  - Use Bayes if you think a parameter is very close to zero
  - Otherwise, with rich data either method is good
  - With poor data, do you really want to be doing this analysis at all?
  - When using Bayes, be careful of your priors!
- If the genealogy search is inadequate, both methods will fail (and fail in similar ways)

Break

Outline

1. Introduction to coalescent theory

2. Genealogy samplers

3. **Survey of samplers**

   (a) **BEAST**
   (b) **Genetree**
   (c) **IM/IMa**
   (d) **Lamarc**
   (e) **Migrate-N**

4. Evolutionary forces

5. Practical considerations

BEAST (http://evolve.zoo.ox.ac.uk/beast/)

- Drummond and Rambaut
- Estimates:
  - Overall population size x mutation rate
  - Overall growth rate
  - With multiple time points, mutation rate and generation time
  - Detailed skyline plots of growth rate
  - Relaxed molecular clock
- Bayesian analysis
- DNA, RNA, amino acids, codon data, continuous and discrete morphological traits

BEAST

- Strengths:
  - Multiple time point data (ancient DNA, microorganisms)
  - Flexible population growth model
  - Highly flexible mutation model
- Weaknesses:
  - Single population
  - No recombination

IM, IMa2 (http://lifesci.rutgers.edu/ heylab/HeylabSoftware.htm#IM)

- Nielsen, Hey, Wakeley
- Estimates:
  - Population size x mutation rate
  - Immigration rates
  - Size of ancestral population
  - Time of divergence
  - Daughter population growth rates (IM only)
- Bayesian analysis
- DNA, RNA, microsatellites, HapSTRs
- IM has the most models; IMa2 has more than two populations

IM/IMa2

- Strengths:
  - Correct analysis of young (less than 4N generations) populations
  - Distinguishing gene flow from common ancestry
- Weaknesses:
  - Single time point only
  - No recombination
  - Exponential growth only

LAMARC (http://evolution.gs.washington.edu/lamarc.html)

- Kuhner, Beerli, Felsenstein et al.
- Estimates:
  - Population size x mutation rate
  - Immigration rates
  - Growth rates
  - Overall recombination rate
- Likelihood or Bayesian analysis
- DNA, RNA, SNPs, microsats, elecrophoretic alleles
- Gene mapping, haplotype inference

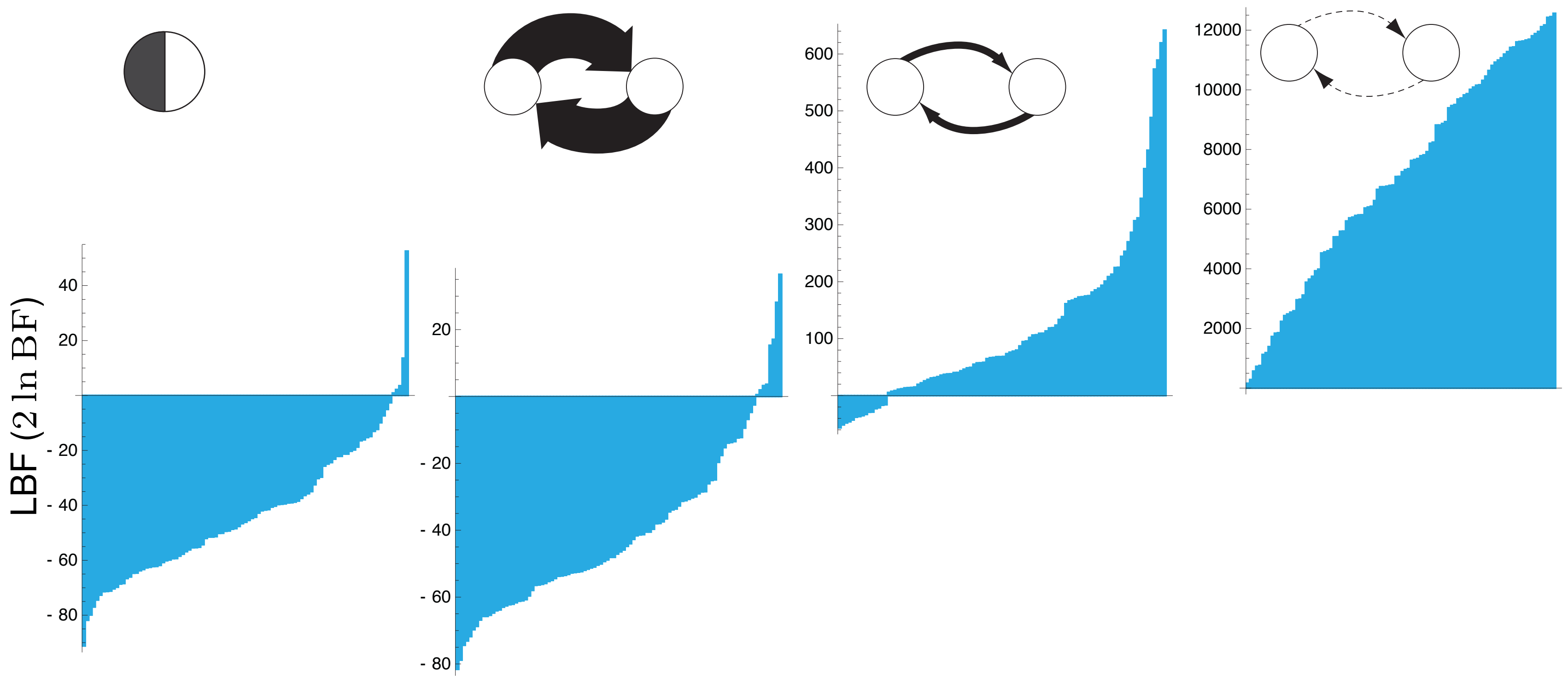
LAMARC

- Strengths:
  - Recombination
  - Data with unknown haplotype phase
  - Combining dissimilar loci
- Weaknesses:
  - Assumes stable population structure (divergence coming soon!)
  - Single time point data only
  - Exponential growth only

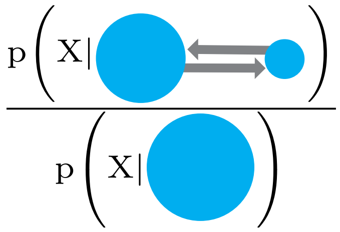
MIGRATE-N (http://popgen.csit.fsu.edu/Migrate-n.html)

- Beerli
- Estimates:
  - Population size x mutation rate
  - Immigration rates
  - Tests among different migration models
- Likelihood or Bayesian analysis
- DNA, RNA, SNPs, microsats, elecrophoretic alleles
- Multiple time points

Bayes factor tests of models

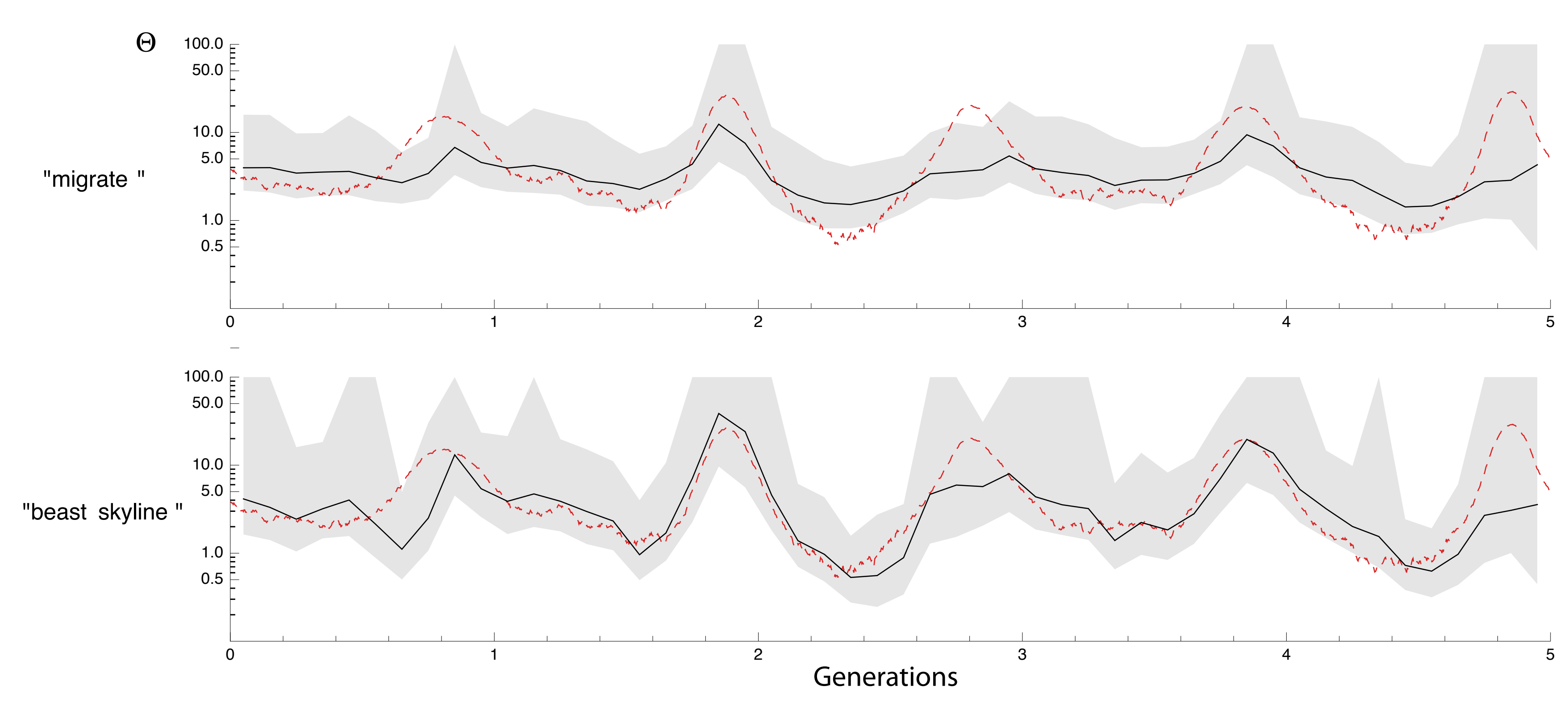


$$\text{LBF} = 2\ln\frac{p(X|M_1)}{p(X|M_2)} = 2\ln$$



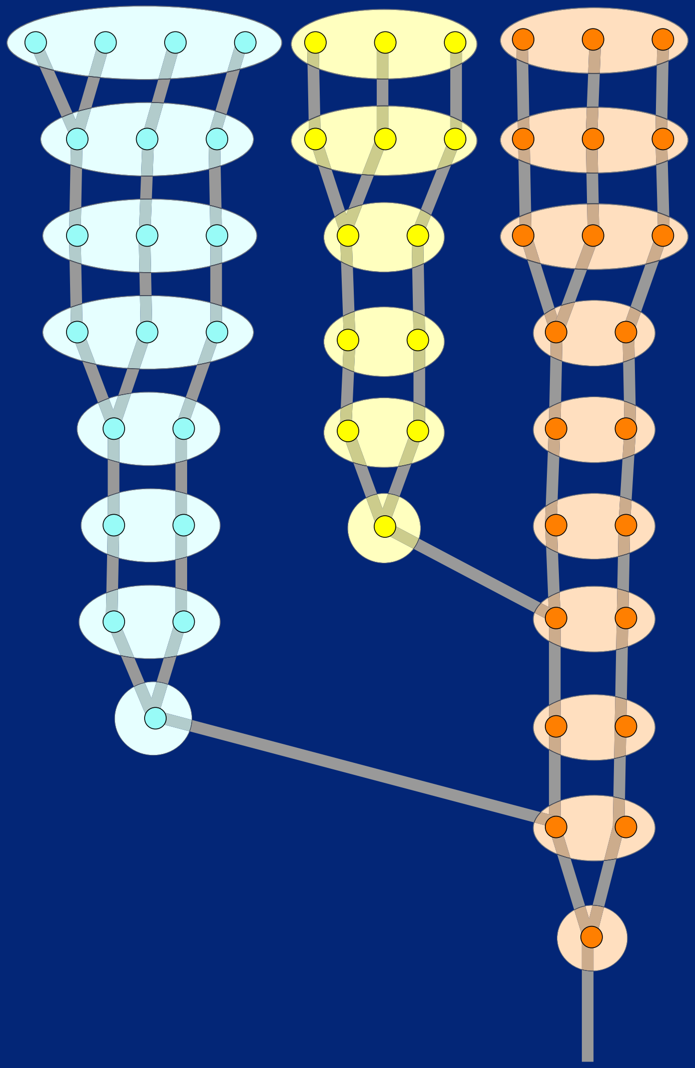
MIGRATE-N

- Strengths:
  - Skyline plots for all parameters
  - Multiple time points
  - Bayes factor tests of different models
- Weaknesses:
  - Assumes stable population structure and size
  - No recombination or growth



