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–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

Variants and extension of the coalescent

- Population growth/shrinkage over time
- Migration between populations
- Population divergence
- Mutation rate variation
- Times of significant events
- Recombination (Philip Thursday morning)
- Selection (Philip Friday morning)

Outline

- What kind of information is available about this evolutionary force or process?
- How is it usually parameterized?
- What algorithms or programs deal with it?
- What special issues arise with this force?

Variable population size

- During times when a population is large, lineages coalesce slowly
- During times when a population is small, lineages coalesce quickly

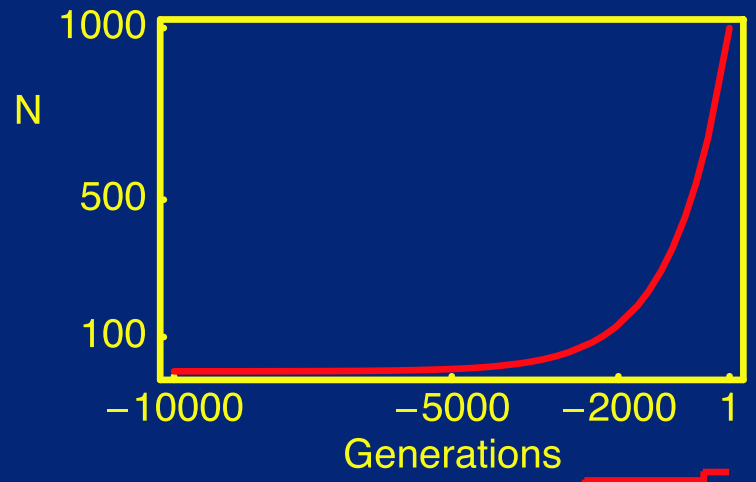
This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the population size Θ.

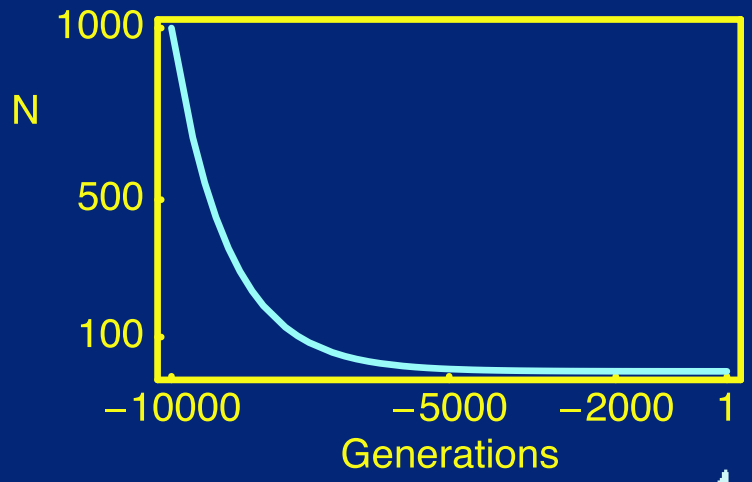
Parameterization of growth

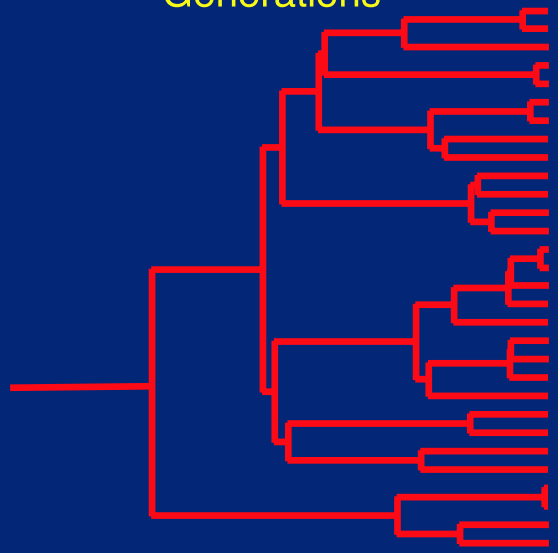
- No growth – Θ is constant
- Exponential growth model – growth rate g
- Logistic growth model – growth rate r, carrying capacity K
- Stepwise growth model – step time, Θ before and after
- Free growth model – a series of steps

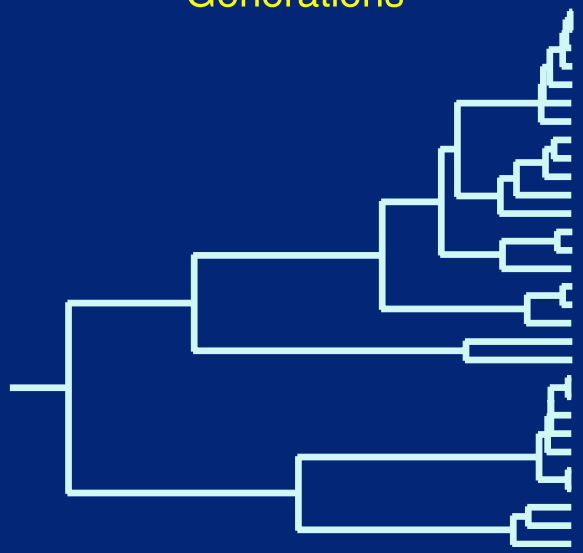
Not easy to tell the models apart!

Exponential population size expansion or shrinkage

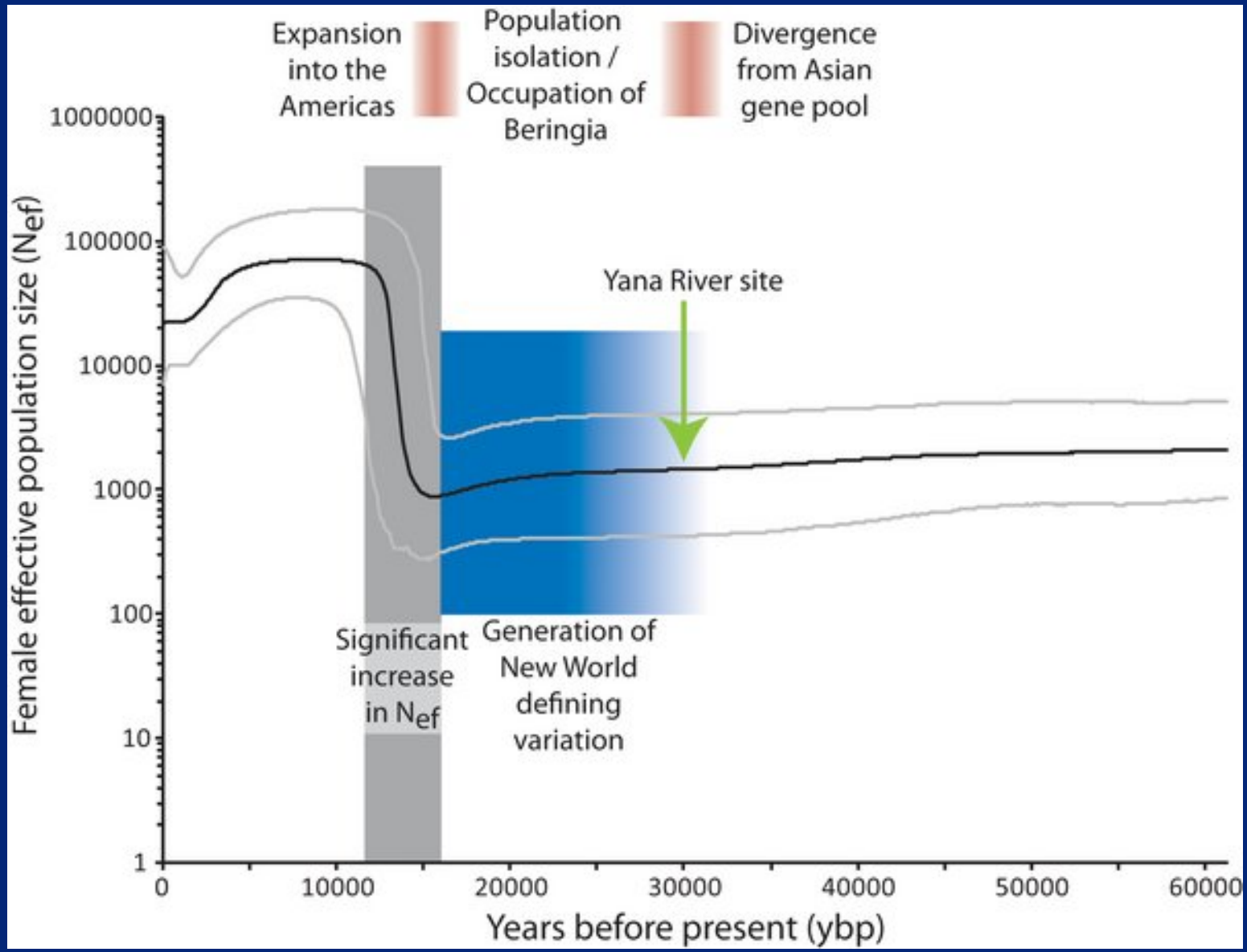




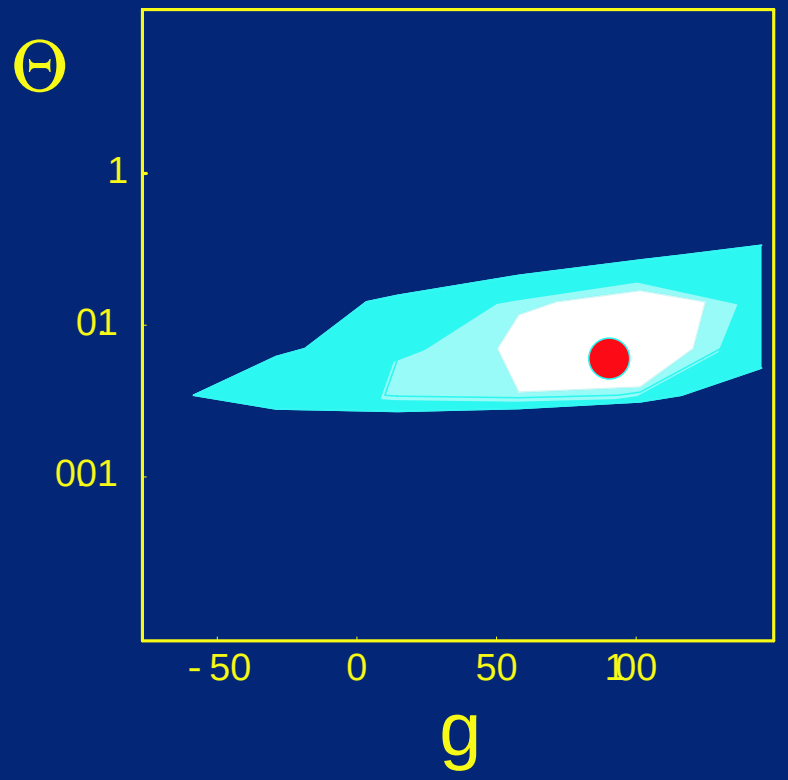




Bayesian skyline plot: Heled and Drummond 2008



Genealogy-sampler inference of exponential growth



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

What does g mean?

