# Plan for Module 16

| | | | |
|---|---|---|---|
| Wednesday 6/22 | 1:30-3:00 | Introduction | Philip |
| | 3:30-4:00 | Introduction (continued) | Philip |
| | 4:00-5:00 | Introduction | Mary |
| Thursday 6/23 | 8:30-10:00 | Recombination | Philip |
| | 10:30-12:00 | Recombination practical | Philip |
| | 1:30-3:00 | Population size and structure | Mary |
| | 3:30-5:00 | Gene flow practical | Mary |
| | 5:00-7:00 | Tutorial | Mary/Philip |
| Friday 6/24 | 8:30-10:00 | Selection | Philip |
| | 10:30-12:00 | Selection practical | Philip |
| | 1:30-3:00 | Applications and study design | Mary |
| | 3:30-5:00 | Coalescent practical | Mary |

## Details–Wednesday

- Wednesday afternoon: Introduction to the Coalescent
  - population genetics, Wright-Fisher model
  - 2-sample coalescent
  - n-sample coalescent
  - Coalescent and sequence variation
  - Parameters of the coalescent
  - Case studies

# Details–Thursday

- Thursday morning: Recombination
  - Genetic recombination
  - Linkage disequilibrium
  - LDhat, RJMCMC, Phase
  - Hands-on recombination exercise
- Thursday afternoon: Growth and Gene Flow
  - Population growth and shrinkage
  - Population subdivision and gene flow
  - Population divergence
  - Genealogy samplers: Migrate-N, Lamarc, Beast, IM
  - Hands-on gene flow exercise

# Details–Friday

- Friday morning: Selection
  - Phylogenetic approaches
  - Population genetics approaches
  - Coalescent approaches
  - Hands-on selection exercise
- Friday afternoon: Applications of the Coalescent
  - Study design
  - Limits of applicability
  - Validation
  - Hands-on study fine-tuning exercise

# Outline

1. What types of questions can the coalescent answer?

2. What approaches are used?

3. Case studies

# What is the coalescent good for?

- We are interested in questions like
  - How big is this population?
  - When did these populations diverge?
  - Are they 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? What kind?

# Coalescent versus traditional population genetics

- Traditional pop gen:
  - Trace the evolutionary process *forward* in time
  - Predict range of outcomes for a giving starting position

- Coalescent analysis:
  - Trace the evolutionary process *backward* in time
  - Predict range of scenarios leading to given final position

- Since we know final position more often than starting position, the coalescent is useful for many questions where traditional population genetics struggles

# Coalescent versus traditional population genetics

- Traditional pop gen: A neutral allele is now at 5% frequency
  - How likely is it to fix?
  - How long will that take?
  - What if it were under selection?
- Coalescent: Ten out of thirty haplotypes surveyed carry a particular variant
  - How old is the variant?
  - Is it under selection?
  - Has it been transferred among populations?

## Range of applicability

The coalescent is appropriate for:

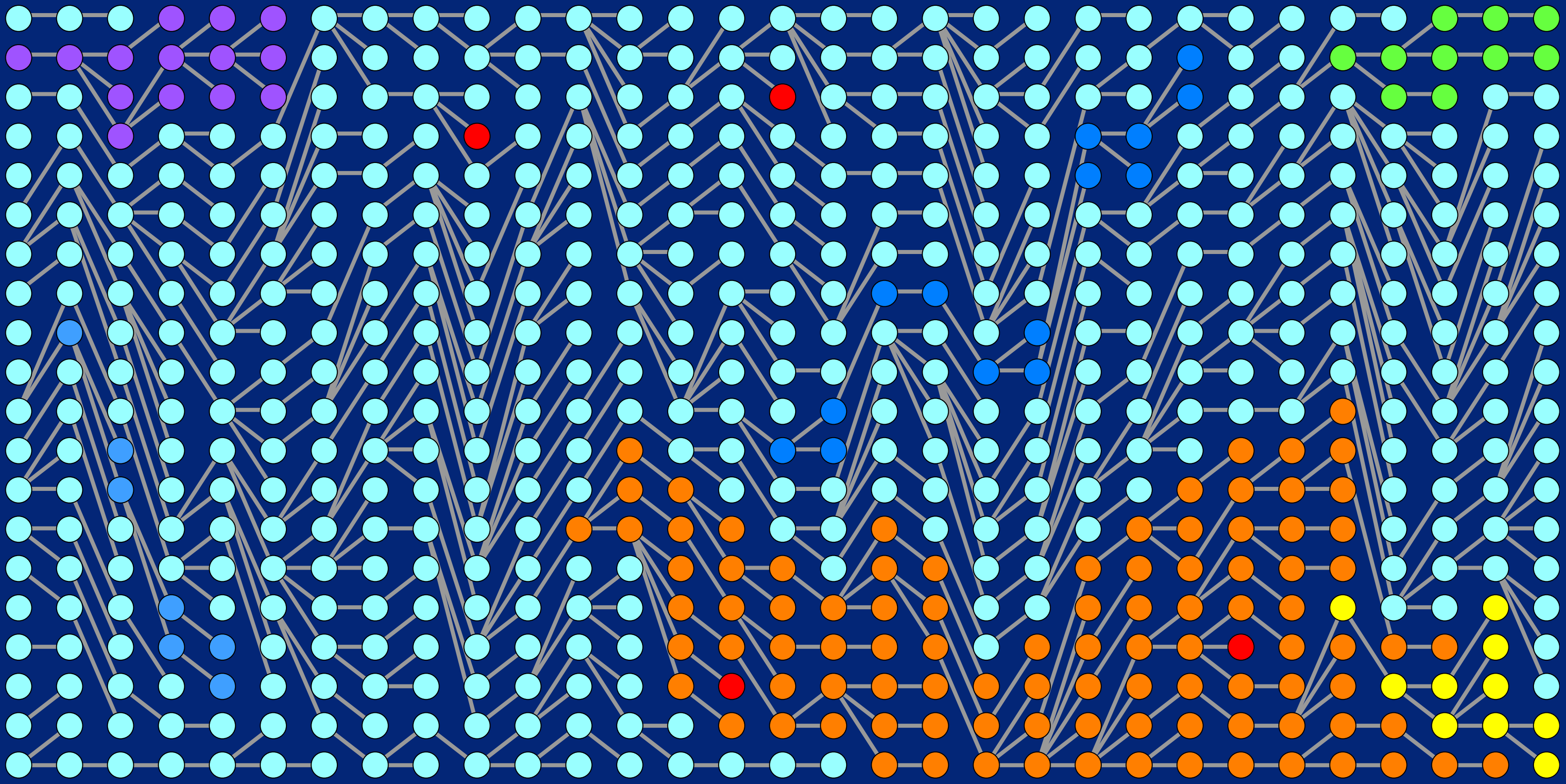
- Single populations
- Interrelated populations
- Recently diverged species

Beyond this level, other processes come into play

# Key concepts in the coalescent

- In an idealized coalescent everything depends on population size
- Deviations from idealized population give us information on other parameters:
  - Difference between census size and effective size
  - Changes in size over time
  - Population subdivision
  - Population splitting (divergence)
  - Recombination
  - Natural selection

# 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 is changing 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
- Social insects have the $N_e$ of their breeding population, not their headcount

## The $\Theta$ 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 $\mu$.
- We can no longer estimate $N$ but only the composite parameter $\Theta$.
- $\Theta = 4N_e\mu$ in diploids and $2N_e\mu$ in haploids.
- External information can allow us to separate $\Theta$ and $\mu$.

