

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, electrophoretic alleles
- Gene mapping, haplotype inference

LAMARC

- Strengths:

- Recombination
- Data with unknown haplotype phase
- Combining dissimilar loci

- Weaknesses:

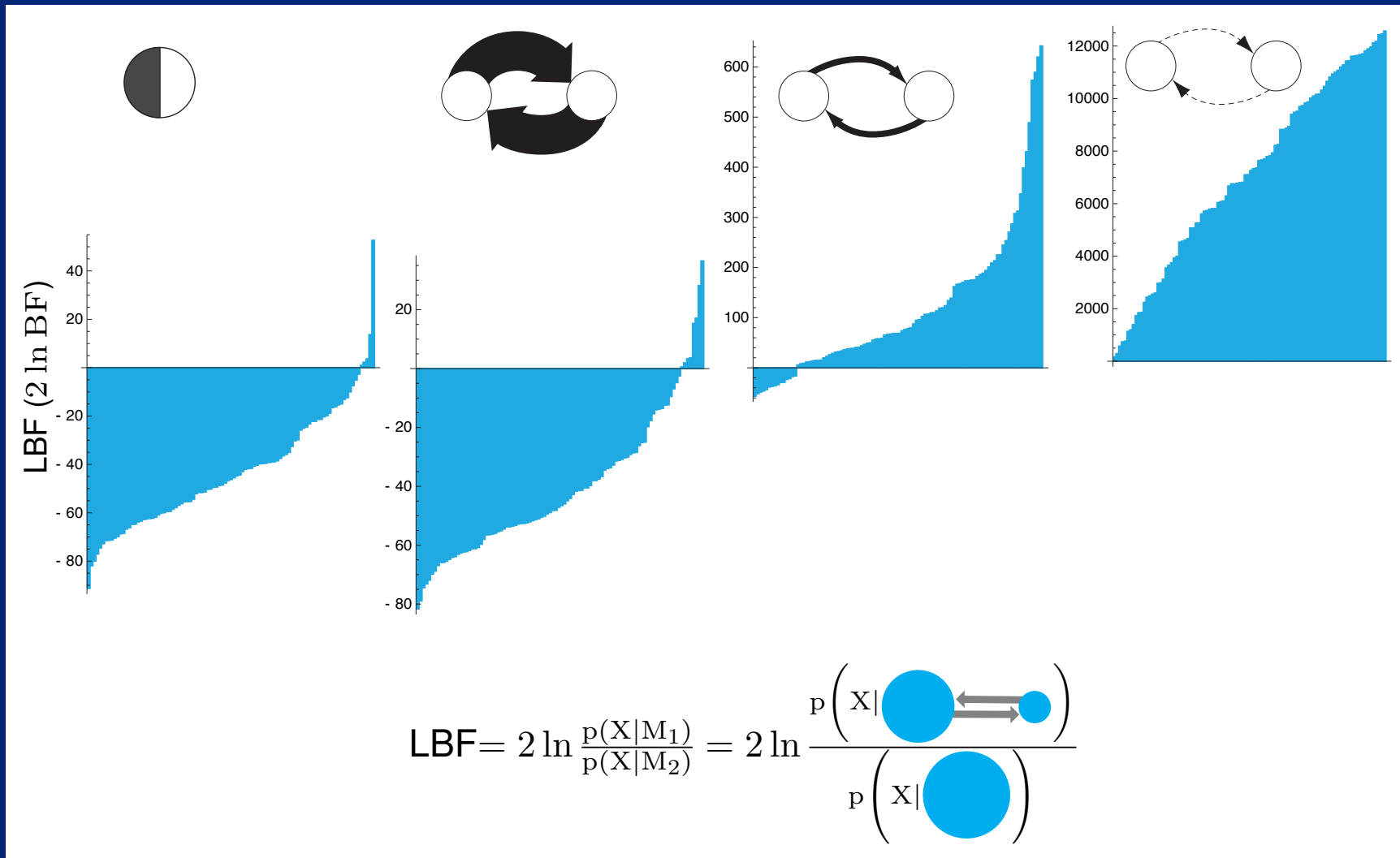
- Assumes stable population structure (divergence coming soon!)
- Single time point data only
- Exponential growth only

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, electrophoretic alleles
- Multiple time points

Bayes factor tests of models



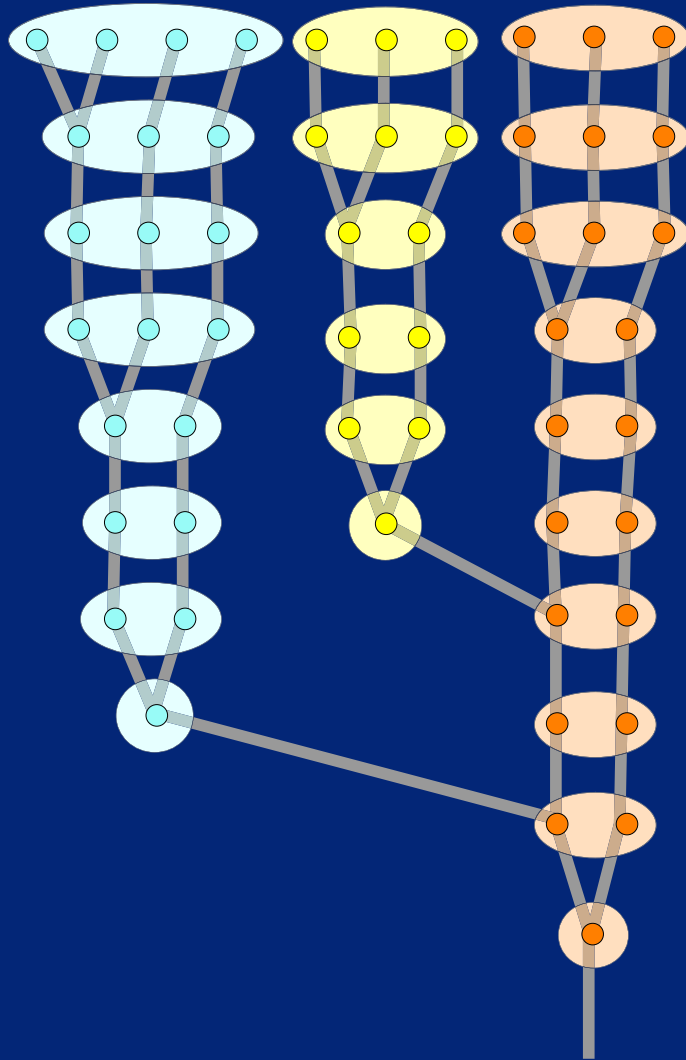
MIGRATE-N

- 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

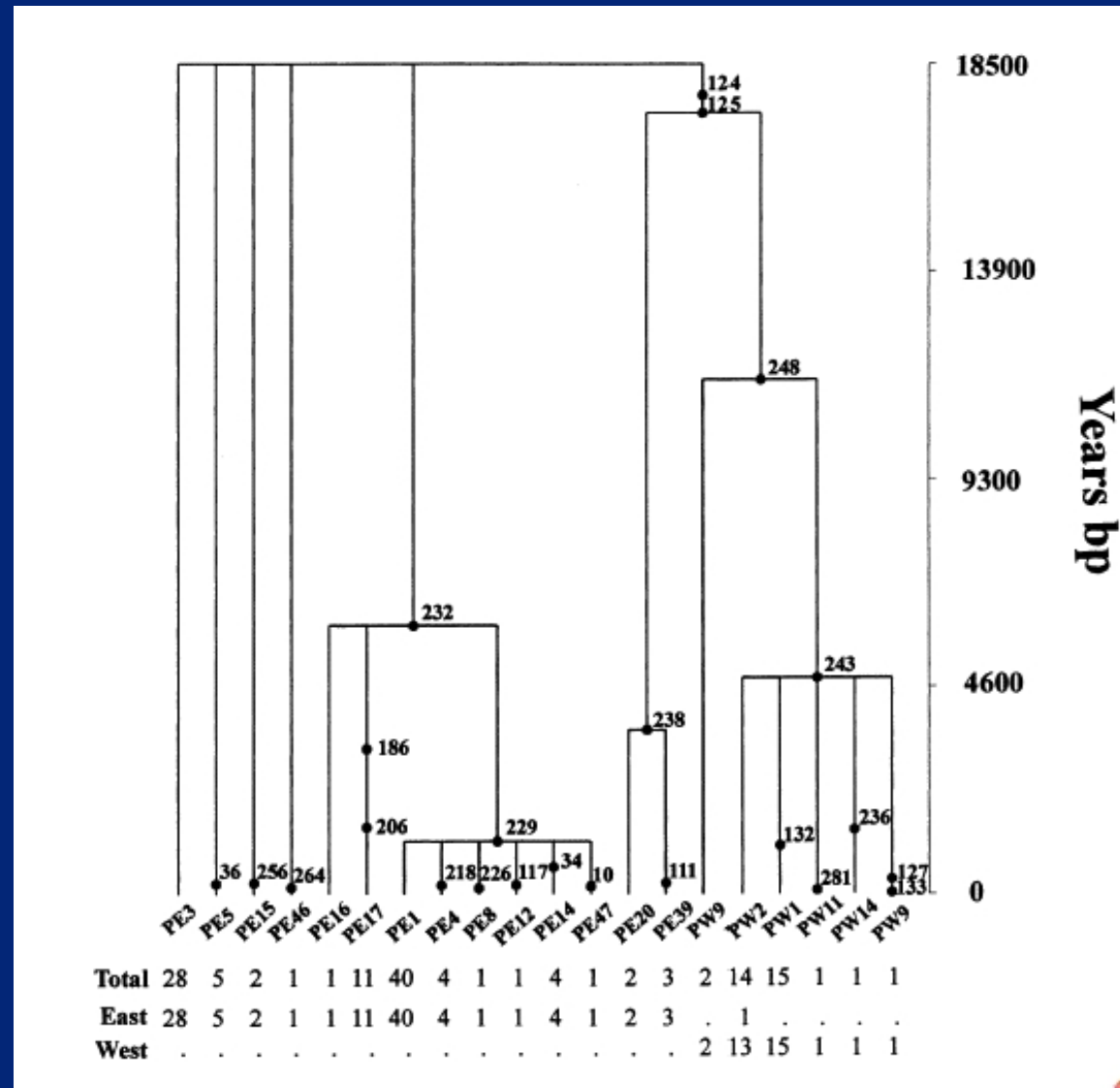
(<http://www.stats.ox.ac.uk/~griff/software.html>)



