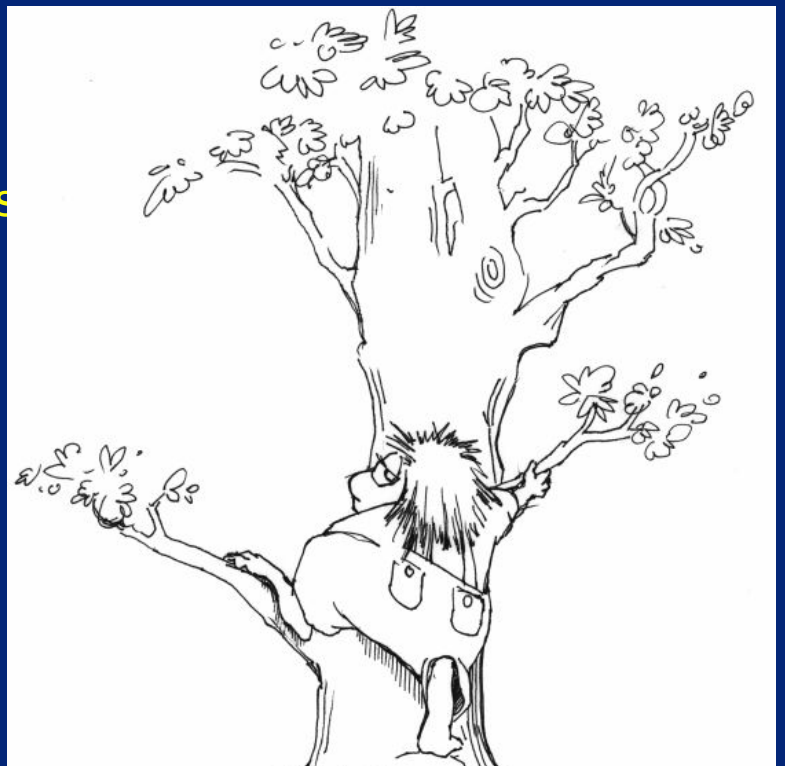
Coalescent Likelihood Methods for Estimating Population Parameters

Mary K. Kuhner
Genome Sciences
University of Washington
Seattle, WA, USA

Summer Institute in Statistical Genetics



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–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

Course business

- We offer signed certificates of completion for this course
- If you want yours, please pick it up from me today

Six genealogy samplers to consider

- 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

Six genealogy samplers to consider

- MIGRATE (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

Six genealogy samplers to consider

- BEAST (http://evolve.zoo.ox.ac.uk/beast/)
  - Drummond and Rambaut
  - Estimates:
    - Overall population size x mutation rate
    - Overall growth rate
    - With sequential samples, mutation rate and generation time
    - Detailed skyline plots of growth rate
    - Relaxed molecular clock
  - Bayesian analysis
  - DNA, RNA, amino acids, codon data

Six genealogy samplers to consider

- IM, IMa, IMa2 (http://lifesci.rutgers.edu/ heylab/HeylabSoftware.htm#IM)
  - Nielsen, Hey, Wakeley et al.
  - Estimates:
    - Population size x mutation rate
    - Immigration rates
    - Size of ancestral populations
    - Times of divergence
    - Daughter population growth rates (IM only)
  - Bayesian analysis
  - DNA, RNA, microsatellites, HapSTRs

- IMa and IMa2 are more efficient; IM has a larger choice of models

Six genealogy samplers to consider

- GENETREE (http://mathgen.stats.ox.ac.uk/software.html)
  - Griffiths et al.
  - Estimates:
    - Population size x mutation rate
    - Exponential growth rate
    - Time of most recent common ancestor
    - Times of significant mutations
  - Likelihood analysis (independent genealogies)
  - DNA (infinite sites)

Useful review paper

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

There have subsequently been major improvements to IM and minor improvements to most other packages.

Designing a study

- What kind of data are available?
- What do you want to know?
- What are the pluses and minuses of available techniques?
- What is expected practice in your subfield?
- How much time do you have?

Designing a genealogy sampler study

- Major questions:
  - Should I be doing this analysis at all?
  - What model should I use?
  - How much data do I need?
  - How long do I have to run the program? (Are we done yet?)

When is a coalescent analysis inappropriate?

Things you can determine in advance:

- Randomly sampled population data are not available
  - A sample of one HIV sequence from each serotype is not usable
  - Data assigned to populations by genetic analysis can't be used to infer migration rates of those populations
- No believable mutational model is available
  - RFLPs
  - AFLPs
  - Insertion/deletion
  - Gene order

When is a coalescent analysis inappropriate?

Things that emerge from analysis:

- Data are too far outside available population models
  - Extremely rapid population change
  - Extremely non-neutral evolution
  - Extremely non-constant gene flow or recombination
- Time-scale of interesting events is much longer or shorter than the organism's coalescent time (approx. $4N_e$ generations)

Some doubtful attempts

- What is the rate of horizontal gene transfer between bacteria and plants?
- How fast did the HIV epidemic spread in the Middle East?
- What is the effective population size of pre-cancer cells in the esophagus?

When is a particular parameter not inferrable?

Θ

- Data should not be invariant
- Data should not be saturated (unalignable)
- Population must be old enough:
  - Expected depth of tree is Θ
  - If population much younger than that, little information on its size
- Unacknowledged recombination or selection can obscure answer
- Don't forget that a big linked locus like mtDNA is still only one locus

When is a particular parameter not inferrable?

Growth rate

- For exponential growth, $4N_e g$ is key parameter
- $4N_e g >> 1$ leads to star phylogenies with little information
- $4N_e g << 0$ can lead to infinite TMRCA! (I don't know what the exact cutoff is)

When is a particular parameter not inferrable?

Migration rate

- If $4N_em$ much greater than 1, populations homogenize and gene flow hard to measure
- If $4N_em$ very low migration events are so rare their frequency can't be estimated well
- Very high recombination weakens evidence of migration (haplotypes are too short)

When is a particular parameter not inferrable?