The parameter g often causes confusion.

$$\Theta_{t\mu} = \Theta_{now} e^{gt\mu}$$

To interpret g we need an external estimate of the mutation rate μ. Given that, we can ask how large the population was a given number of generations or years (depending on the units of our mutation rate estimate) in the past, using this equation.

Positive g is a growing population, negative g is a shrinking one.

Population growth: non-sampler approaches

- Summary statistics based on:
  - Mismatch distribution
  - Between-locus variability
  - Allele size distribution (microsatellites)
  - Imbalance between variance and heterozygosity
- Nested clade analysis
- Skyline plots
- Approximate Bayesian Computation bottleneck detection

Population growth: sampler approaches

- Exponential growth: LAMARC, IM, BEAST
- Growth estimation is biased upwards with one locus
  - Confidence intervals are more reliable than maxima
  - Multiple loci help a lot
  - Multiple time points are even better
- BEAST offers Bayesian skyline plots for more detail on growth

A cautionary tale

- Mismatch distribution–distribution of number of differences between haploptypes
- Theoretical distribution is exponential when no growth
- With growth, it has a peak
- Score all pairwise mismatches from human data–peak is seen
- This makes sense as human population has grown....

BUT!

A cautionary tale

- Simulate non-growing populations
- Distribution NEvER looks exponential
- (next slide from Slatkin and Hudson 1991)
- Why?

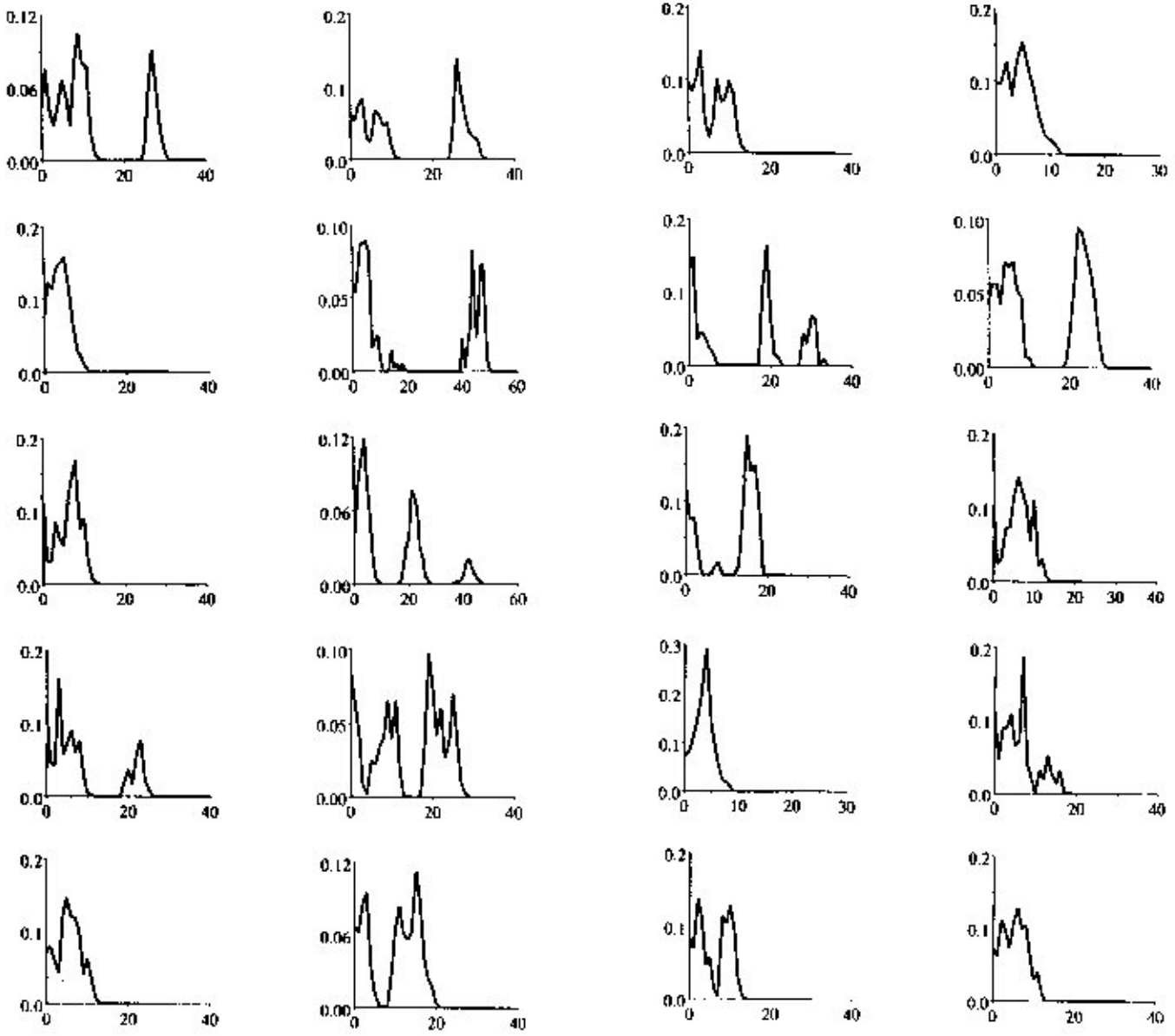
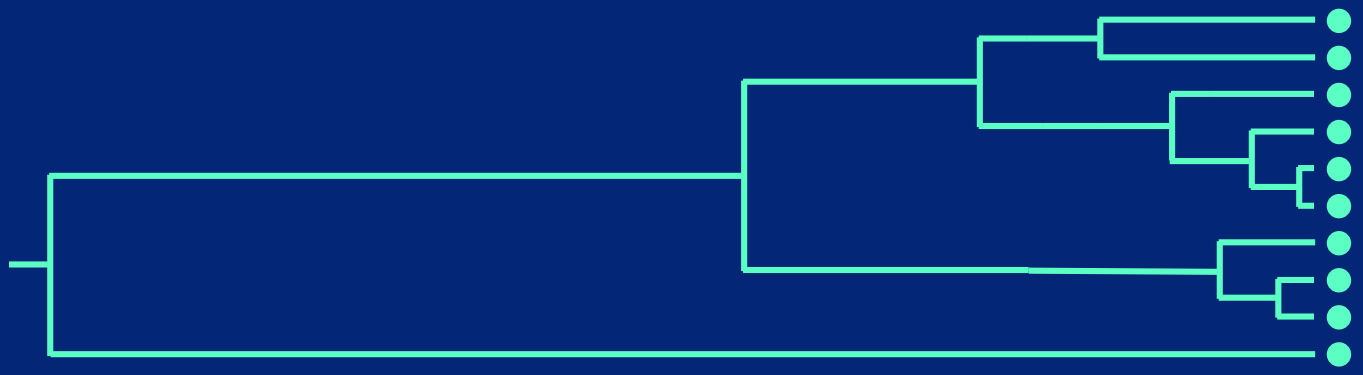


FIGURE 3. Frequency distributions of pairwise differences for 20 replicate simulations. The distributions of all 1225 pairs in samples of 50 genes from a panmictic population are plotted. The data were generated using a simulation program described in the text. In each graph, the abscissa is the number of sites at which two samples differ and the ordinate is the fraction of pairs that differ.

A cautionary tale

- Expected distribution of ONE pairwise mismatch is exponential
- Multiple draws from the same population are not independent
- The peaks come from deep coalescences in the tree

A cautionary tale



A cautionary tale

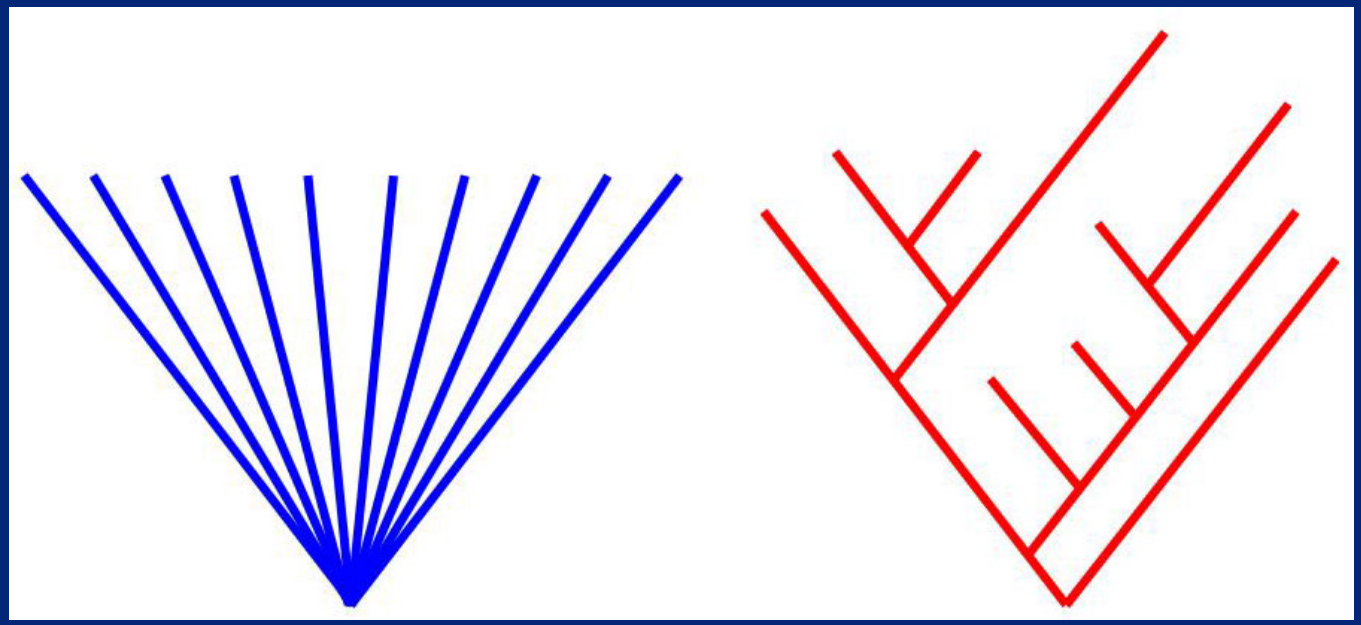
Are mismatch distributions still useful?

