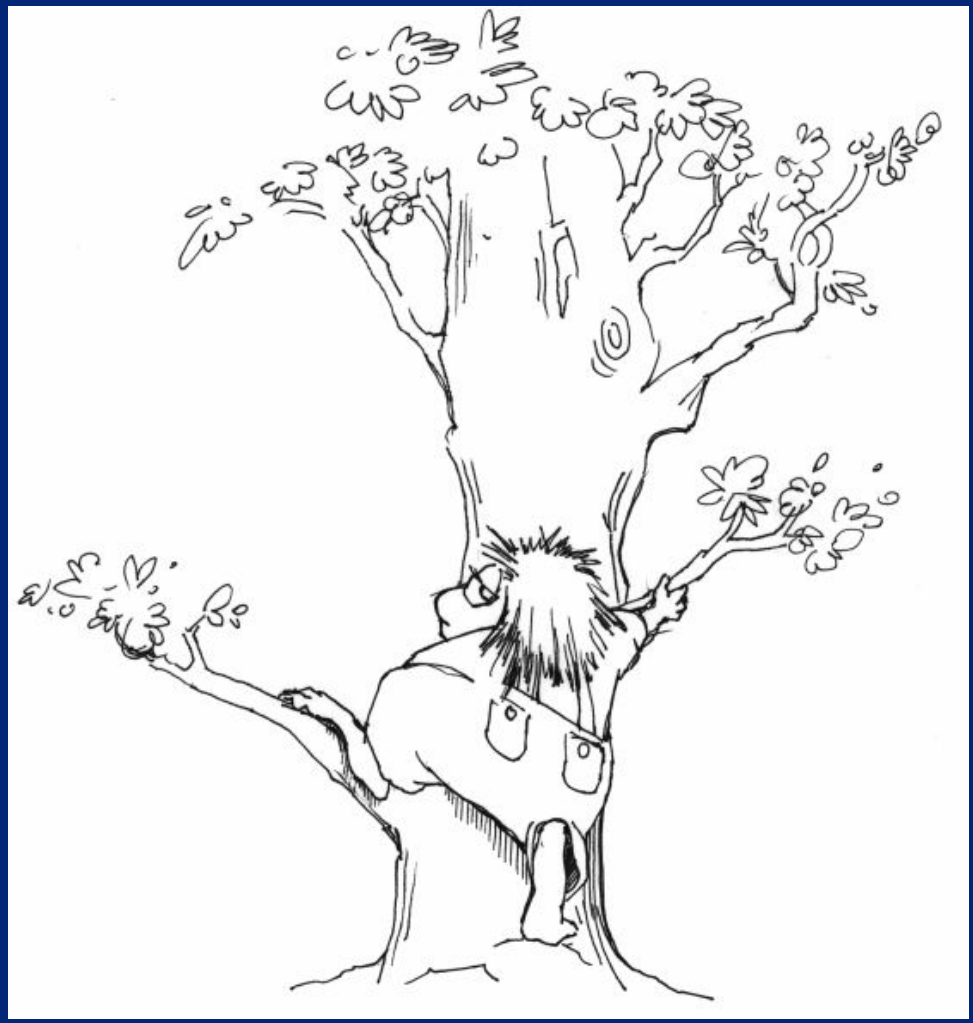
Extensions of the coalescent

Peter Beerli
Genome Sciences
University of Washington
Seattle WA



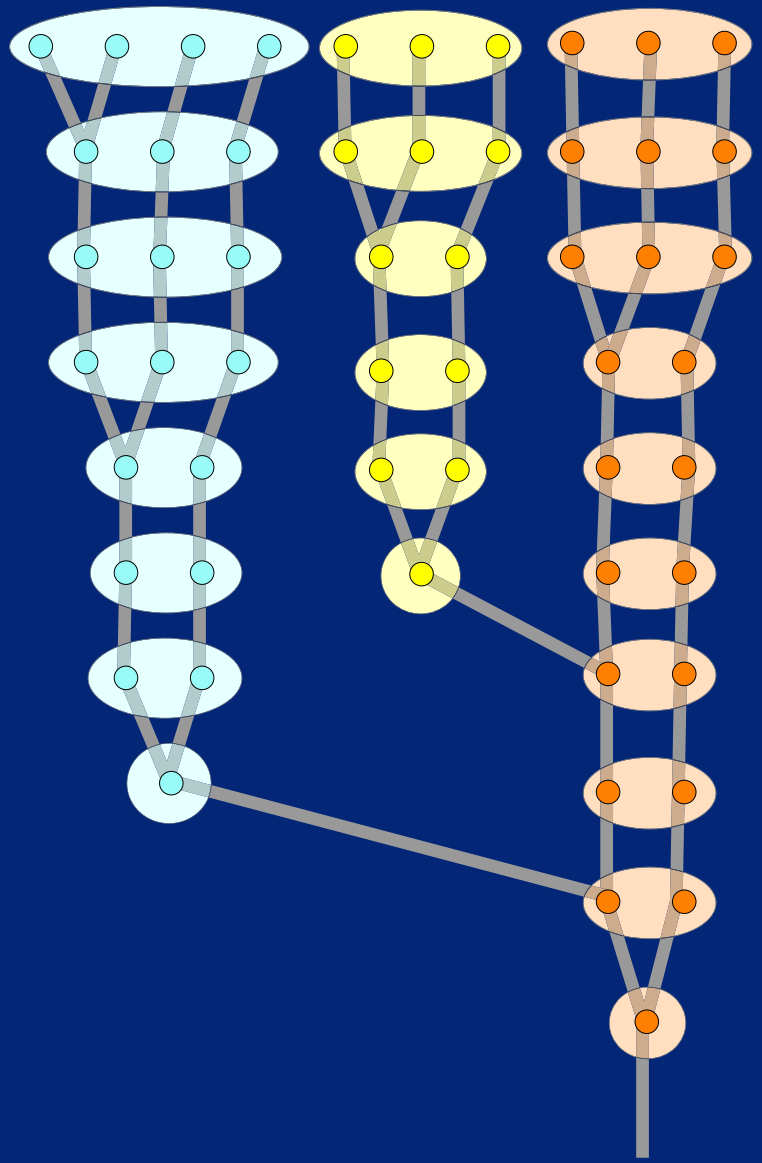
Outline

1. **Introduction to the basic coalescent**

2. **Extensions and examples**

   - Different approaches to estimate parameters of interest
   - Population growth
   - Gene flow
   - Recombination
   - Population divergence
   - Selection
   - Combination of some of the population genetics forces