Divergence time

- Needs to be more recent than MRCA (approx $4N_e$ generations
- Very recent divergence not visible (less than 1/10 of this?)
- High gene flow destroys ability to infer divergence (certainly $4N_e > 1$ will have little or no power)

A cautionary tale

Abdo, Crandall and Joyce 2004

- Simulation studies to test inference of migration
- Three Θ values, four M values
- Under many circumstances inference was very poor

A cautionary tale

I resimulated Abdo et al's data:

- Low Θ with high M had no variable sites
- Low M never had more than one (obligatory) migration per tree
- High Θ with low M had mutationally randomized data
- Only a few parameter combinations led to data that could be analyzed at all

A cautionary tale

- Easy mistakes to make (I have made them too)
- Meaningful biological range of these parameters can be narrow
- Bear this in mind when:
  - Choosing types of data
    - If DNA sequences nearly invariant, consider microsatellites
  - Choosing priors
  - Designing simulations

A caveat

- The rest of this talk will focus on genealogy samplers
- Data demands are different for other types of analysis:
  - Allele frequency estimation needs a bigger sample
  - Inference based on infinite-sites needs a low mutation rate
- In general, if data are not rich enough for a genealogy sampler, they are not very informative with any method
- Using both sampler and non-sampler methods is good (remember the red drum study)

What model should I use?

- Mutational models
  - Nucleotide sequences
  - Microsatellites
  - Others
- Population models
  - Growth
  - Migration and subpopulation structure
  - Recombination

What mutational model should I use?

- Nucleotide sequences
  - Optimize model using MODELTEST (Posada and Crandall)
  - Use most nearly optimal model available in your chosen software
  - Using a more complex model will probably not help
  - If sequences are short:
    - Fix mutational parameters at published values
    - Or values from other samples from your organism
    - Or, failing that, from closely related organisms

What mutational model should I use?

- Microsatellites
  - Single-step model probably best available
  - K-Allele model overstates chance of large changes
  - LAMARC offers a mixed model but it is not validated well yet
- Others
  - BEAST offers codon and protein models
  - Codon model best for coding sequence–but SLOW
  - K-Allele model generally useful for unusual types of data

What population model should I use?

- Growth
  - Several programs offer exponential growth
  - Real populations do not grow exponentially forever
  - BEAST offers Bayesian skyline plots, but poor resolution without multiple time-point sampling
  - If growth is very recent, a no-growth analysis will perform better

What population model should I use?

- Defining populations
  - Programs do not perform well unless populations have some structure
  - STRUCTURE (Pritchard) is useful in deciding whether to pool populations
  - Do not use STRUCTURE to assign individuals to subpopulations and then analyze them as if they belonged there!
  - (This has the effect of sending migrants back home....)
- How many populations?
  - More than 2-3 populations too many unless many loci available
  - For cases with many populations, try symmetrical migration rates and/or constrain unneeded rates to 0
- How many parameters to estimate
  - MIGRATE offers tests based on AIC to help weed out unneeded parameters

What population model should I use?

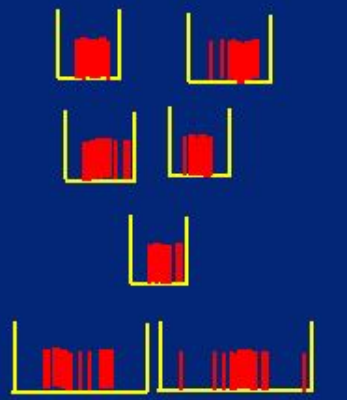
- Recombination
  - Risky to ignore recombination if it is present
  - "Casting out" apparent recombinants biases Θ downward
  - Four-gamete test can tell whether it is dangerous to disregard recombination
  - If combining recombining and non-recombining loci (eg mtDNA and nuclear DNA) prefer a recombinant analysis
  - May be able to ignore recombination for very short sequences

Bayesian versus likelihood samplers

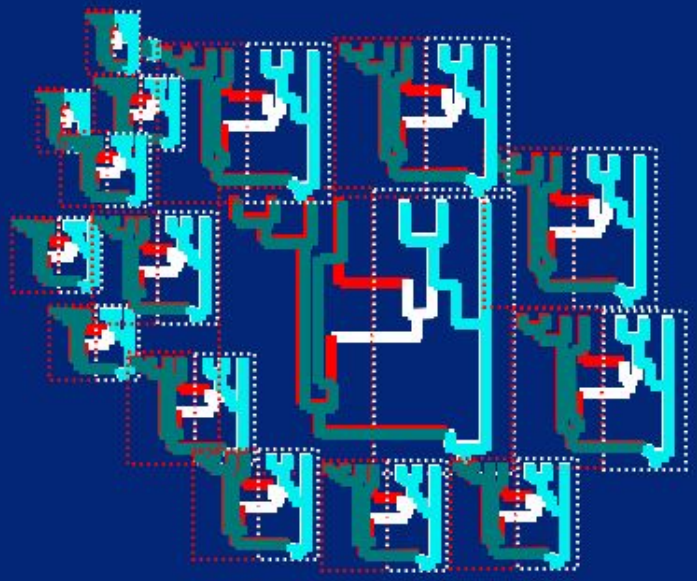
- Likelihood sampler:
  - Sample genealogies according to driving values
  - Estimate parameter values from stored genealogies
  - Replace driving values with estimates and repeat until satisfied
- Bayesian sampler:
  - Sample genealogies according to current parameter values
  - Sample parameters from prior according to current genealogies
  - Estimate parameter values from histogram of values visited

New search scheme for Bayes

Parameter space
(determined by priors)



Tree space



Keep a list of all accepted parameters

Bayesian versus likelihood samplers

Which to prefer?