# How to separate $N_e$ and $\mu$?

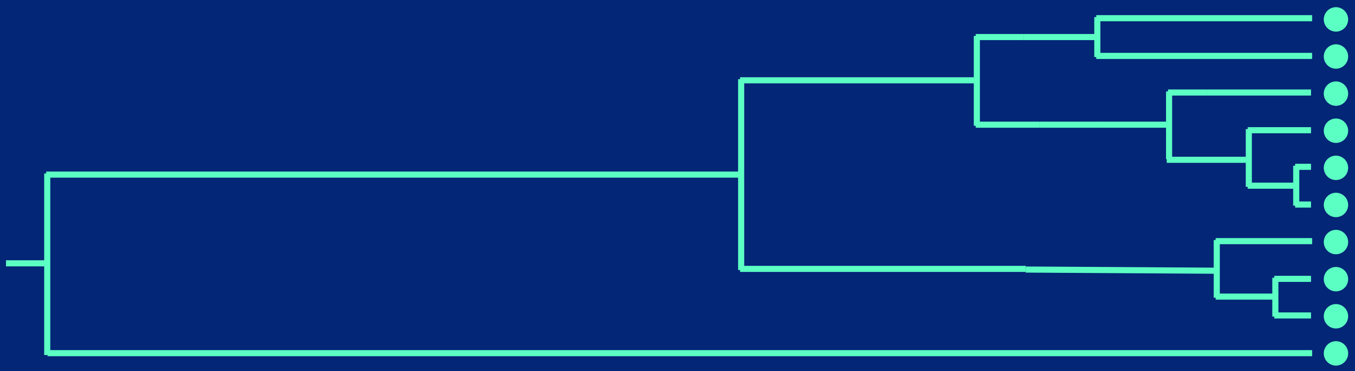
Given an estimator of $\Theta = 4N_e\mu$:

- If one is known, the other naturally follows
- $N_e$ is hard to estimate
- $\mu$ can be estimated from dated fossils
- Multiple observations with significant evolution between them allow separation of $N_e$ and $\mu$
- These could come from ancient DNA
- In fast-evolving species like viruses they can be directly observed

# Utopian coalescent population size estimator

1. We get the correct genealogy from an infallible oracle

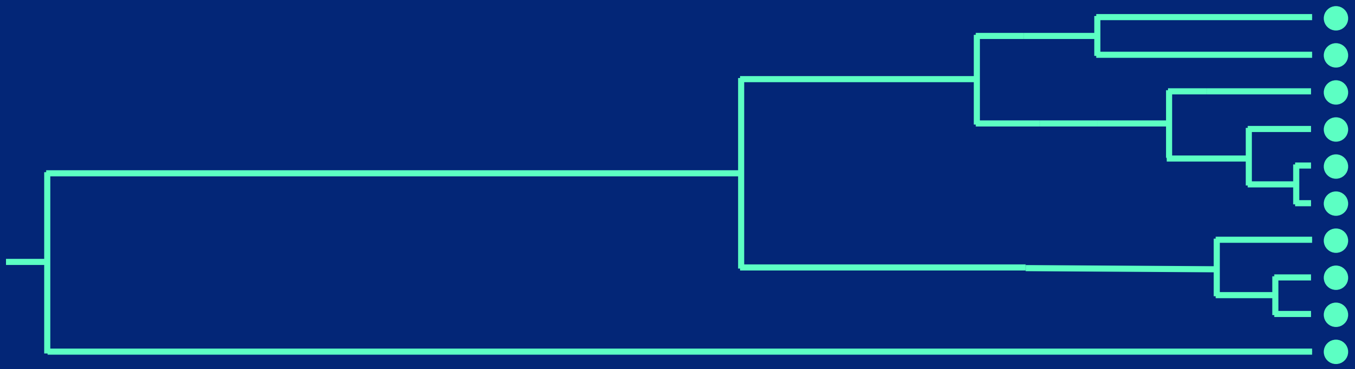
2. We know that we can calculate $p(\text{Genealogy}|N_e)$



# Utopian population size estimator

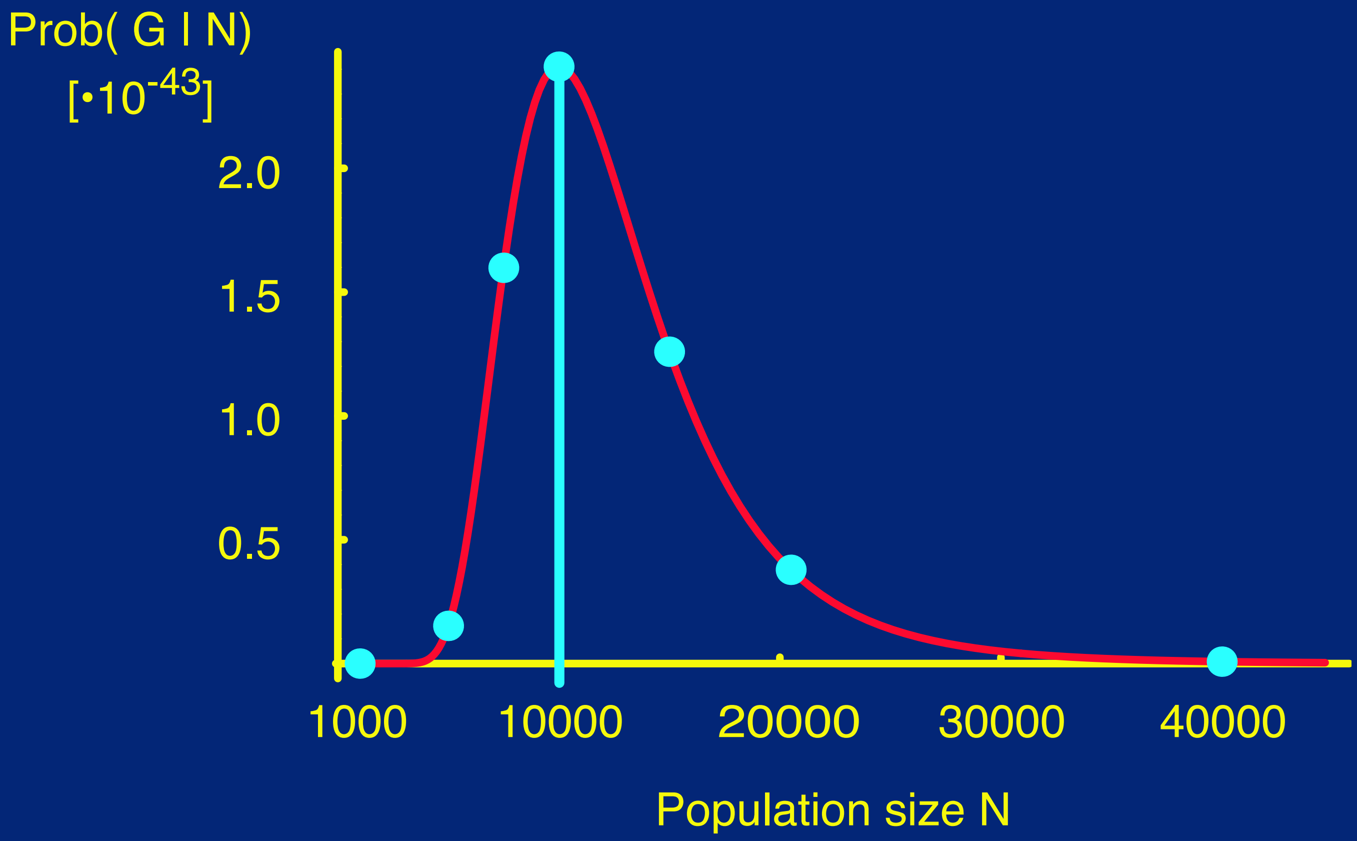
1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation