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

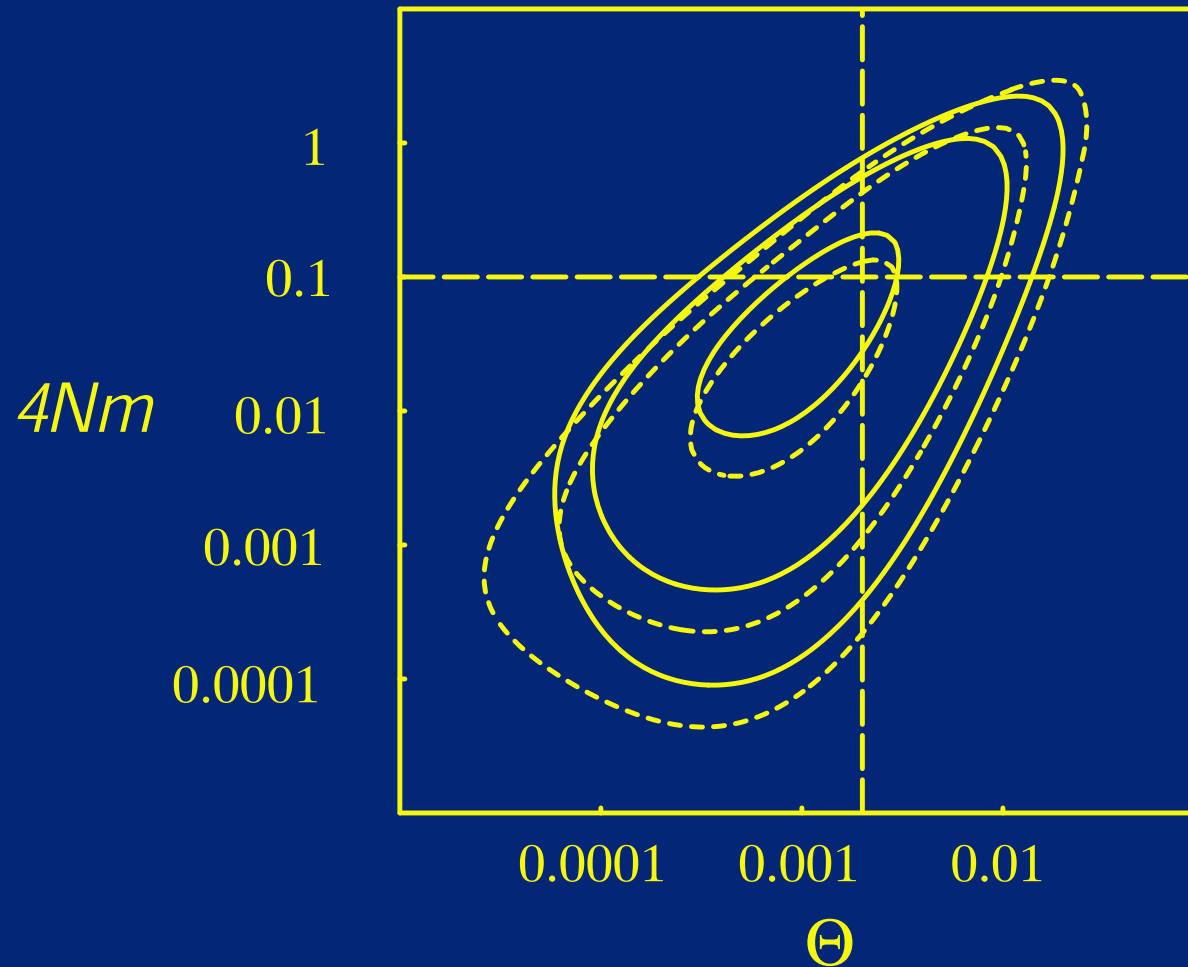
Dating mutations events using *Genetree*

Milot et al. (2000)



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*)

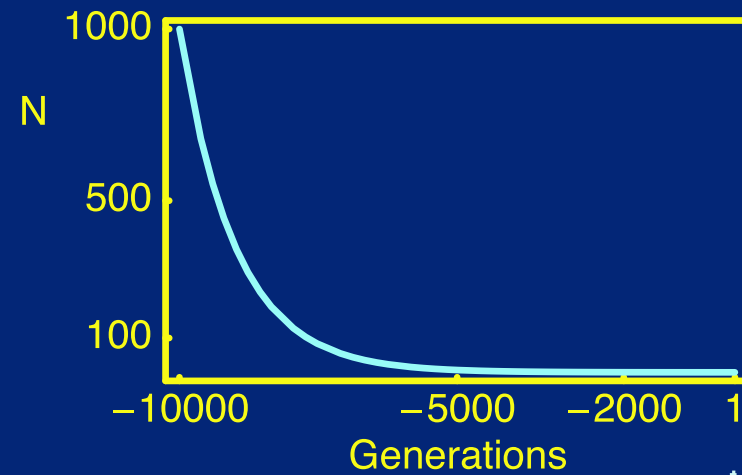
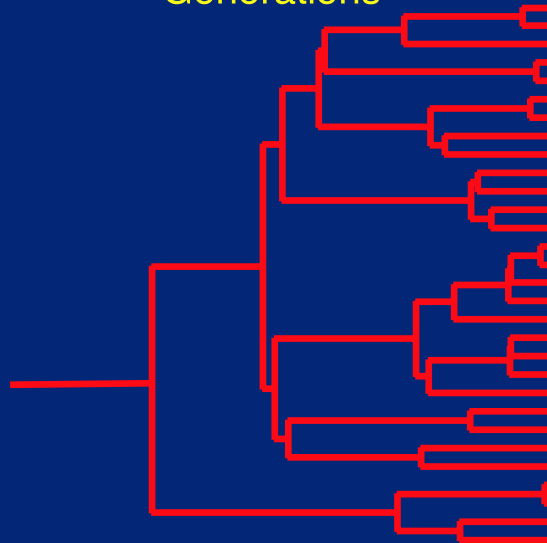
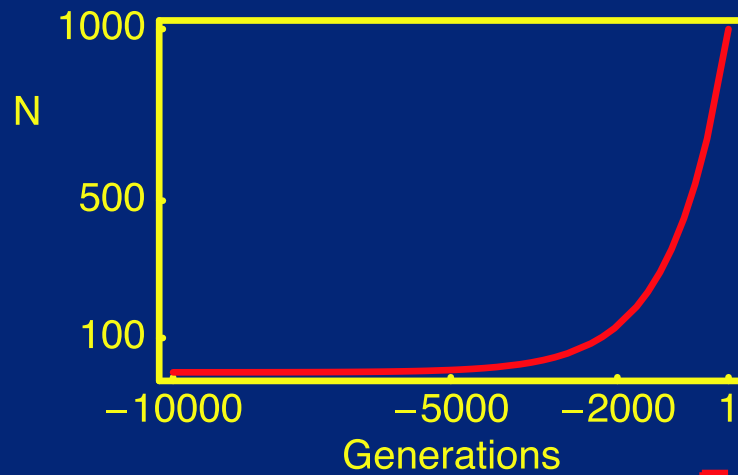
- With one time point, we estimate $\Theta = 4N_e\mu$ in diploids
- The number estimated is $2N_e\mu$ in haploids or $N_e\mu$ in mtDNA
- Two ways to separate N_e and μ :
 - Dated historical data (ancient DNA, etc.)
 - External estimate of mutation rate
- For most organisms, N_e is less than N
- Demographic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