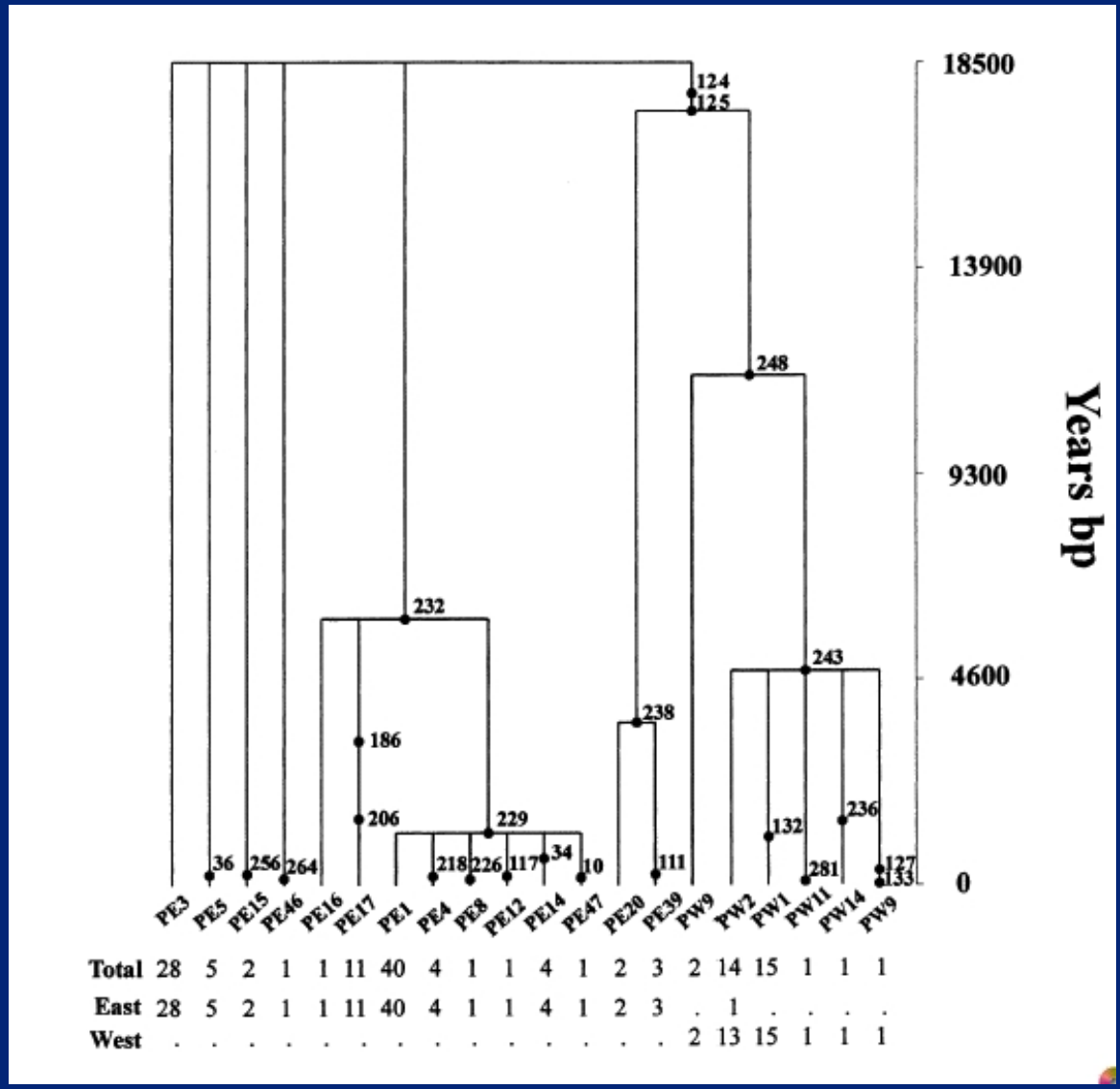
An alternative approach: Griffiths-Tavaré algorithm



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

Dating mutations events using *Genetree*

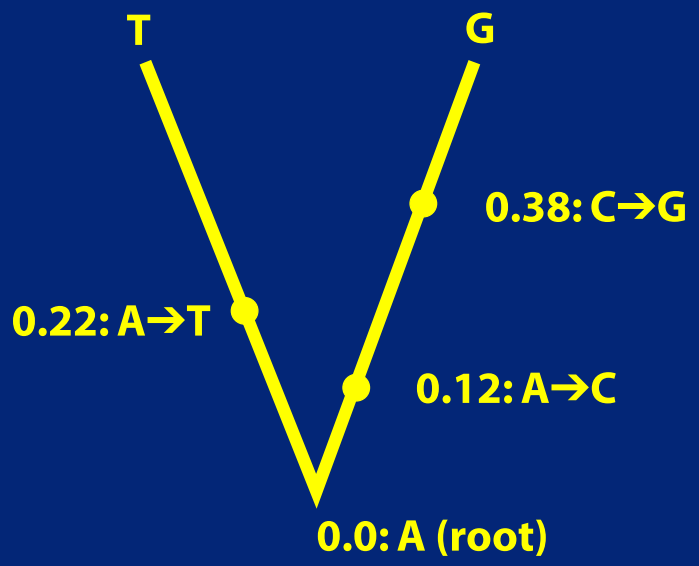
Milot et al. (2000)



Other alternatives

- Wilson and Balding (1998), Beaumont (1999): microsatellite specific
- Nielsen (2001): mutations as missing data

Both methods treat mutations as additional events on the genealogies.



Variants and extension of the coalescent

- Population size changes over time
- Gene flow among multiple population
- Recombination rate
- Selection
- Divergence

A general approach to these extensions

$$\mathrm{p}(G|\text{parameter}) = \prod_{\text{All time intervals } u_j} f(\text{waiting time})\ \mathrm{Prob}\ (\text{event happens})$$

A general approach to these extensions

$$\mathrm{p}(G|\text{parameter}) = \prod_{\text{All time intervals } u_j} f(\text{waiting time})\ \mathrm{Prob}\ (\text{event happens})$$

$$\mathrm{p}(G|\Theta) = \prod_{u_j} f(\Theta)\ \frac{2}{\Theta}$$

A general approach to these extensions

$$\mathrm{p}(G|\text{parameter}) = \prod_{\text{All time intervals } u_j} f(\text{waiting time})\ \mathrm{Prob}\ (\text{event happens})$$

$$\mathrm{p}(G|\Theta) = \prod_{u_j} f(\Theta)\ \frac{2}{\Theta}$$

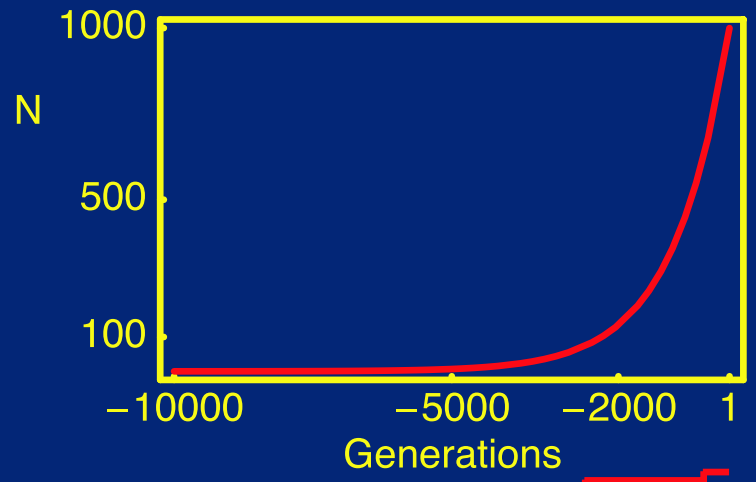
$$\mathrm{p}(G|\Theta,\alpha,\beta) = \prod_{u_j} f(\Theta)\ f(\alpha)f(\beta) \begin{cases} \frac{2}{\Theta} & \text{if event is a coalescence,} \\ \mathrm{Prob}\ (\alpha) & \text{if event is of type } \alpha, \\ \mathrm{Prob}\ (\beta) & \text{if event is of type } \beta. \end{cases}$$