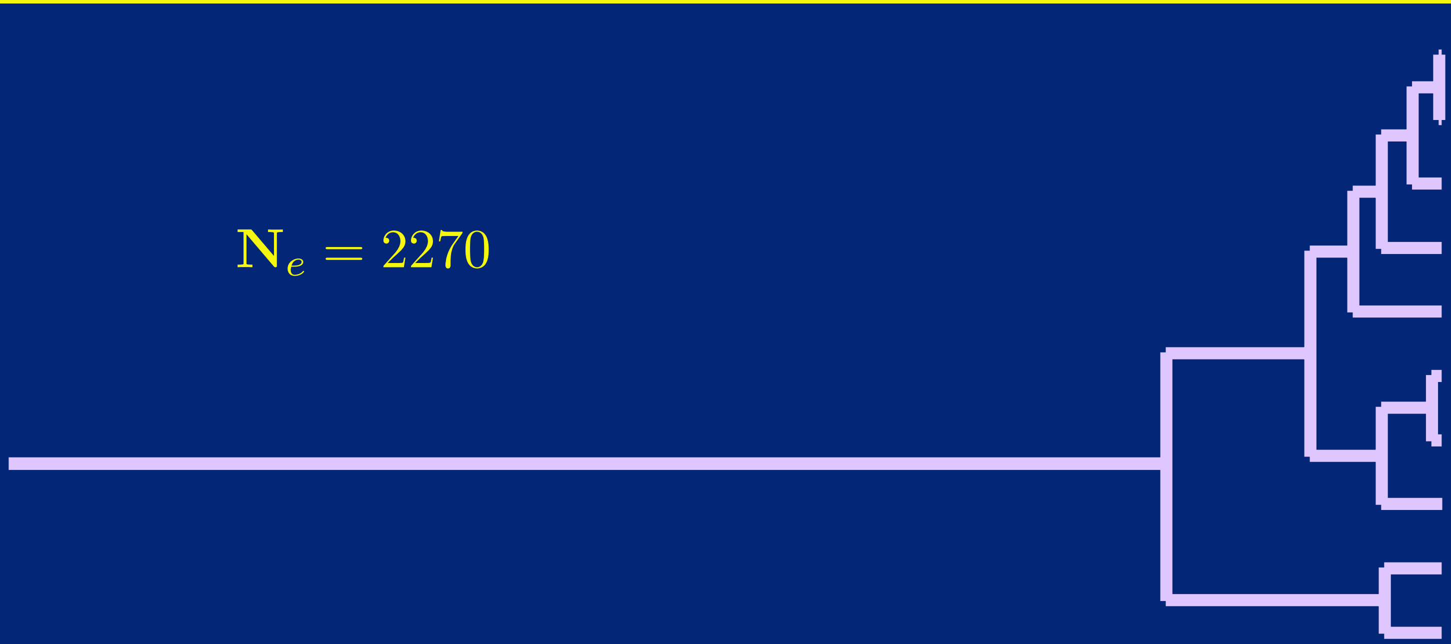
$$p(\text{Genealogy}|N_e) = \prod_{j}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N_e}} \frac{1}{2N_e}$$

# Utopian population size estimator

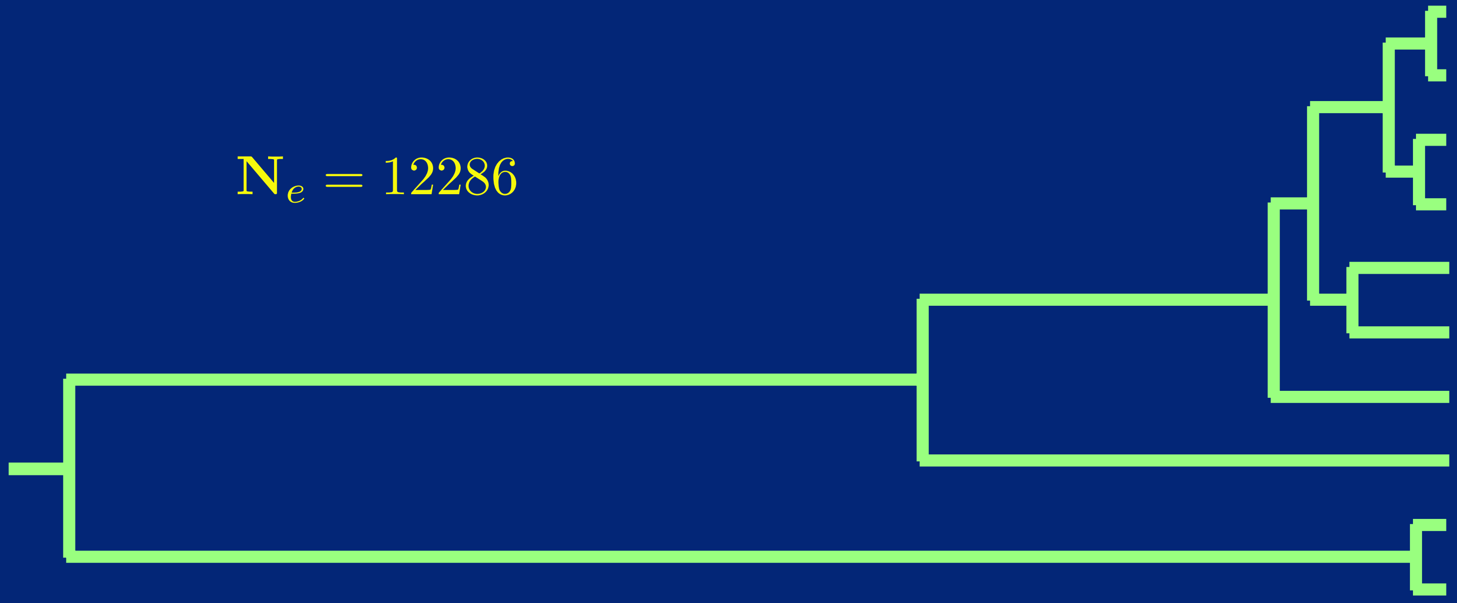


# Utopian population size estimator

$\mathrm{N}_e = 2270$



$\mathrm{N}_e = 12286$



## 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 doesn't have enough information for good tree inference

# Ways to use the coalescent

- Summary statistics
  - Watterson's estimator of $\theta$
  - FST (estimates $\theta$ and/or migration rate)
  - Hudson's and Wakeley's estimators of recombination rate
- Known-tree methods
  - UPBLUE (Fu)
  - Skyline plots
- Few-tree methods
  - Nested clade analysis (Templeton)

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

# More ways to use the coalescent

- Approximate Bayesian Computation (ABC)
  - Simulate random coalescent genealogies with a particular set of parameters
  - Simulate data on those genealogies
  - Adjust parameters until results are “similar” to real data
  - Summary statistics define “similar”
- Full genealogy samplers
  - Find many genealogies which fit the data well
  - Make a joint estimate of the parameters from all of these genealogies

These methods are more powerful and flexible, but challenging to design and use

## Use of the coalescent: two case studies

Things to look for:

- What questions were being addressed?
- What types of data were used?
- How were coalescent methods used?
- How were non-coalescent methods used?
- How were the results validated?

