MONASH University

Science

Genealogy samplers

Anders Gonçalves da Silva

Population and landscape genomics workshop – Day 2

25 March 2014

CBA – ANU

Outline

- Motivate the need for genealogy samplers
- Examine the role of mutations in the coalescent
- A quick tour of genealogy samplers
- Genealogy sampler in action: *Aedes aegypti* gene flow networks
- Setting a genealogy sampler to run – Migrate-N
- Some practical considerations
- A future prospect

Question?

- Why are so many coalescent parameters scaled by mutation rate?

Probability of a genealogy

$$Coalescent = P(G|\theta)$$

Estimating *G…*

HARD!!!

What do we do?

- We are certain that:
  - There is a genealogy
  - We have no way of knowing it with certainty…
- We have computers:
  - We can simulate genealogies…

Which genealogies do we simulate?

Number of genes	Number of genealogies
2	1
3	3
4	18
5	180
6	2700
8	1.6×10^{6}
20	5.6×10^{29}

Borrow from phylogenetics

$$Phylogeny = P(Data|G)$$

Use MCMC to focus on high value trees

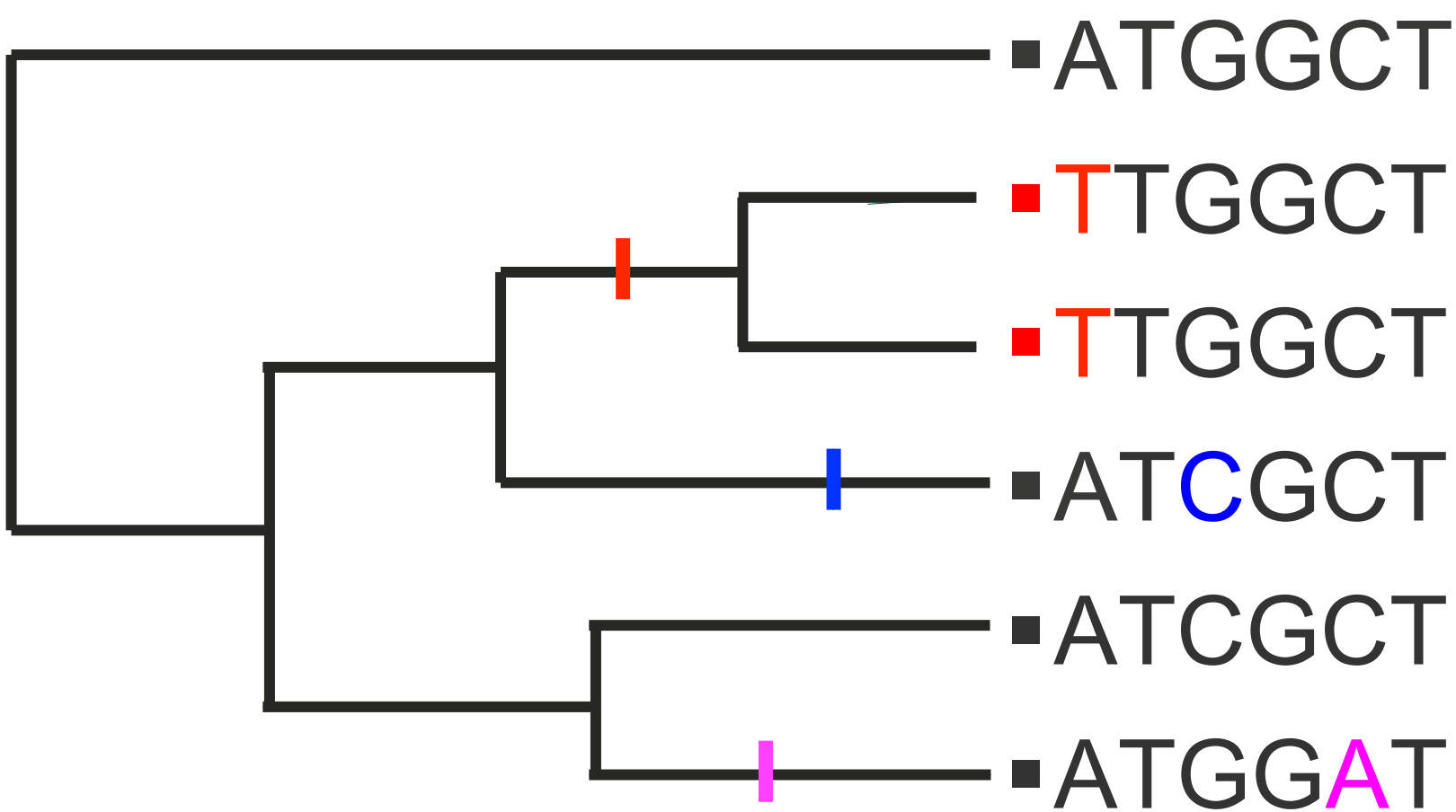
What is the data?