- One pair per locus would give "expected" distribution
- Raggedness of distribution can suggest lack of growth
- Not an efficient way to use the data

Bottom line: locus history is a TREE

General issues with growth estimation

- Low statistical power–multiple loci needed
- Too-fast growth turns the tree into a star
  - Star shows fast growth but can't pin down rate
  - Star-like data have little power for inferring other parameters



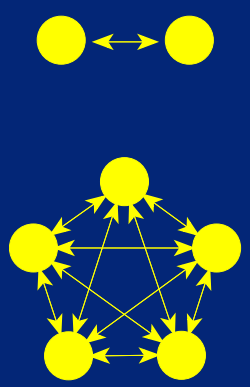
General issues with growth estimation

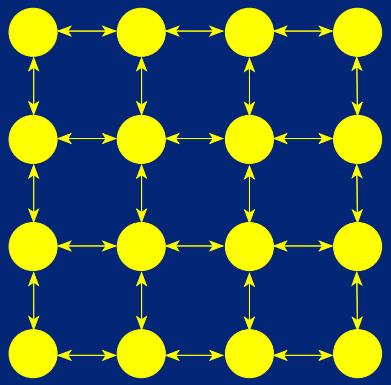
- If growth is fast, all growth models look similar
- Very recent growth has little effect on the coalescent, even if extreme
- Only most recent episode of growth likely to be visible
- Ancient DNA gives MUCH more power for growth inference than single time point samples

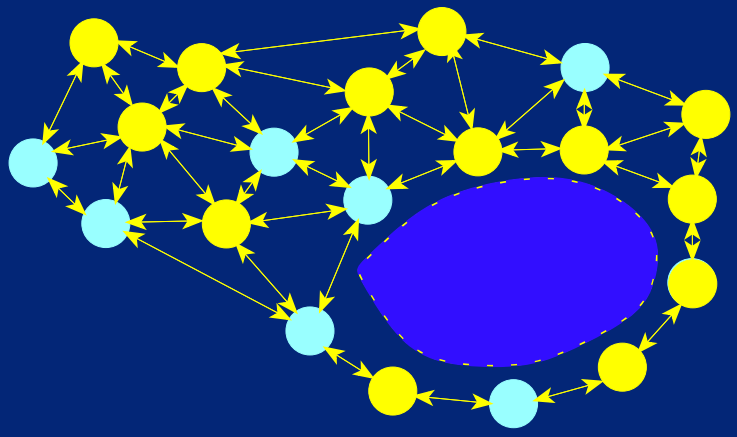
Specific issues with genealogy-sampler growth estimation

- Multiple unlinked loci not always available
  - LAMARC can compensate by using partially linked loci
  - BEAST can compensate by using multiple time points
- Growth rates so high that co-estimation of Θ and g not possible
  - If one parameter is held constant, the other can be estimated
  - Multiple time point samples in BEAST can resolve this problem
- Growth estimate contains unknown μ
  - Multiple time point samples in BEAST can resolve this problem
  - Mutation rate can be measured experimentally or inferred from phylogenetic data plus fossil dates

Migration among stable populations







Parameterization of migration rates

- m – chance that a lineage migrates each generation
- $M = \frac{m}{\mu}$ – scaled by mutation rate
- $4N_e m$ – scaled by population size (migrants per generation)

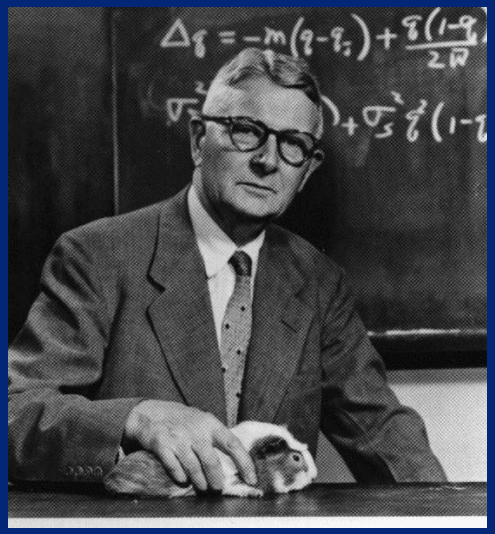
Non-sampler approaches

- F_{ST} summary statistics (contrast within-population and between-population variation)
- Haplotype sharing

Arlequin is a widely used program which carries out these (and many other) tests:

http://cmpg.unibe.ch/software/arlequin35/

FST in practice



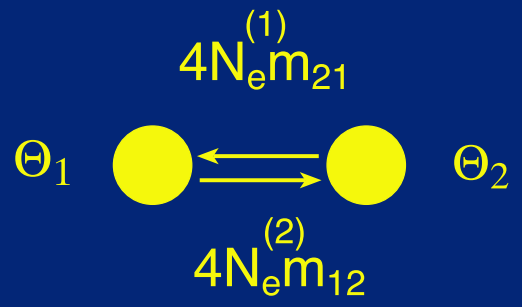
Sewall Wright showed that

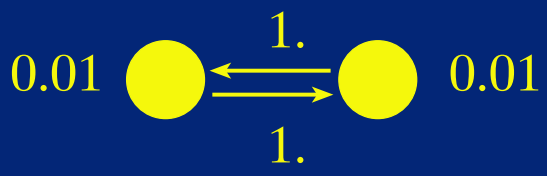
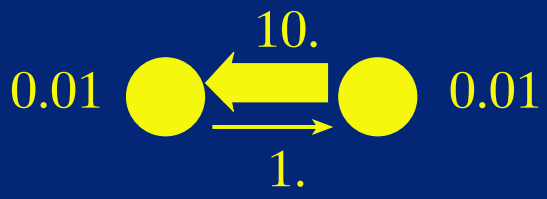
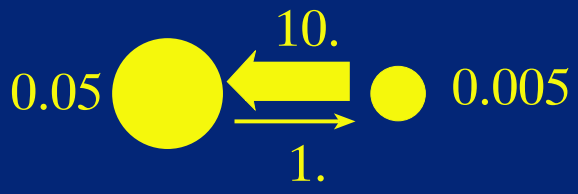
$$F_{ST} = \frac{1}{1 + 4Nm}$$

and that it assumes

- All migration rates are the same
- All subpopulation sizes are the same

FST in practice



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.

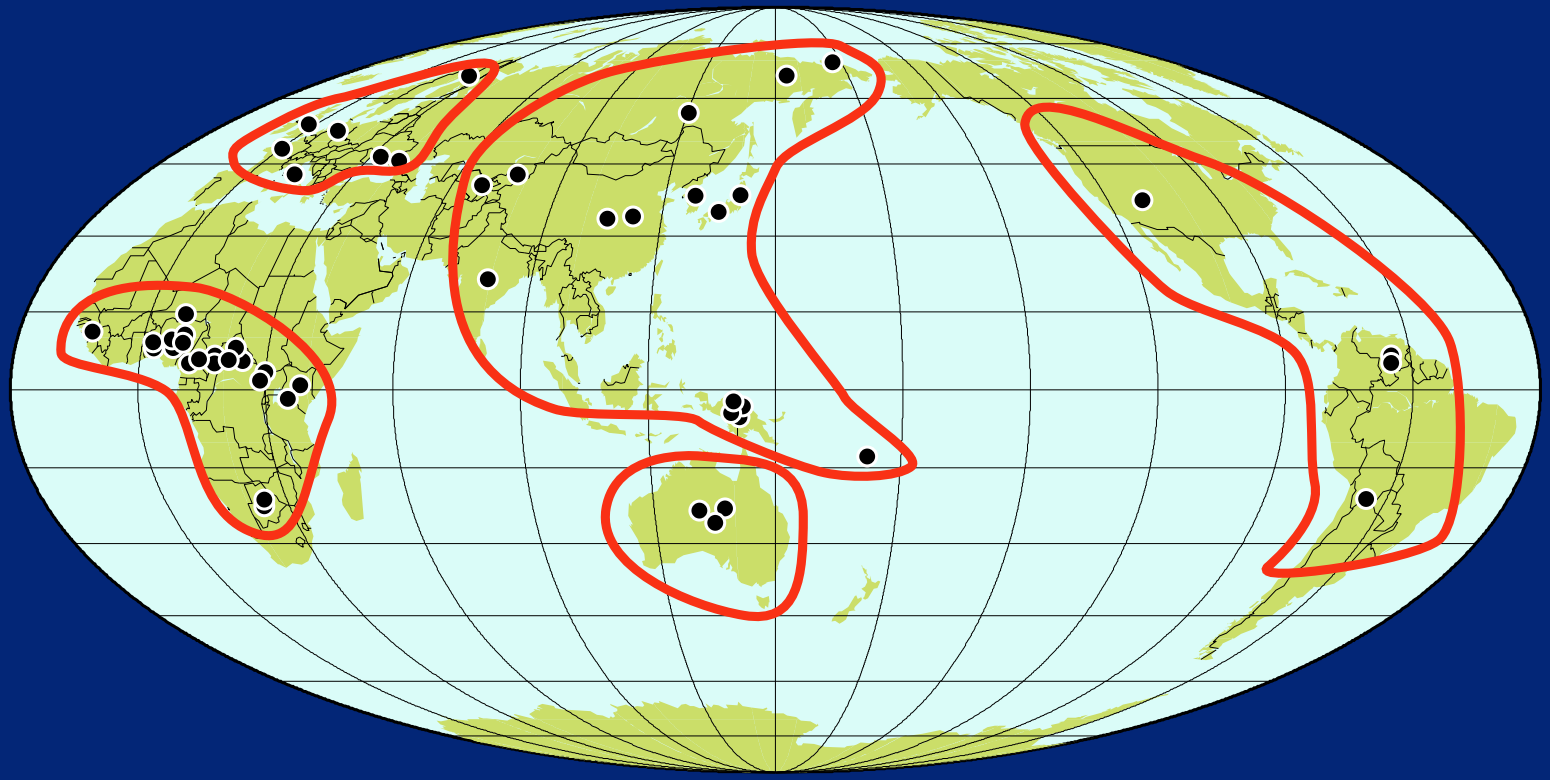
	Population 1		Population 2	
	Θ	$4N_e^{(1)}m_1$	Θ	$4N_e^{(2)}m_2$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