Comparison of skyline plots between MIGRATE-N and BEAST for simulated influenza data with multiple time points

Genetree (http://www.stats.ox.ac.uk/g̃riff/software.html)

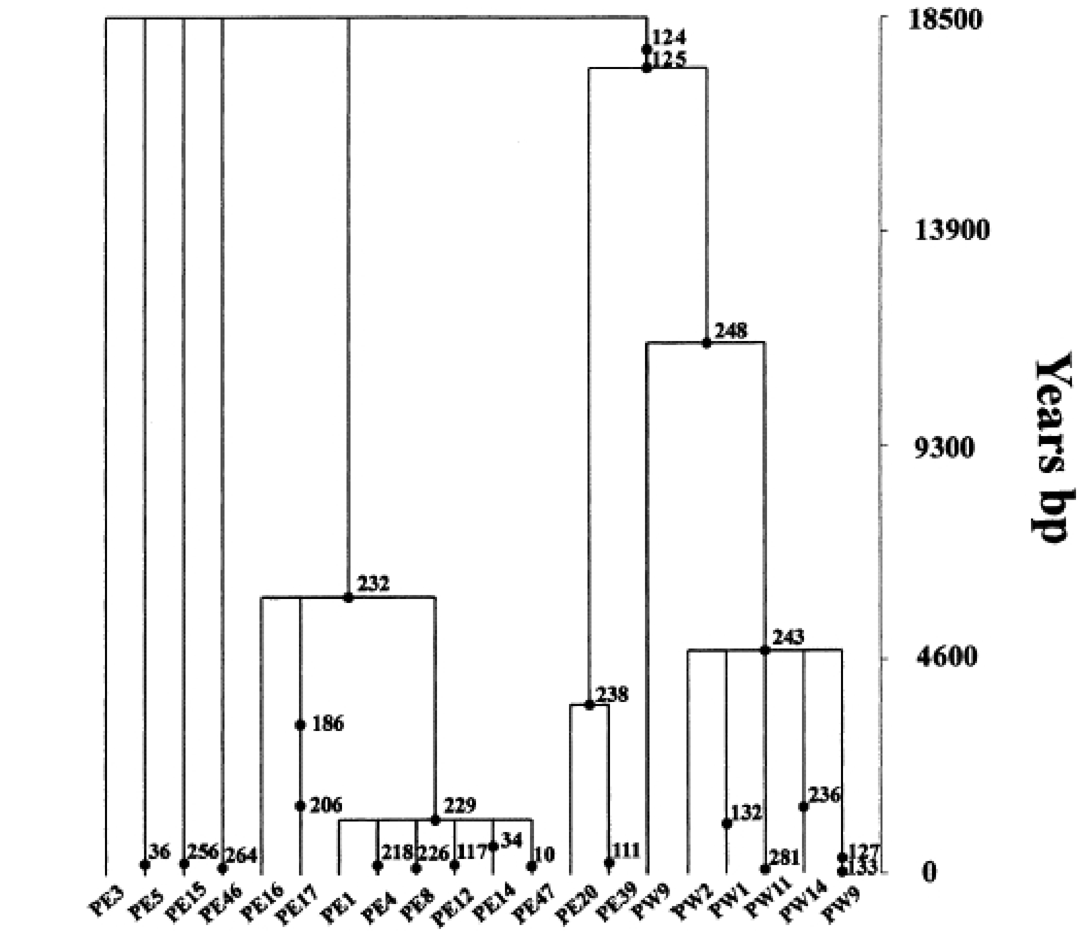


- Infinite sites model
- Use MCMC to sample a path through the possible histories
- Sample many different possible histories

Dating mutations events using *Genetree*

Milot et al. (2000)

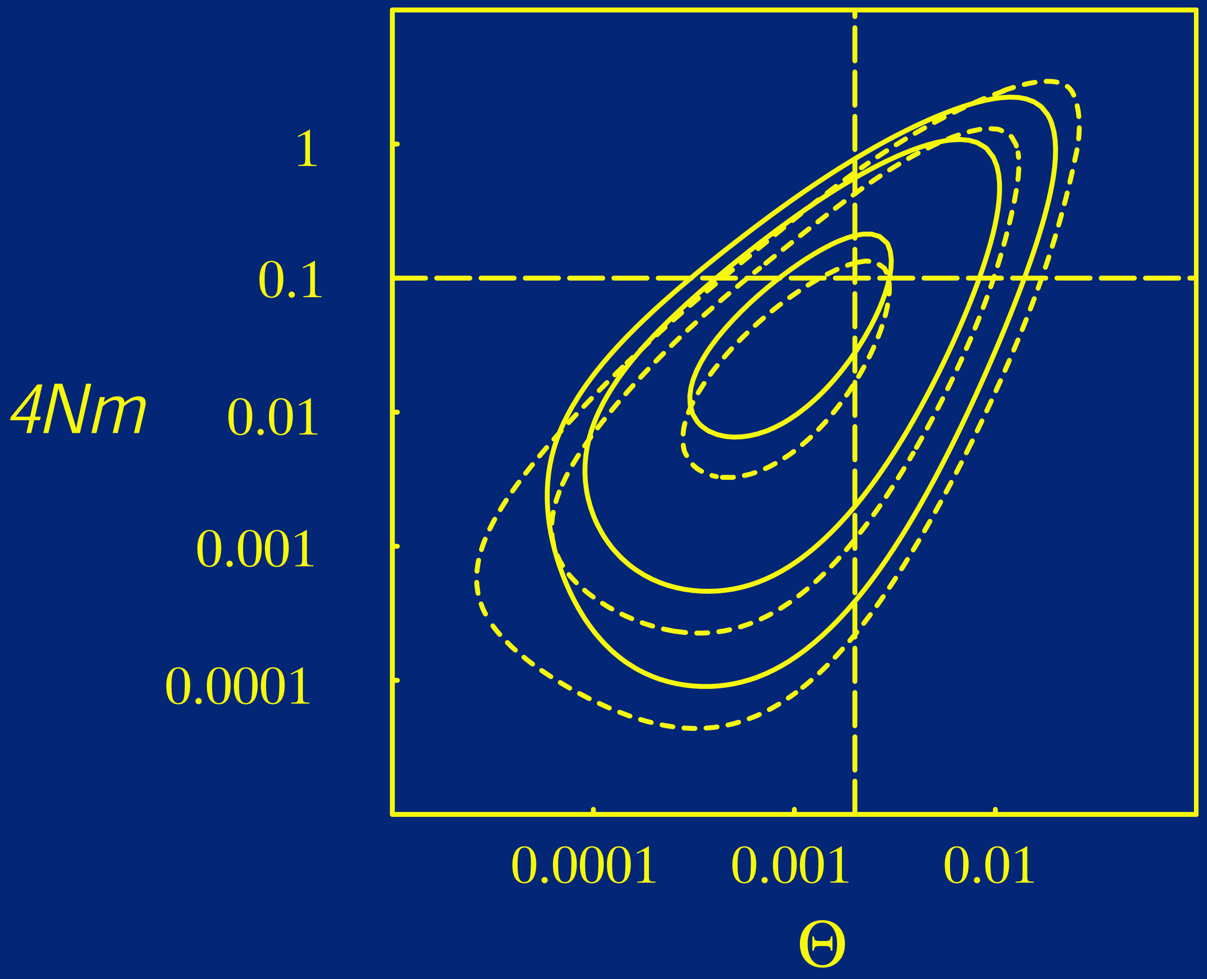
Photo by
Stephen J. Lang



	PE3	PE5	PE15	PE46	PE16	PE17	PE1	PE4	PE8	PE12	PE14	PE47	PE20	PE39	PW9	PW2	PW1	PW11	PW14	PW9
Total	28	5	2	1	1	11	40	4	1	1	4	1	2	3	2	14	15	1	1	1
East	28	5	2	1	1	11	40	4	1	1	4	1	2	3	.	1	.	.	.	.
West	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	13	15	1	1	1

Comparison between *Migrate-N* and *Genetree*

(Beerli and Felsenstein 2001)



Genetree

- Strengths:
  - Efficient search
  - Dating of specific mutations
  - Dating of the common ancestor
- Weaknesses:
  - Infinite-sites mutational model only
  - No recombination
  - Exponential growth only
  - Single time point
  - Less developed user interface

Outline

1. Survey of samplers

2. **Evolutionary forces**

   - Genetic drift (Θ)
   - Population growth/shrinkage
   - Migration
   - Recombination
   - Population divergence
   - Multiple time points
   - Haplotype uncertainty
   - Disequilibrium mapping

3. Practical considerations

Genetic drift ($Theta$)