- Genetic data:
  - –Sequences
  - –Microsatellites
  - –SNPs
  - –Allozymes

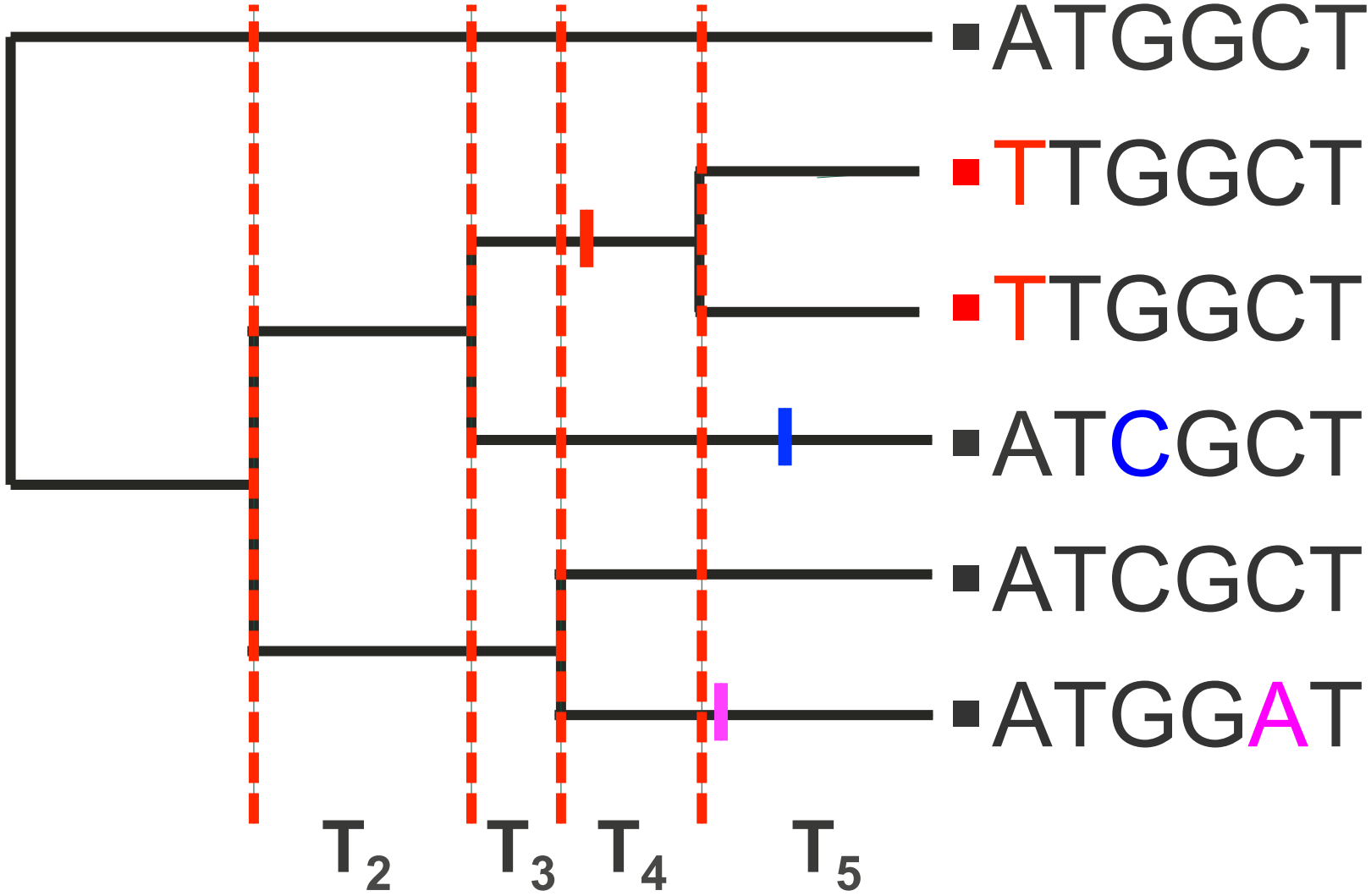
We need a mutation model…

1. Infinite alleles model
2. Infinite sites model
3. Finite sites model

With a mutational model we can…



Consequence: Time measured in mutations



If we put it together…

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

Genealogy samplers: Genetree

- Data:
  - Sequences
- ML estimates of parameters:
  - Mutation rates
  - Migration rates
  - Population rates
  - Distribution of TMRCA
  - Age of mutations
- By RC Griffiths:
  - http://www.stats.ox.ac.uk/~griff/software.html