AMOVA

Analysis of Molecular Variance

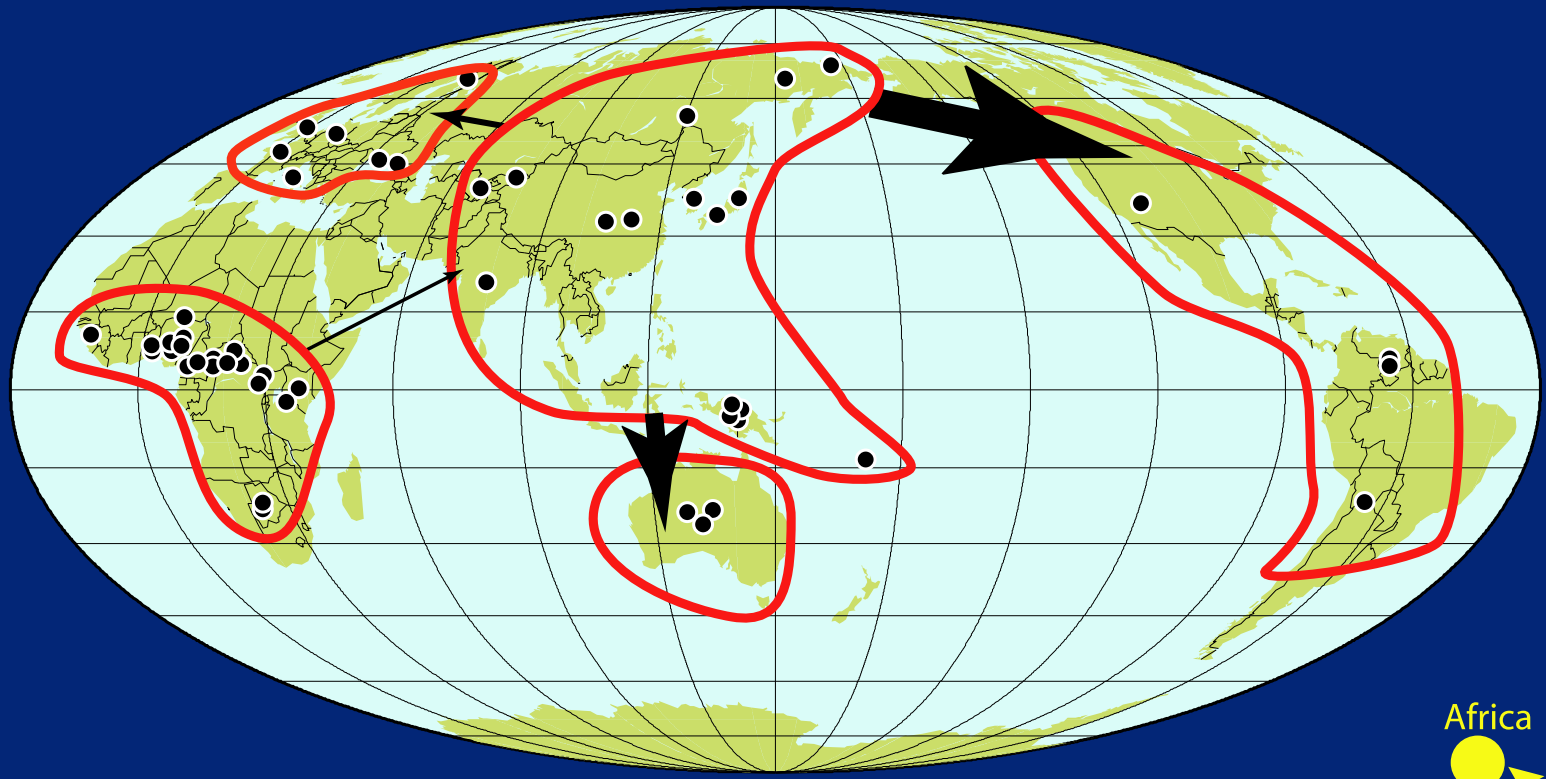
- Count differences among haplotypes within and between populations
- Compare to a null expectation from permuted data
- Infers degree of population subdivision; can be mined for estimates of specific migration rates
- More flexible than F_{ST}
- Commonly done with Arlequin

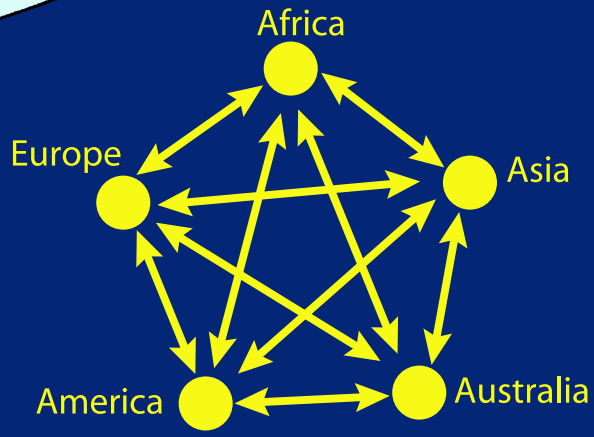
Complete mtDNA from 5 human "populations"



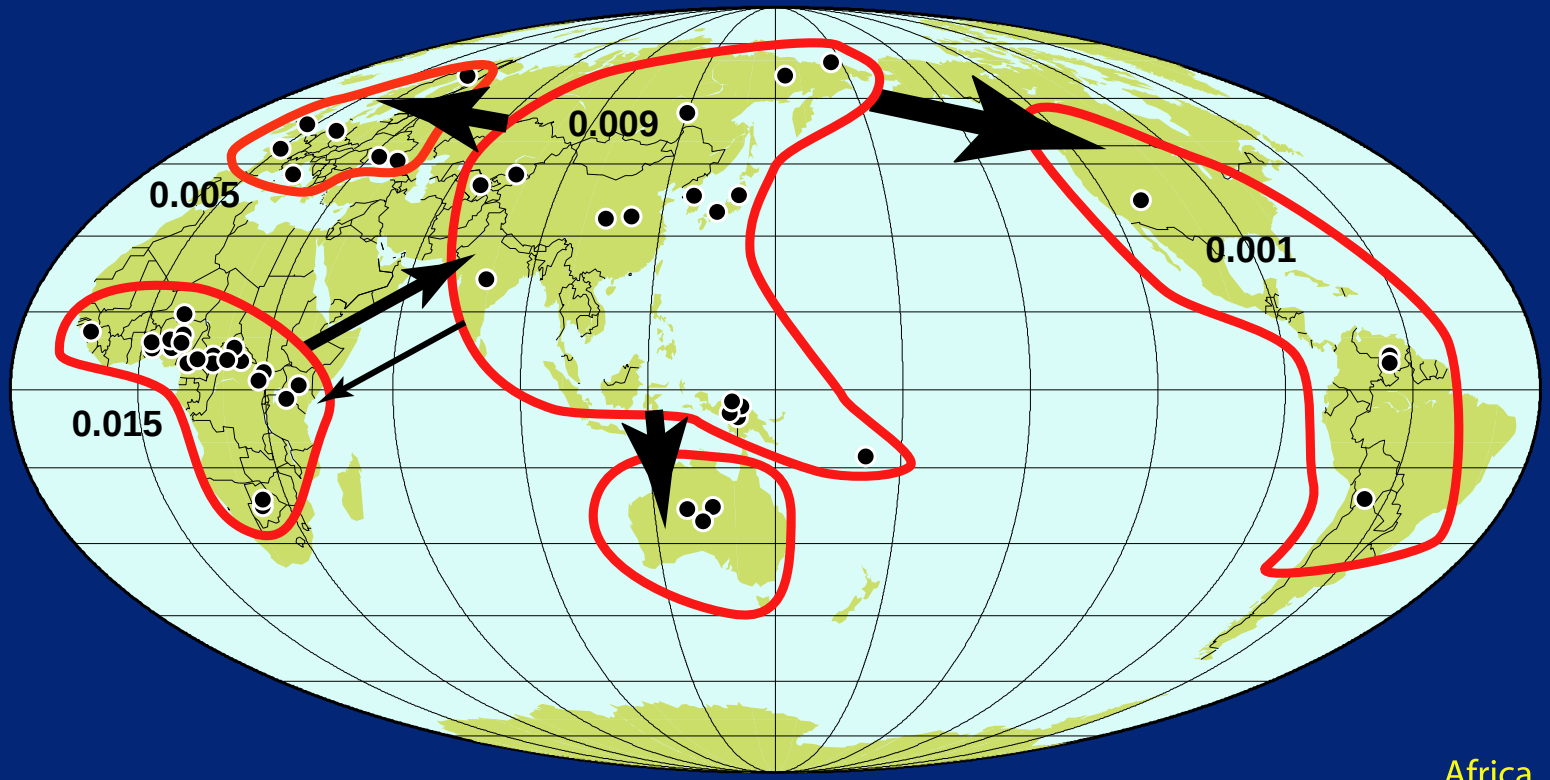
A total of 53 complete mtDNA sequences ($\sim$ 16 kb):
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.
Assumed mutation model: F84+Γ

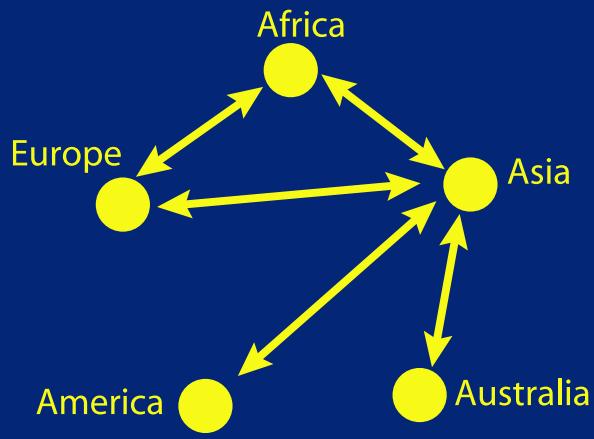
Full model: 5 population sizes + 20 migration rates