- Kuhner MK, Smith LP, 2007. Comparing likelihood and Bayesian coalescent estimation of population parameters. Genetics 175: 155-165.
- Conclusion: no substantial difference
- Beerli P, 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22: 341-345
- Conclusion: Bayesian is superior when data are sparse and number of parameters is high

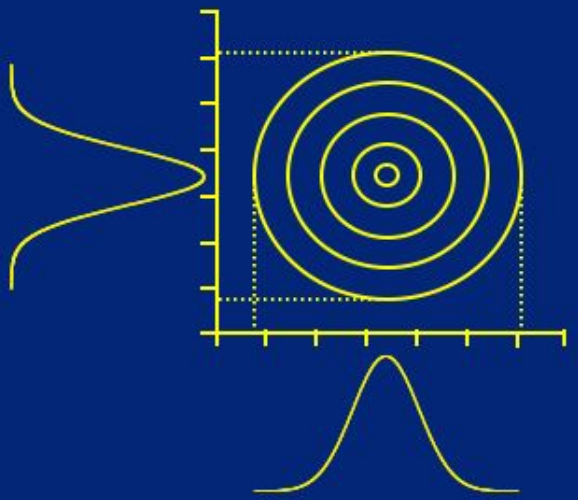
Bayesian versus likelihood samplers

- Likelihood method may have biased (too narrow) confidence intervals when:
  - True parameter value very close to zero
  - Driving values far from truth (run more chains!)
  - Sample of trees inadequate (run more steps!)
  - Data are sparse

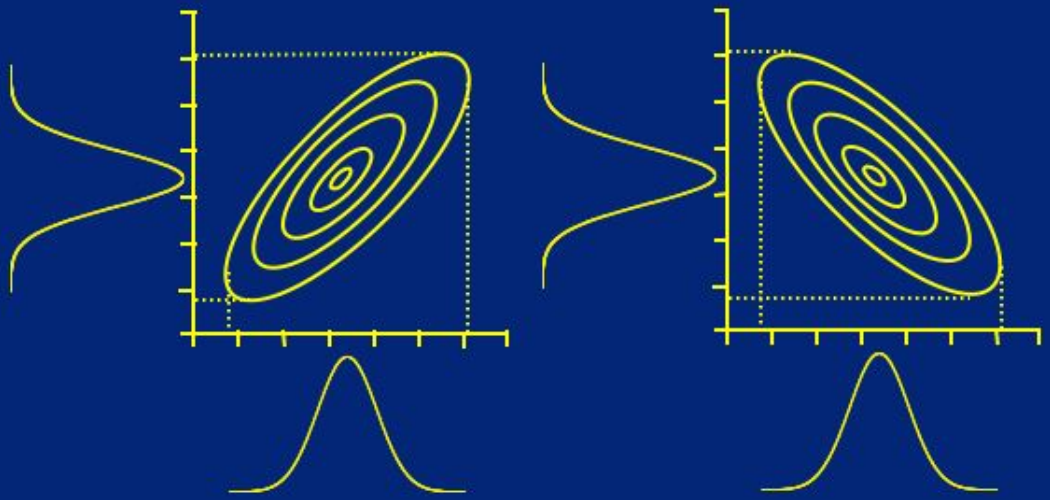
Bayesian versus likelihood samplers

- Bayesian method may be biased when:
  - Prior not appropriate: too narrow, too wide, excludes truth
- In sparse data cases you may appear to get more information from Bayesian than likelihood **because of information in your prior**
- This is only good if your prior is well-founded
- Current Bayesian implementations lose information about correlation among parameters which is available with likelihood