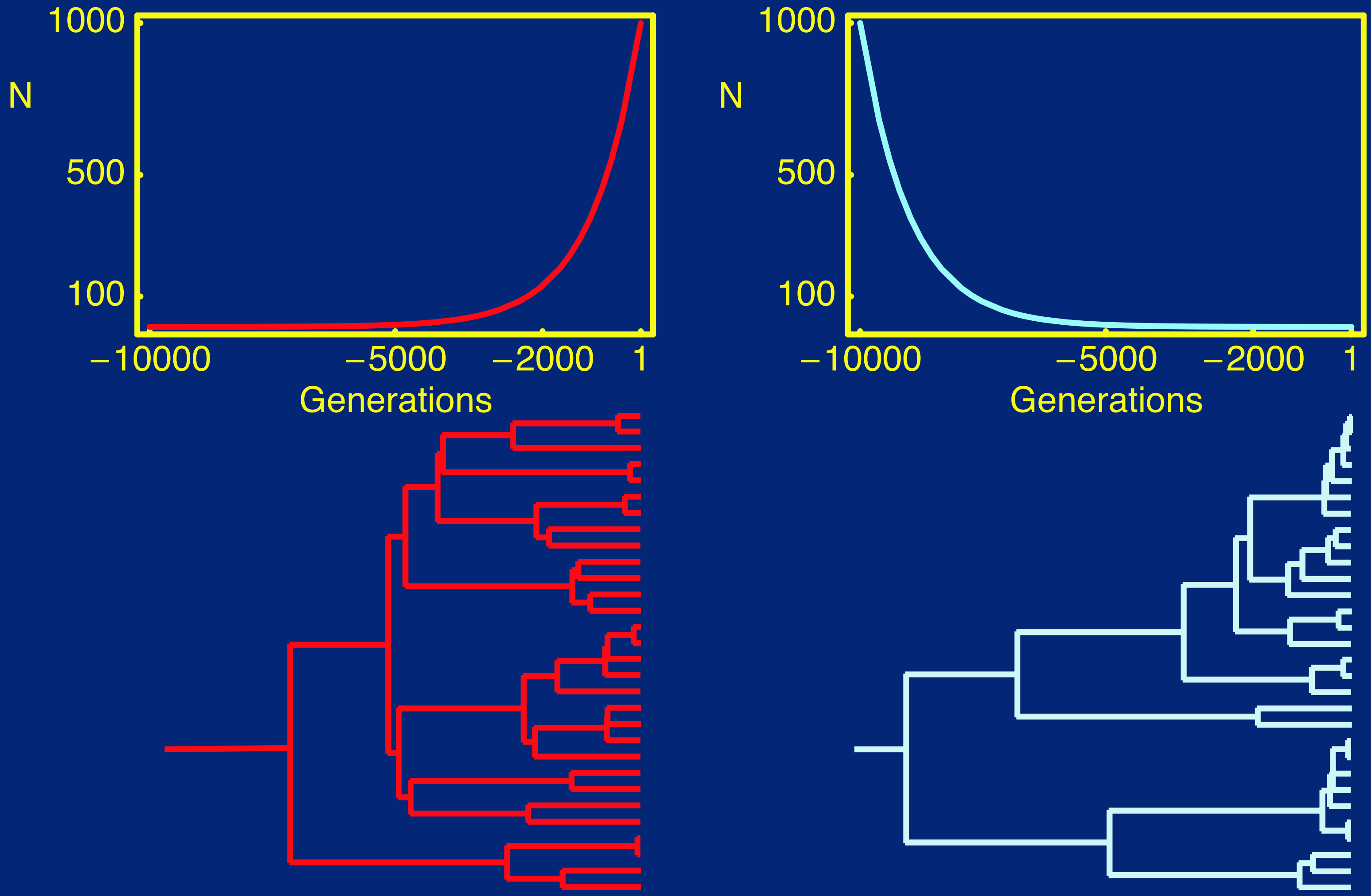
- With one time point, we estimate $\Theta = 4N_e\mu$ in diploids
- The number estimated is $2N_e\mu$ in haploids or $N_e\mu$ in mtDNA
- Two ways to separate N_e and μ:
  - Dated historical data (ancient DNA, etc.)
  - External estimate of mutation rate
- For most organisms, N_e is less than N
- Demographic models can help resolve this

Variable population size

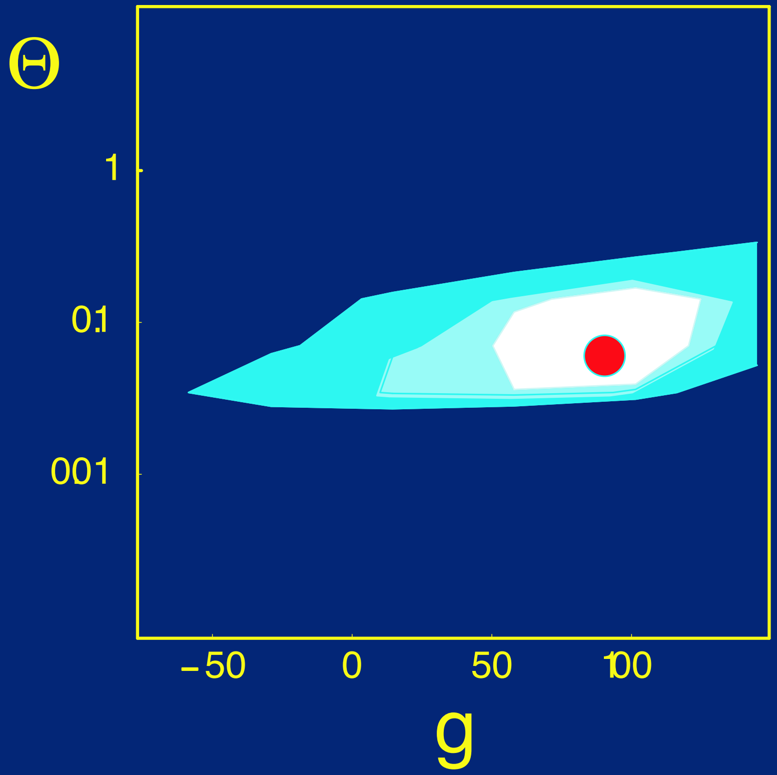
- In a small population lineages coalesce quickly
- In a large population lineages coalesce slowly

This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the current population size Θ.

Exponential population size expansion or shrinkage

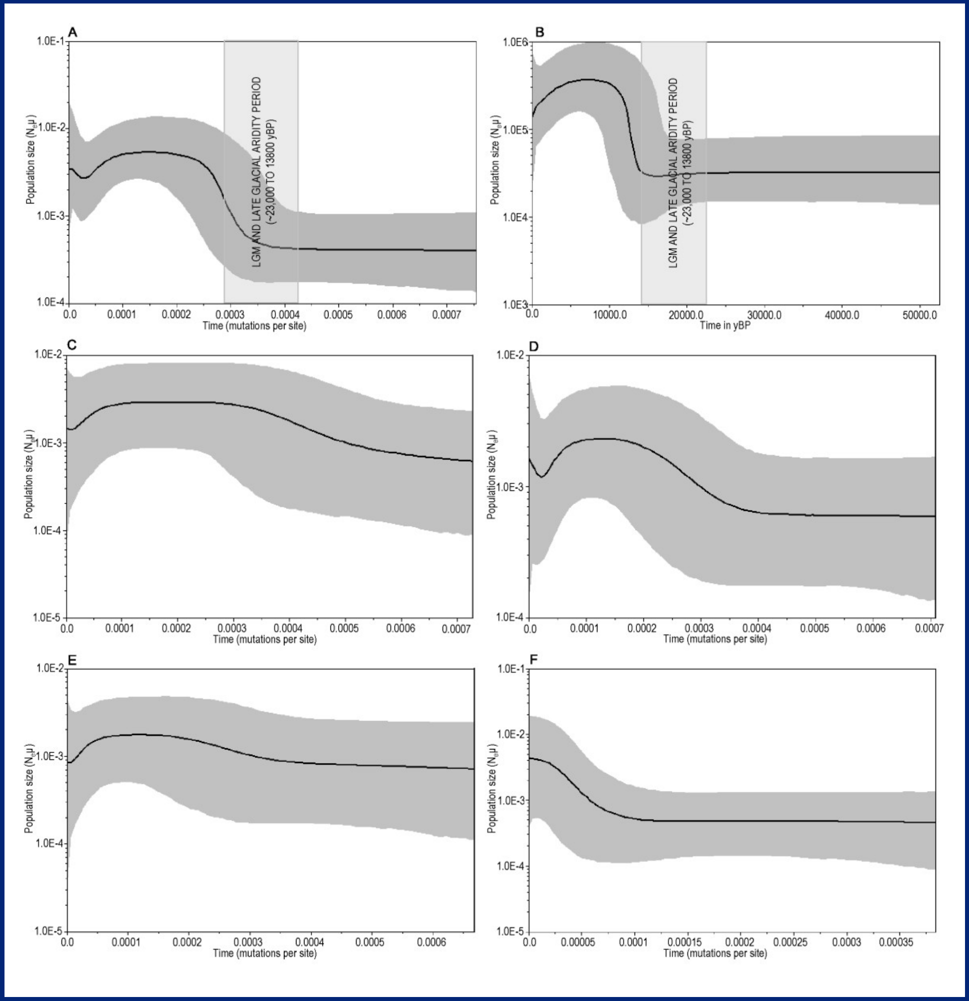


Grow a frog



Mutation Rate	Population sizes	
	-10000 generations	Present
10^{-8}	$8,300,000$	$8,360,000$
10^{-7}	$780,000$	$836,000$
10^{-6}	$40,500$	$83,600$

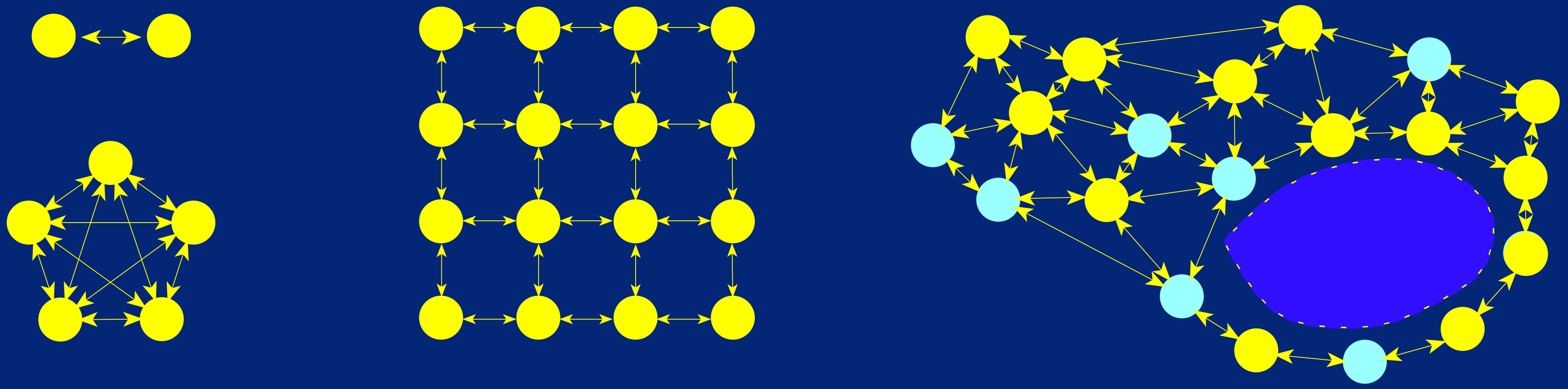
Bayesian skyline plots



Growth estimation software

- Currently done with *Lamarc* or *Beast*
- Statistically weaker than estimation of Θ:
  - Biased upwards with one locus/one timepoint
  - Reasonable results with multiple unlinked loci
  - Even better results with multiple timepoints
- *Lamarc* assumes exponential growth/shrinkage
- *Beast* has a generalized model