Restricted model: only migration into neighbors allowed

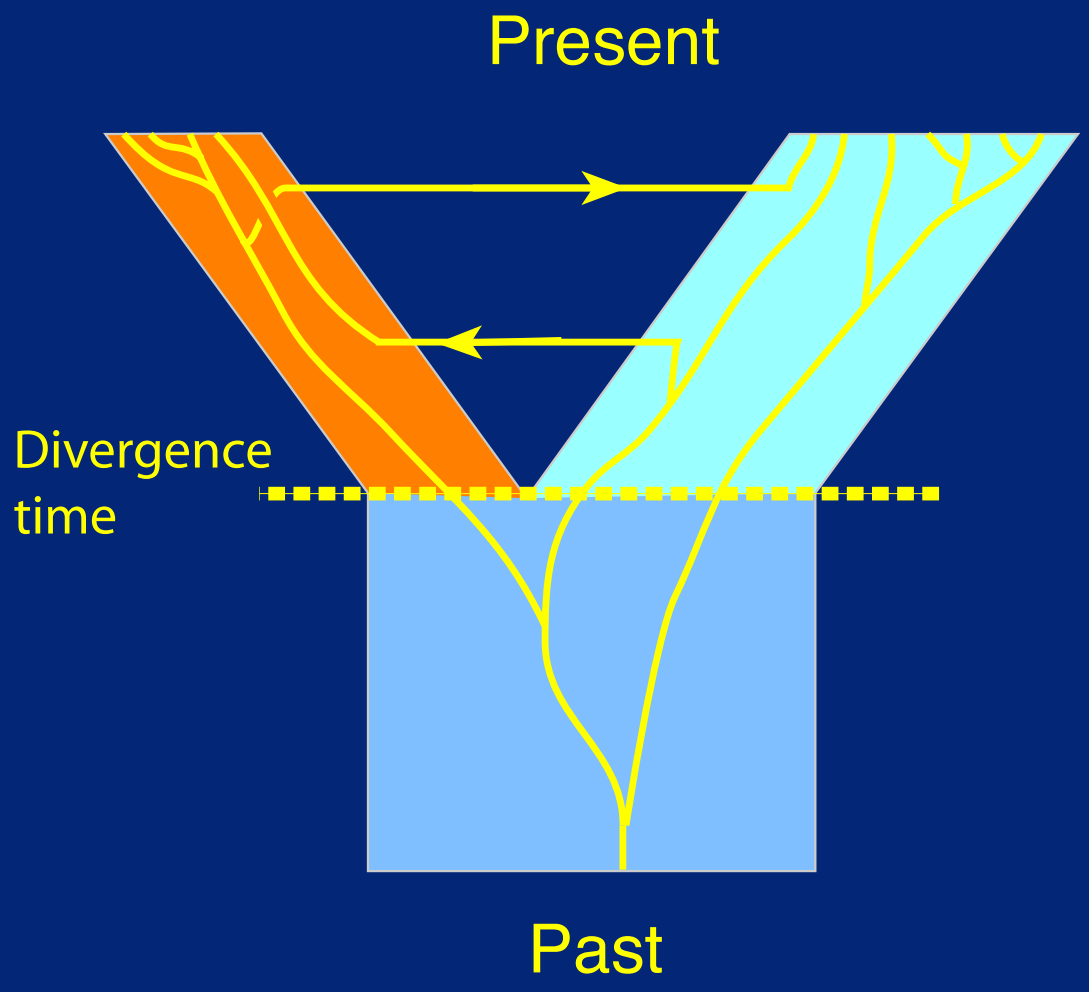




Migration in stable populations

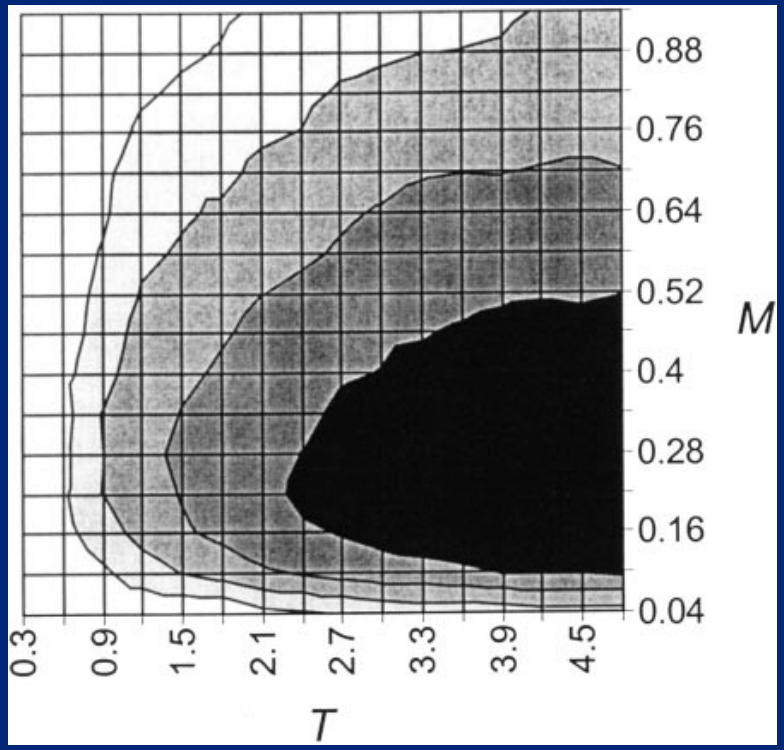
- MIGRATE and LAMARC have similar capabilities
- These programs estimate:
  - Θ per subpopulation
  - Immigration from each subpopulation into each of the others
  - Can use a restricted migration matrix
- Assumptions: no selection, stable population structure
- Unlimited populations in theory but huge data sets needed for more than 2-3 populations
- MIGRATE offers migration skyline plots for additional detail

Migration with divergence



Migration with divergence

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.



Migration with divergence

- IM/IMa/IMa2 estimate:
  - Θ per subpopulation (now up to ten subpopulations)
  - Θ of ancestral populations
  - Immigration in each direction
  - Times of divergence
- IM can also estimate growth or shrinkage of daughter populations
- IMa is a newer program with a more efficient algorithm
- IMa2 is a variant of IMa with support for more than 2 populations
- Assumptions: no selection, no recombination, single time point

General issues with migration estimation

- Migration too high–can't infer divergence time
- Divergence too recent–can't infer migration rate
- All methods assume constant rate, not bursts
- Multiple loci needed for good power

Specific issues with genealogy sampler migration estimation

- Multiple unlinked loci not available
  - LAMARC may help here by using partially linked loci
- Population structure changing too fast
  - For LAMARC and MIGRATE population structure must be stable
  - Population divergence masquerades as excess migration
  - IM and IMa handle divergence
- Think of "migration" broadly:
  - Movement between geographic regions
  - Movement between host types
  - Movement between partitions within a patient

Migration skyline plots

- MIGRATE can produce skyline plot of migration events over time
- Non-quantitative way to detect divergence
- Can also show other violations of homogeneity

When to use which sampler?

- LAMARC or MIGRATE
  - Populations stable for approximately $4N_e$ generations
  - As many populations as your data can stand
  - LAMARC handles data with recombination
  - MIGRATE offers skyline plots of migration rate
- IM or IMa
  - Populations arose less than $4N_e$ generations ago
  - Up to 10 populations
  - Relationships among populations known (if more than 2)
  - No recombination

Nested clade analysis

NPA or NPCA (Templeton et al. 1995)

- Infers assorted forces based on shape of haplotype network
- Limited support for recombination based on multiple alternative networks
- In simulation studies with simple scenarios, fairly good recovery of real effects
- However, up to 75% false positives (claims of effects that are not there)

Nested clade analysis

An opinion statement:

- It's often true that if force A is in effect, symptom B results
- NPCA is fairly good at identifying these
- But–how often does symptom B occur WITHOUT force A?
- Massive false positives arise if this is ignored

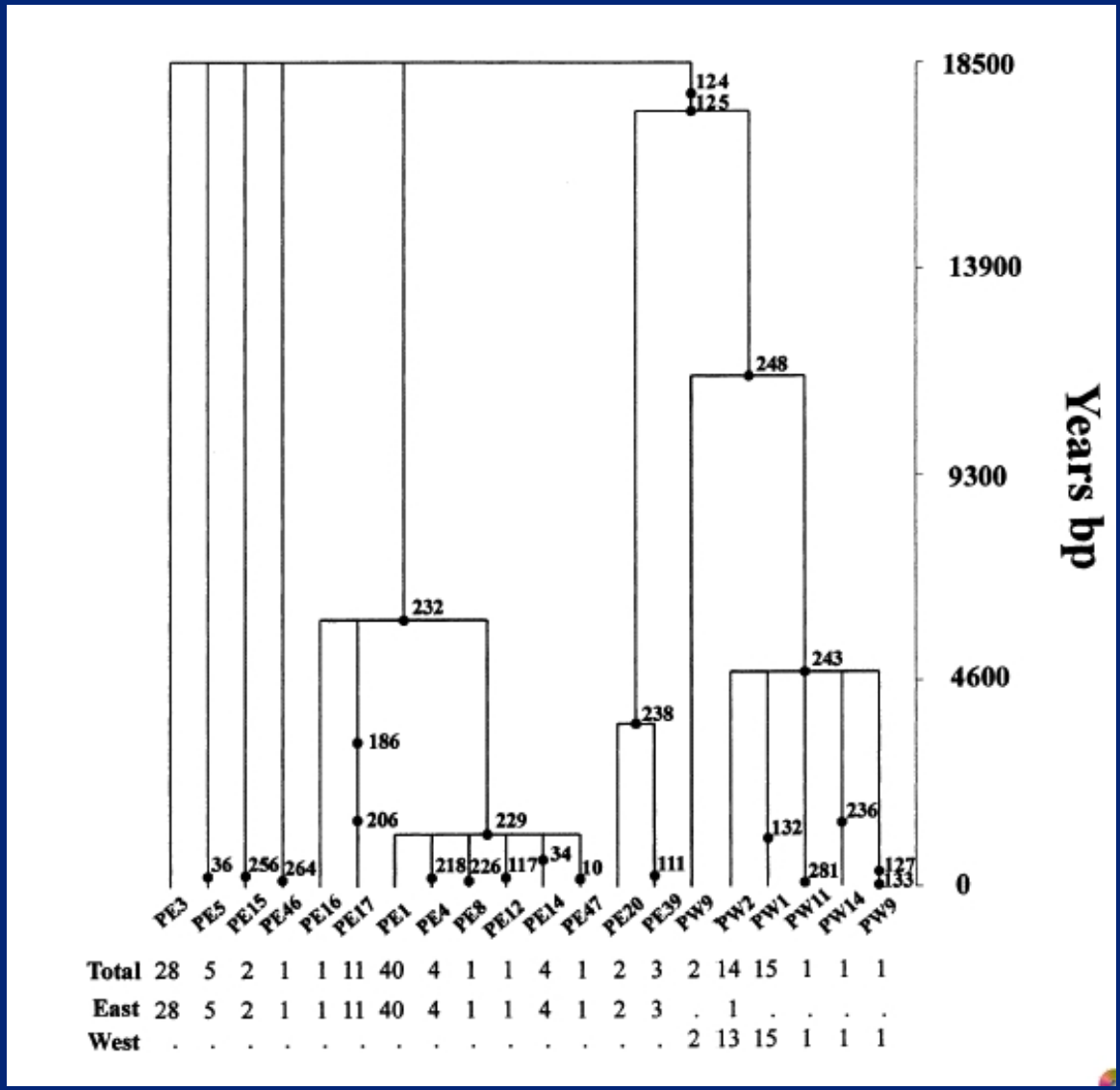
Nested clade analysis

An opinion statement:

- NPCA promises more than **any** method can now deliver
- Biologists want what it offers, so they use it
- 88% of NCPA results based on one locus; not enough
- Any result from NPCA needs validation by another method
- That other method **may not exist;** this is a weakness in the field (and grounds for more research)

Times of significant events

Milot et al. (2000)



Specific issues for times of significant events

- GENETREE can estimate the times of specific mutations or coalescences
- Always pay attention to the error bars of these estimates
- Assumes infinite-sites model:
  - No multiple hits to the same site
  - No back mutation
- This model is reasonable for low Theta (example: human SNPs)
- Not applicable to high Theta (example: HIV virus)

Recombination rate estimation

Recombination rate estimation

- LAMARC estimates per-site recombination rate
- Assumptions:
  - Equal recombination rate at every site
  - No changes in recombination rate over time
  - Single time point
- Accurate (but slow) for high recombination rate
- Some difficulties with low recombination rate
- Even if recombination rate not of interest, including recombination can improve estimation of other forces

Specific issues for recombination estimation

- LAMARC estimates a rate; it does not identify individual recombinants
- It is complimentary to methods such as bootscanning
- If recombinations occur in bursts (during a viral co-infection, for example), rate will be overestimated
- Recombination hot spots not handled: LDhat is preferable for this

Getting ready for the practical

- In the practical we will run a genealogy sampler
- Questions to bear in mind:
  - What are we trying to find out? How is our answer parameterized?
  - What assumptions are we making?
  - How do we know if we've run long enough?
  - How could this result be validated?