Variable population size

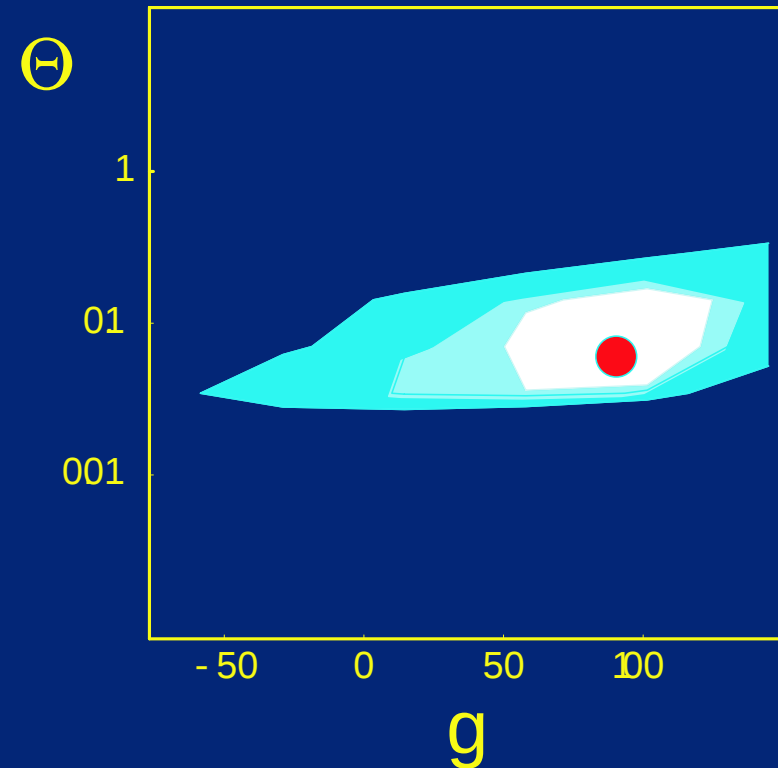
- In a small population lineages coalesce quickly
- In a large population lineages coalesce slowly

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

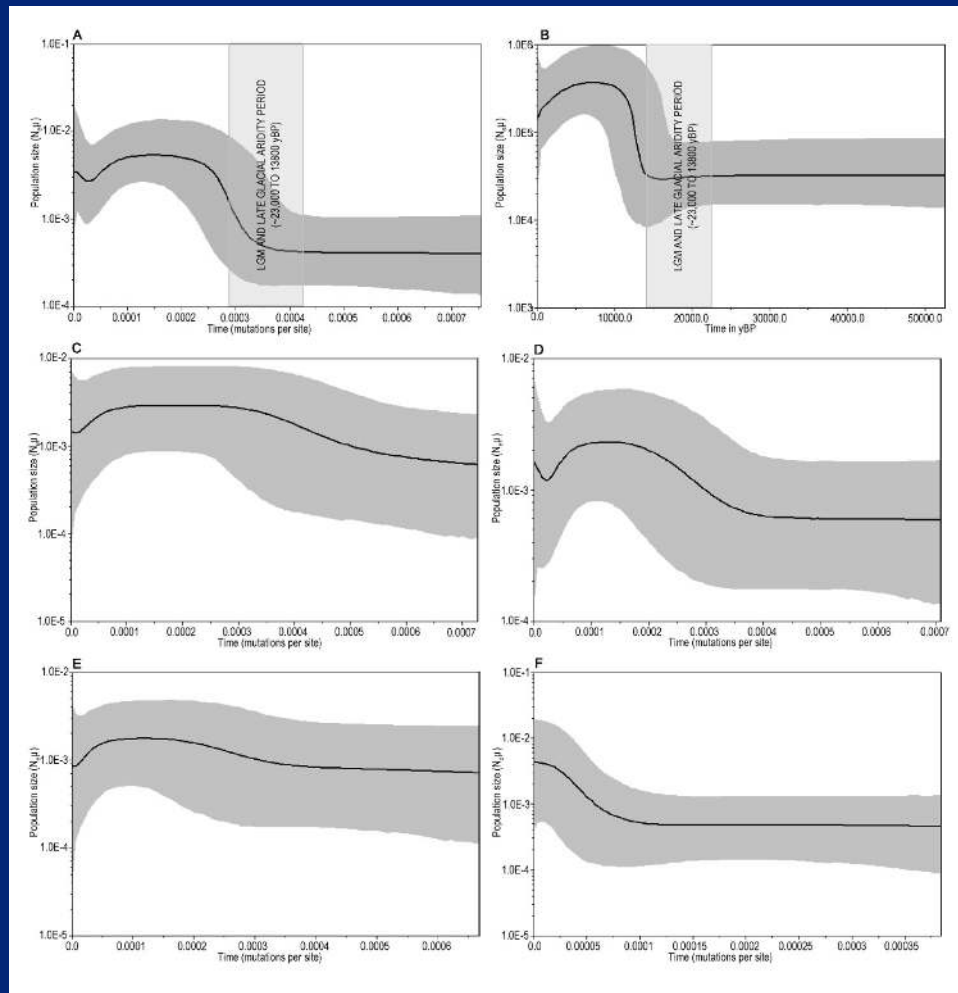


Grow a frog



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

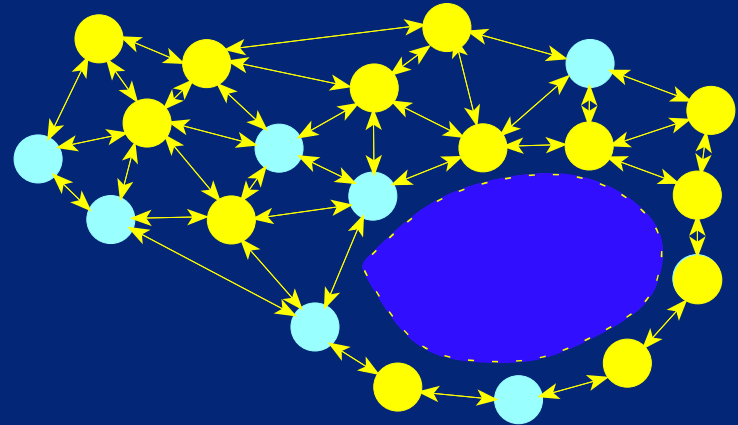
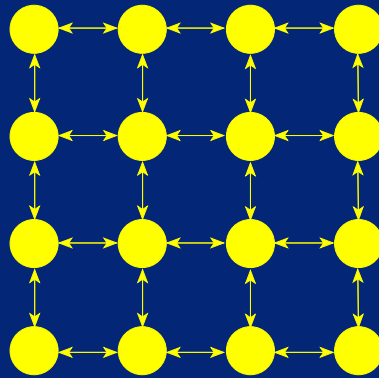
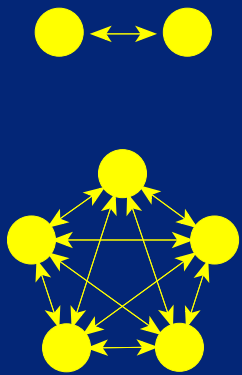
Bayesian skyline plots



Growth estimation software

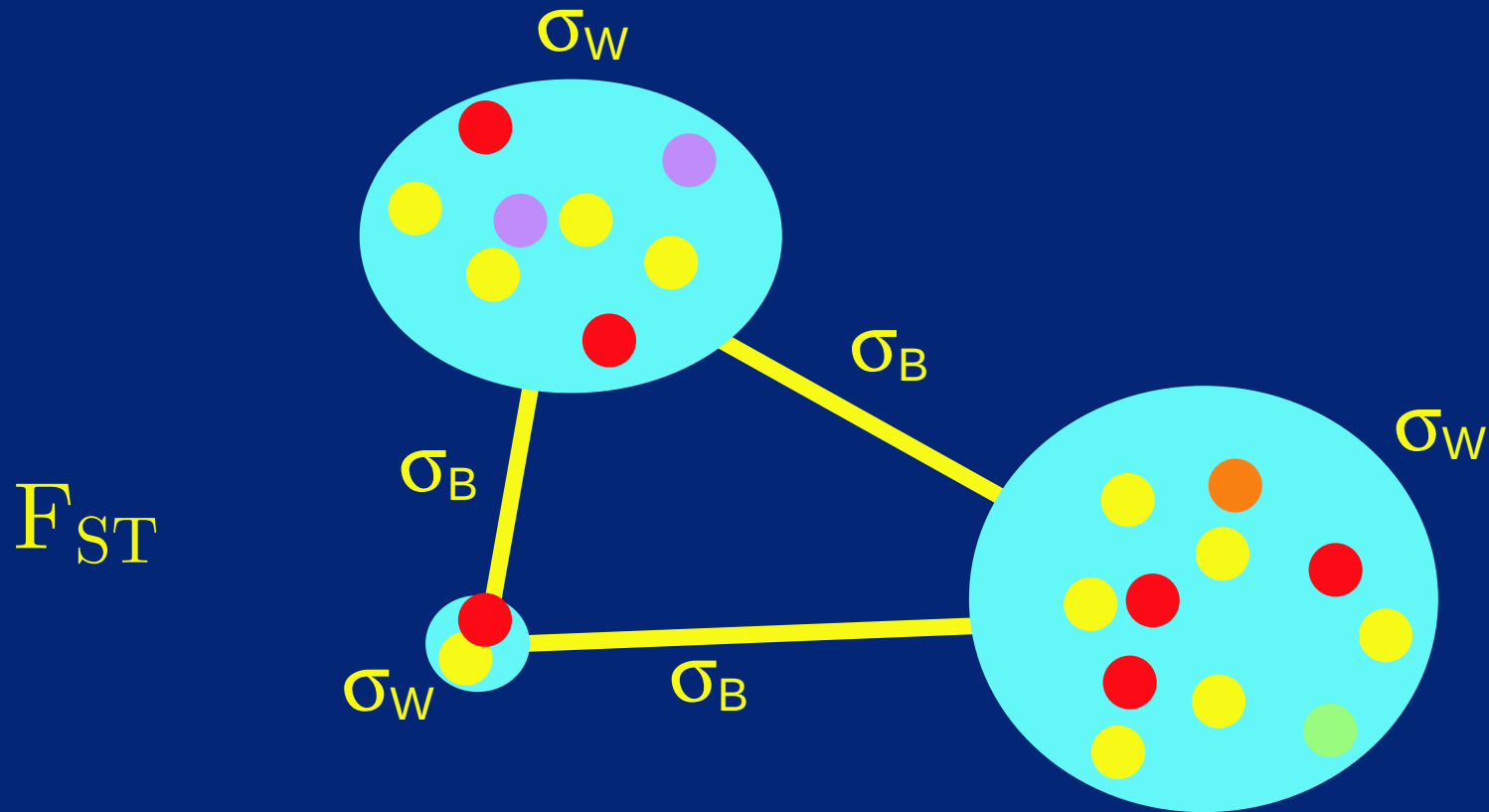
- Currently done with *Lamarc* or *Beast*
- Statistically weaker than estimation of Θ :
 - Biased upwards with one locus/one timepoint
 - Reasonable results with multiple unlinked loci
 - Even better results with multiple timepoints
- *Lamarc* assumes exponential growth/shrinkage
- *Beast* has a generalized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