# 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 $(N)$: 3,400,000

- Effective population size $(N_e)$: ?

- 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

## What is the genetic population size of red drum?

Assumptions of the coalescent analysis:

- Constant population size
- Mutation rate $10^{-3}$ to $10^{-5}$
- No selection

## 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
- "Estuary lottery" may explain results

# Where to go with this finding?

- Check it experimentally–maternity testing of young fish?
- Try to find reasons for the high reproductive variance
- Be careful of this species as it may be fragile
  - Red drum were once commercially fished
  - The population responded poorly and the fishery was closed
  - Despite large numbers the species may be vulnerable
  - Are there other species like this?

# 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.

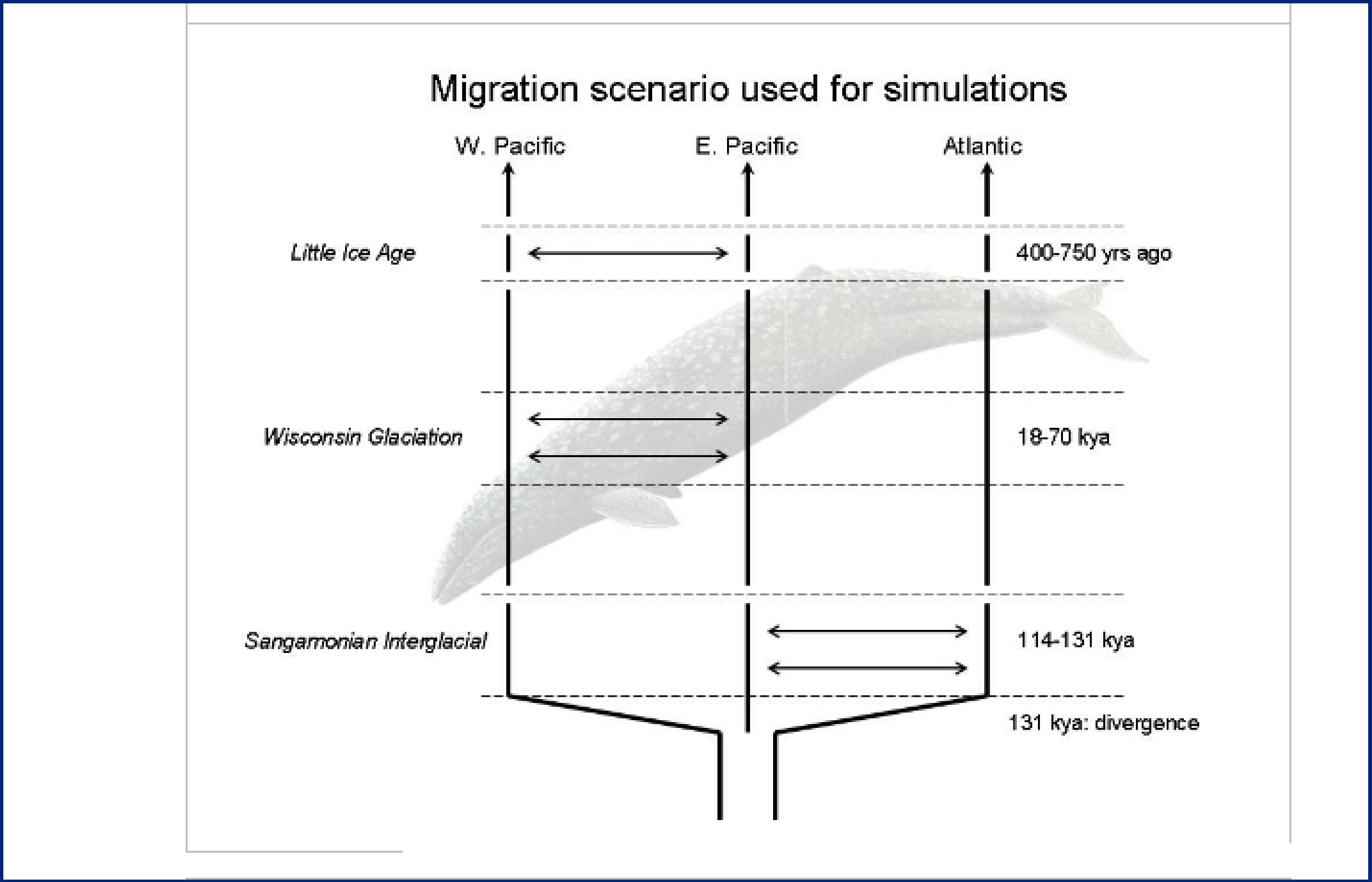
## What was the long-term population size of gray whales?

- 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 does this imply?

- 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

# Should we believe this result?

- Strengths:
  - Multiple loci improve power and avoid distortions
  - Population structure taken into account
  - Does not rely on whalers' records, which may be falsified
- Weaknesses:
  - Interpretation relies on external estimate of mutation rate
  - Selection on coding loci could distort results
  - Relies on model of relationship between $N$ and $N_e$

## Where to go with this finding?

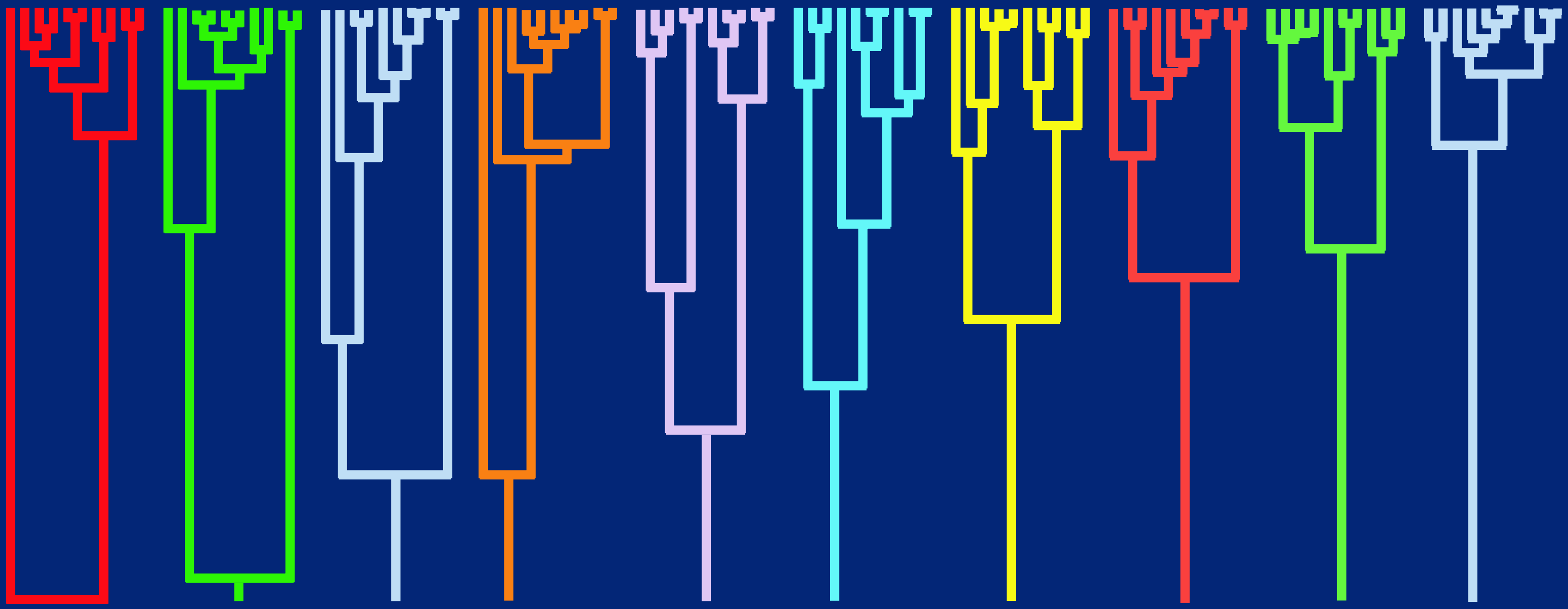
- Non-coding sequences
- Whale lice as corroboration? (Jon Seger's work)
- Ancient DNA?
- More sophisticated demographic models?
- Not time to de-list gray whales yet