Getting ready for the practical

- Running the genealogy sampler LAMARC
- Inference of: Θ, growth rate, migration rates
- Short description of sampler follows

Likelihood version: Driving value

- To sample trees, we need a distribution
- We don't know the true distribution, so we assume one
- The assumed parameter is called the $driving\ value$

Driving value

- 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

How this works in practice

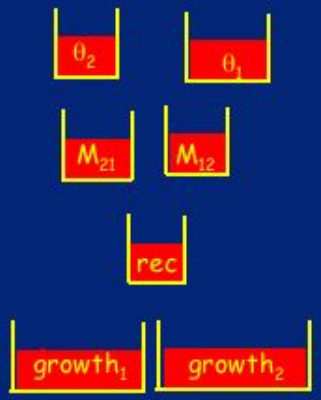
- Run multiple consecutive "initial chains"
- These allow the sampler to get good starting values
- When the values have settled down, run one "final chain"
- This will produce the final estimate
- Behavior of the initial chains is a clue:
  - Did we run long enough?
  - Do we need more chains or longer chains?
  - How long is this going to take, anyway?

Bayesian version: Priors

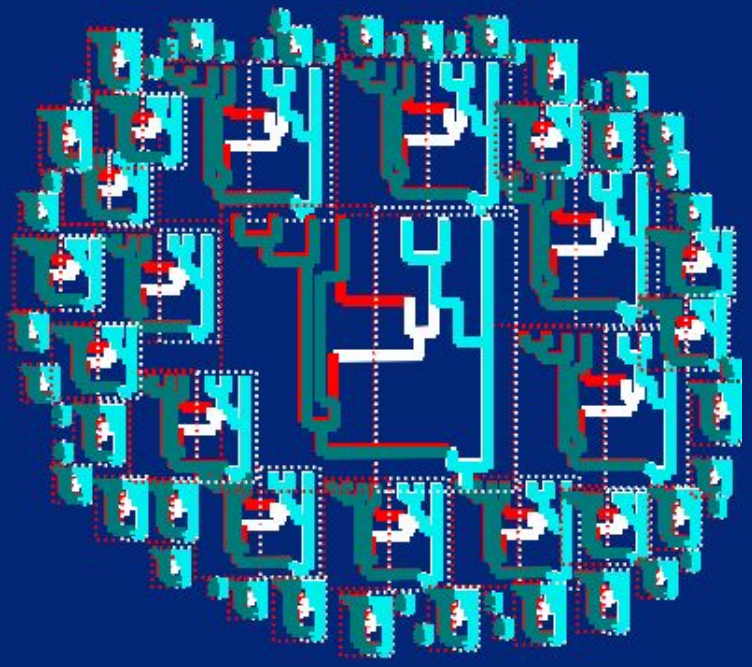
- Alternatively, we can use a Bayesian algorithm
- We place priors on each parameter
- New values are sampled from the prior
- We tabulate accepted values and form a curve
- Multiple cycles are not necessary

New search scheme for Bayes

Parameter space
(determined by priors)



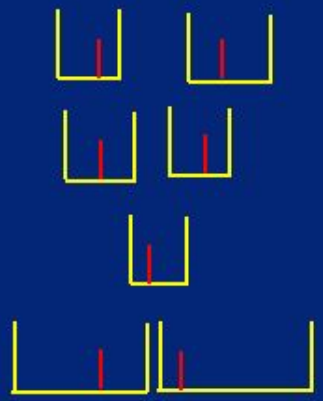
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

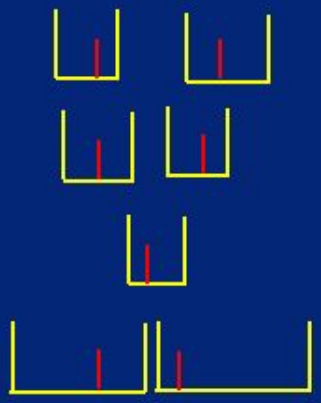
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

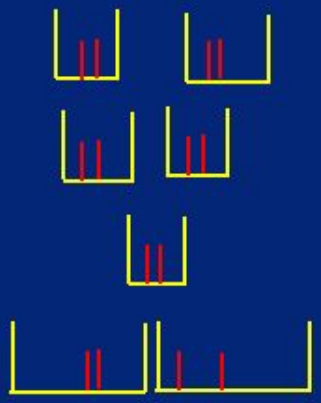
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

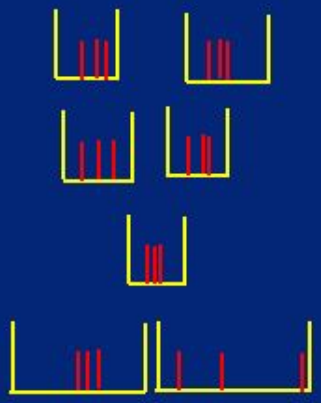
Tree space



New search scheme for Bayes

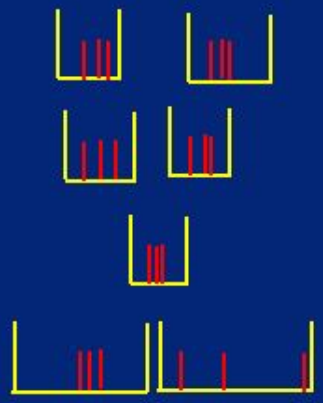
Parameter space
(determined by priors)

Tree space



New search scheme for Bayes

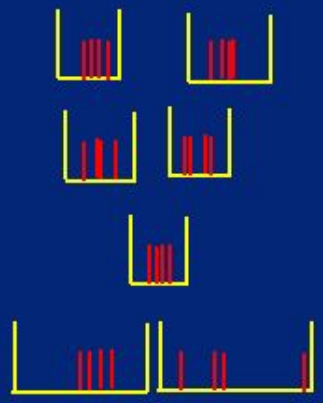
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

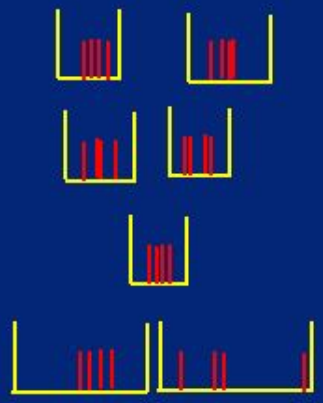
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

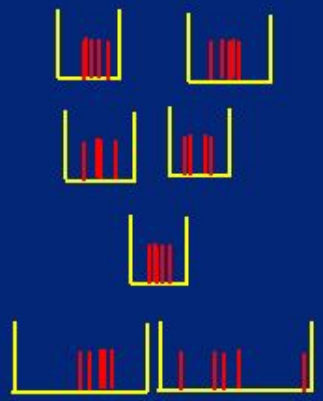
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

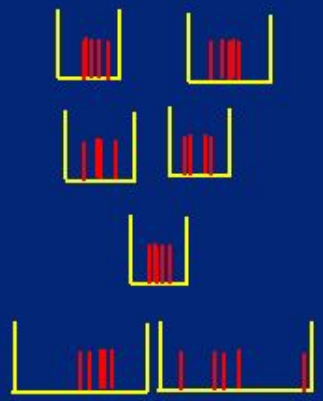
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

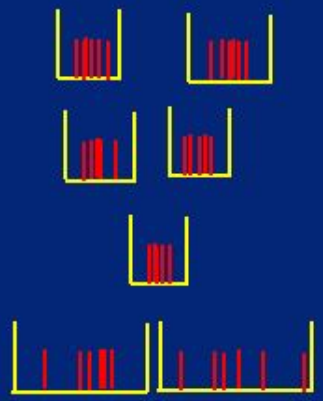
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

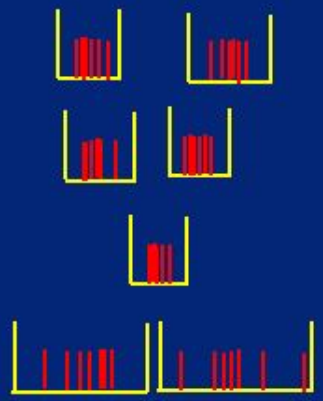
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

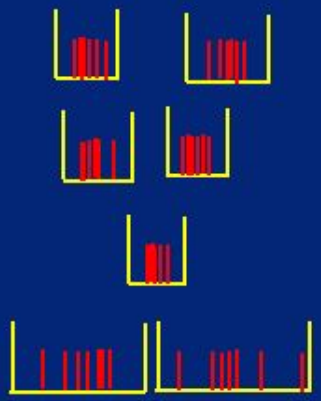
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

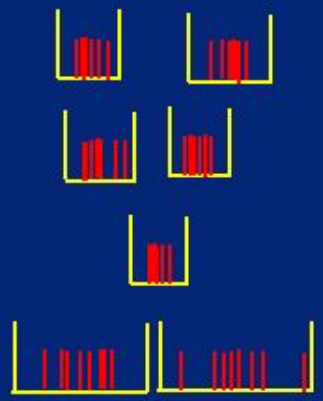
Parameter space
(determined by priors)