Genealogy Samplers: BEAST

- Data:
  - Sequences, codons, morphological traits
- Bayesian estimates of parameters:
  - Population size (scaled by mutation rate)
  - Population growth
  - Mutation rate (with multiple time points)
  - Skyline plots
  - Relaxed molecular clock
- A Rambaut and A Drummond:
  - http://beast.bio.ed.ac.uk

Genealogy samplers: IM/IMa2

- Data:
  - Sequences, microsatellites
- Bayesian estimates of parameters:
  - Population size (scaled by mutation rate)
  - Immigration rates (scaled by mutation rate)
  - Size of ancestral population
  - Time of divergence between populations
- J Hey:
  - https://bio.cst.temple.edu/~hey/software/software.htm

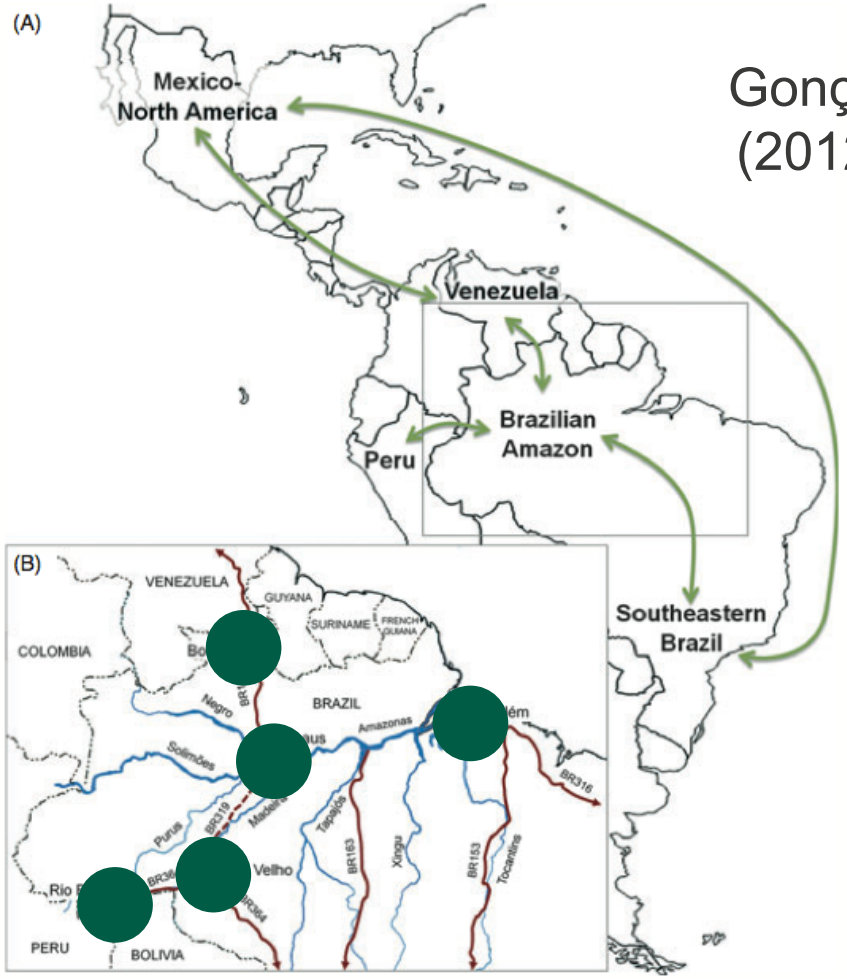
Genealogy samplers: LAMARC

- Data:
  - SNPs, microsatellites, sequence data, allozymes
- Bayesian and ML estimates of parameters:
  - Population size (scaled by mutation rate)
  - Migration (scaled by mutation rate)
  - Population growth rates
  - Recombination rates
- M Kuhner et al.:
  - http://evolution.genetics.washington.edu/lamarc/lamarc_download.html