# Information content of the coalescent

What can best give us more information?

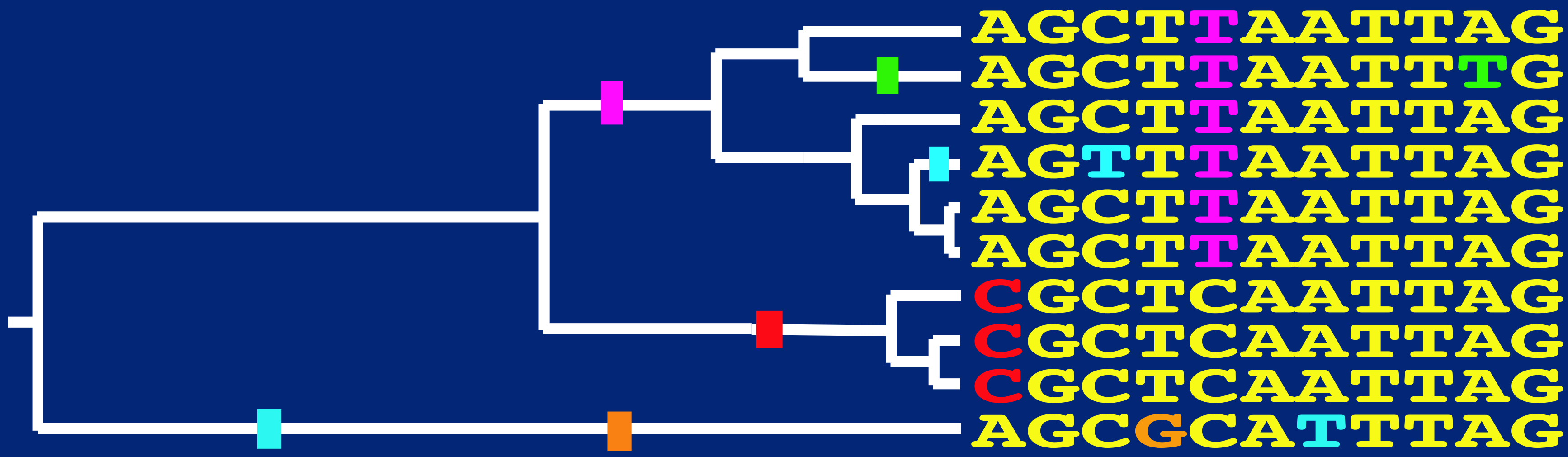
- More individuals?
- More base pairs?
- More loci?

# Variability of the coalescent



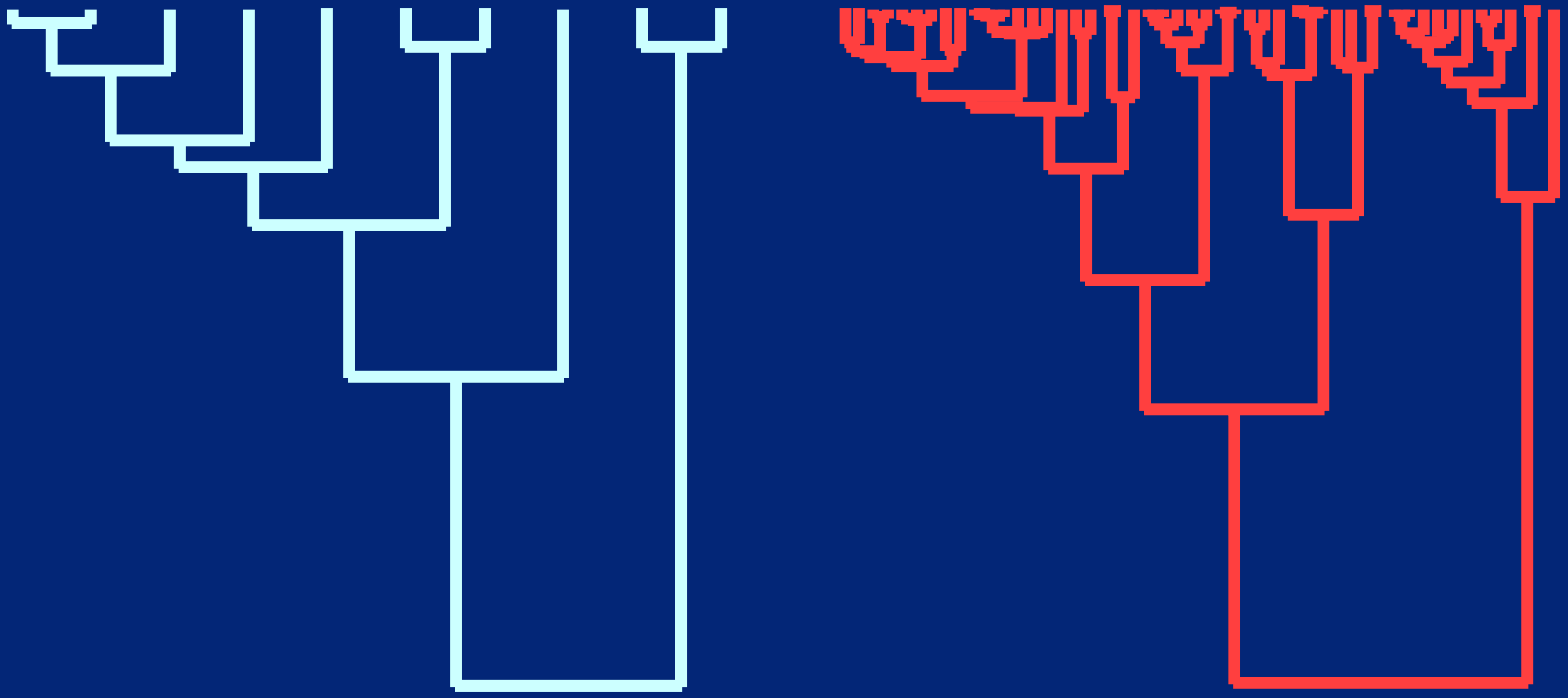
10 coalescent trees generated with the same population size, $N = 10,000$