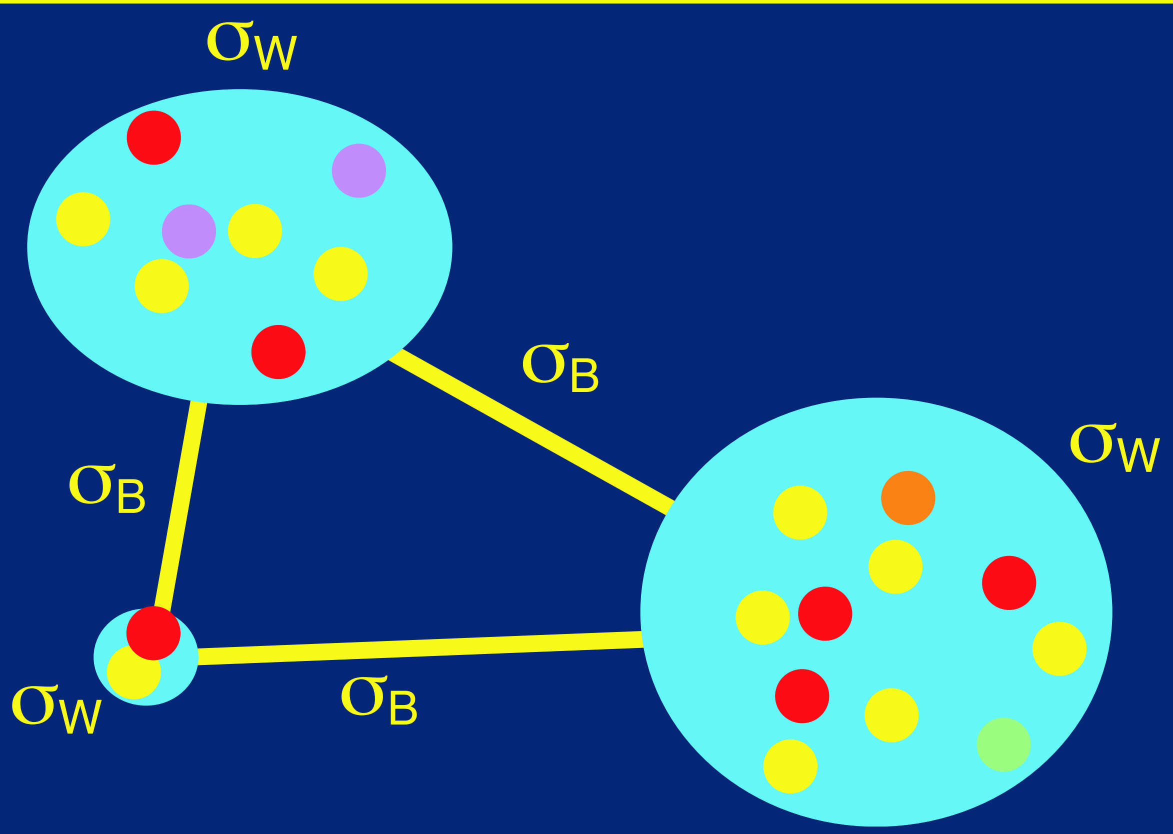
Gene flow



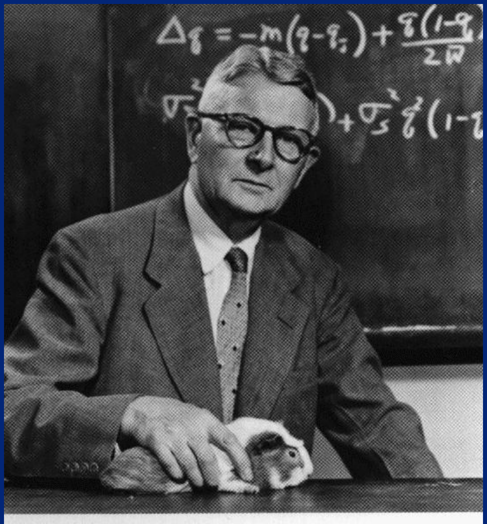
$$\mathrm{p}(G|\boldsymbol{\Theta},\mathbf{M}) = \prod_{u_j} \left(\prod_{i}^{\text{pop.}} g(\Theta_i, \mathbf{M}_{.i}) \right) \begin{cases} \frac{2}{\Theta} & \text{if event is a coalescence,} \\ M_{ji} & \text{if event is a migration from } j \text{ to } i. \end{cases}$$

Gene flow: What researchers used (and still use)

F_{ST}



What researchers used (and still use)



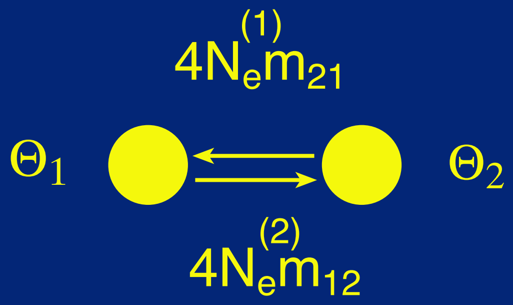
Sewall Wright showed that

$$\mathrm{F}_{\mathrm{ST}} = \frac{1}{1 + 4Nm}$$

and that it assumes

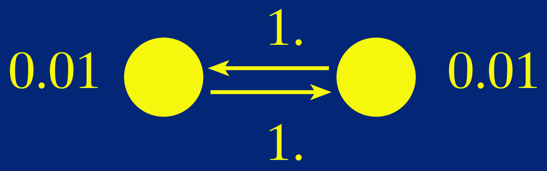
- migration into all subpopulation is the same
- population size of each island is the same

Simulated data and Wright's formula

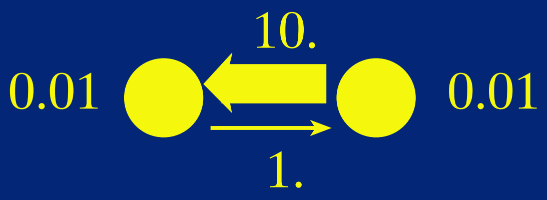


True values

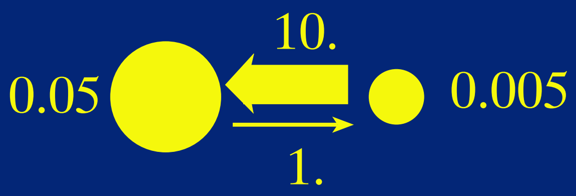
Estimated values



1.14±0.77



7.80±22.20



11.46±18.54

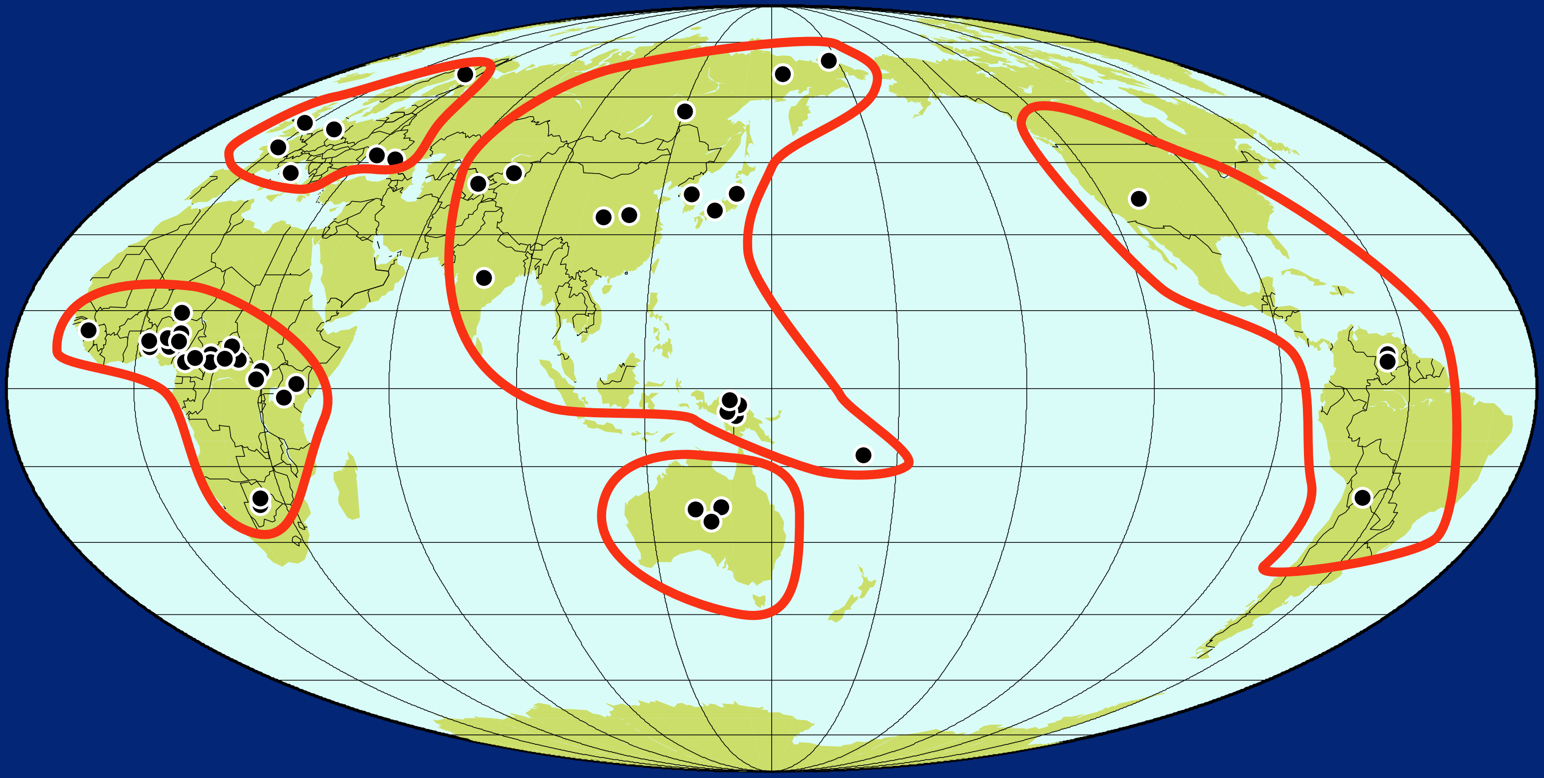
Maximum Likelihood method to estimate gene flow parameters

(Beerli and Felsenstein 1999)

100 two-locus datasets with 25 sampled individuals for each of 2 populations and 500 base pairs (bp) per locus.