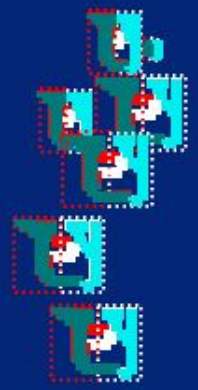
Tree space

New search scheme for Bayes

Parameter space
(determined by priors)

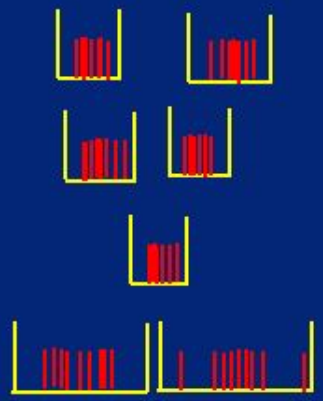


Tree space

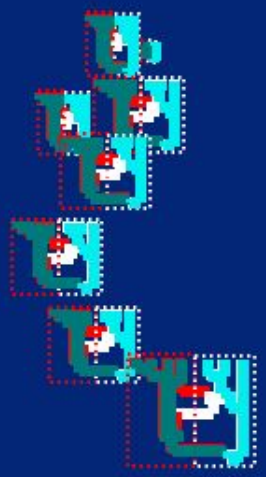


New search scheme for Bayes

Parameter space
(determined by priors)

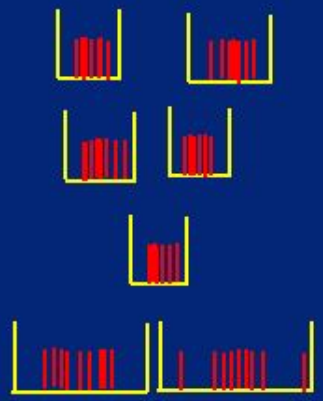


Tree space

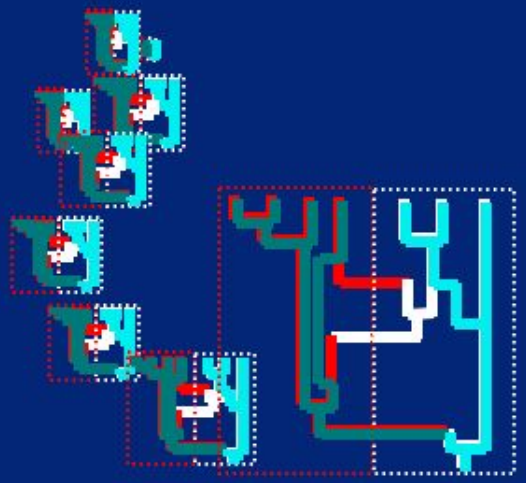


New search scheme for Bayes

Parameter space
(determined by priors)

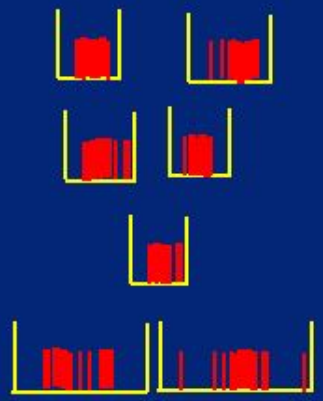


Tree space

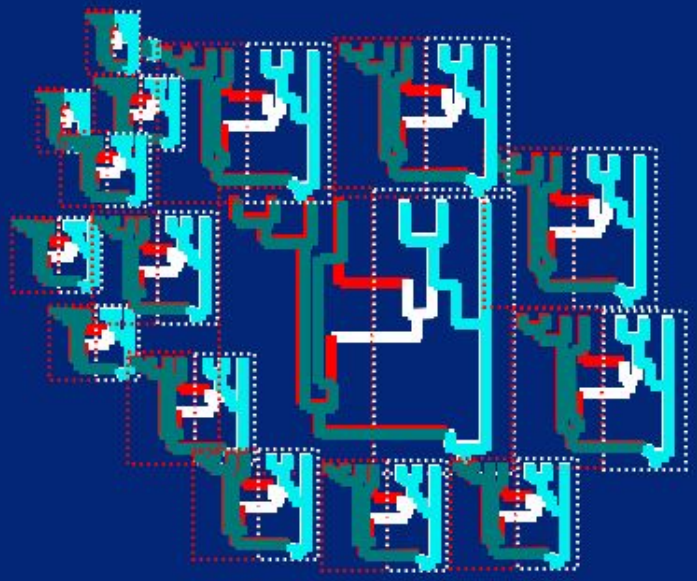


New search scheme for Bayes

Parameter space
(determined by priors)

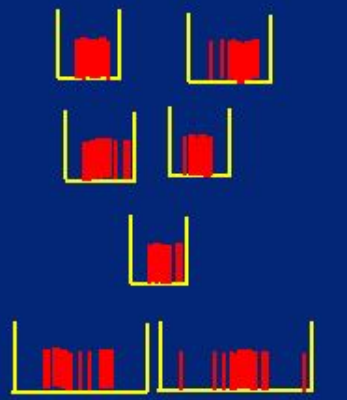


Tree space

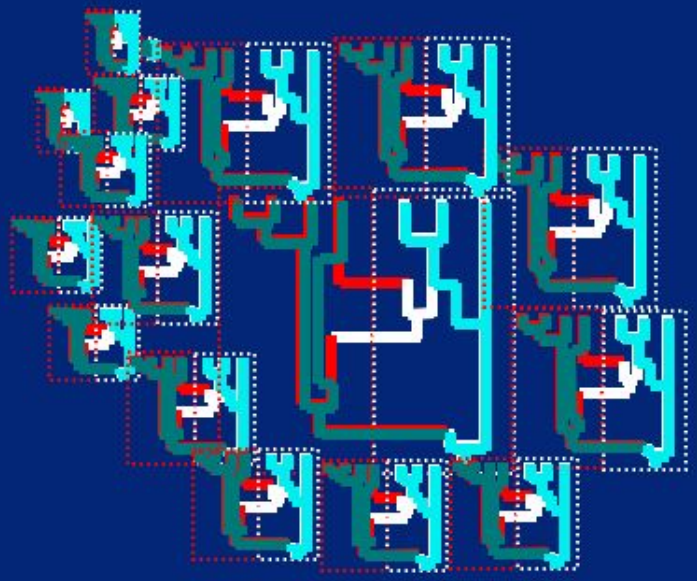


New search scheme for Bayes

Parameter space
(determined by priors)