# Variability of mutations



AGCTTTAATTAG
AGCTTTAATTTG
AGCTTTAATTAG
AGTTTTAATTAG
AGCTTTAATTAG
AGCTTTAATTAG
CGCTCAATTAG
CGCTCAATTAG
CGCTCAATTAG
AGCGCATTTAG

# 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

## Focus: genealogy samplers

- My practicals will focus on genealogy samplers
- Most statistically powerful way to extract information from coalescent genealogies
- Challenging to design and use
- Brief overview now, practical details later

# Parameter estimation by genealogy sampling

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

## Parameter estimation by genealogy sampling

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

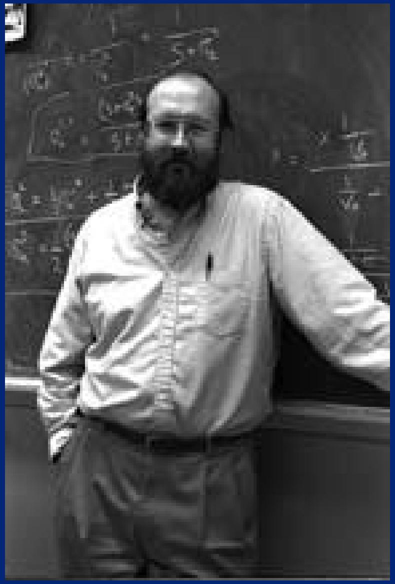
## Parameter estimation by genealogy sampling

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

## Parameter estimation by genealogy sampling

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

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



## Parameter estimation by genealogy sampling

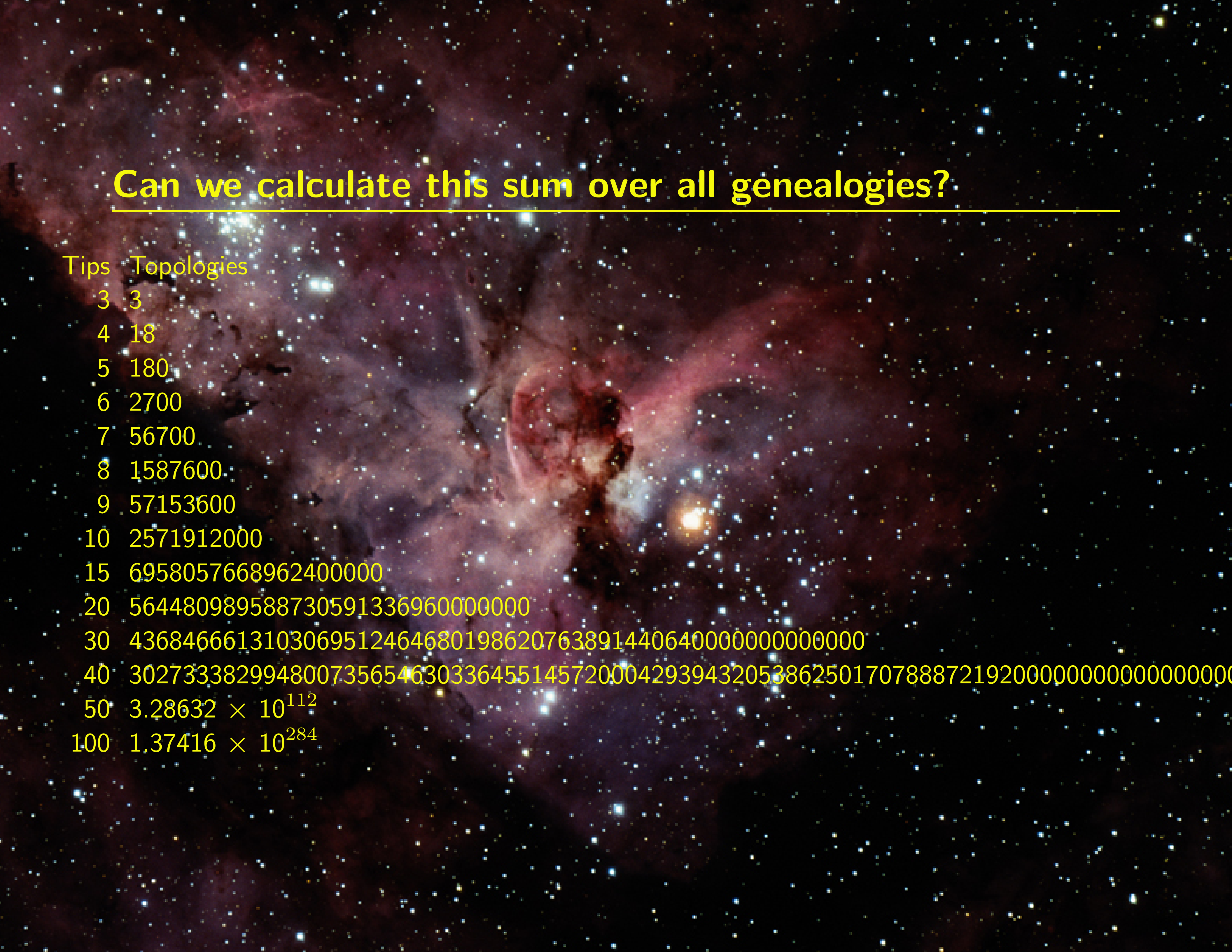
$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

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

# Parameter estimation by genealogy sampling

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

$\sum_G$ is a problem



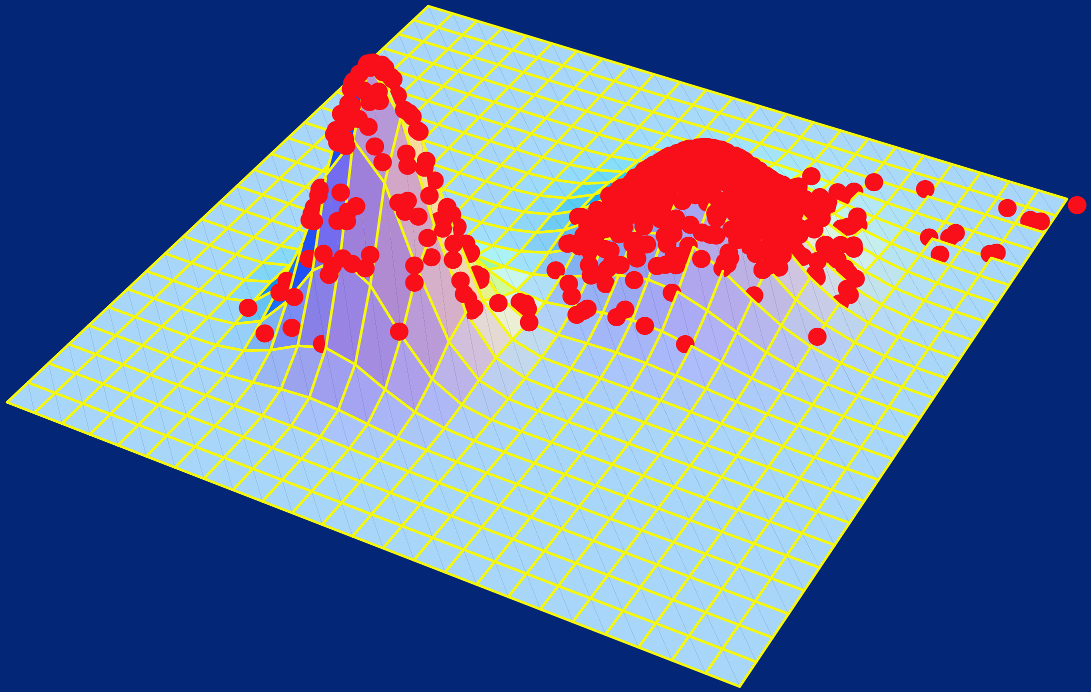
## Can we calculate this sum over all genealogies?