Gene flow

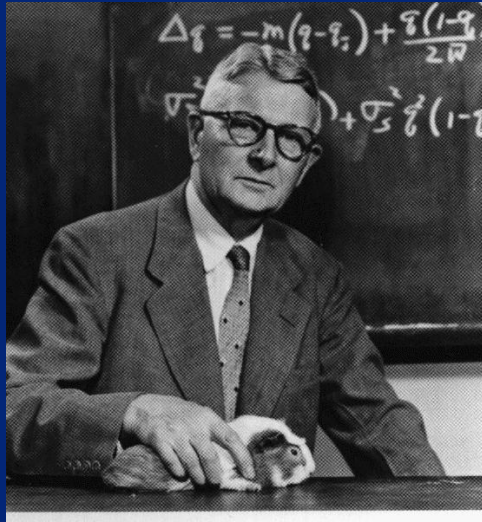


$$p(G|\Theta, \mathbf{M}) = \prod_{u_j} \left(\prod_i^{\text{pop.}} g(\Theta_i, \mathbf{M}_{.i}) \right) \begin{cases} \frac{2}{\Theta} & \text{if event is a coalescence,} \\ M_{ji} & \text{if event is a migration from } j \text{ to } i. \end{cases}$$

Gene flow: What researchers used (and still use)



What researchers used (and still use)



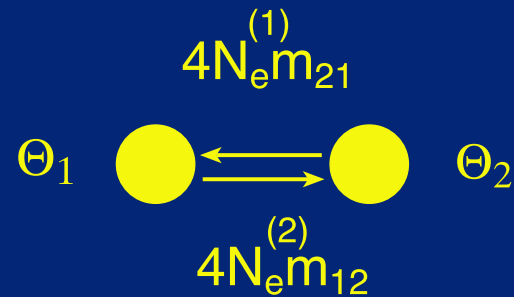
Sewall Wright showed that

$$F_{ST} = \frac{1}{1 + 4Nm}$$

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
0.01		1.14±0.77
0.01		7.80±22.20
0.05		11.46±18.54

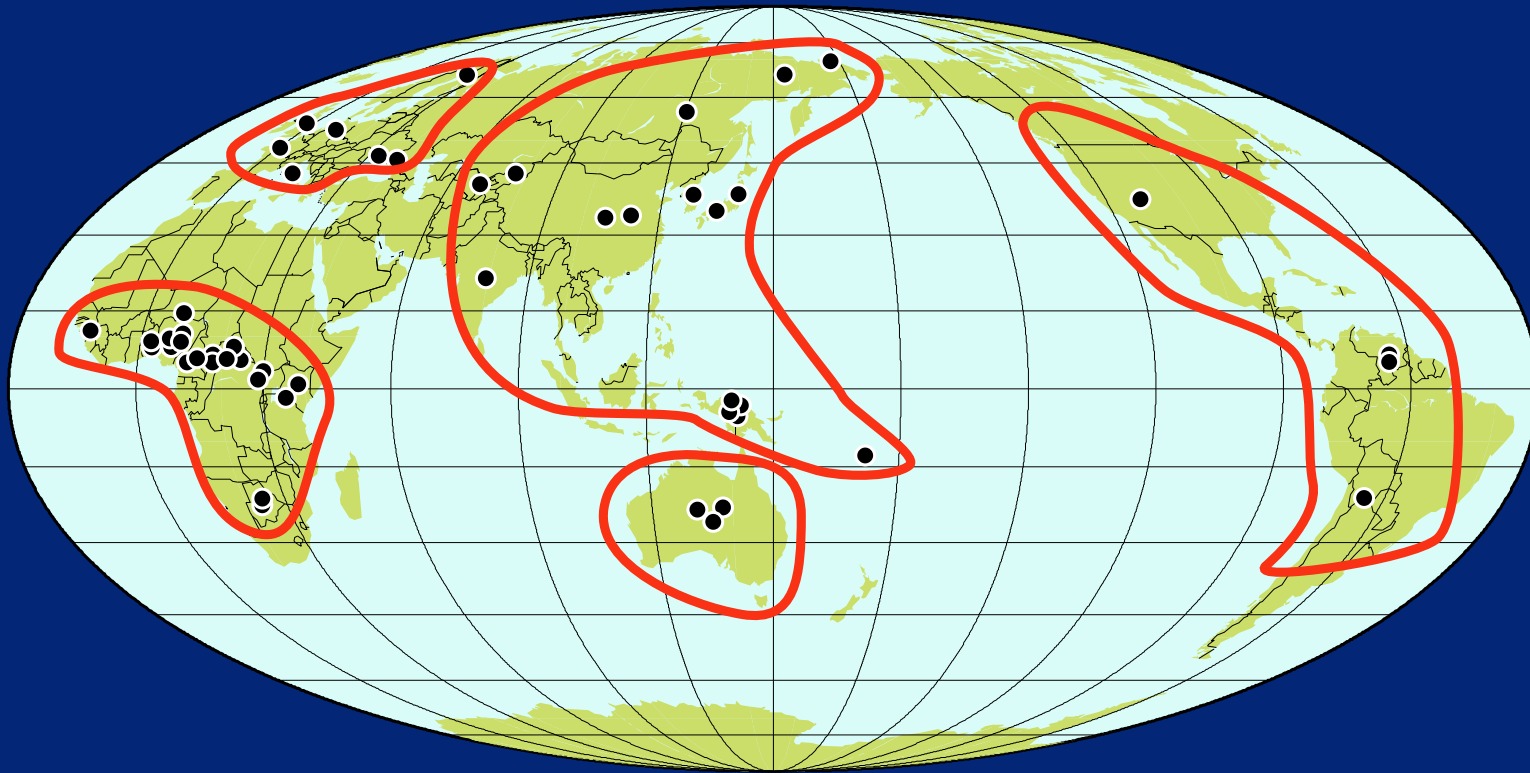
Maximum Likelihood method to estimate gene flow parameters

(Beerli and Felsenstein 1999)