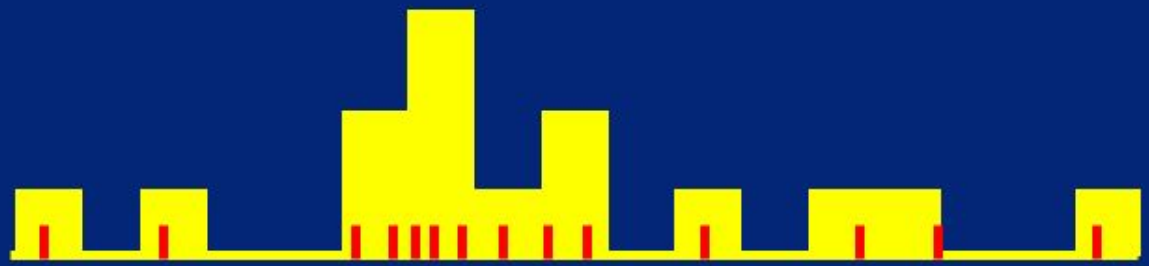
Tree space



Keep a list of all accepted parameters

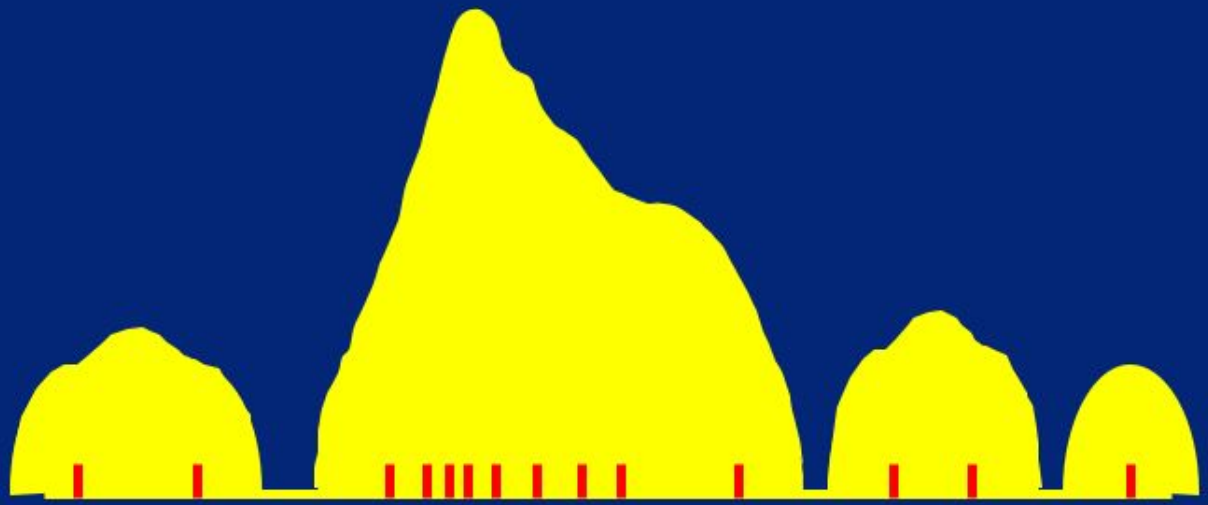
Data collection and curve smoothing

Data collection and curve smoothing

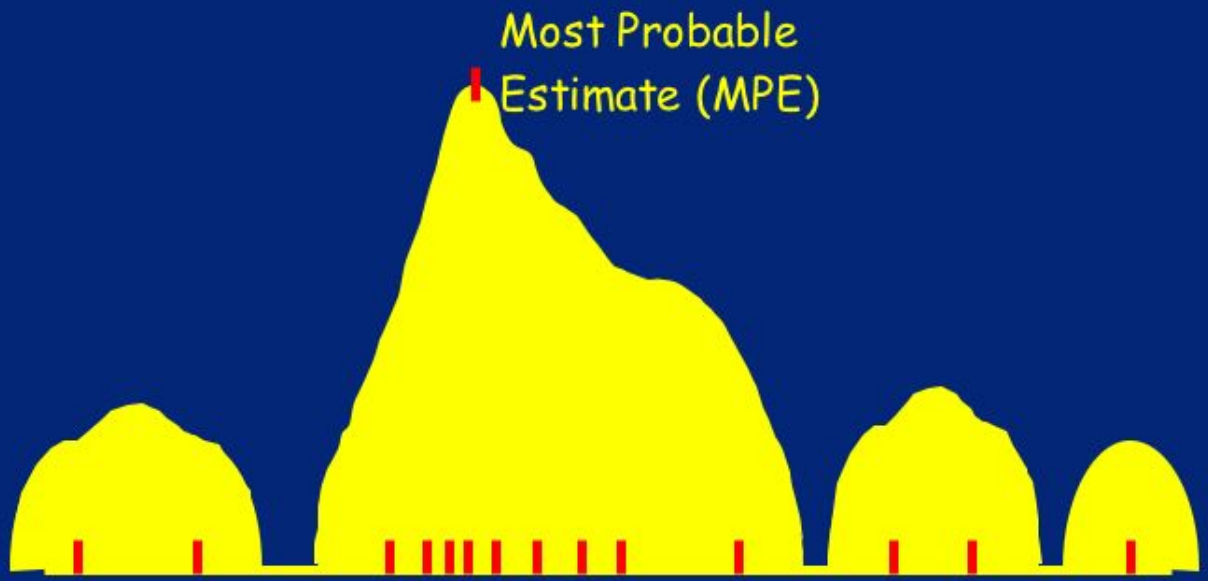


Data collection and curve smoothing

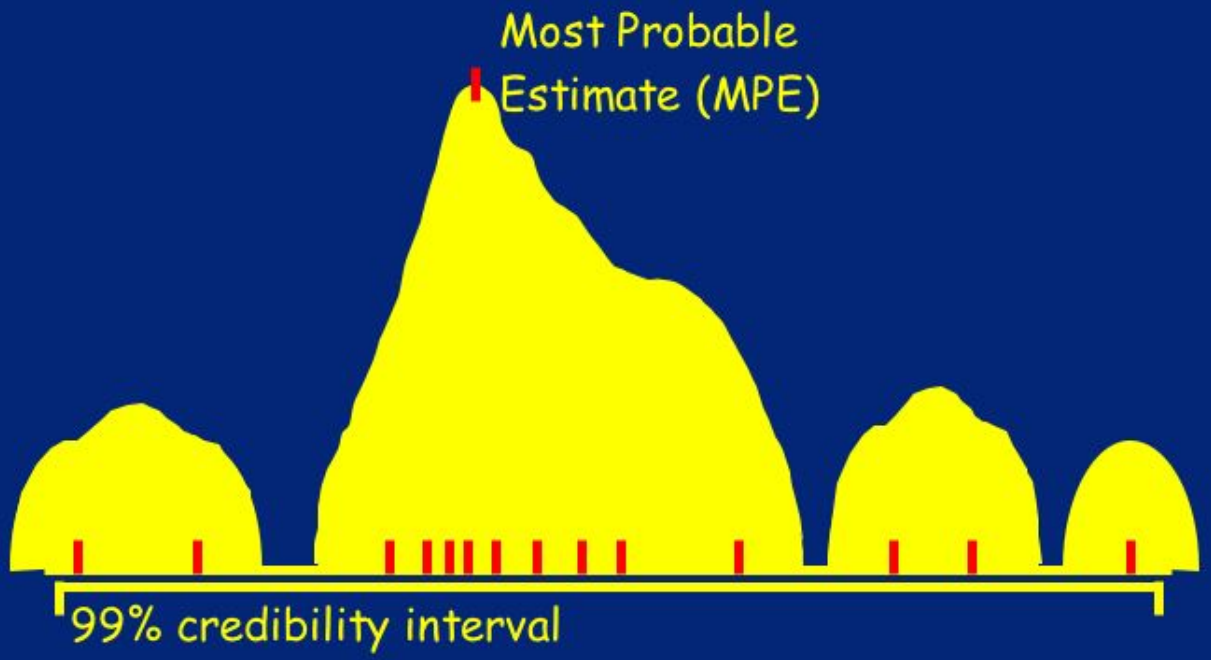
Data collection and curve smoothing



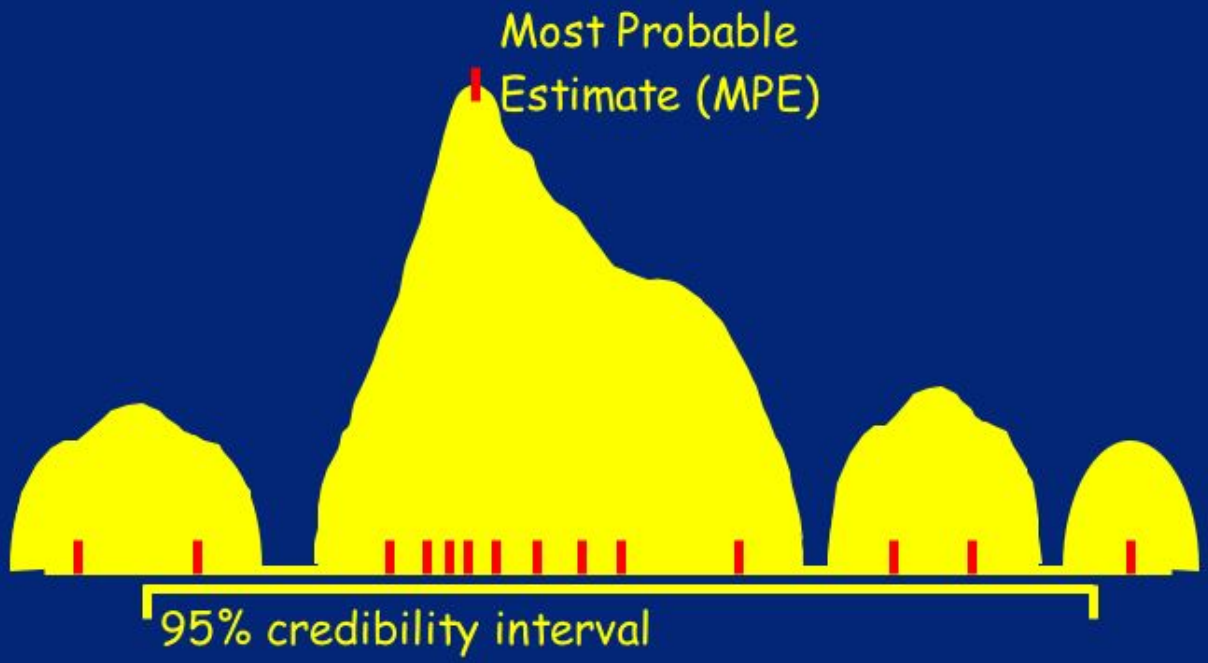
Data collection and curve smoothing



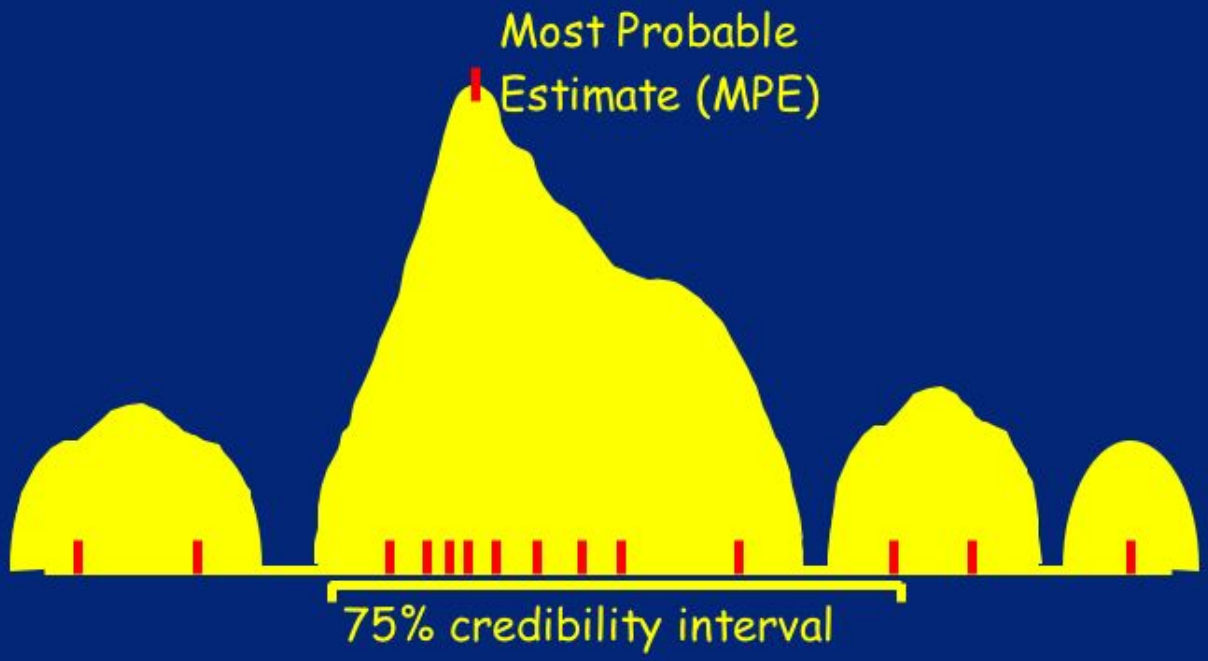
Data collection and curve smoothing



Data collection and curve smoothing



Data collection and curve smoothing



Bayes versus Likelihood

- Overall performance of likelihood and Bayesian samplers is similar
  - (Beerli 2006, Kuhner 2007)
  - Bayesian has edge when data are sparse
  - Likelihood gives more information about correlation of parameters

How this works in practice

- A Bayesian run is usually just one "final chain"
- (You could do an initial chain to get a time estimate)
- Starting values are not so important
- Good priors are essential!

Bayesian priors

- A prior should represent our pre-existing knowledge of a parameter
- Often biologists cannot quantify this
- Instead, we use "non-informative" priors and hope it won't matter
- No prior is really uninformative:
  - If it is too narrow, it may exclude the truth and will certainly overstate confidence
  - If it is too wide it will drastically slow the search and may understate confidence
  - If it is not roughly centered around the truth it may introduce bias
- To know the "right" prior we would need to know the answer....

Bayesian priors

Advice:

- Use all available information to set the priors:
  - Previous studies
  - Analogies with similar systems
- Give a little fudge room but do NOT be extremely conservative
- None of the samplers work well with extremely wide priors (they don't focus the search)
- If your results are piled up on one side of the prior, widen the prior next time
  - This is totally inappropriate in Bayesian theory
  - None the less, it improves results....

General advice

- Don't be afraid to try things
- Interrupt any run that will take too long for a 90 minute practical
- After completing one run, move or rename the results file or it may be overwritten
- Draw a picture of your results
- Ask questions
- Talk to your neighbors, TAs and instructors

Preparation for practicals

- 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
- This will save time and pain during the demo. Thank you!