| 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 | 3027333829948007356546303364551457200042939432053862501707888721920000000000000000 |
| 50 | $3.28632 \times 10^{112}$ |
| 100 | $1.37416 \times 10^{284}$ |

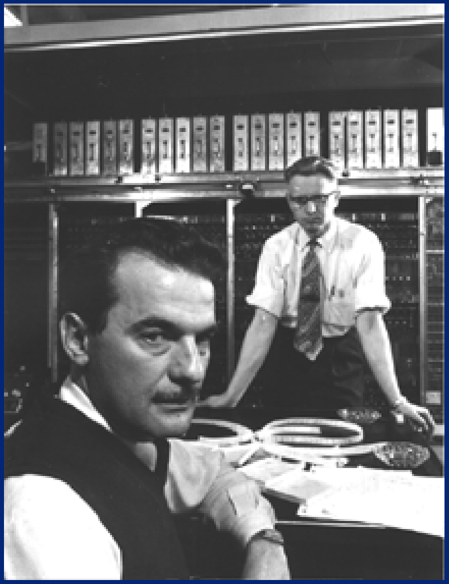
# A solution: Markov chain Monte Carlo

- 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

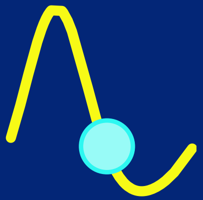


# A solution: Markov chain Monte Carlo

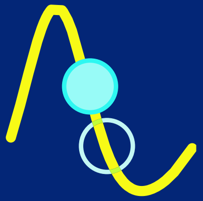


Metropolis recipe

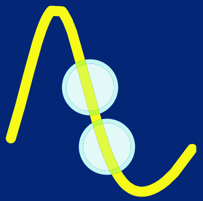
0. first state



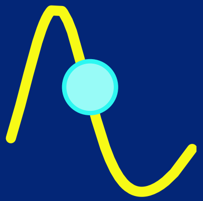
1. perturb old state and calculate probability of new state



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.