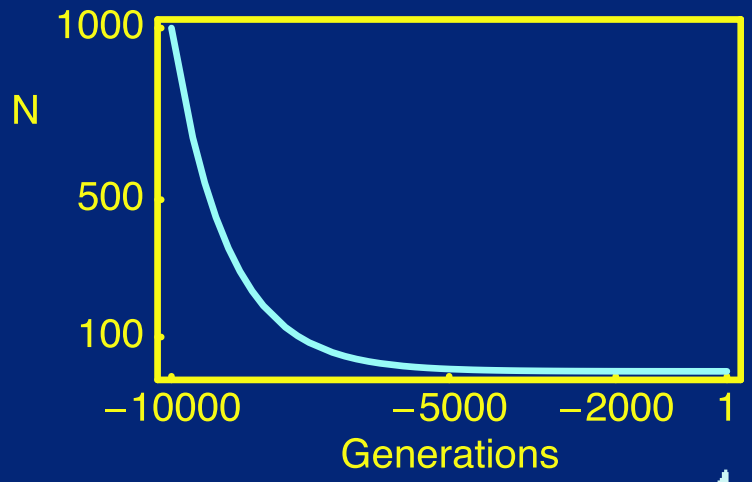
Variable population size

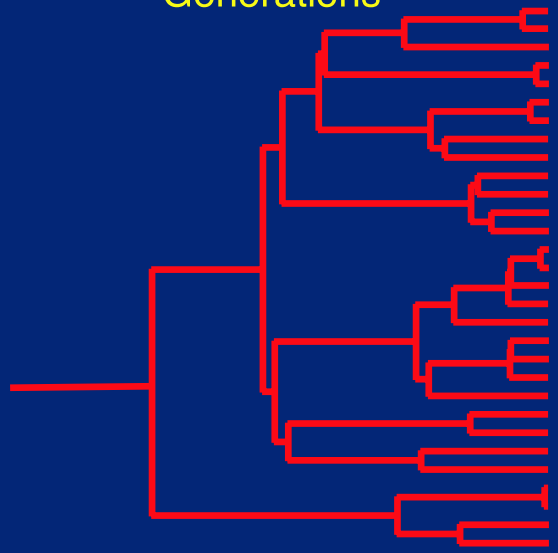
- In a small population lineages coalesce quickly
- In a large population lineages coalesce slowly

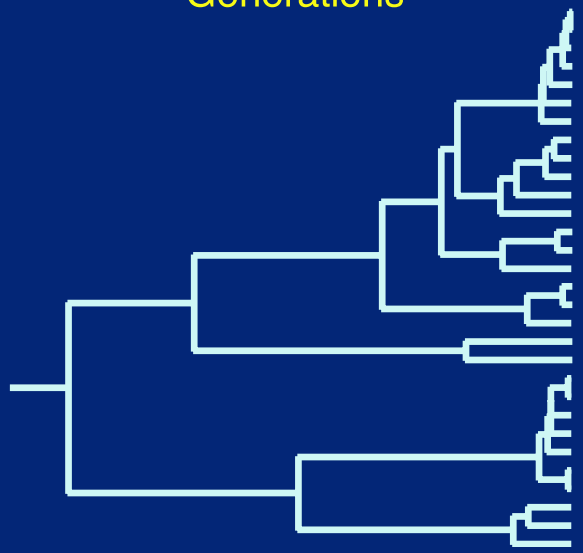
This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the population size Θ.

Exponential population size expansion or shrinkage

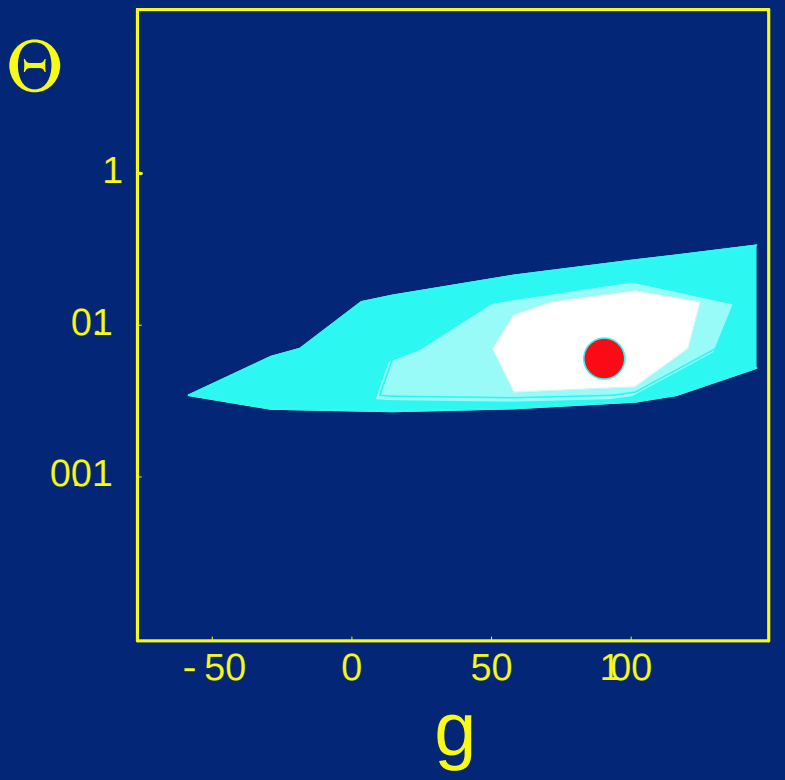






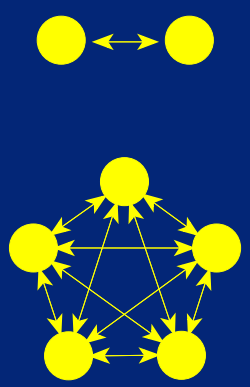


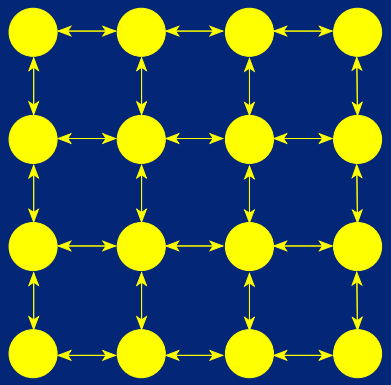
Grow a frog

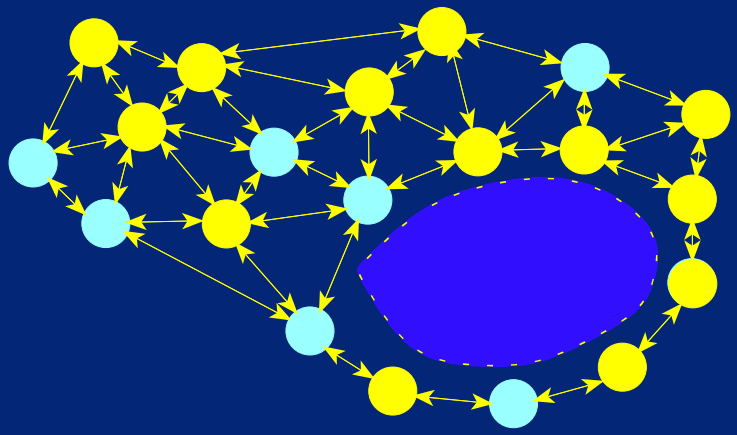


Mutation Rate	Population sizes	
	-10000 generations	Present
10^{-8}	$8,300,000$	$8,360,000$
10^{-7}	$780,000$	$836,000$
10^{-6}	$40,500$	$83,600$