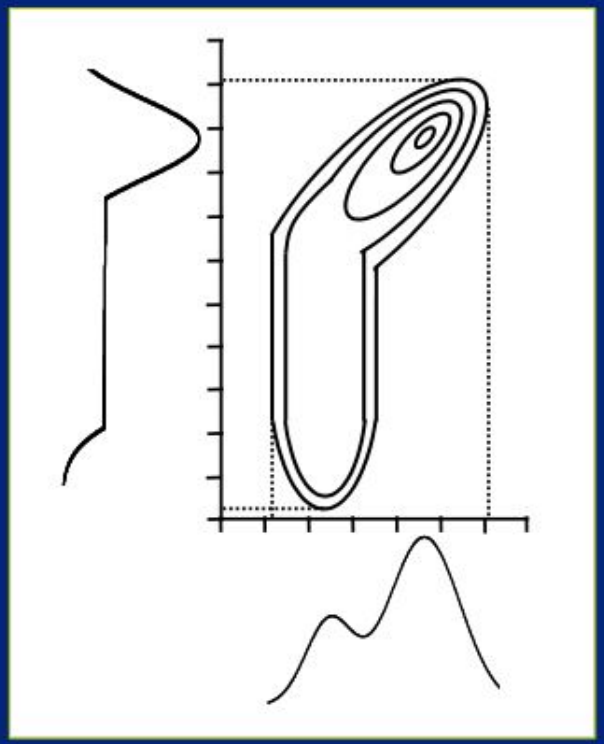
Loss of Correlation Information



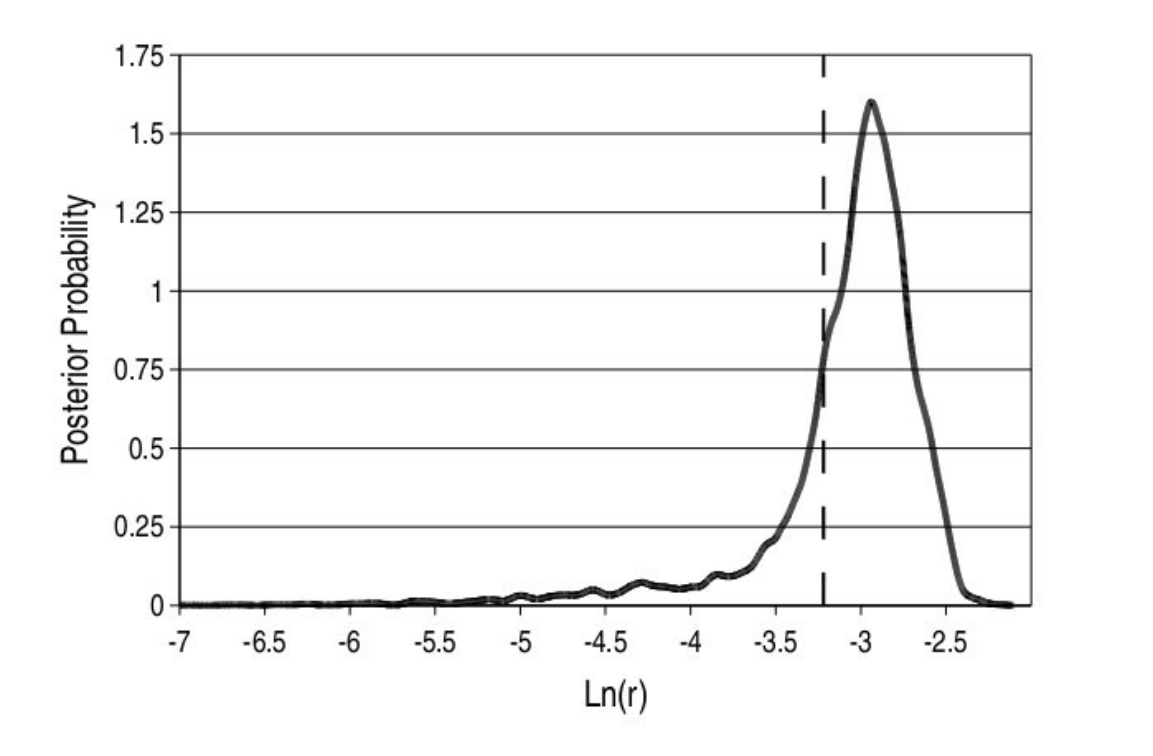
Loss of Correlation Information



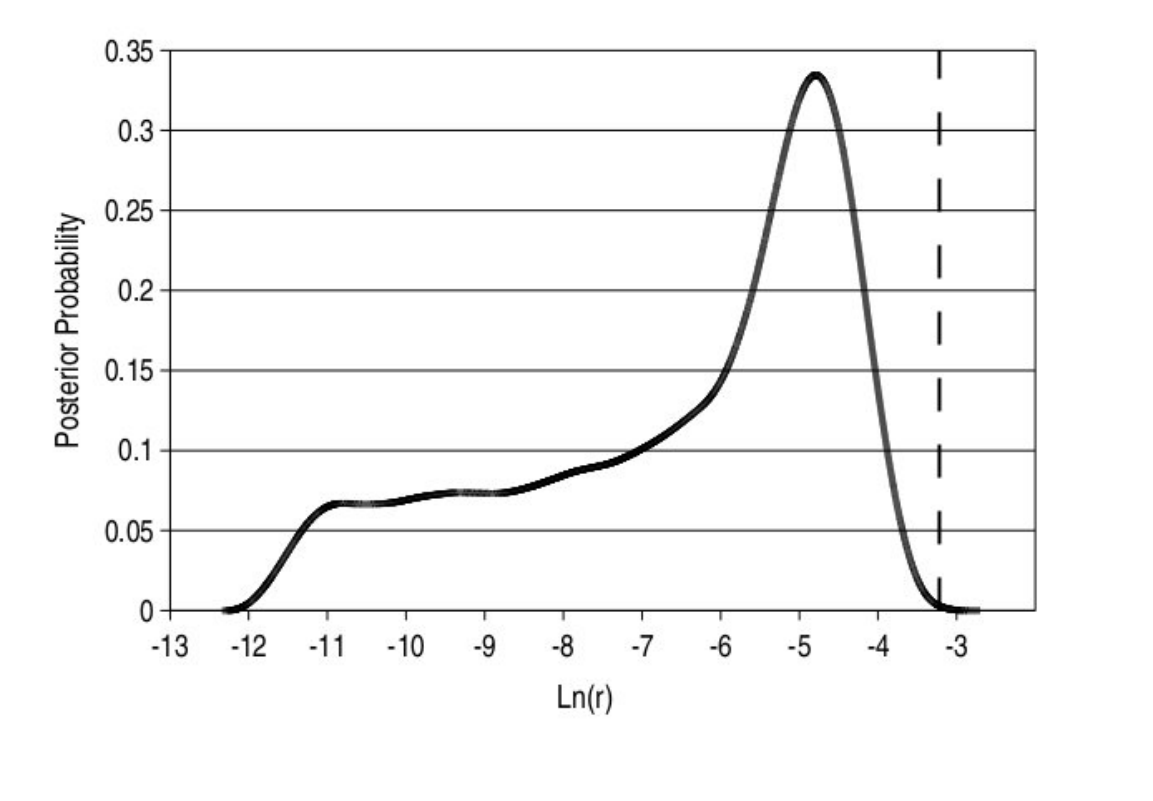
Loss of Correlation Information



Nice outcome: Curve mainly reflects underlying data



Not so nice outcome: Curve strongly influenced by prior



Bayesian versus likelihood samplers

Which to use?

- When things are working well, methods are very similar
- Practical choice often based on software availability
- A Bayesian analysis with well founded prior is probably best
- If priors very unclear, prefer likelihood
- Both methods need adequate number of sampled genealogies!
- Speed difference not substantial

How much data do you need?

- For analyses without recombination:
  - Several unlinked loci are best
  - 2-3 unlinked DNA loci or 5-10 microsats can give reasonable results
  - Multiple time points or long sequences with recombination can compensate for lack of unlinked loci
  - No more than 20-25 samples per population needed
  - Ideally DNA sequences should have at least 10-15 variable sites
  - If polymorphism is low, longer sequences are needed

How much data do you need?

- For analyses with recombination:
  - A single locus can work if it's long
  - Length needed depends on polymorphism level
  - For human DNA levels, 20 KB is a good size
  - Multiple loci are still good if not too short
  - No more than 20-25 samples per population
  - This will take **much longer**

How much data do you need?–Citations

- Pluzhnikov, A. and Donnelly, P. (1996) Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144, 1247-1262

- Felsenstein, J. (2006) Accuracy of coalescent likelihood estimators: do we need more sites, more sequences, or more loci? Mol. Biol. Evol. 23, 691-700.

How much data do you need?

- If you are data-starved:
  - Reduce the number of parameters
  - Do several runs and compare results
  - Pay careful attention to confidence intervals
  - Don't expect the world!