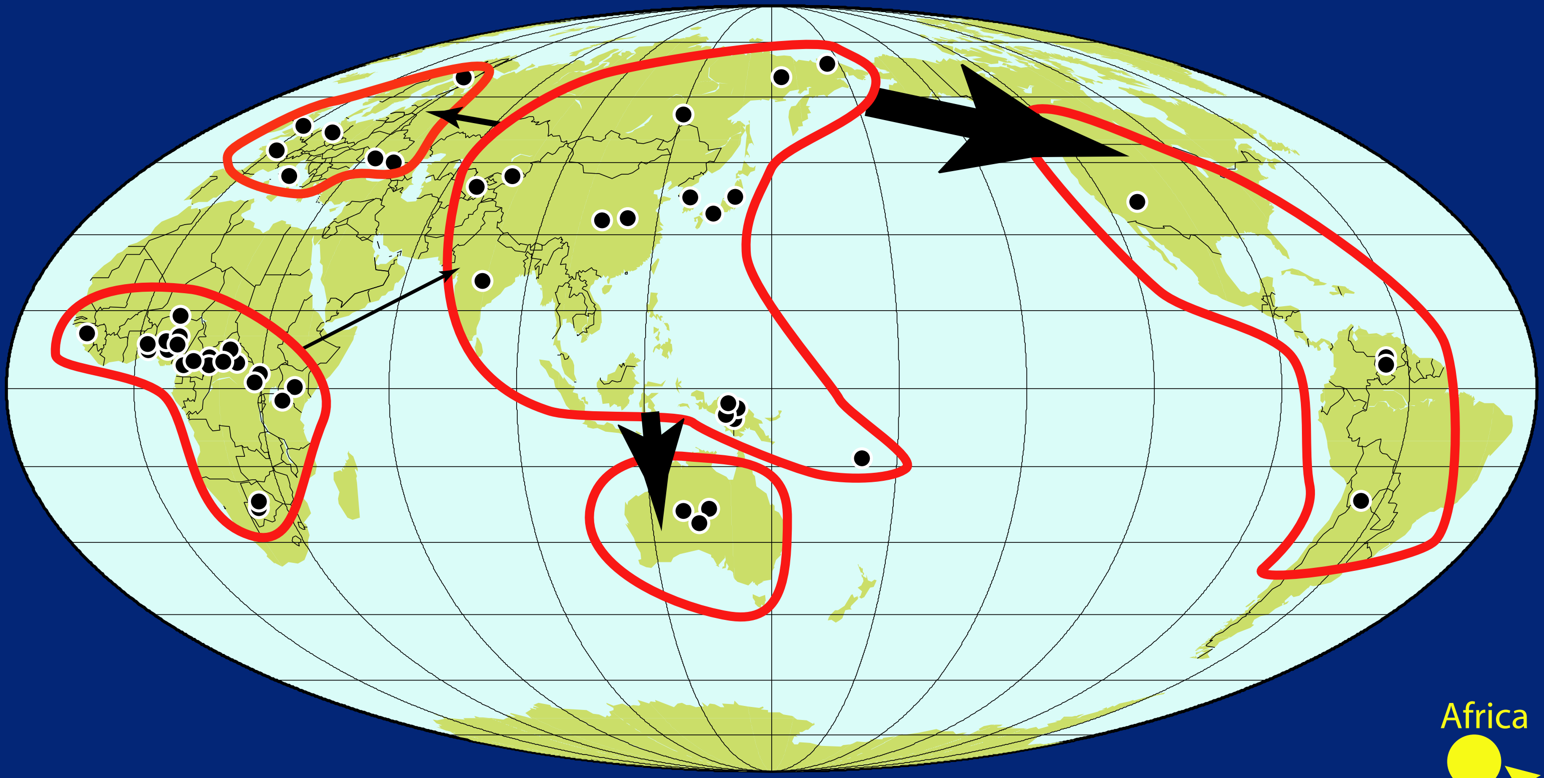
	Population 1		Population 2	
	Θ	$4N_e^{(1)}m_1$	Θ	$4N_e^{(2)}m_2$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

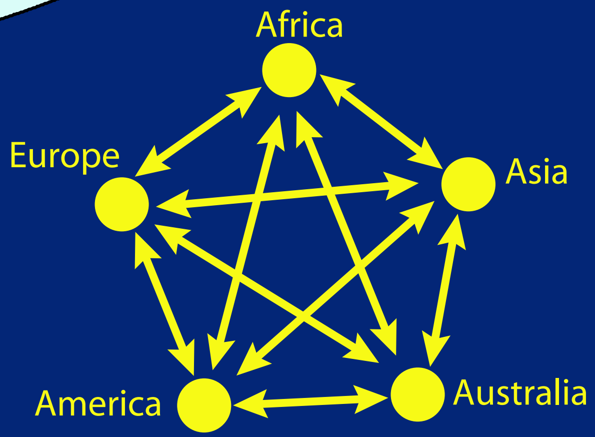
Complete mtDNA from 5 human "populations"



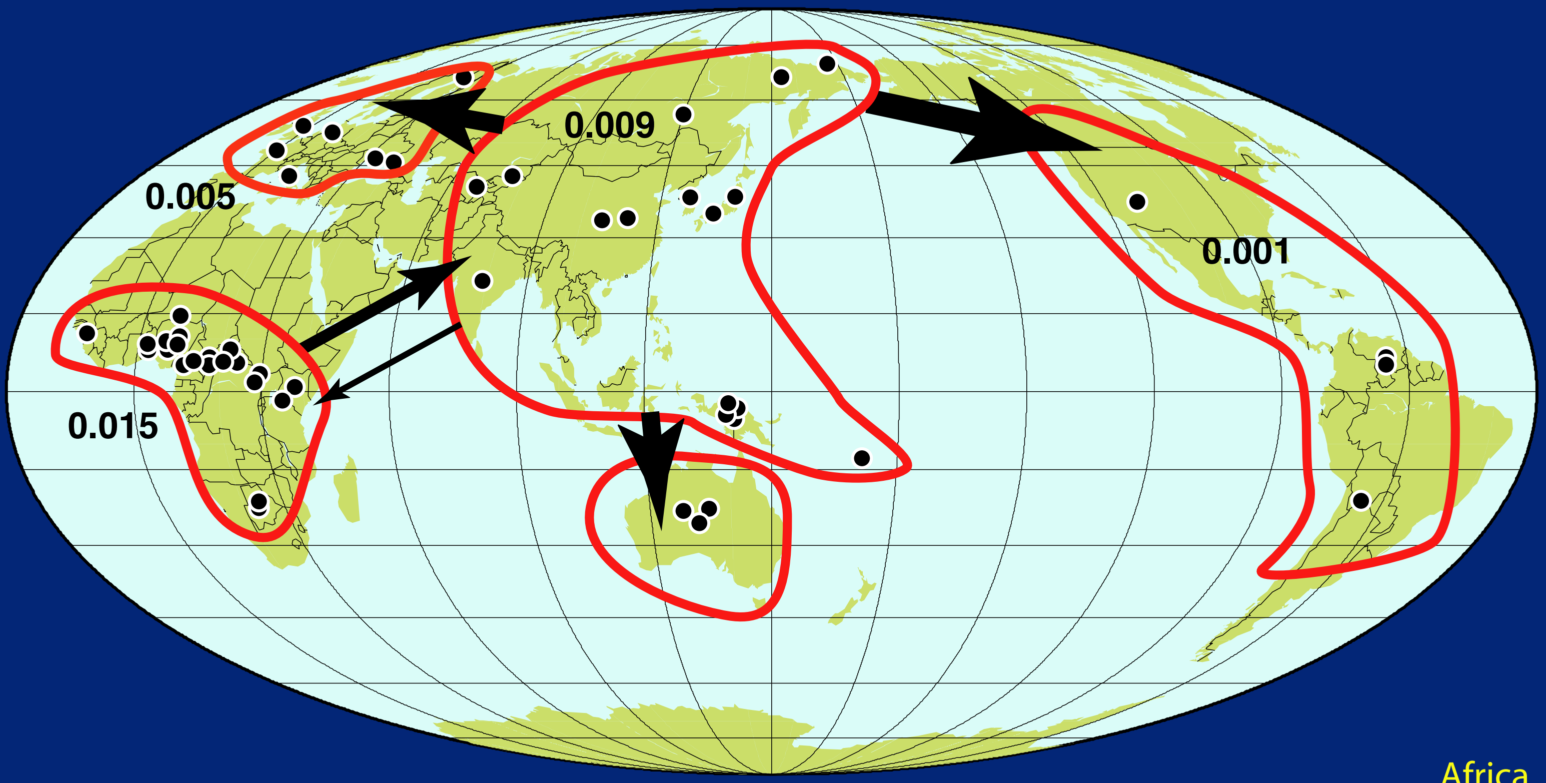
A total of 53 complete mtDNA sequences ($\sim$ 16 kb):
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.
Assumed mutation model: F84+Γ

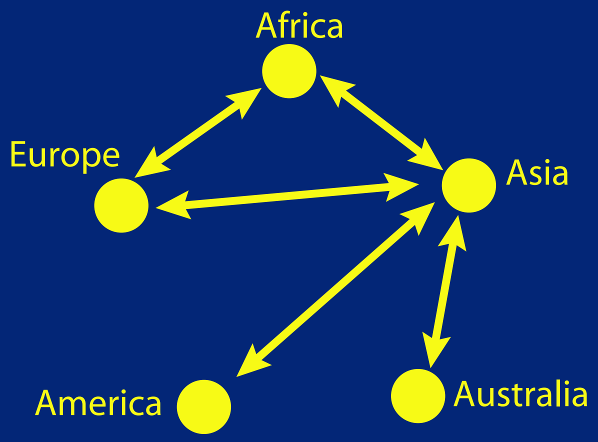
Full model: 5 population sizes + 20 migration rates





Restricted model: only migration into neighbors allowed

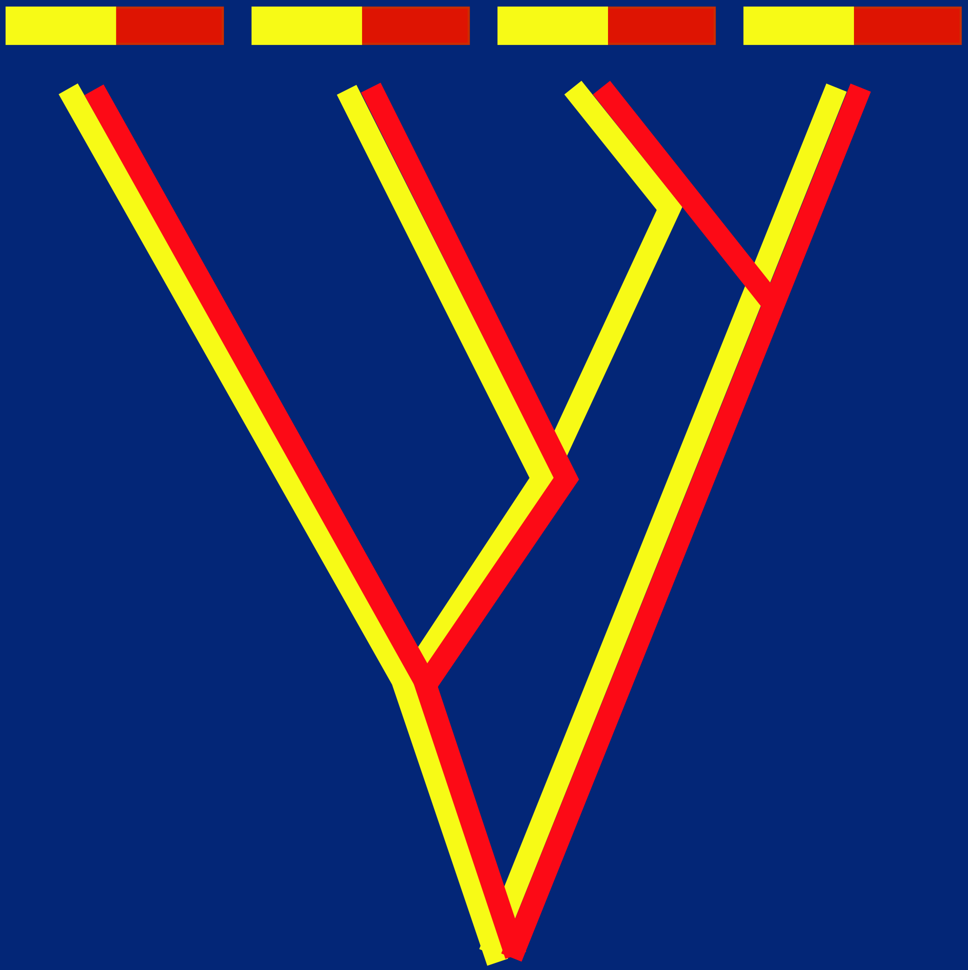




Coalescent migration estimation

- Done by *Lamarc*, *Migrate-N*, *IM/IMa* estimating:
  - Θ per subpopulation
  - Immigration from each subpopulation into each of the others
- *Lamarc* and *Migrate-N* assume stable population structure
- *IM/IMa* assume divergence of two or more populations from a common ancestor