Gene flow



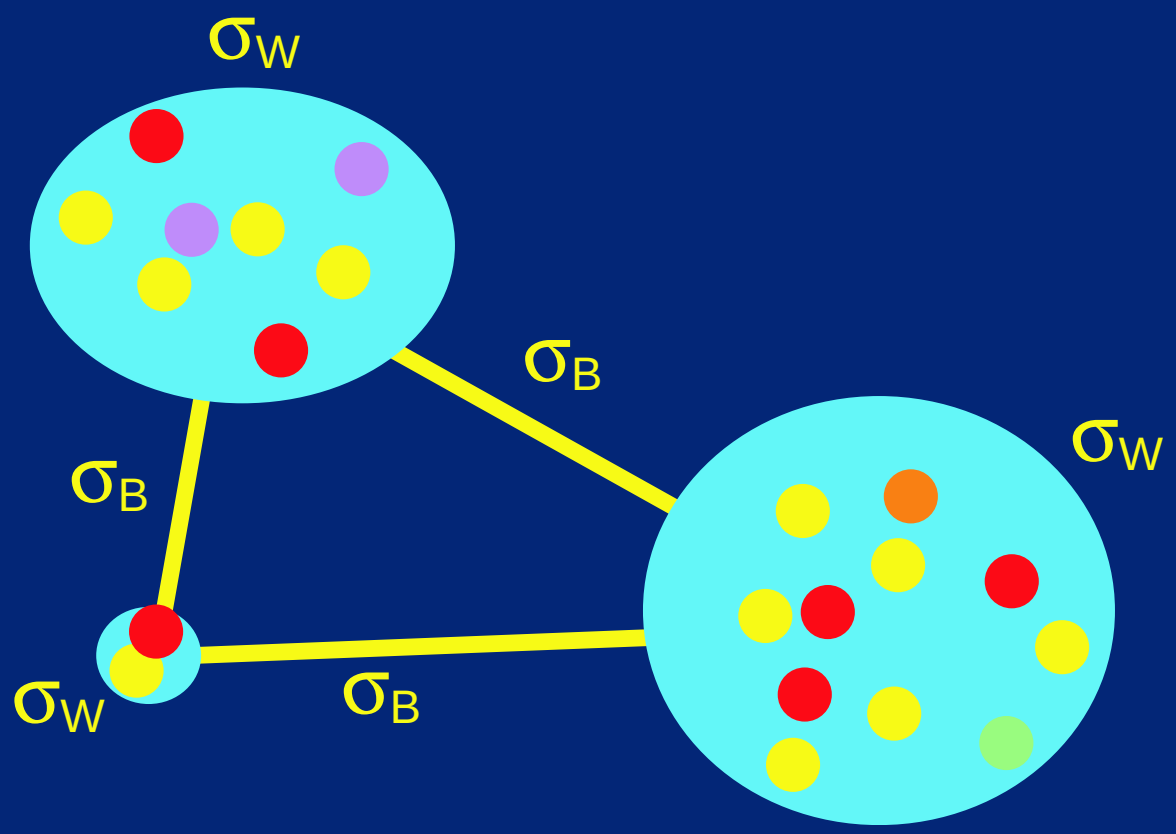




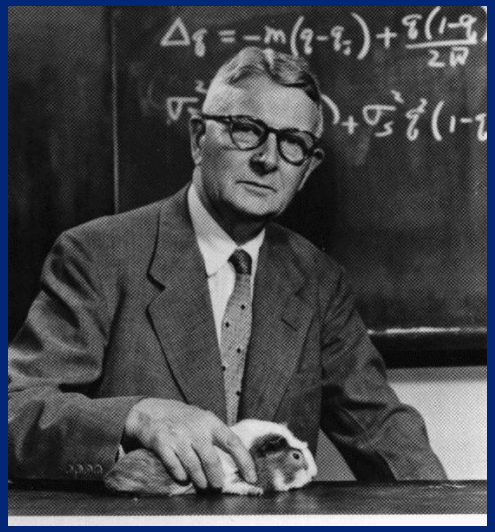
$$\mathrm{p}(G|\boldsymbol{\Theta},\mathbf{M}) = \prod_{u_j}\left(\prod_{i}^{\text{pop.}} g(\Theta_i,\mathbf{M}_{\cdot i})\right)\begin{cases}\frac{2}{\Theta} & \text{if event is a coalescence,}\\ M_{ji} & \text{if event is a migration from } j \text{ to } i.\end{cases}$$

Gene flow: What researchers used (and still use)

F_{ST}



What researchers used (and still use)



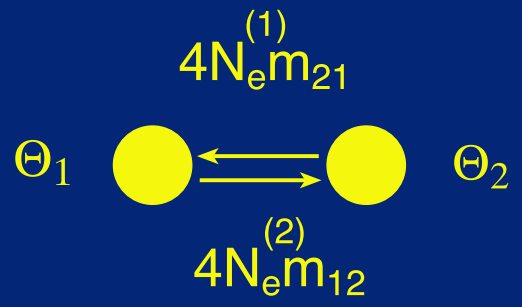
Sewall Wright showed that

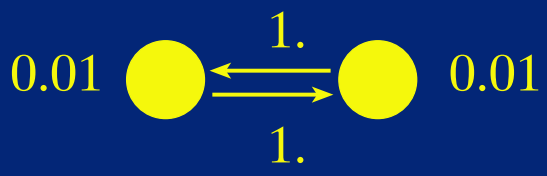
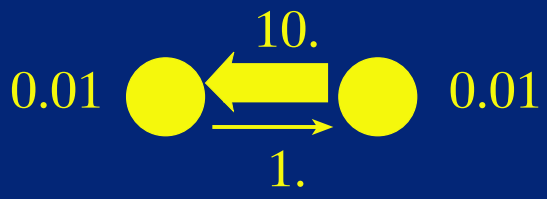
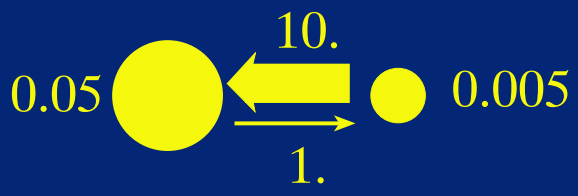
$$F_{ST} = \frac{1}{1 + 4Nm}$$

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



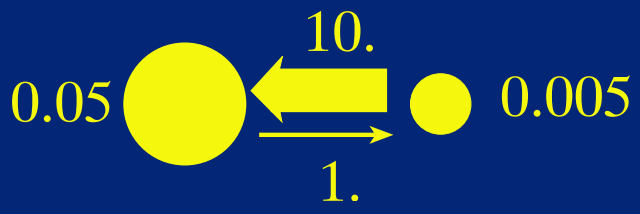
True values	Estimated values
	1.14±0.77
	7.80±22.20
	11.46±18.54

Maximum Likelihood method to estimate gene flow parameters

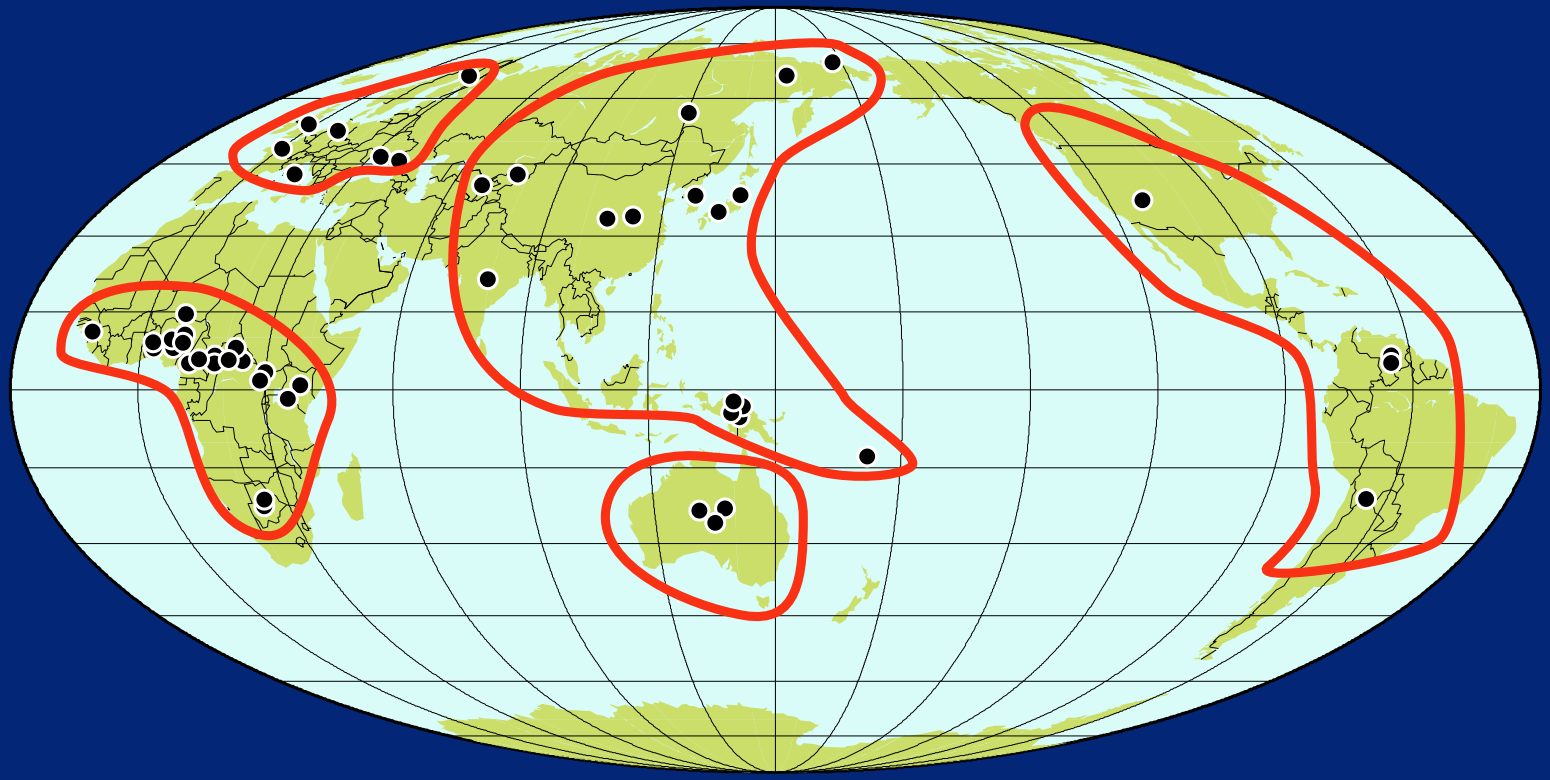
(Beerli and Felsenstein 1999)