How long to run?

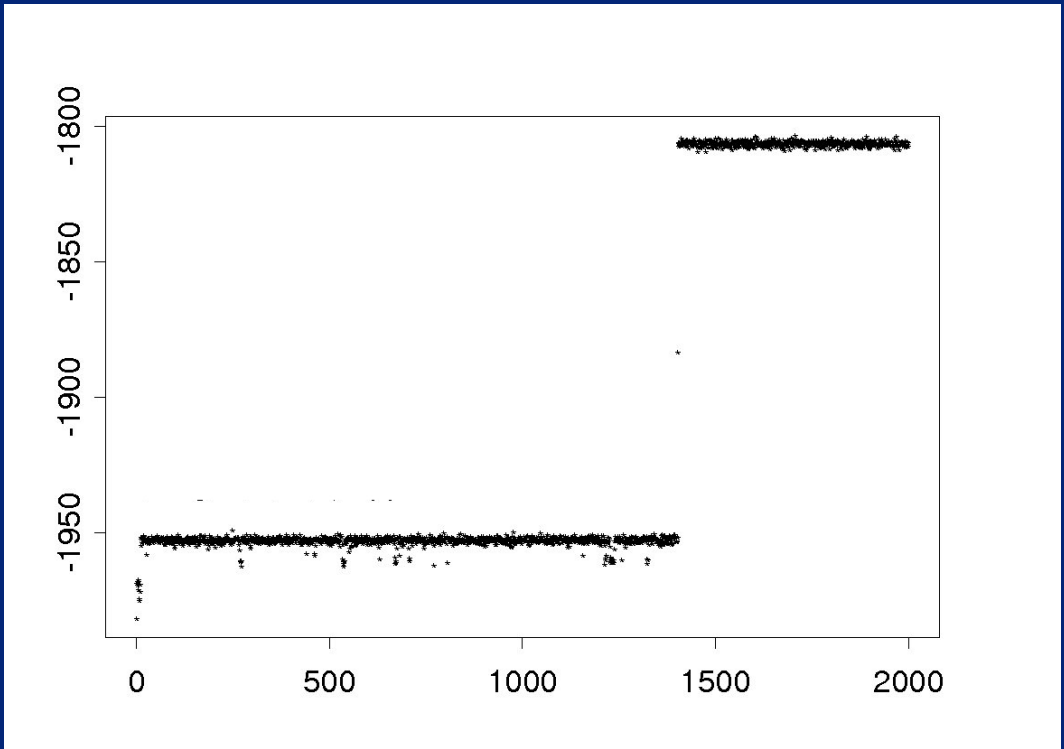
- Some general principles:
  - Results should be broadly similar if program is re-run
  - Longer runs needed for good confidence intervals
  - If run is too short, confidence intervals may exclude the truth
- These programs require informed use
- "Black box" application will lead to misleading results
- Publications must give details of run conditions

Program takes forever to run

- You may be asking too much
- Try restricting your migration model
- 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
- Break analysis into parts that can be run separately
- (MIGRATE only) Use several computers in parallel
- Future direction: run calculations on graphics card!

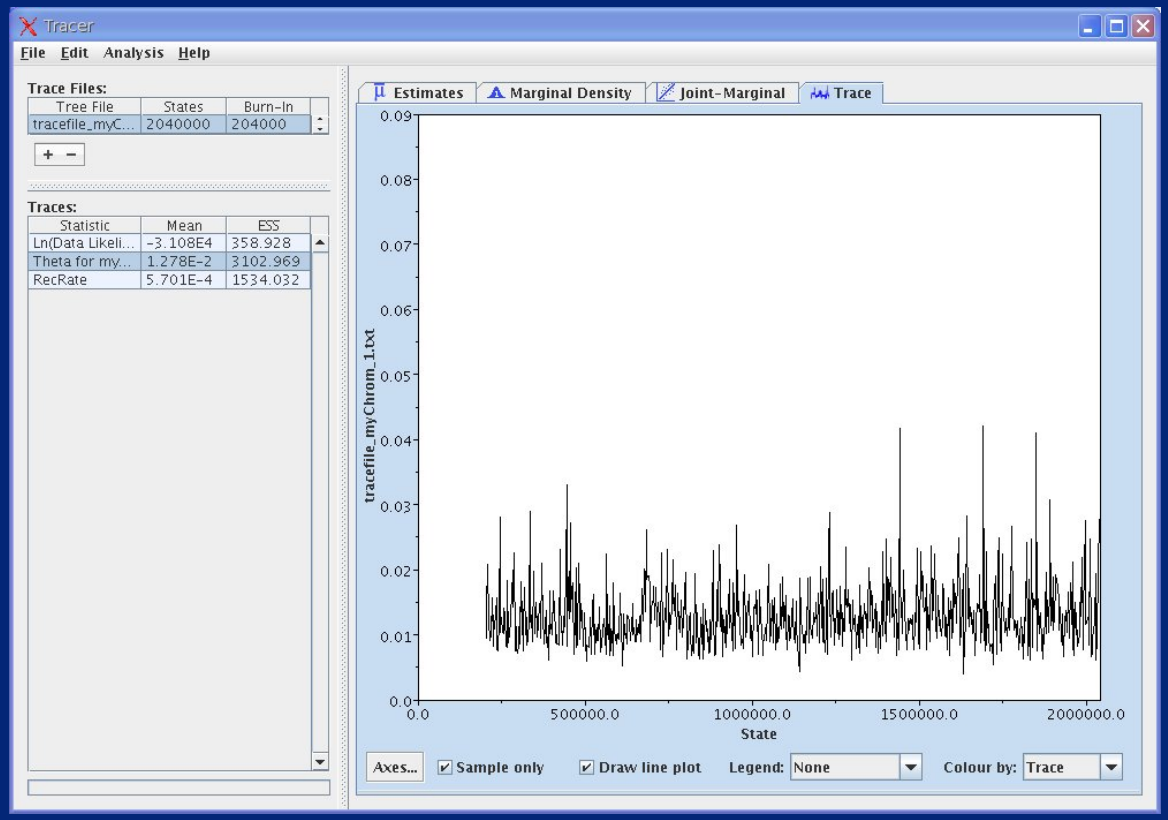
Has the run converged?