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

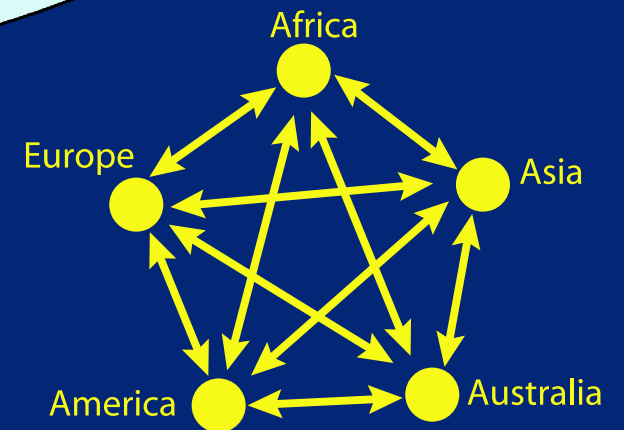
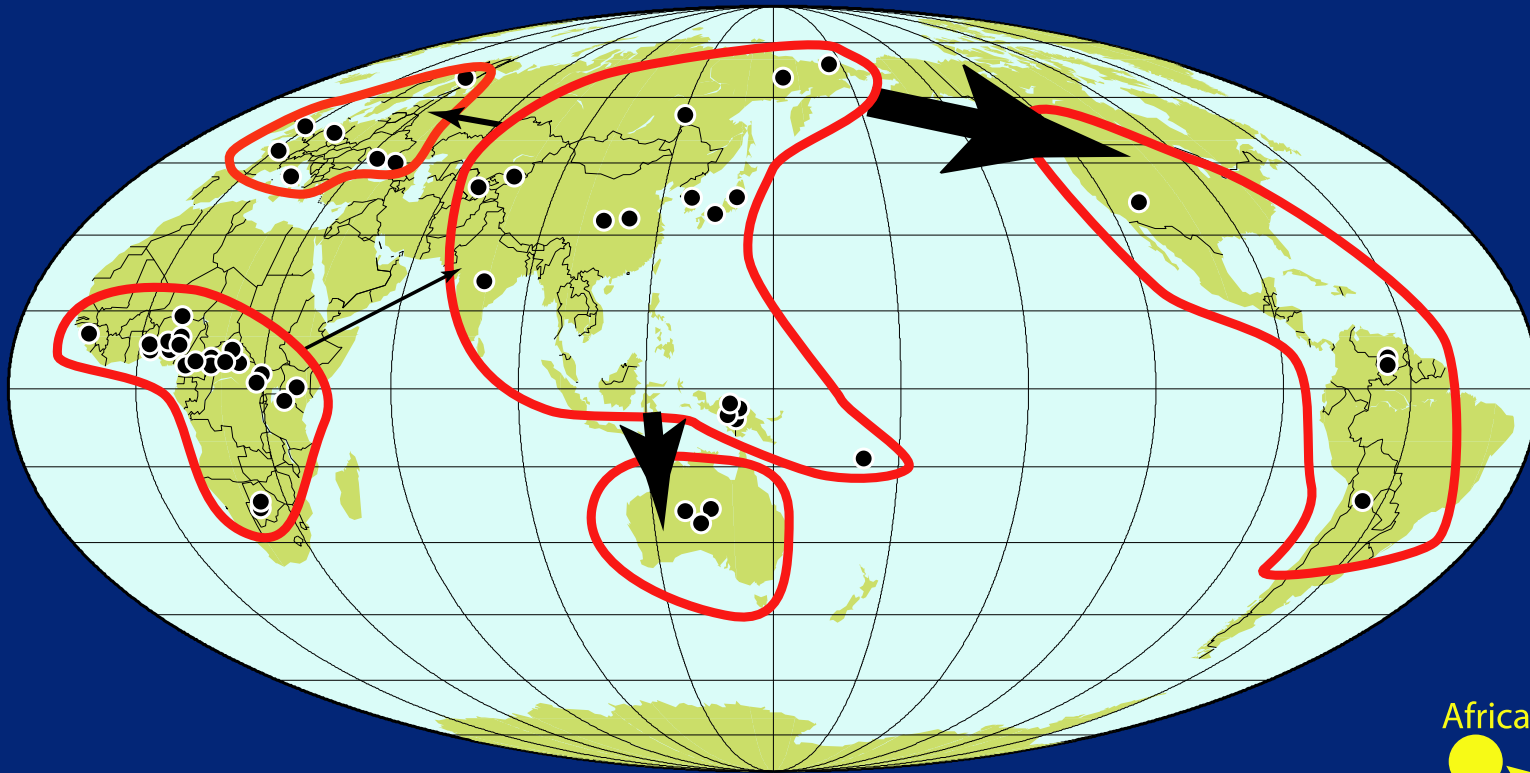
	Population 1		Population 2	
	Θ	$4N_e^{(1)}m_1$	Θ	$4N_e^{(2)}m_2$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

Complete mtDNA from 5 human “populations”

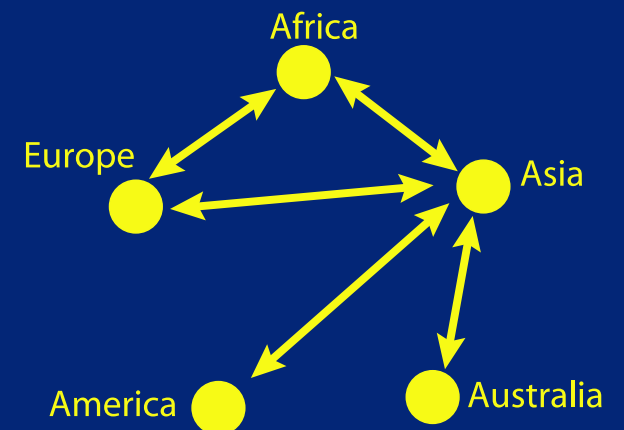
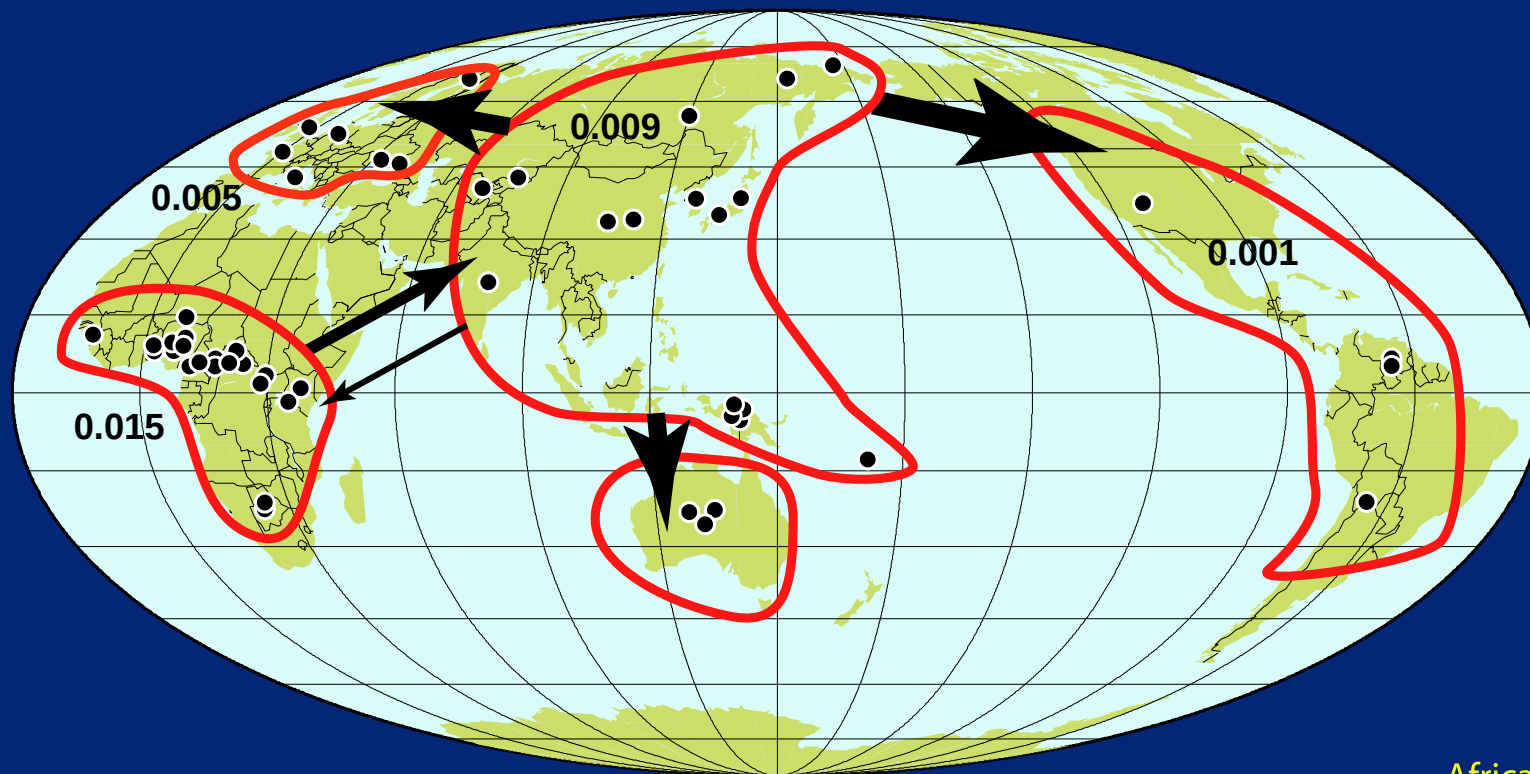


A total of 53 complete mtDNA sequences (~ 16 kb):
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.
Assumed mutation model: F84+ Γ