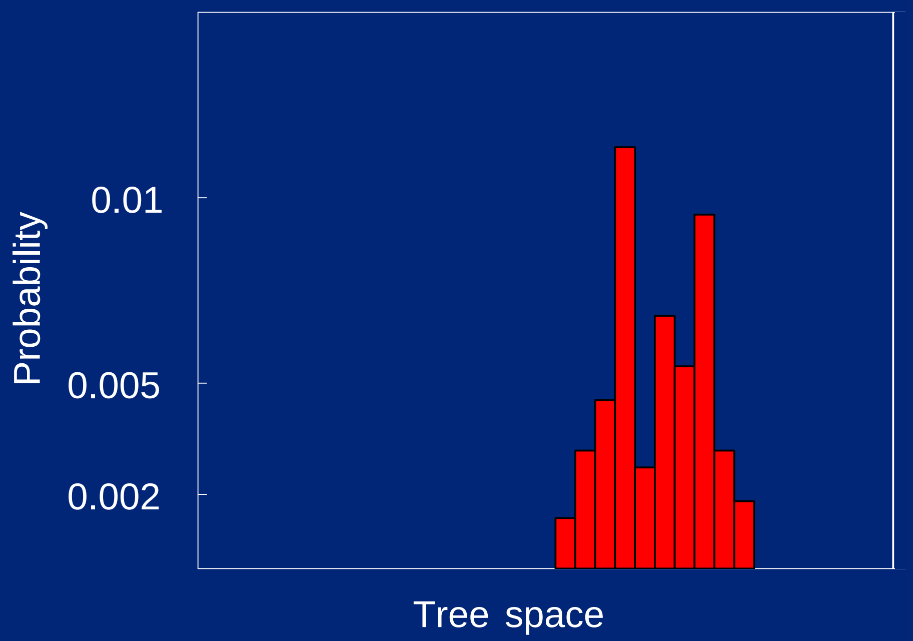


3. move to new state if accepted otherwise stay at old state

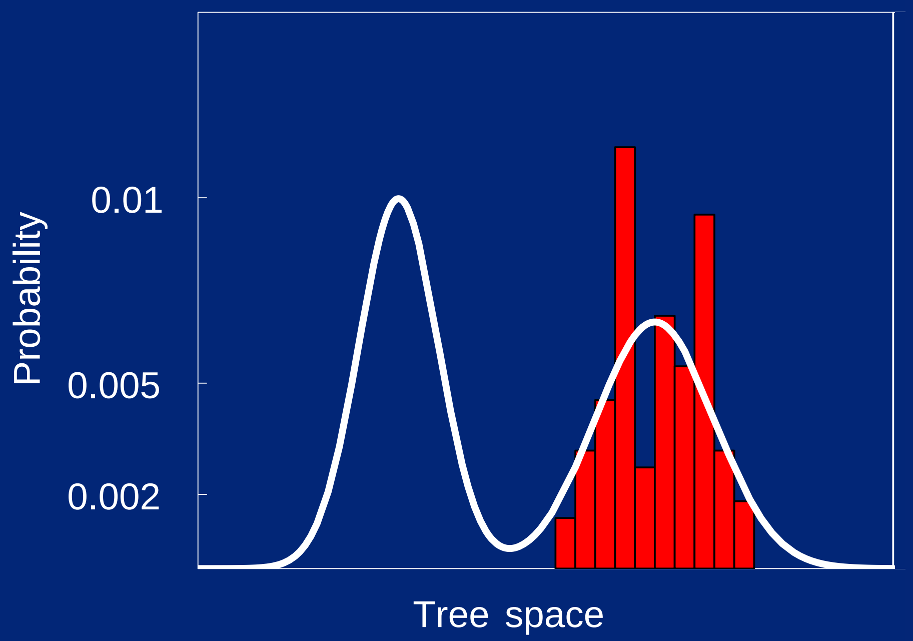


4. go to 1

# MCMC walk result



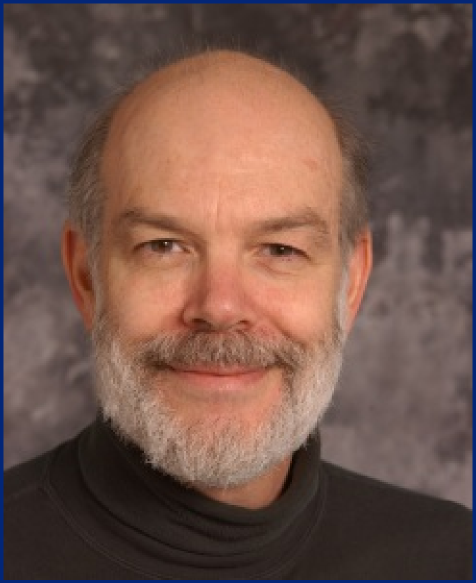
# MCMC walk result



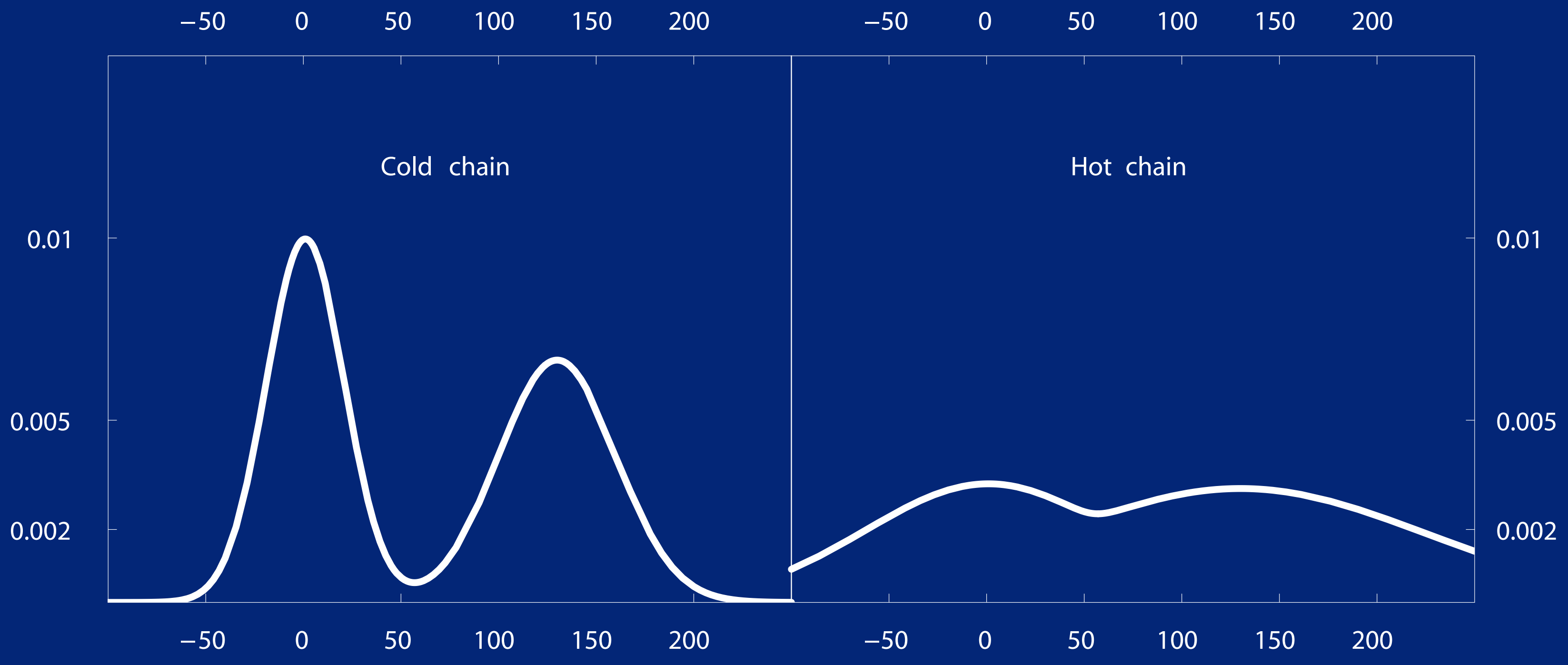
# Improving our MCMC walker: MCMCMC or MC$^3$

Metropolis Coupled Markov chain Monte Carlo

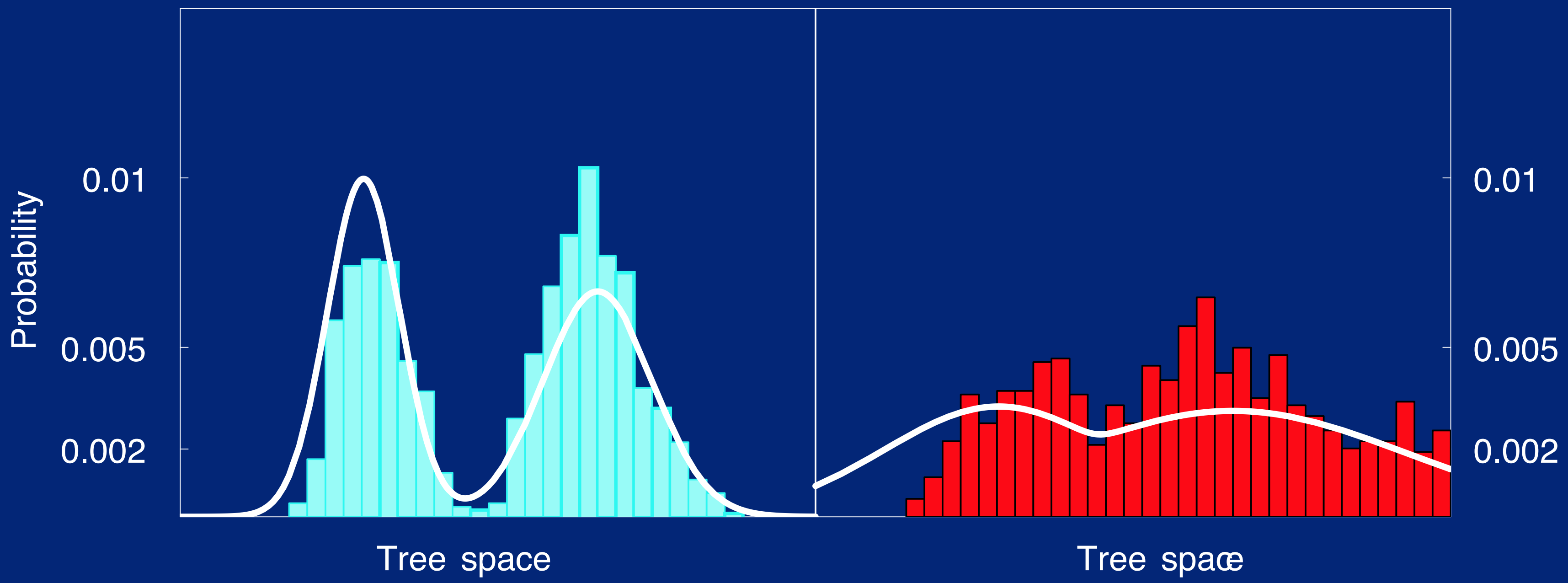
- The “hot” searches see a flattened landscape
- Only the “cold” search is used to make the estimate
- Good solutions found by a “hot” search can be imported into the “cold” search



# Improving our MCMC walker: MCMCMC or MC$^3$



# better MCMC walk result



## Paul Lewis' McRobot

This program can be found at

http://lewis.eeb.uconn.edu/lewishhome/software.html

It carries out a Markov Chain Monte Carlo search on a simple surface.

## McRobot Experiment

- How well does the robot search:
  - A single hill?
  - Two hills close together?
  - Two hills far apart?
- Does heating help with the far-apart case?
- Try 2, 3 and 4 chains: which seems best?

# Preparation for Thursday sessions

- Data files for demonstration are not on Zip disk
- They can be downloaded from:
  - http://evolution.gs.washington.edu/lamarc/sisg-2011/demo/
- Please download these before the demonstration Thursday
- This will save time and pain during the demo. Thank you!