- Success can be measured as convergence
- However, a stuck search may appear to converge



Courtesy of Elizabeth Thompson

TRACER analysis



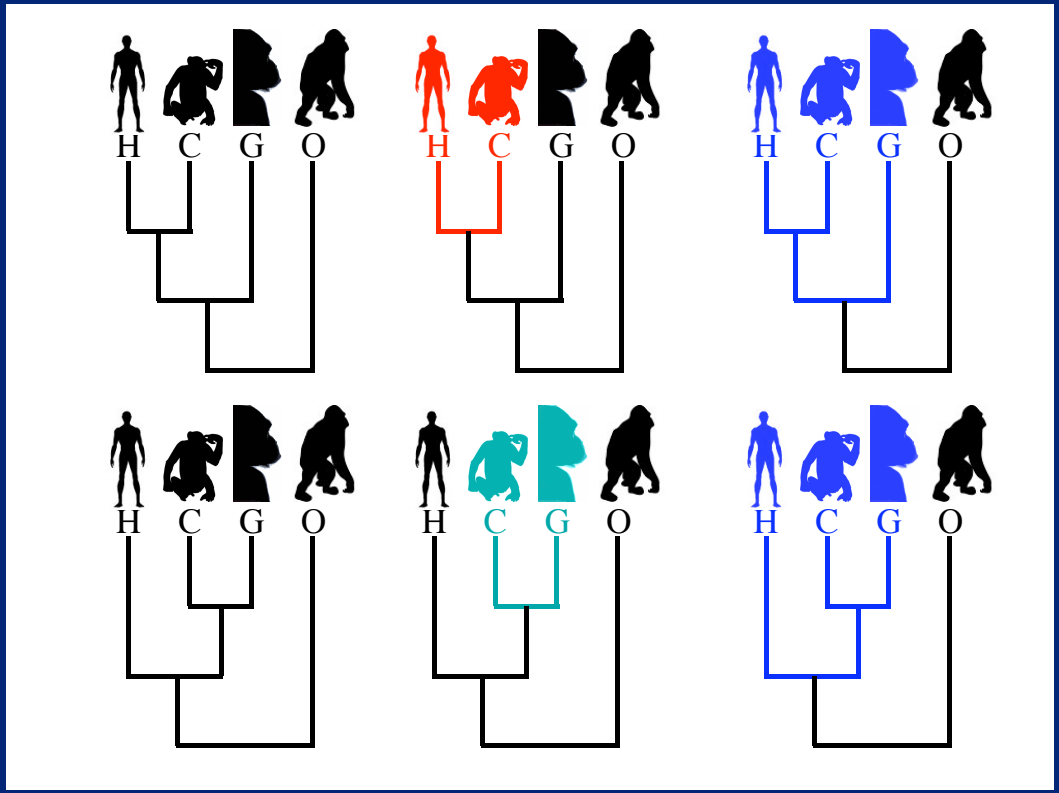
TRACER analysis

- TRACER program of Rambaut and Drummond
- Traces of parameter values over time
- Histograms of posterior probabilities
- ESS (Effective Sample Size) statistic
- Compatible with BEAST, LAMARC, MIGRATE
- IM/IMa have similar functions built in