Genealogy samplers: Migrate-N

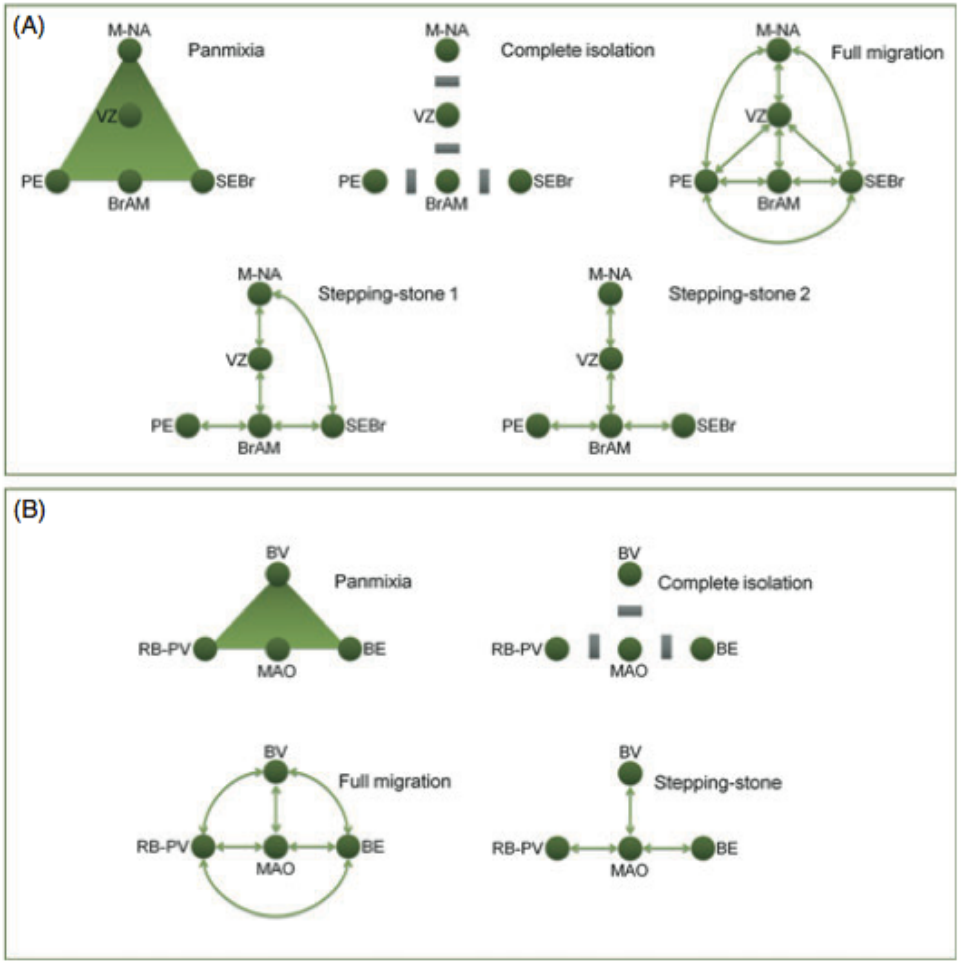
- Data:
  - SNPs, sequence data, microsatellites, allozymes
- Bayesian and ML estimates of:
  - Population size (scaled by mutation rate)
  - Immigration rates (scaled by mutation rate)
- *Bayesian estimates of migration networks*
- P Beerli:
  - http://popgen.sc.fsu.edu/Migrate/Migrate-n.html