Recombination rate estimation

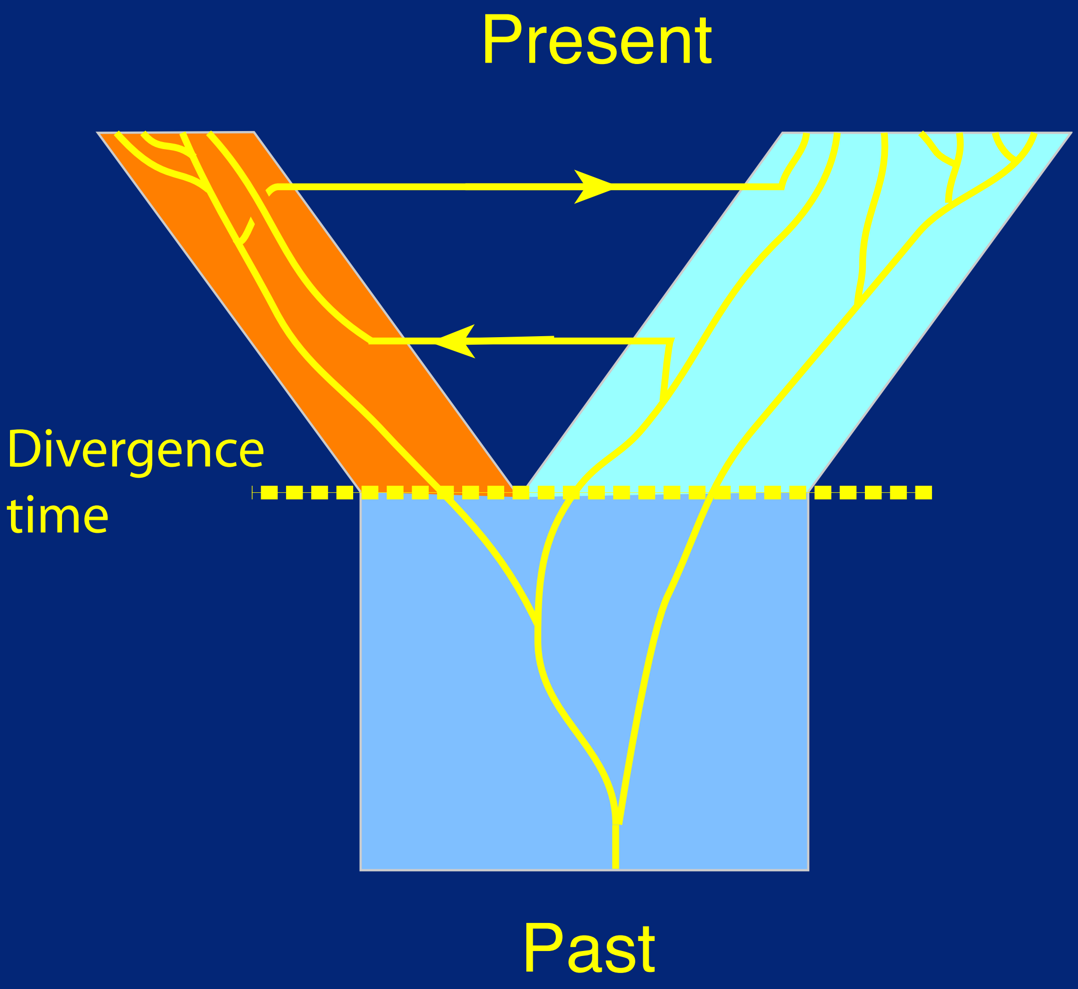


Coalescent recombination estimators

- Previously done with *Recombine*
- Currently done with *Lamarc*
- Assumptions:
  - No gene conversion
  - Equal recombination rate at every site
- Allows correct use of data with recombination to estimate other parameters
- Use of recombining data in a non-recombination-aware algorithm leads to bias

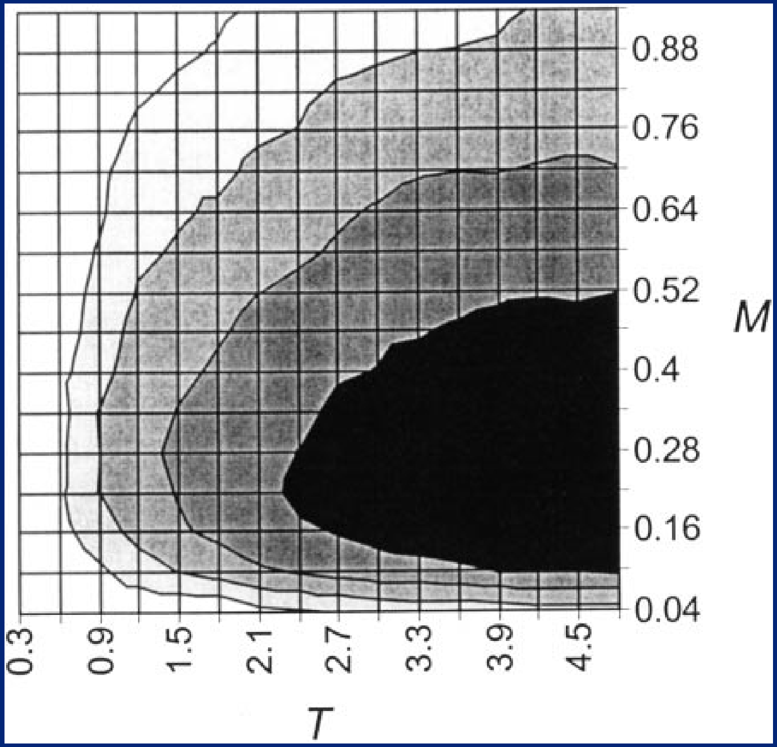
Estimation of divergence time

Wakeley and Nielsen (2001)



Estimation of divergence time

Wakeley and Nielsen (2001) Figure 7. The joint integrated likelihood surface for T and M estimated from the data by Orti et al. (1994). Darker values indicate higher likelihood.



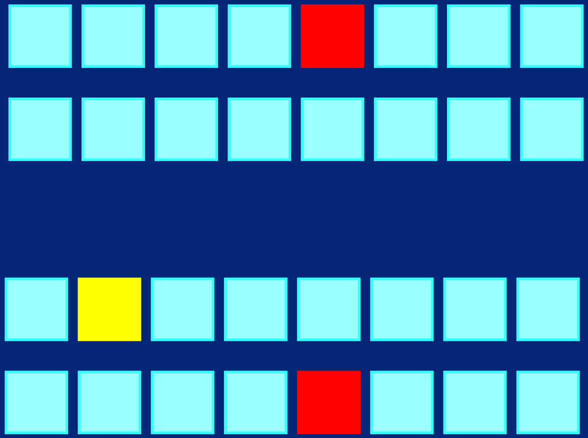
Coalescent divergence estimators

- Done with IM/IMa
- Up to 10 populations
- Co-estimates divergence time, migration rates and populations sizes
- Not all data sets can separate migration from divergence
- Multiple loci are helpful

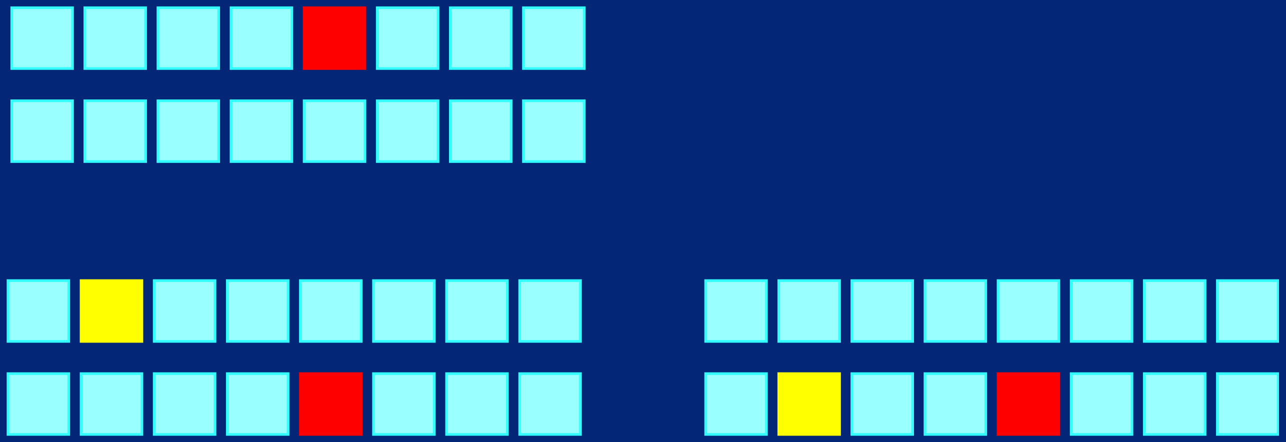
Multiple time points

- Ancient DNA or historical samples of fast-evolving organisms
- Done with *Beast* or *Migrate-N*
- Points must be:
  - Dated
  - Far enough apart for measurable evolution
- Advantages:
  - Separation of Θ into N_e and μ
  - Much better resolution of growth rates

Haplotype uncertainty



Haplotypes



Either haplotypes must be resolved or the program must integrate over all possible haplotype assignments.

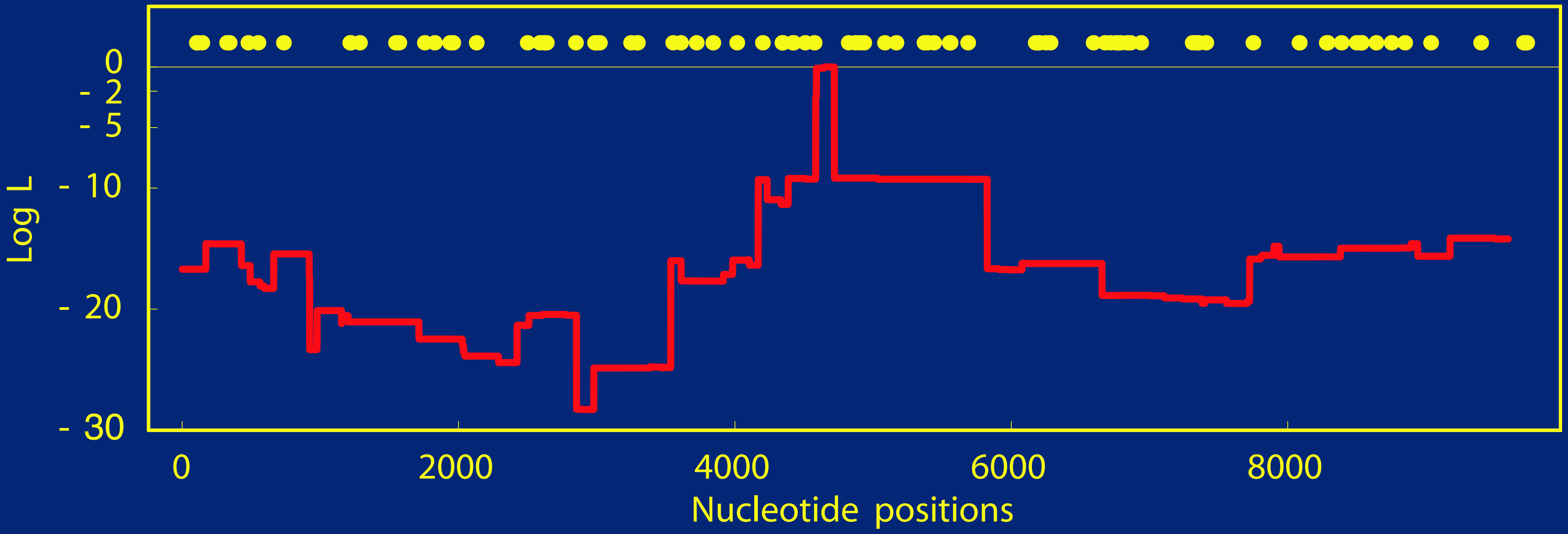
Currently only *Lamarc* can do the latter.