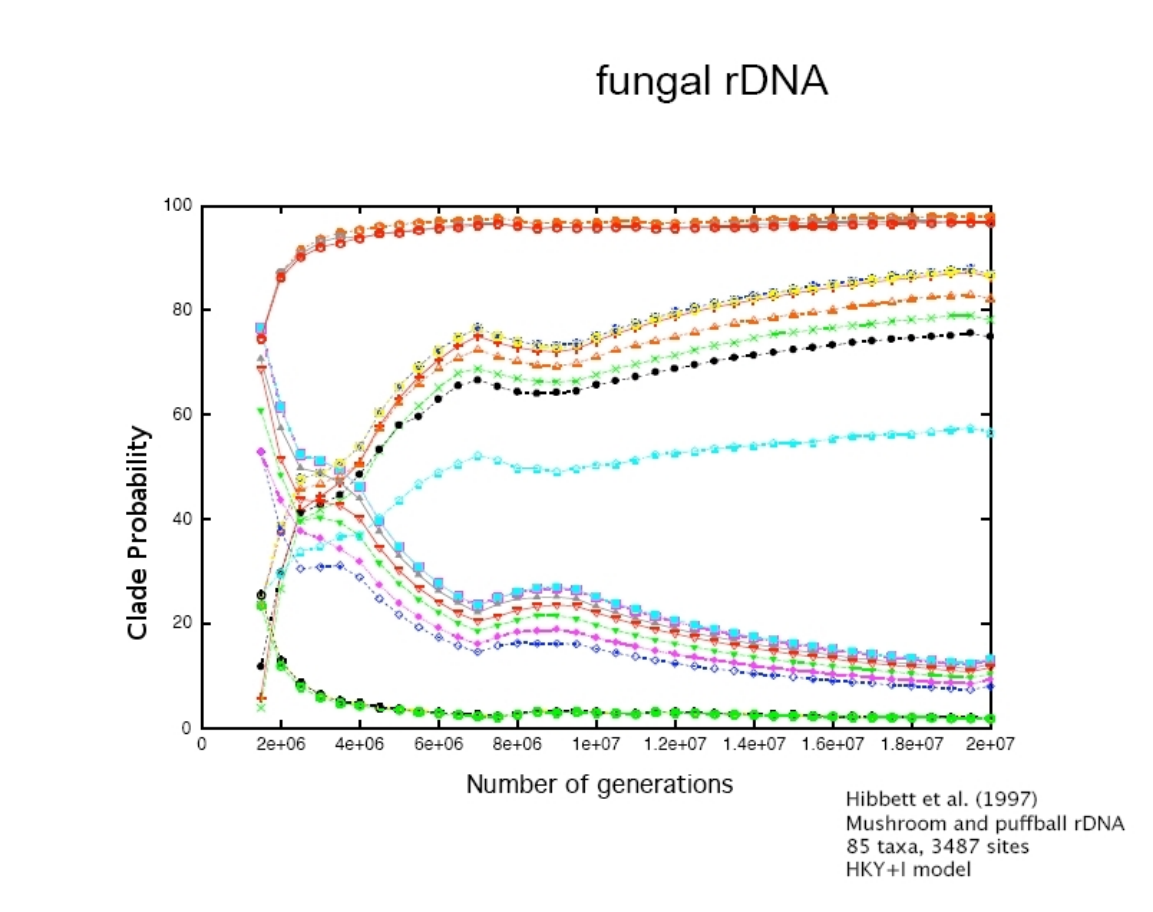
Effective Sample Size

- Effective Sample Size (ESS) corrects sample size for autocorrelation
- ESS = runlength divided by autocorrelation time
- Low ESS is strong evidence of a too-short run
- Unfortunately, high ESS does not guarantee convergence

Clade probabilities with AWTY



Convergence for clade probability

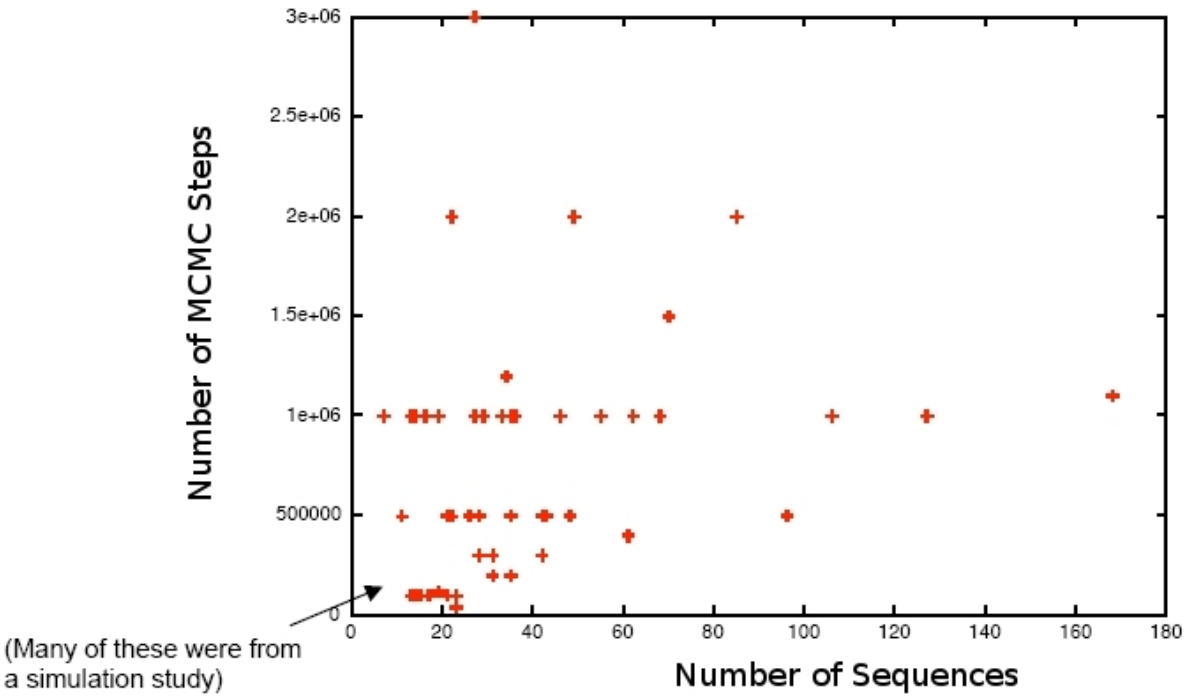


Courtesy of David Swofford, AWTY program

How long are people running their chains?

Literature search for chain lengths used with MrBayes:

- Molecular Biology and Evolution (17 papers)
- Molecular Phylogenetics and Evolution (33 papers)
- Taxon (4 papers)



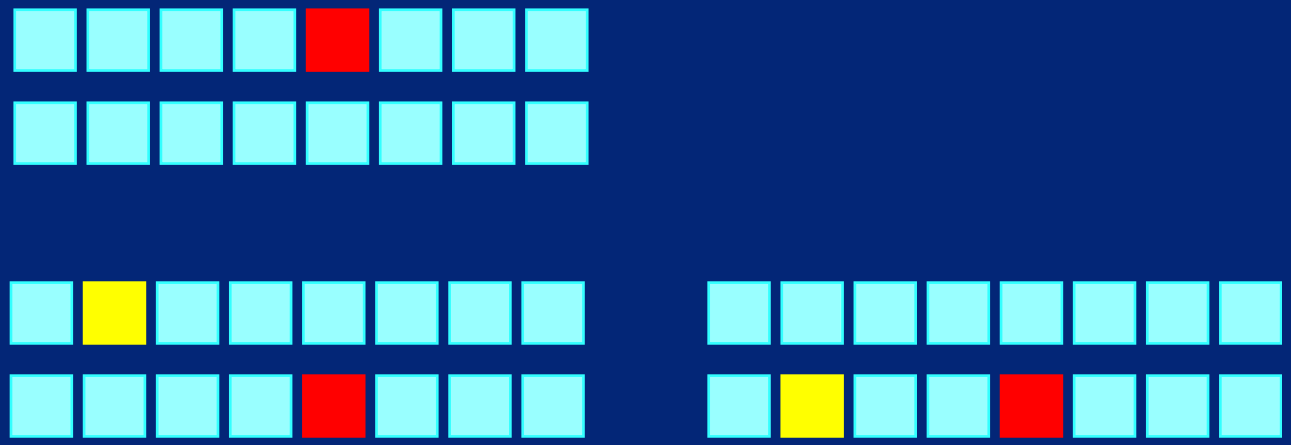
(Many of these were from a simulation study)