Bayesian estimate of migration networks

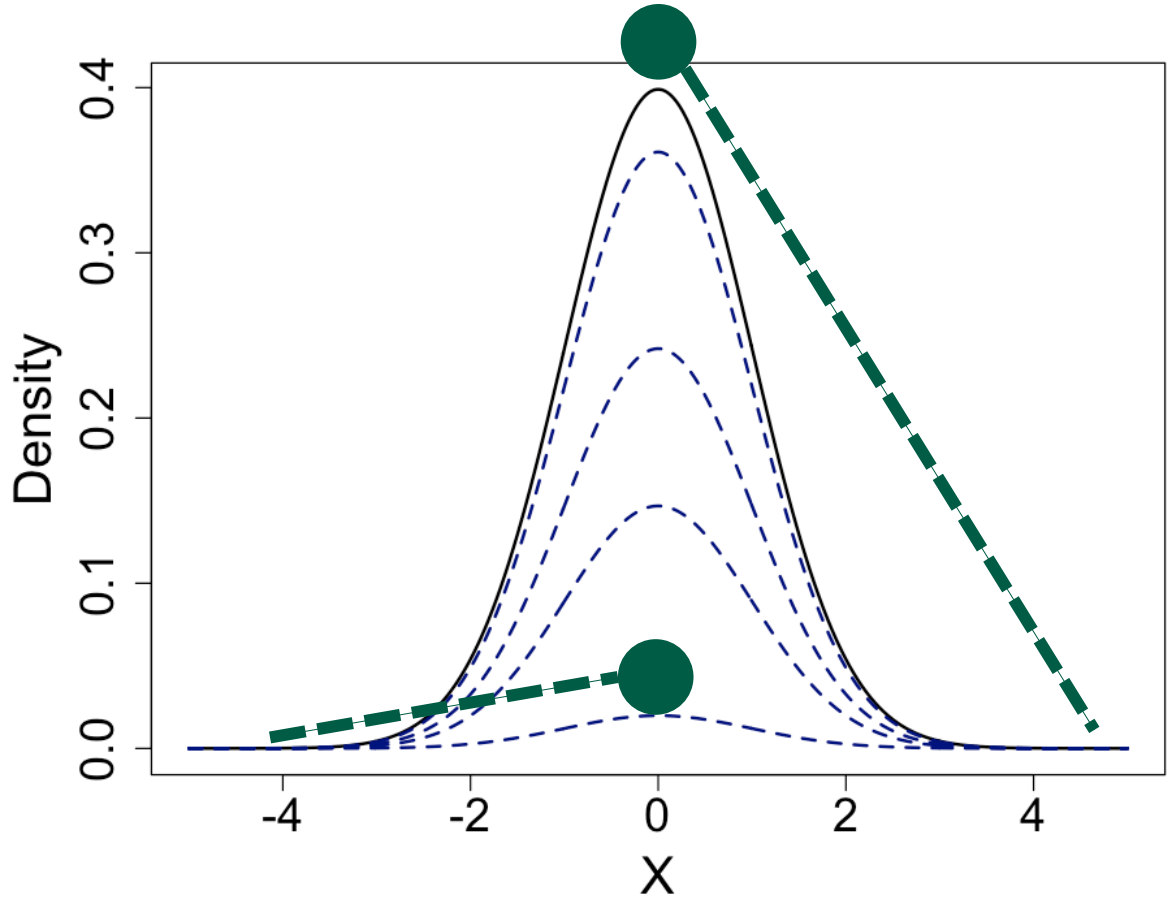


Gonçalves da Silva et al. (2012). *Evol Appl* 5: 664

Bayesian estimate of migration networks



MC^3 or Metropolis-Coupled MCMC



Posterior support for migration network

Spatial scale	Model	Marginal LogLike	LogBF	Posterior Probability	Rank
Continent	Stepping-stone (15)	-692.40	0.00	0.99	1
	Full migration (25)	-709.19	-33.58	10^{-10}	3
	Panmixia (1)	-725.38	-65.96	10^{-15}	4
Amazon	Full-migration (16)	-548.71	0.00	0.94	1
	Panmixia (1)	-551.59	-5.76	0.05	2
	Isolation	-657.24	-217.06	10^{-48}	4

Calculating log(Bayes Factor)

$$log(BF) = 2(MLogLike_{M1} - MLogLike_{M2})$$

$$log(BF) = 2(-709.19 - (-692.40))$$

Calculating posterior probability

$$P(Model_1|Data) = \frac{exp(MLogLike_{M1})}{\sum_{i=1}^{N} exp(MLogLike_{Mi})}$$

$$P(StepStone|Data) = \frac{exp(MLogLike_{StepStone} - MLogLike_{max})}{\sum_{i=1}^{N} exp(MLogLike(M_i - MLogLike_{max}))}$$

$$P(StepStone|Data) = \frac{exp(-692.40 - (-692.40))}{\sum_{i=1}^{N} exp(0 + (-16.79) + (-22.00) + (-32.98) + (-254.71)))}$$

$$P(StepStone|Data) \sim 0.999$$

http://popgen.sc.fsu.edu/Migrate/Tutorials/Entries/2010/7/12_Day_of_longboarding.html