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

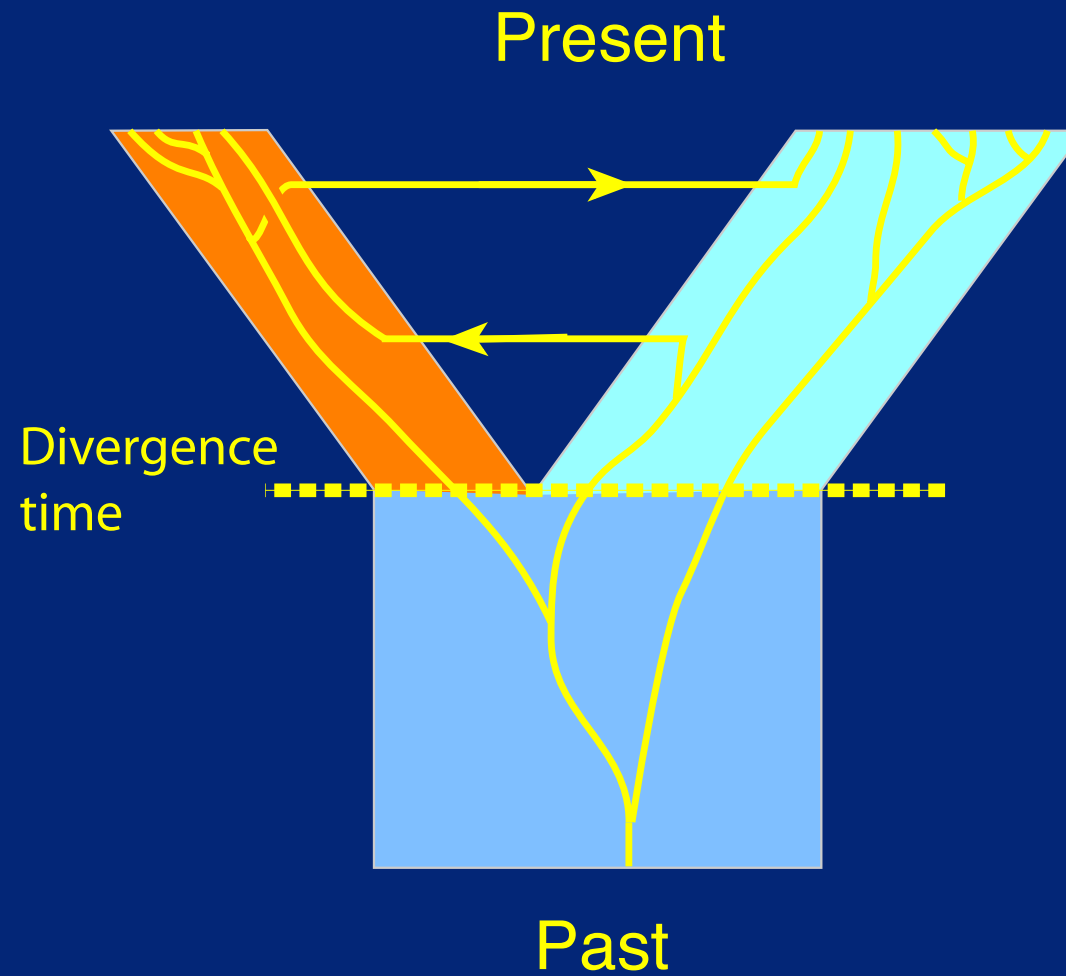


Coalescent recombination estimators

- 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

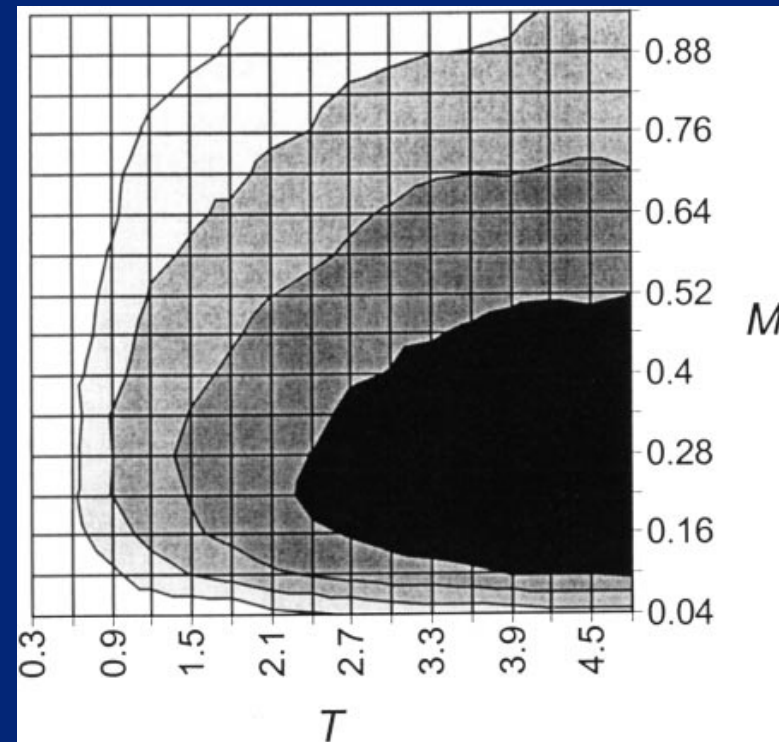
Estimation of divergence time

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.



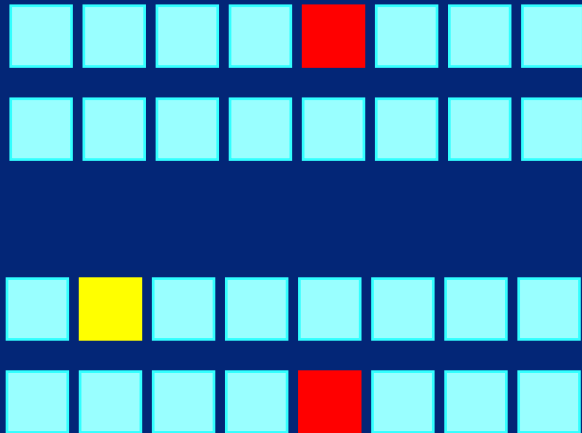
Coalescent divergence estimators

- 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

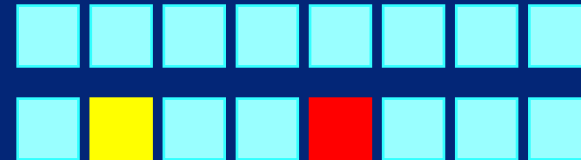
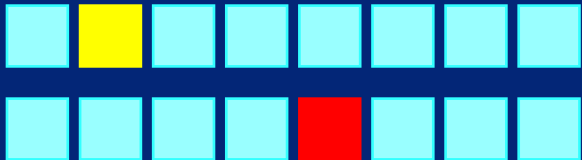
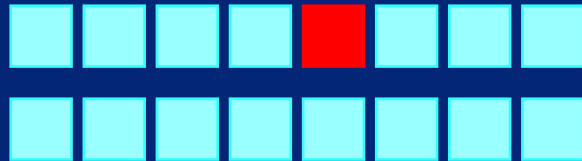
Multiple time points

- 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.

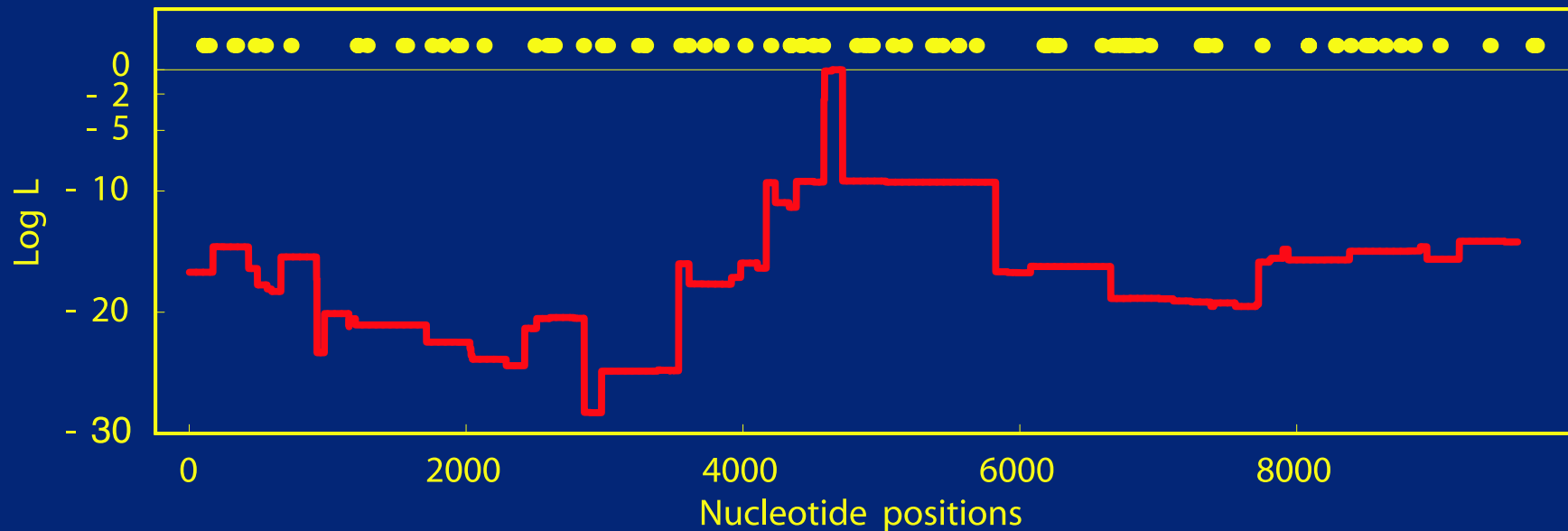
Currently only *Lamarc* can do the latter.