Courtesy of David Swofford, circa 2004

boa: R package for MCMC convergence assessment

- Tests for convergence of Bayesian analyses
- Not yet in common use for geneology samplers, but probably should be!
- Smith, BJ (2007) J Statistical Software 21.
- http://www.public-health.uiowa.edu/boa/

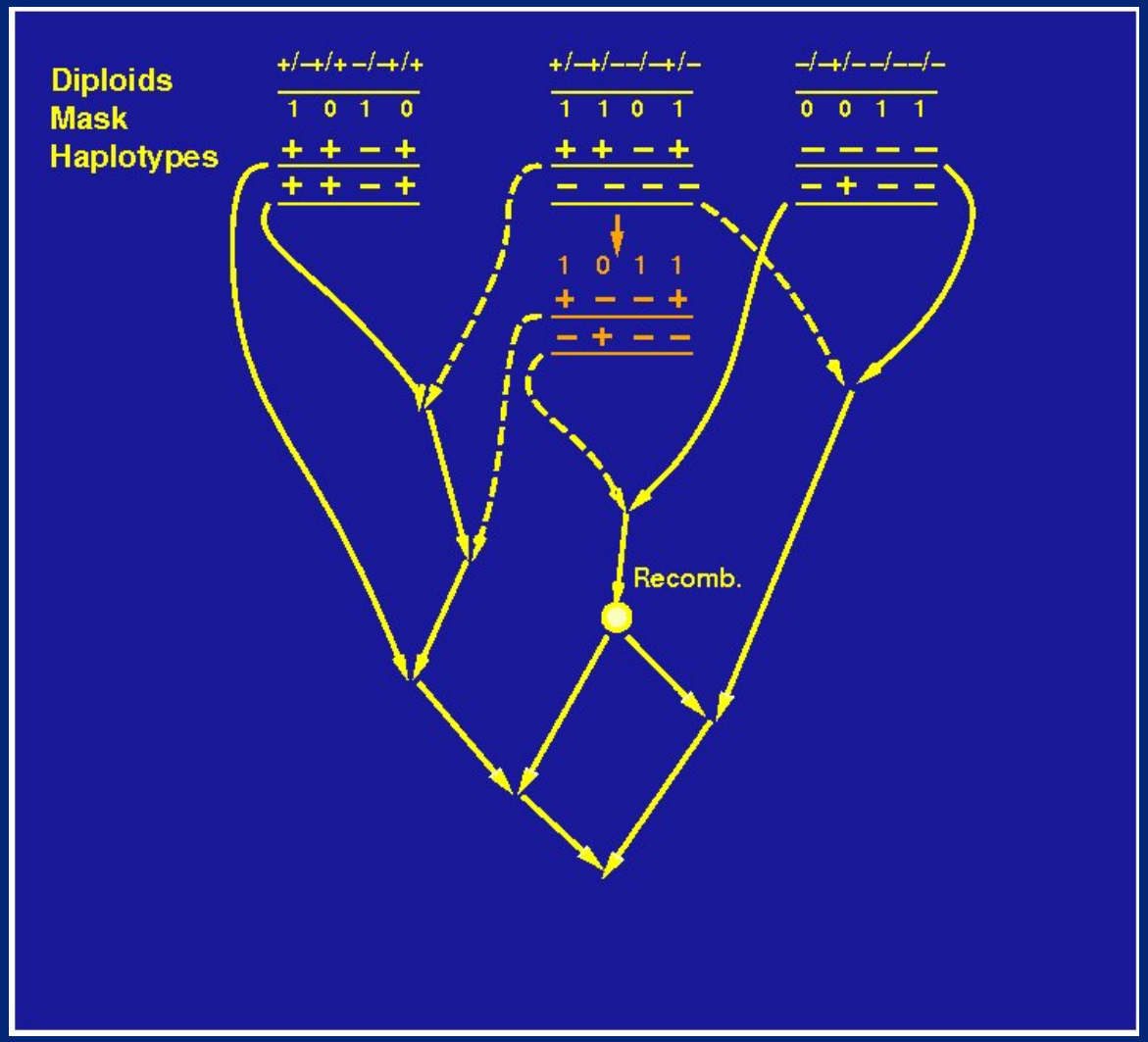
A troublesome example: phase inference



A troublesome example: phase inference

- Some data lack phase information
- Inferring "one best phase" may lead to bias
- MCMC can search simultaneously over:
  - Trees based on current phase assignment
  - Phase assignment based on current tree

A troublesome example: phase inference



A troublesome example: phase inference

Strategy	$\frac{Estimated\Theta}{True\Theta}$
Correct haplotypes	1.0
No haplotype inference	1.65
Haplotype reassignment 10%	1.28
Haplotype reassignment 20%	1.23
Haplotype reassignment 50% (10x search)	1.15
Reassignment with rearrangement	1.33

A troublesome example: phase inference

Strategy	$\frac{Estimated\Theta}{True\Theta}$
Correct haplotypes	1.0
No haplotype inference	1.65
Haplotype reassignment 10%	1.28
Haplotype reassignment 20%	1.23
Haplotype reassignment 50% (10x search)	1.15
Reassignment with rearrangement	1.33
Haplotype reassignment 20%, heated	1.03

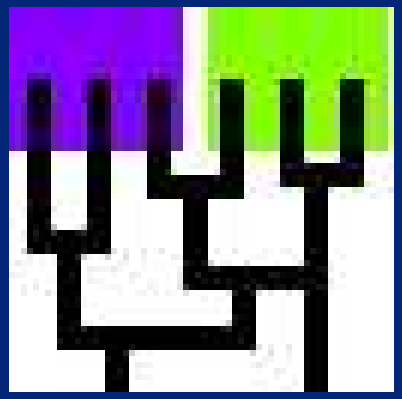
Final thoughts

- Coalescent studies should be carefully designed:
  - Data collection
  - Mutational model
  - Population model
  - Details of analysis
- The strongest studies combine multiple approaches
- Pay as much or more attention to error bars as point estimates

Thanks to

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

Web site



http://evolution.gs.washington.edu/lamarc.html