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

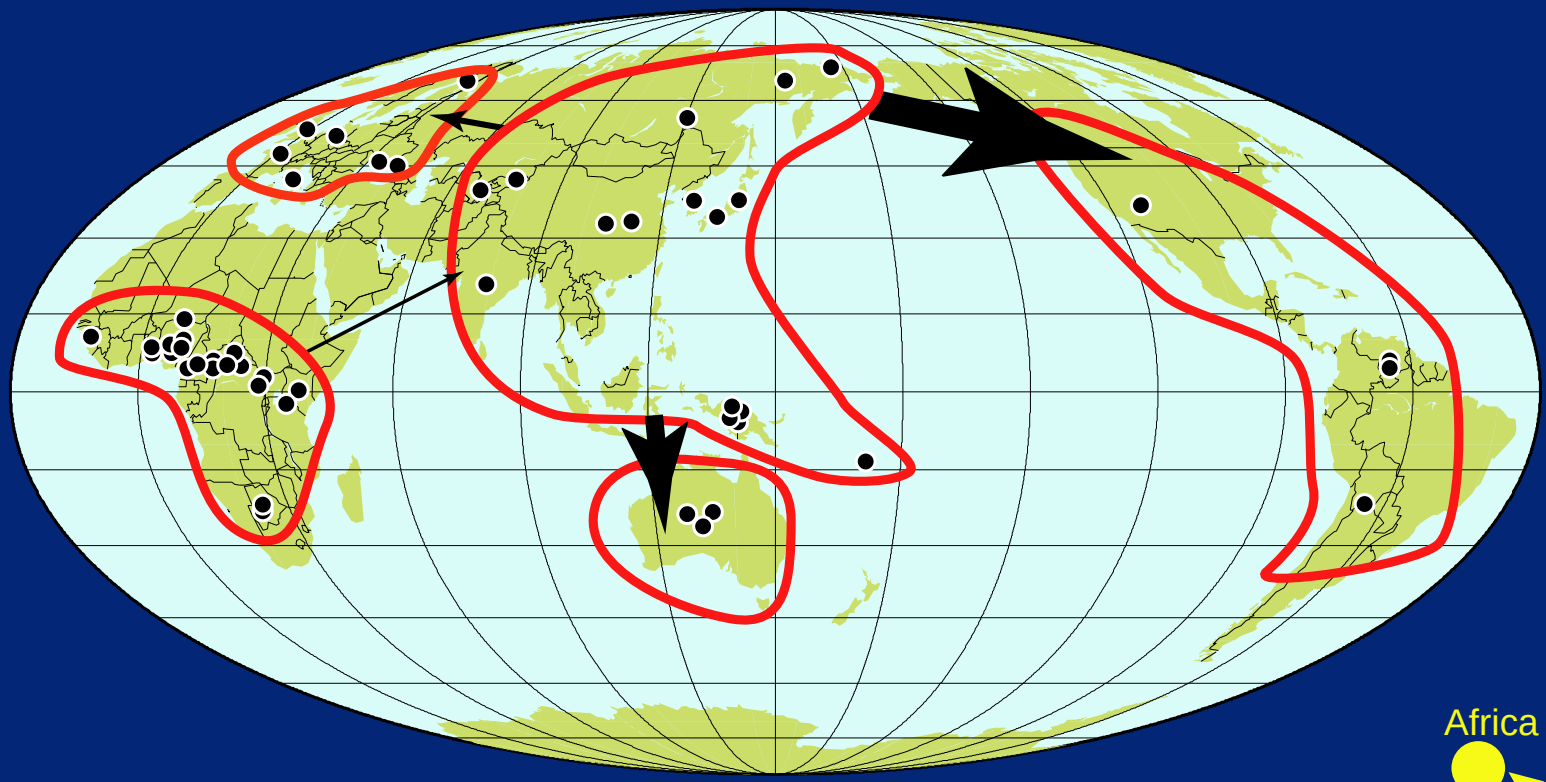


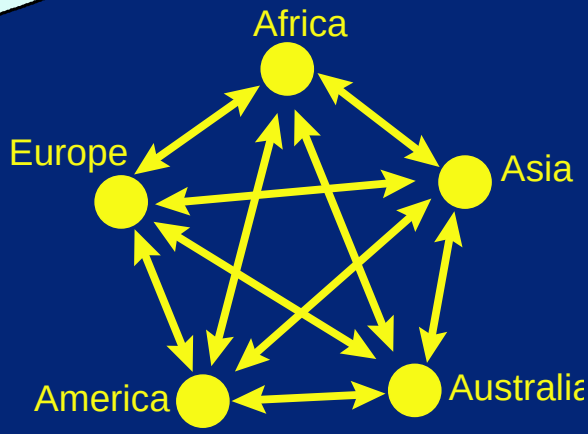
Complete mtDNA from 5 human "populations"



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

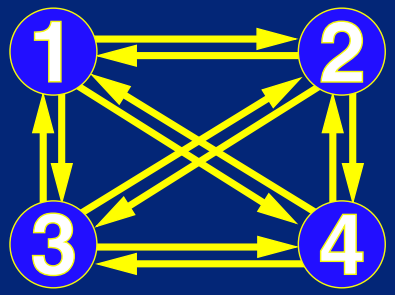
Full model: 5 population sizes + 20 migration rates