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 were used to estimate r and find the location of the disease mutation on the DNA.



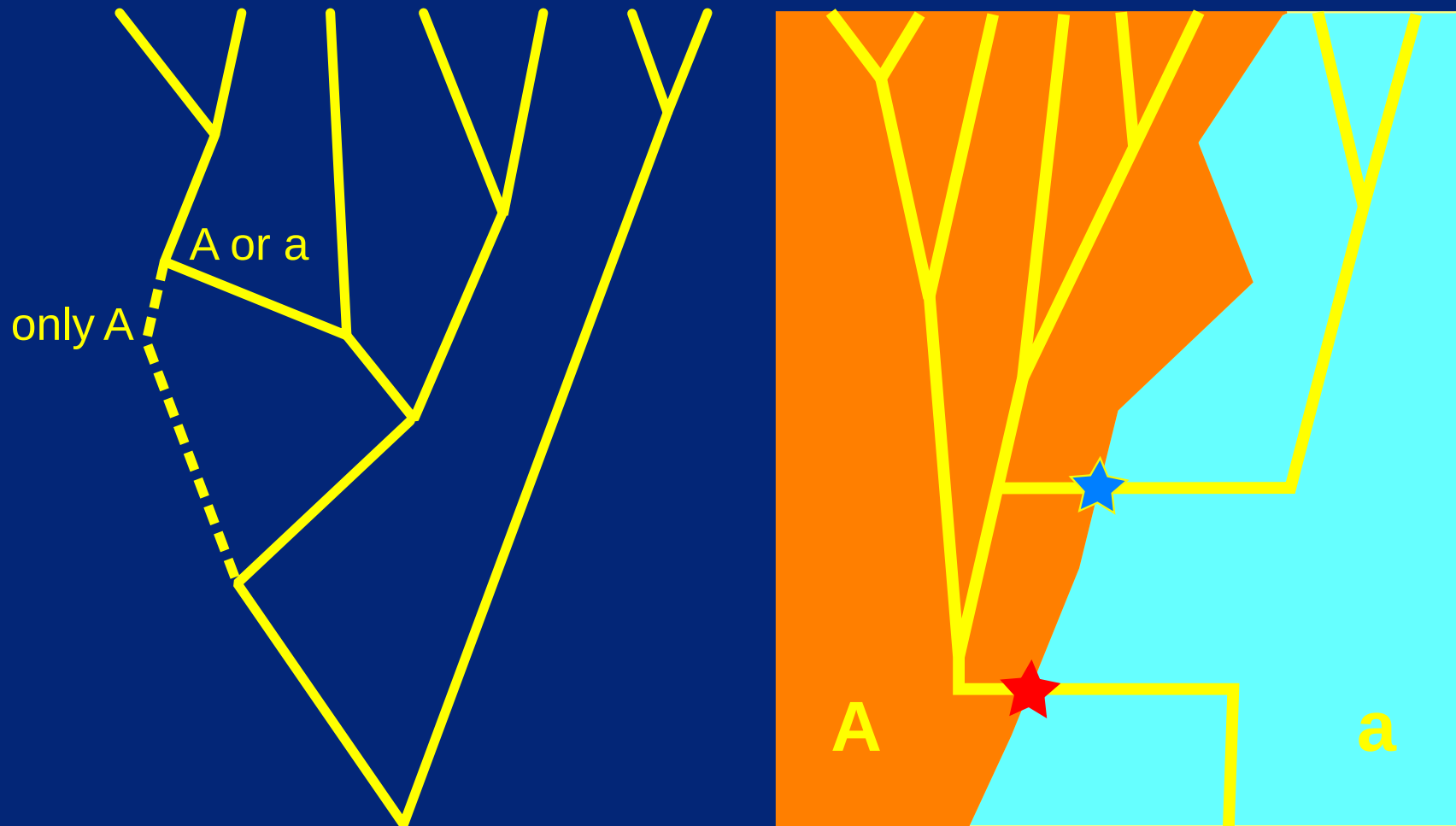
Linkage disequilibrium mapping

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

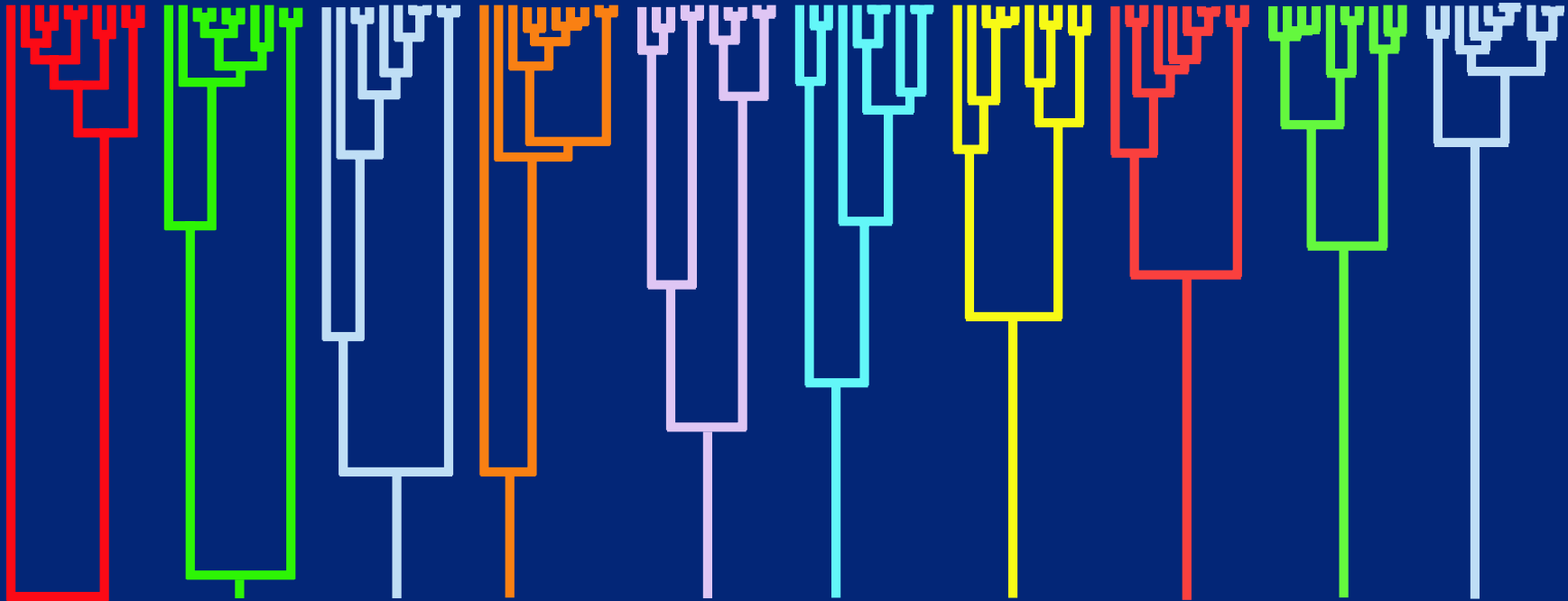
- Introduction to coalescent theory
- Genealogy samplers
- Survey of samplers
- Evolutionary forces
- **Practical considerations**

Information content of the coalescent

What can best give us more information?

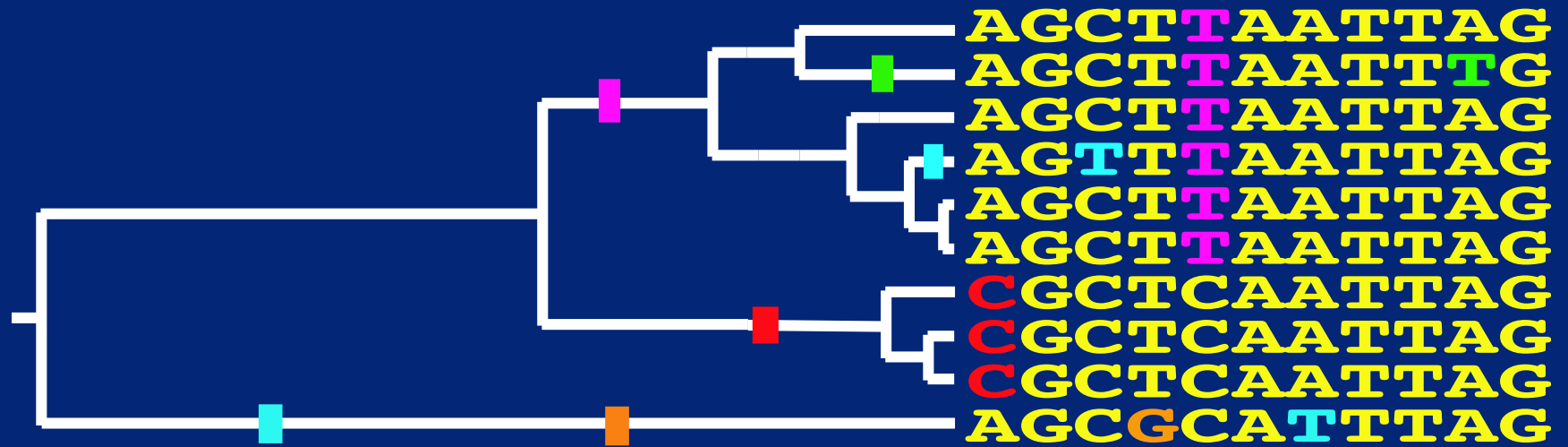
- More individuals?
- More base pairs?
- More loci?