LET'S GIVE MIGRATE-N A GO

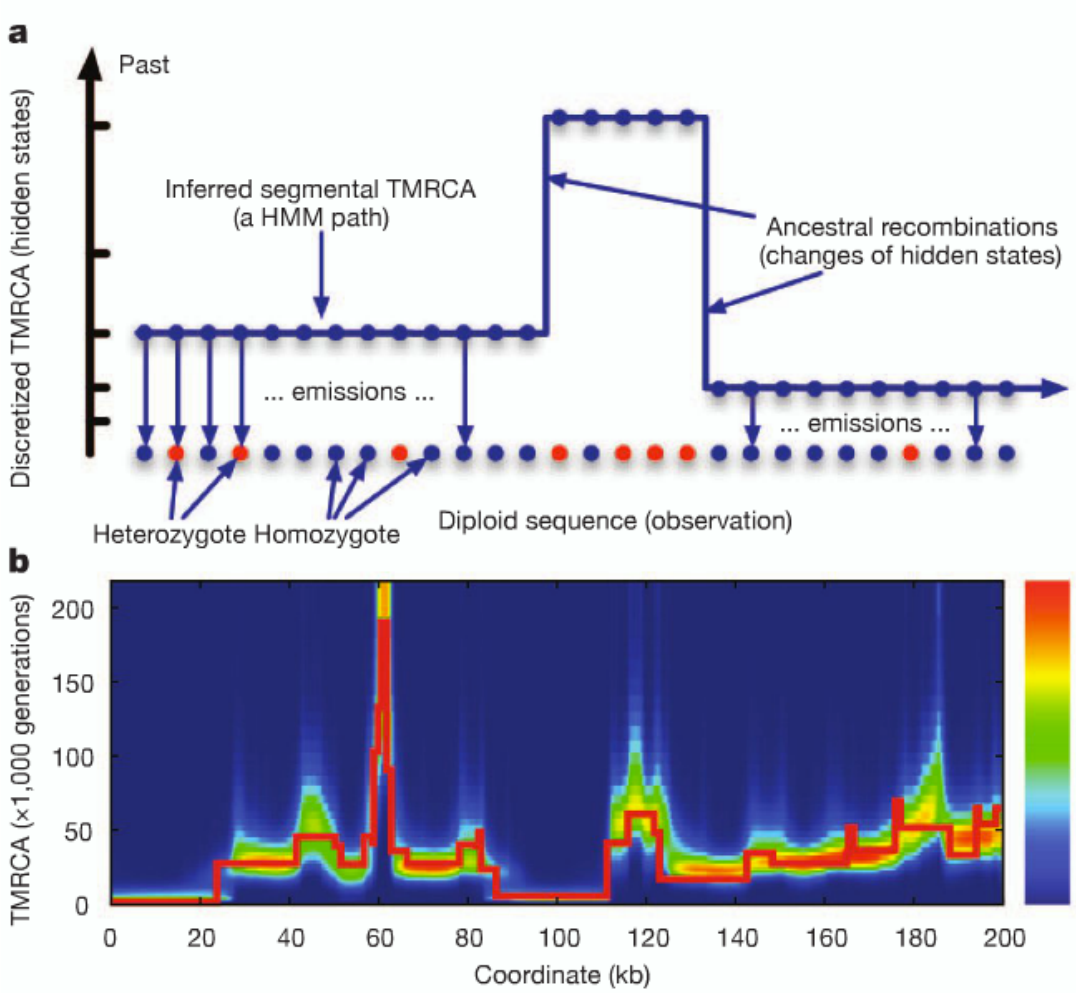
Basic recommendations: Mary Kuhner

- Use Bayesian approach if parameter is close to 0
- Rich data either ML or Bayes approach works well
- If you have poor data: do you want to be doing analysis anyway?
  - Although, Bayes performs better under sparse data
- When using Bayes, must be careful with priors.
  - Choose them carefully
  - Justify them thoroughly!

Major practical problems

- How long to run MCMC for?
- How long is the burnin?
- How many replicates?
- Should I use heating?
- How good are the defaults?

What the future holds



Li and Durbin (2011) *Nature* 475: 493

Further reading

- Kuhner, M. K. (2009). Coalescent genealogy samplers: windows into population history. *Trends in Ecology & Evolution*, *24*(2), 86–93.
- Beerli, P., & Palczewski, M. (2010). Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics*, *185*(1), 313–463.
- Beerli, P. (2009). How to use Migrate or why are Markov chain Monte Carlo programs difficult to use? In *Population genetics for animal conservation* (pp. 42–74). Cambridge, UK: Cambridge University Press.
- 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.