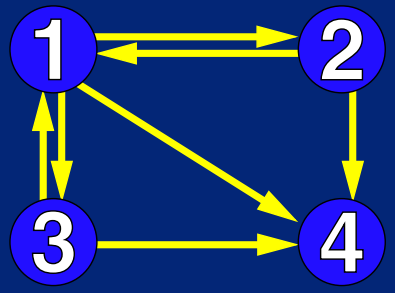


Some examples of possible migration models

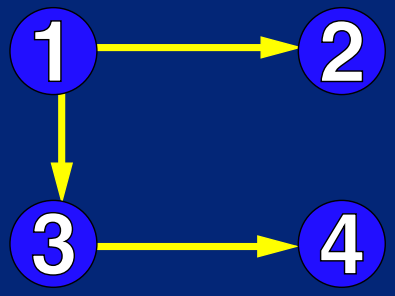
Full sized



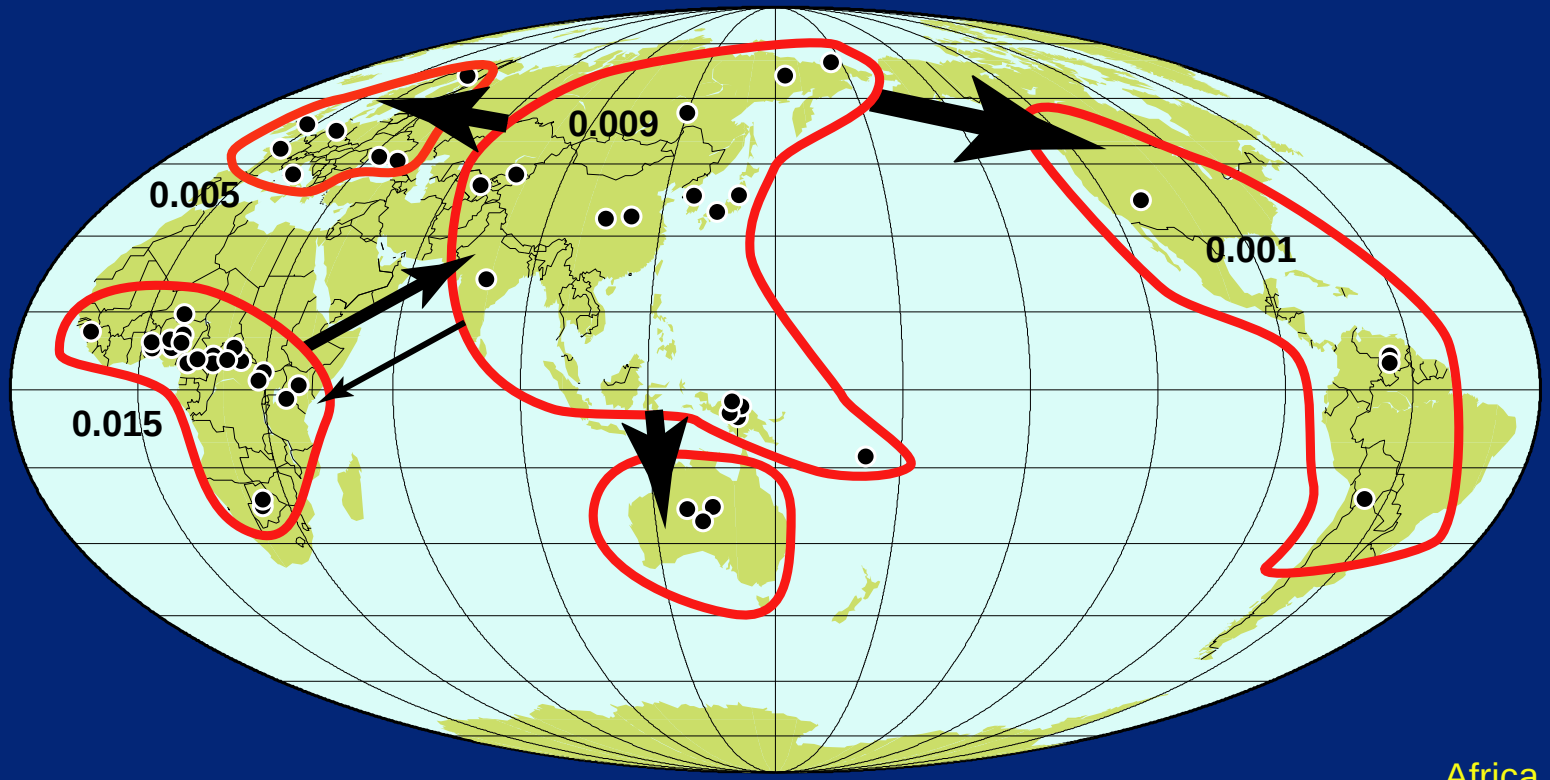
Mid sized

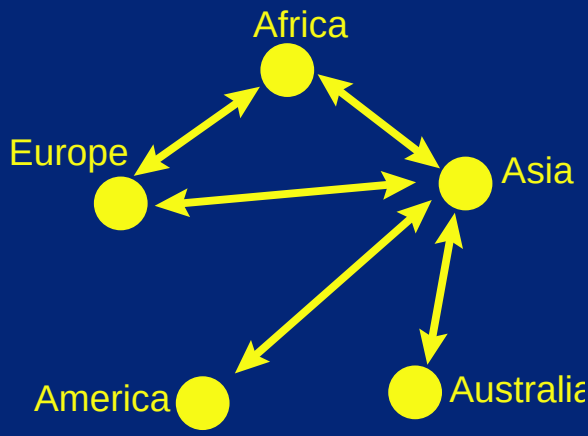


Economy



Restricted model: only migration into neighbors allowed



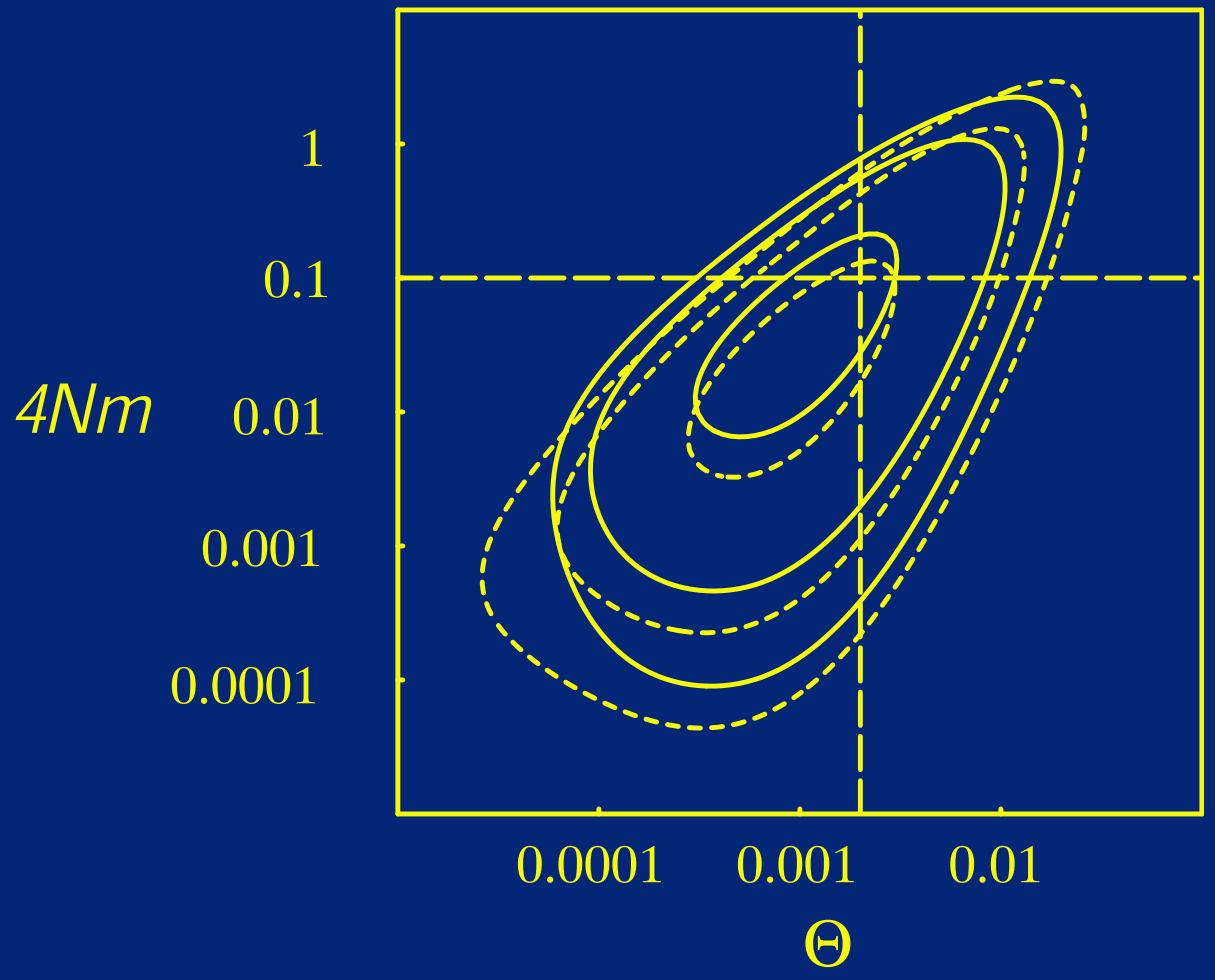


Comparison of approximate confidence intervals

Population pair	$\mathcal{M} = m/\mu$			Model
	2.5%	MLE	97.5%	
Africa $\rightarrow$ Asia	300	490	2190	Full
	30	590	2590	Restricted
Asia $\rightarrow$ Africa	0	0	650	Full
	20	360	1600	Restricted