MCMC versus best-fit haplotypes

- Advantages of MCMC:
  - Avoids bias of ”too good” best fit
  - Incorporates error of haplotypes into error estimates
- Advantages of best-fit haplotyping:
  - Much faster
  - Avoids MCMC search failure issues
  - Can use external evidence about best haplotypes

Linkage disequilibrium mapping

With a disease mutation model we can use the recombination estimator to post-analyze the sampled genealogies that where used to estimate r and find the location of the disease mutation on the DNA.



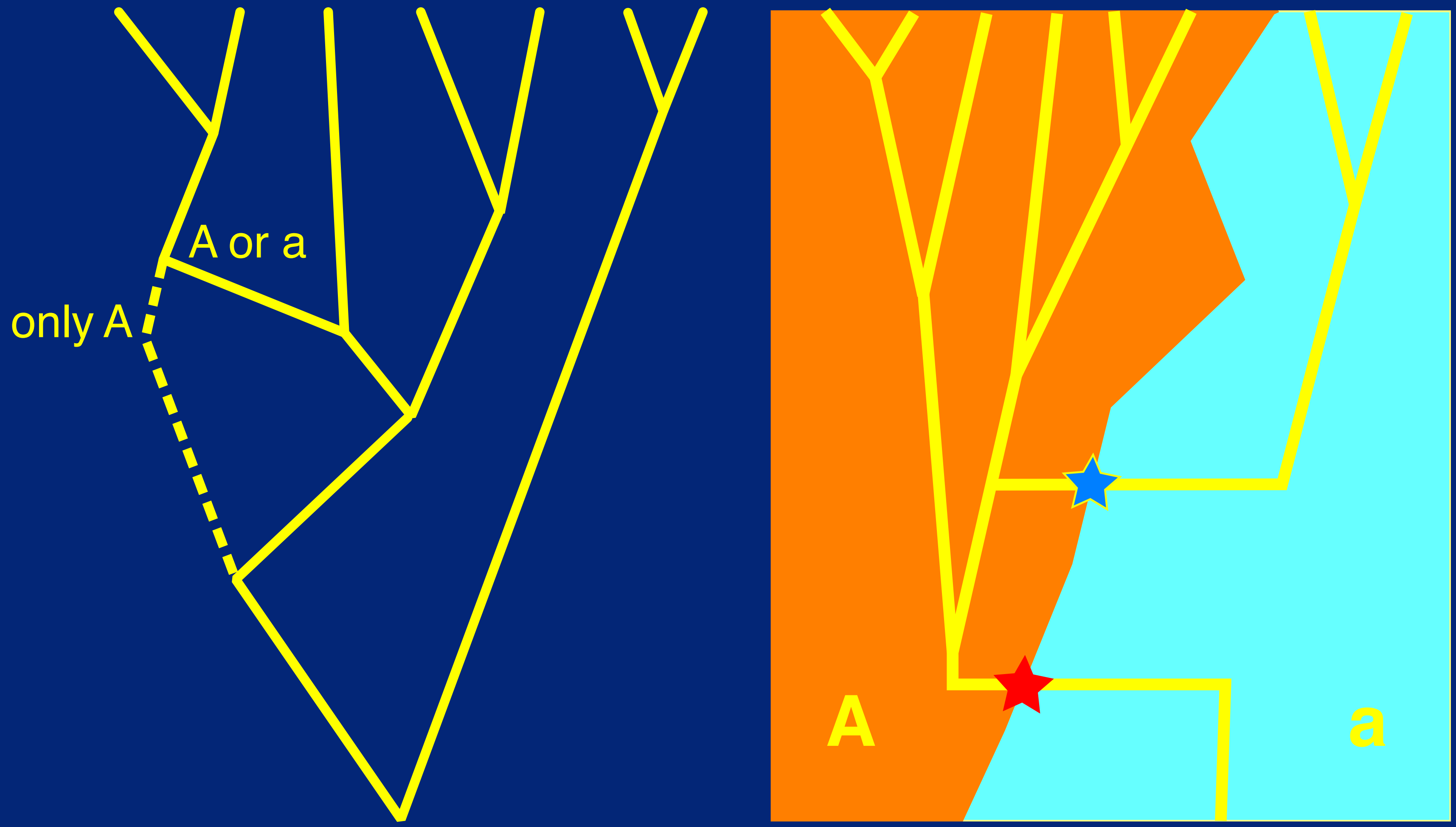
Linkage disequilibrium mapping

Lamarc can perform this type of mapping.

- Takes phenotype data with penetrance model
- Handles haplotype uncertainty
- Currently limited in the size of case it can handle
- We hope to relax this limitation soon

Selection coefficient estimation

Krone and Neuhauser (1999), Felsenstein (unpubl)



Outline

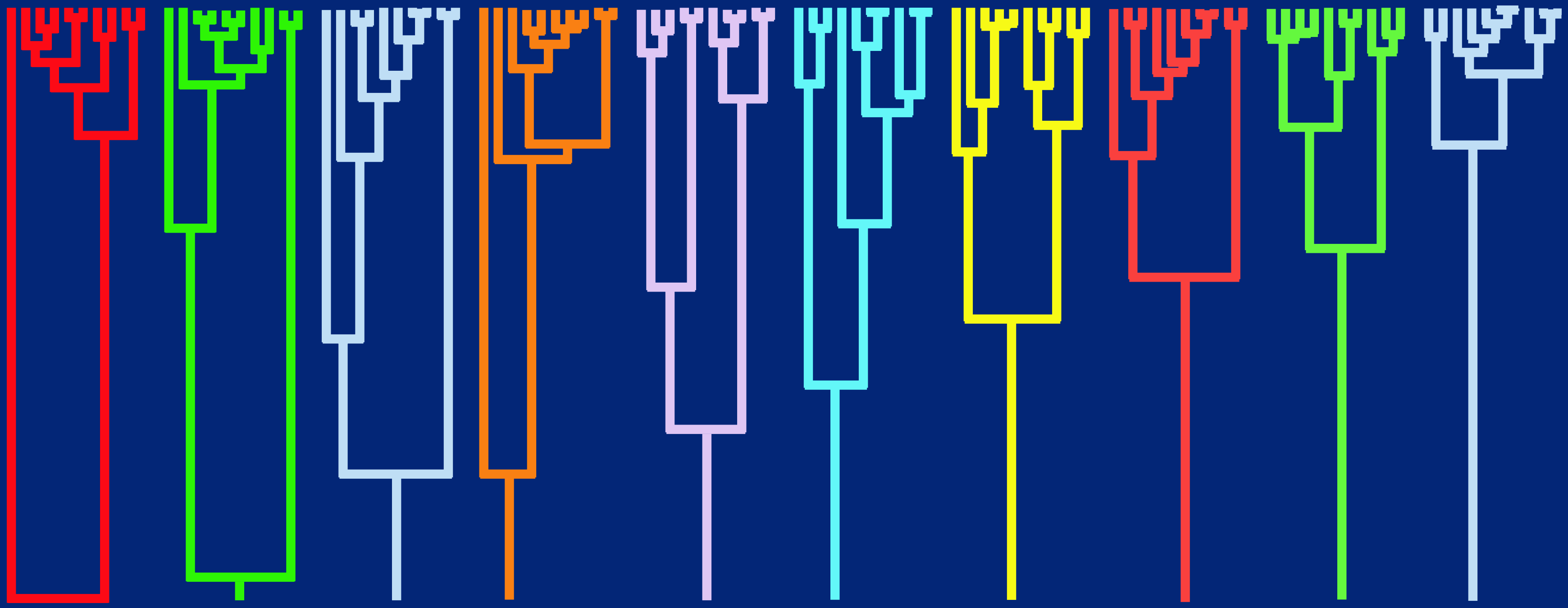
- Introduction to coalescent theory
- Genealogy samplers
- Survey of samplers
- Evolutionary forces
- **Practical considerations**

Information content of the coalescent

What can best give us more information?

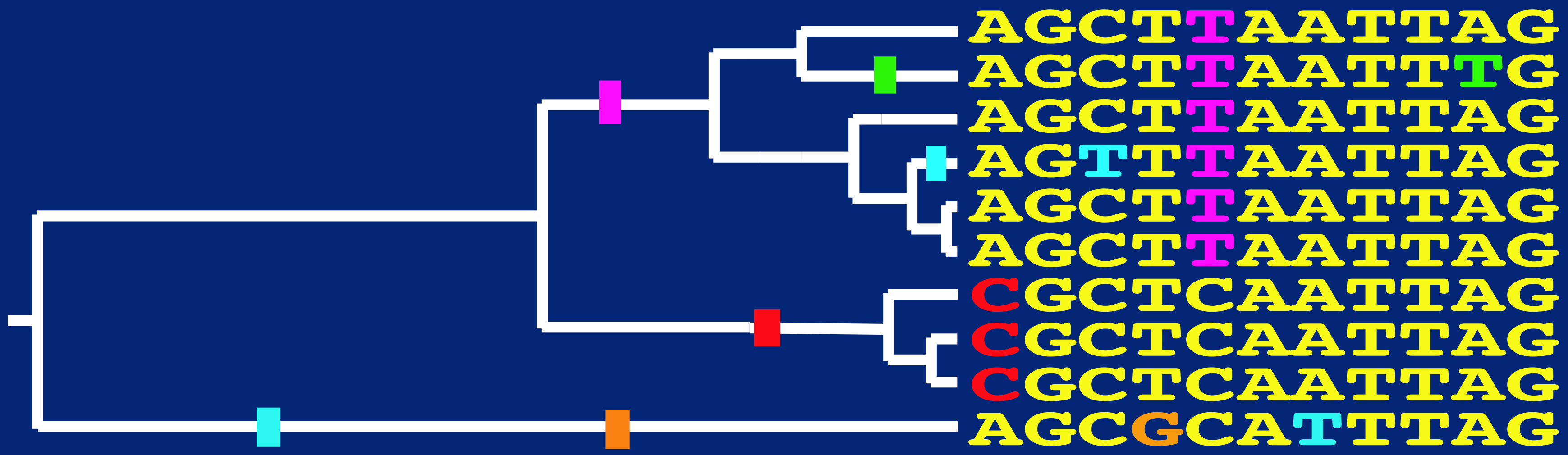
- More individuals?
- More base pairs?
- More loci?