Variability of the coalescent

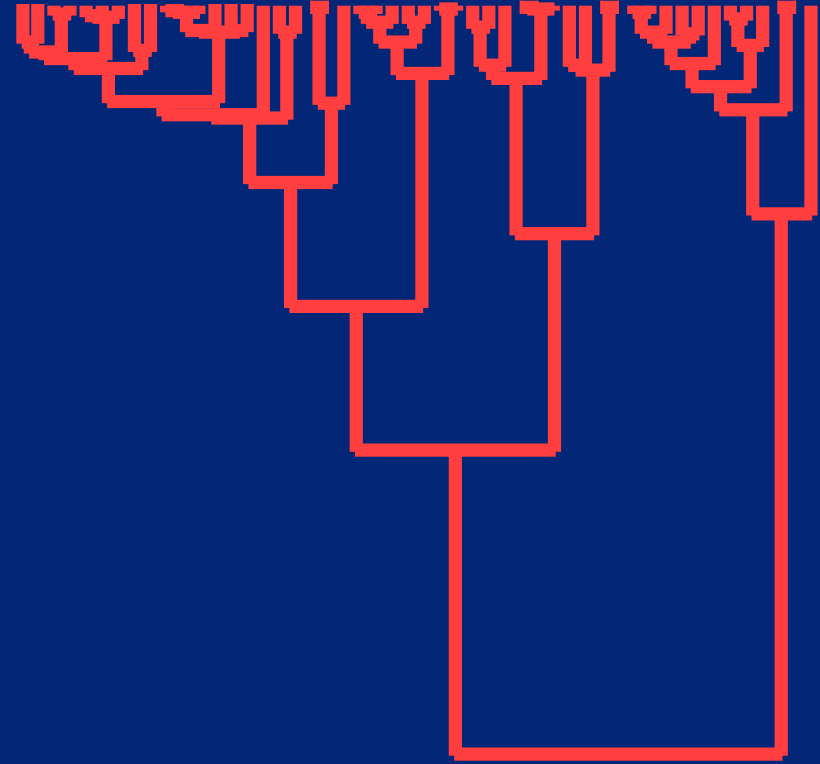
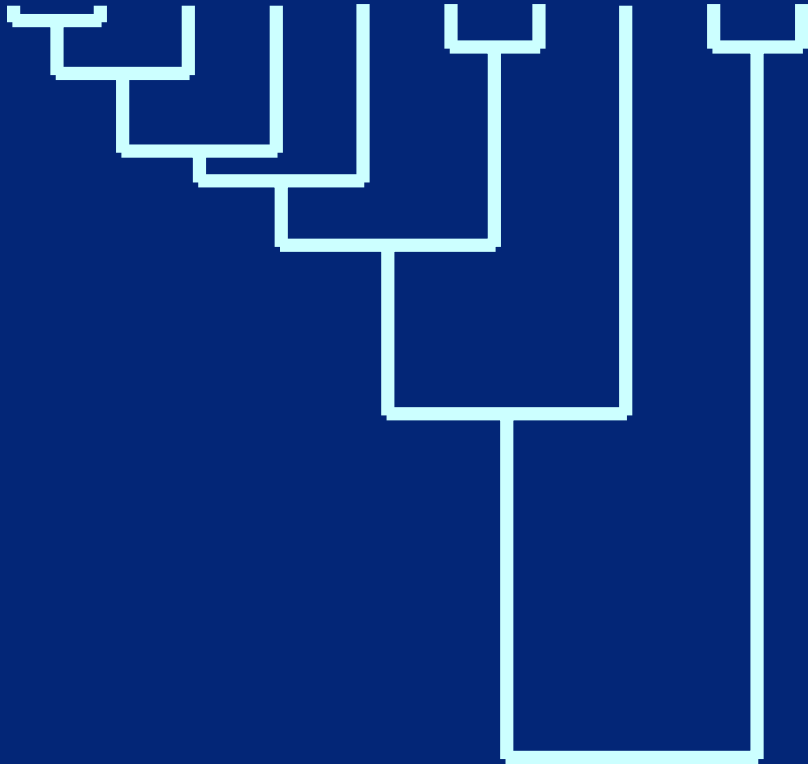


10 coalescent trees generated with the same population size, $N = 10,000$

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. (In a Bayesian run the same problem appears as a trace that is still trending up or down at the end of the run.)

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.

Posterior looks like prior

- Posterior should be prior $\times$ effect of data
- If posterior resembles prior, data are not contributing much!
- This can mean:
 - Not enough data (especially, not enough loci)
 - Non-identifiable parameters (for example, population size of a very young population)
 - Inappropriate prior (much too narrow, much too broad, not containing truth)
- Do not ignore this problem!

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.

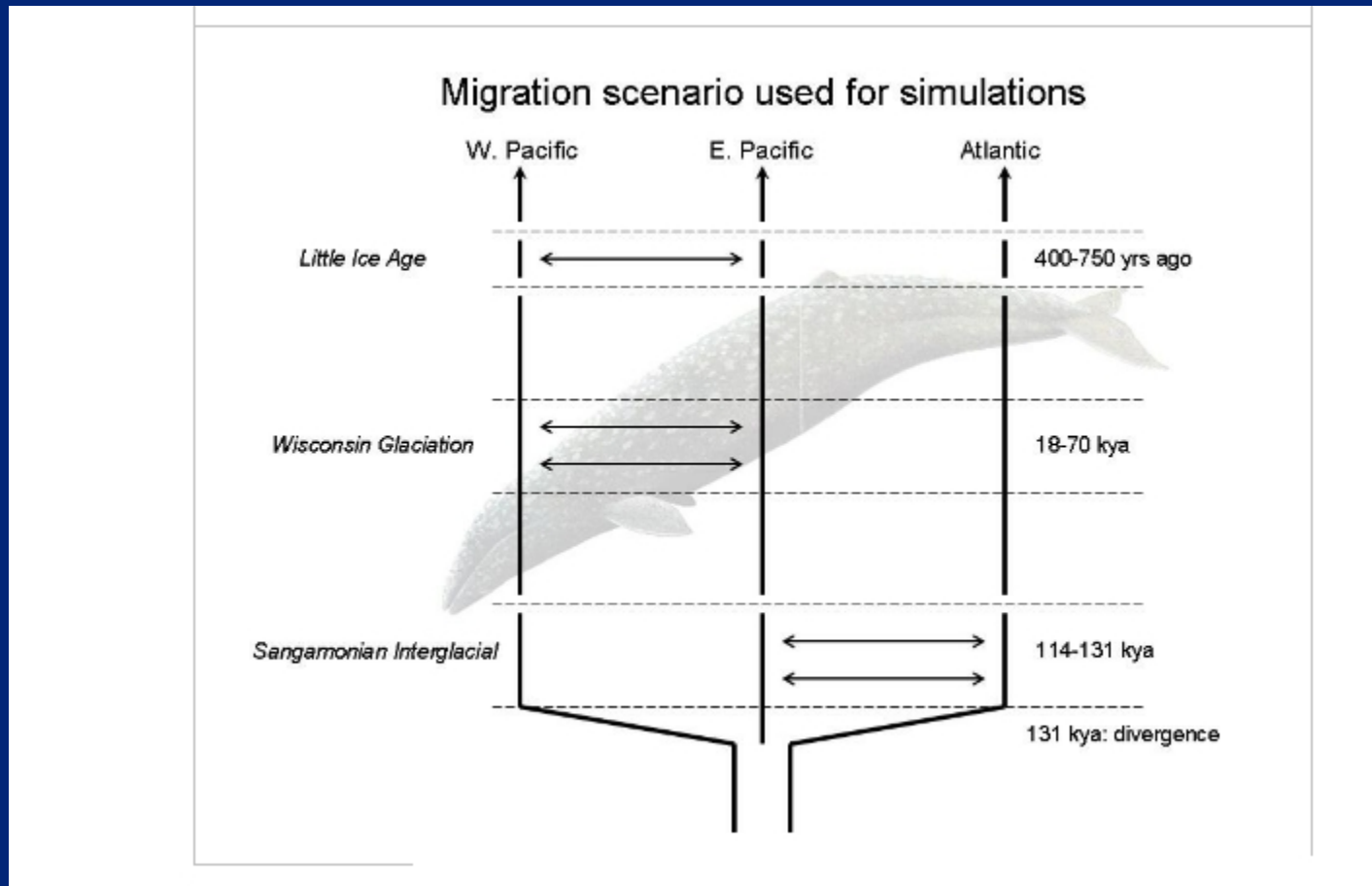
What was the long-term population size of gray whales?

- 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