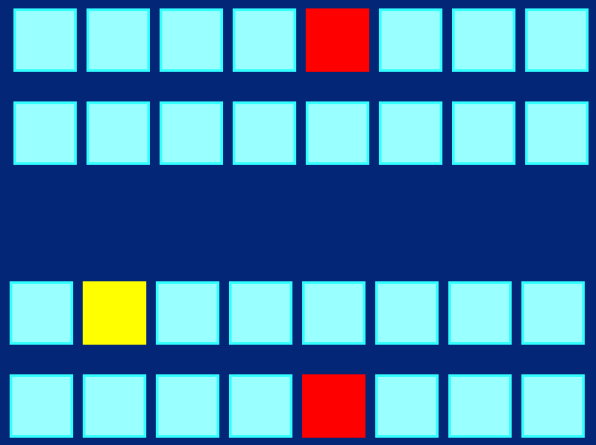
Comparison between *migrate* and *genetree*

(Beerli and Felsenstein 2001)

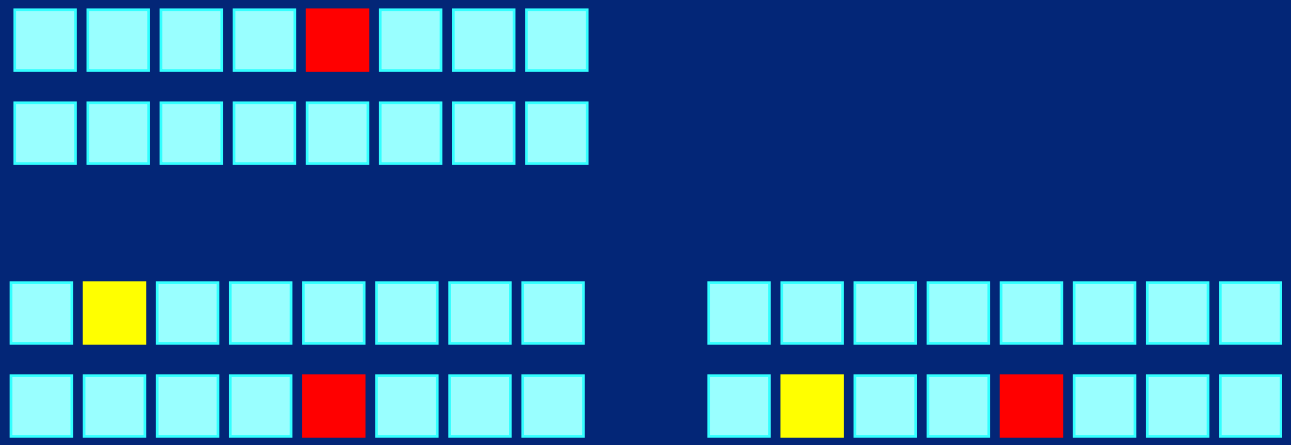


Recombination rate estimation

Haplotypes



Haplotypes



Either haplotypes must be resolved or the program must integrate over all possible haplotype assignments.

Recombination rate estimation (Kuhner et al. 2000)

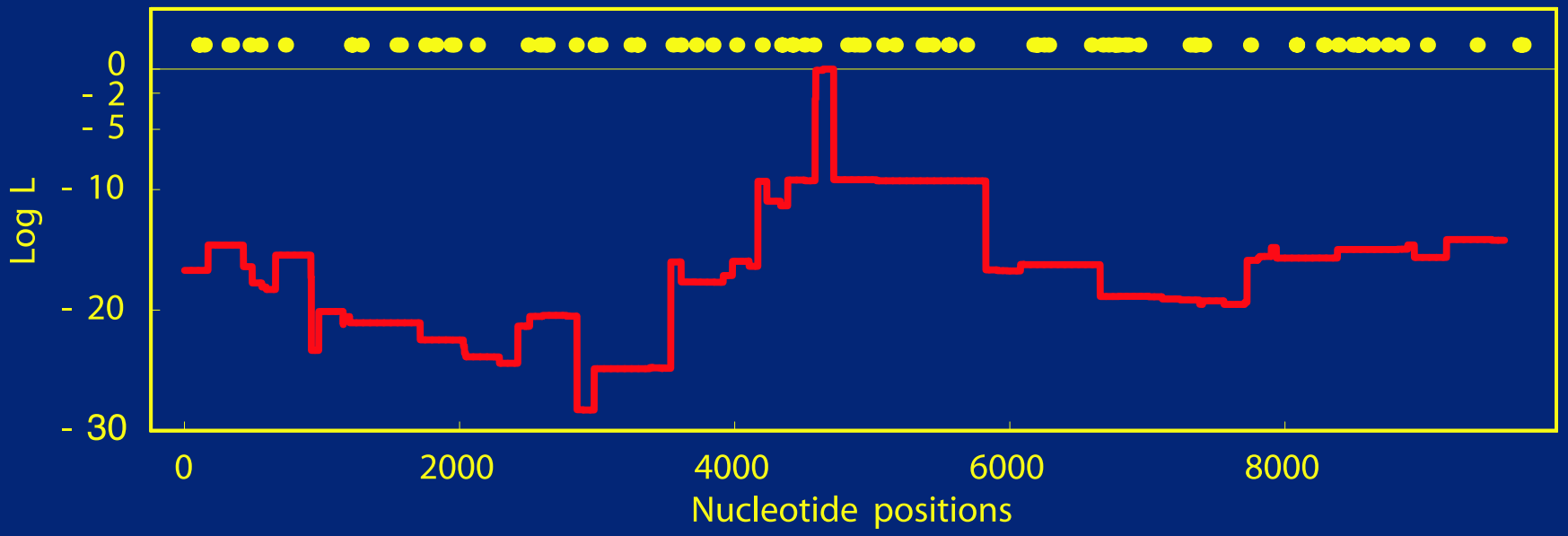
Human lipoprotein lipase (LPL) data: 9734 bp of intron and exon data derived from three populations: African Americans from Mississippi, Finns from North Karelia, Finland, and non-Hispanic Whites from Minnesota.

Population	Haplotypes	Θ_W	Θ_K	r_H	r_K
Jackson	48	0.0018	0.0072	1.4430	0.1531
North Karelia	48	0.0013	0.0027	0.3710	0.3910
Rochester	46	0.0014	0.0031	0.3350	0.2273
Combined	142	0.0016	0.0073	0.6930	0.1521

Shown are the number of haplotypes in each section of the data set, $\hat{\Theta}$ from Wattersons estimator Θ_W and RECOMBINE (Θ_K), and $\hat{r}$ from Hudsons estimator (r_H) and RECOMBINE (r_K).