Variability of the coalescent

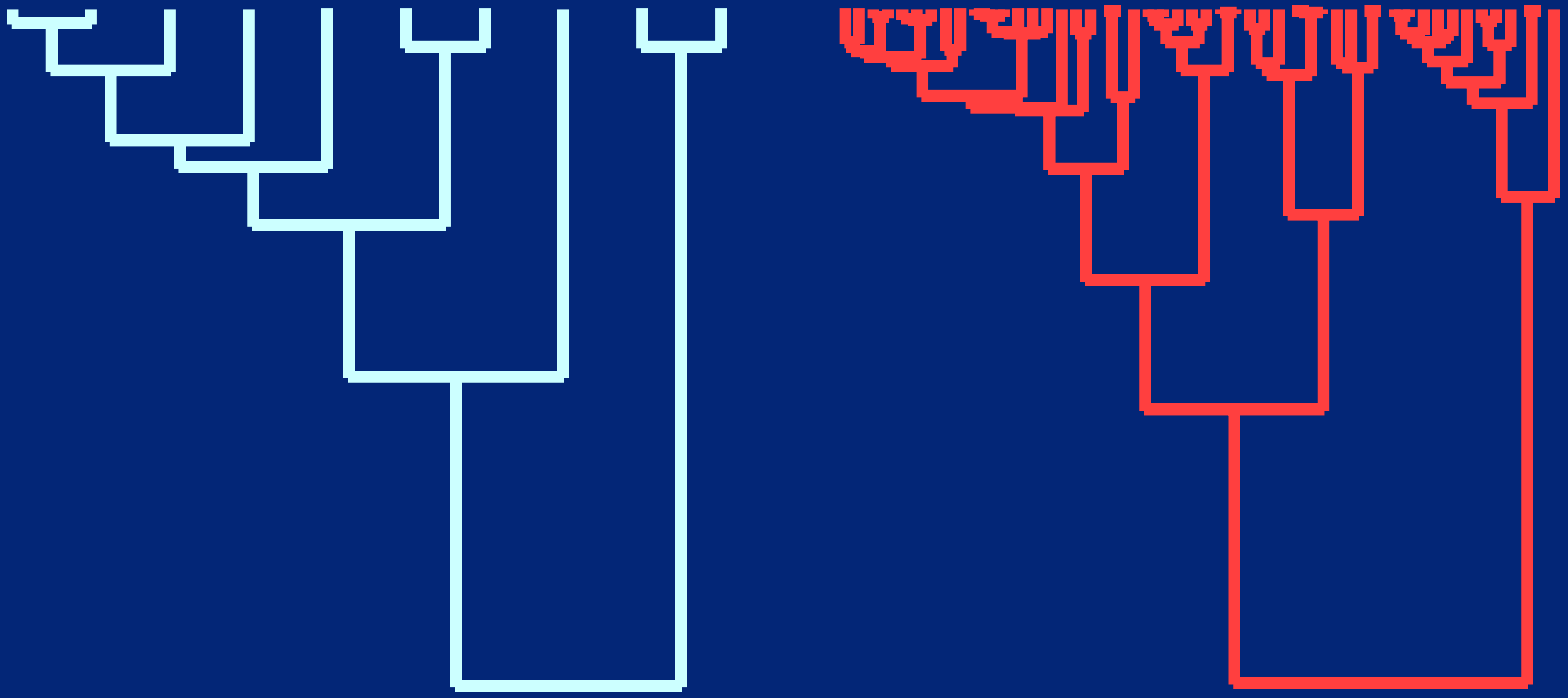


10 coalescent trees generated with the same population size, $N = 10,000$

Variability of mutations



Does adding more individuals help?



The bottom line

- The information content of a single locus is limited
- Additional sequence length or individuals are only mildly helpful
- Multiple loci allow the best estimates
- If recombination is present, long sequences can partially substitute for multiple loci
- Multiple time points can also help, if significant evolution happens between them

Two publications supporting this conclusion

- Felsenstein, J (2005) Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? MBE 23: 691-700.
- Pluzhnikov A, Donnelly P (1996) Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144: 1247-1262.

Practical advice

- The major practical problem: how long to run the program?
- Additionally: how many chains, how many steps per chain?

The problem of defaults

- Length of run varies hugely with data and model
- There are no good defaults
- Programs normally ship with defaults which let you see results quickly
- *These are not suitable for publication runs!*

Parameter estimates are still changing

If your estimate of a parameter looks like this:

Chain	Θ
1	0.0035
2	0.0047
3	0.0088
4	0.0105
5	0.0121

you have not run the program long enough. It's probably best to increase the number of steps in each chain.

Parameter estimates are still changing

If your estimate of a parameter looks like this:

Chain	Θ
1	0.0035
2	0.0047
3	0.0088
4	0.0105
5	0.0121

you have not run the program long enough. It's probably best to increase the number of steps in each chain.

You would prefer to see this:

Chain	Θ
1	0.0056
2	0.0098
3	0.0110
4	0.0107
5	0.0109

Trees aren’t being accepted

If almost all trees are being rejected, the sampler obviously cannot move well.

- This might be due to a bad starting value
- More likely it shows a need for heating

Parameter values leap around

If your estimate of a parameter looks like this:

Chain	r
1	0.0005
2	0.0047
3	0.0001
4	0.1105
5	0.0021

- Your chains may be too short. (Each visits only one of multiple peaks.)
- Your data may have no power.

Program takes forever to run

- You may be asking too much
- If estimating migration, try restricting your migration model
- Disable or fix at constant values parameters you aren't interested in
- Try randomly removing some individuals
  - More than 20 individuals per population doesn't help much
  - Don't systematically remove similar sequences!
- Borrow a faster computer with lots of memory

Error bars too wide

- Particularly common with growth and recombination estimates
- Usually not an error in your run
- Badly performing genealogy samplers get estimates that are TOO NARROW
- If yours are too wide:
  - Limit the number of parameters being inferred
  - Add unlinked loci
  - Add time points
  - Add sequence length, if recombination present
- Always publish error bars; point estimates have no meaning without them

Validating genealogy samplers

Two useful tools:

- TRACER (Drummond and Rambaut)
  - ESS statistic
  - Traces of parameters throughout the run
  - Histograms of parameter values
- AWTY (Swofford)
  - Traces of clade probabilities throughout the run

Review paper

Kuhner MK (2008) Coalescent genealogy samplers: windows into population history. TREE 24:86-93.

Thanks to

Joe Felsenstein
Peter Beerli
Jon Yamato
Lucrezia Bieler
Elizabeth Thompson
Eric Rynes
Lucian Smith
Elizabeth Walkup

What was the long-term population size of gray whales?

Alter, Rynes and Palumbi (2007) DNA evidence for historic population size and past ecosystem impacts of gray whales. PNAS 104: 15162-15167.

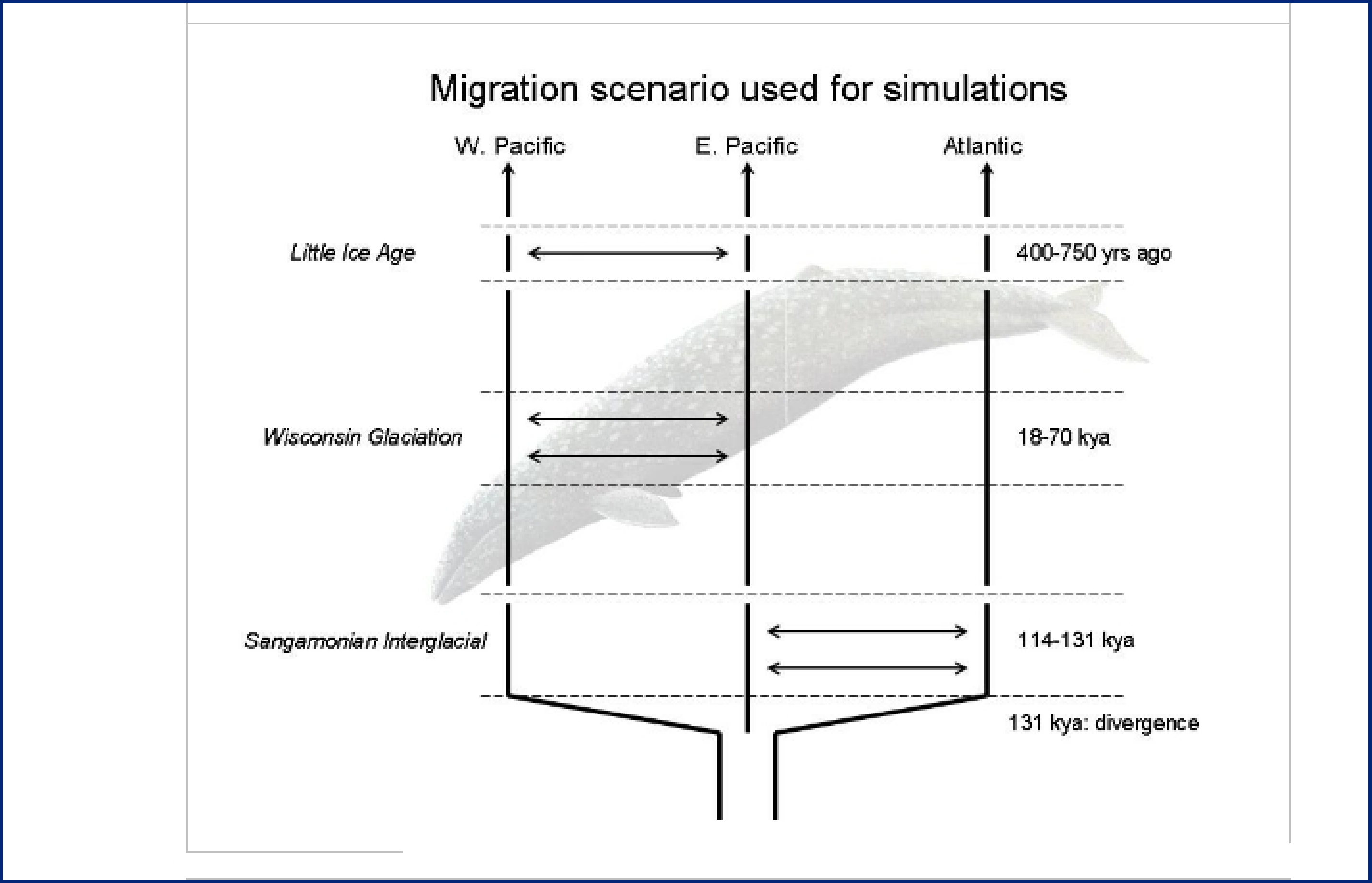
What was the long-term population size of gray whales?

- How many gray whales pre-whaling?
- Whaling ship records not conclusive
- Recent slowing of the observed growth rate may suggest recovery
- Molecular data an alternative source of information

What was the long-term population size of gray whales?

- 10 loci:
  - 7 autosomal
  - 2 X-linked
  - 1 mtDNA
- Complex mutational model with rate variation among loci
- Complex population model with subdivision and copy number
- Complex demographic model relating N_{census} to N_e

What was the long-term population size of gray whales?



What was the long-term population size of gray whales?

	Locus	n	Estimated N
Aut	ACTA	72	162,625
	BTN	72	76,369
	CP	76	77,319
	ESO	72	272,320
	FGG	72	180,730
	LACTAL	72	44,410
	WT1	80	51,972
X	G6PD	30	2,769
	PLP	52	92,655
mtDNA	Cytb	42	107,778
	All data		96,400 (78,500-117,700)
	Current census		18,000-29,000
	Previous models		19,480-35,430

What was the long-term population size of gray whales?

- Important conservation implications
- Effect on ecosystem significant:
  - Resuspension of up to 700 million cubic meters sediment
  - (12 Yukon Rivers worth)
  - Food for 1 million sea birds
- If accepted, result suggests halving gray whale kill rate
- Broadly similar results for minke, humpback, and fin whales