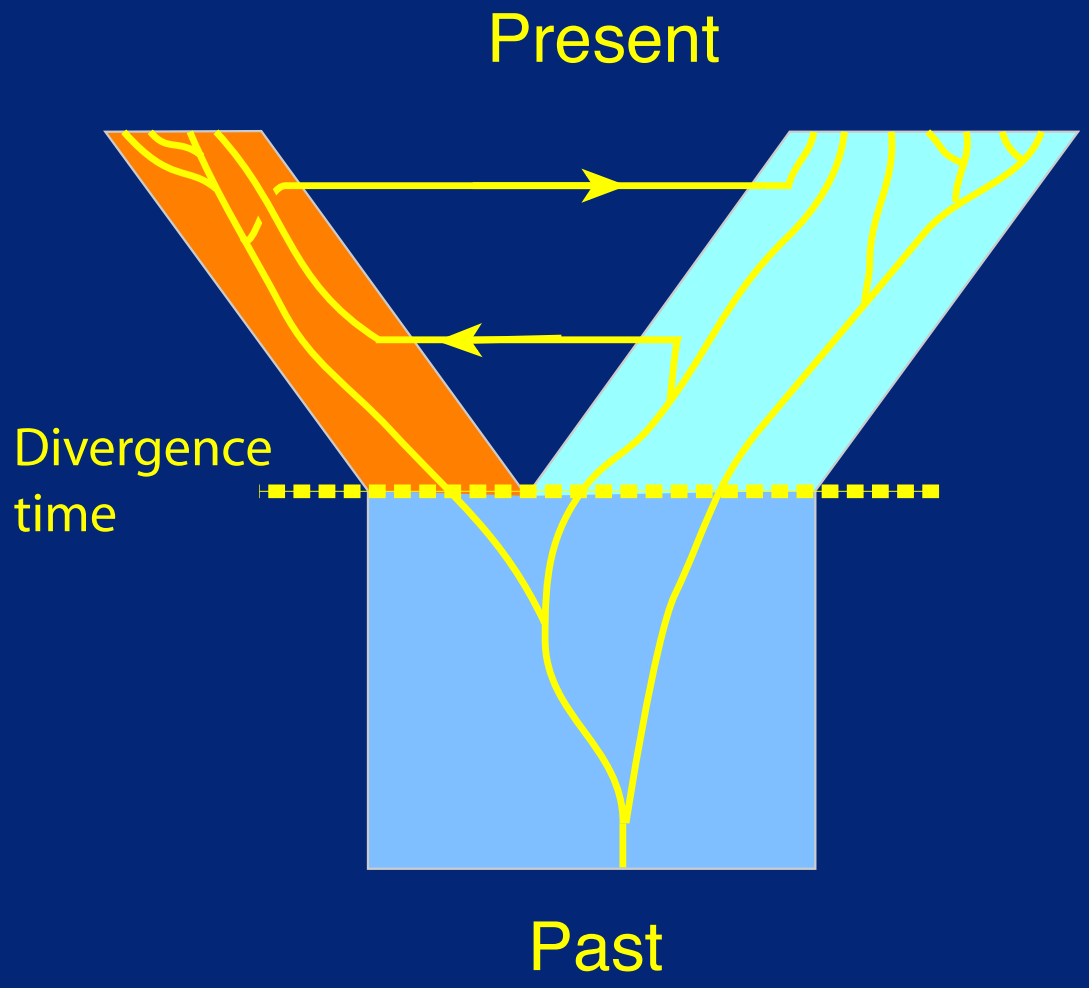
Linkage disequilibrium mapping (Kuhner et al., in prep.)

With a disease mutation model we can use the recombination estimator to post-analyze the sampled genealogies that where used to estimate r and find the location of the disease mutation on the DNA.



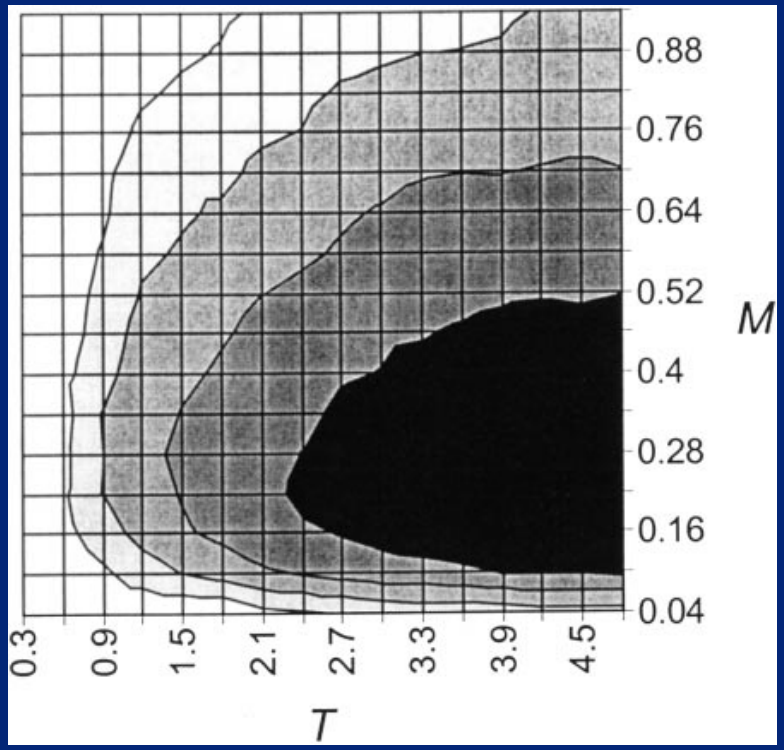
Estimation of divergence time

Wakeley and Nielsen (2001)



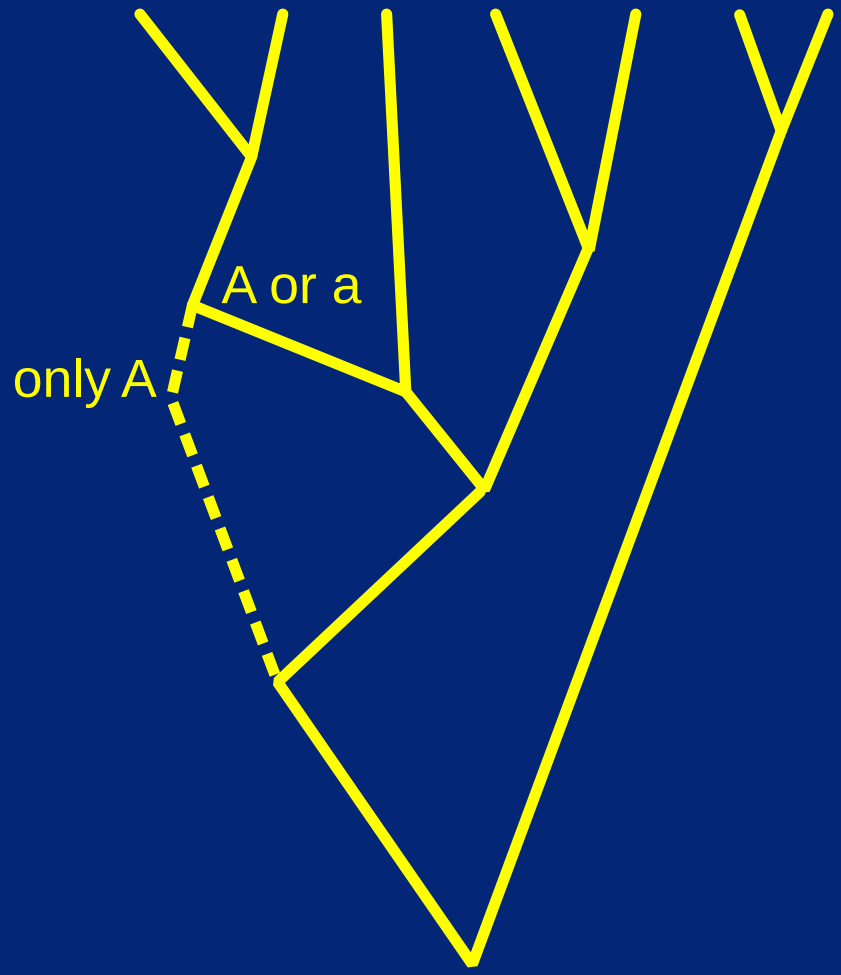
Estimation of divergence time

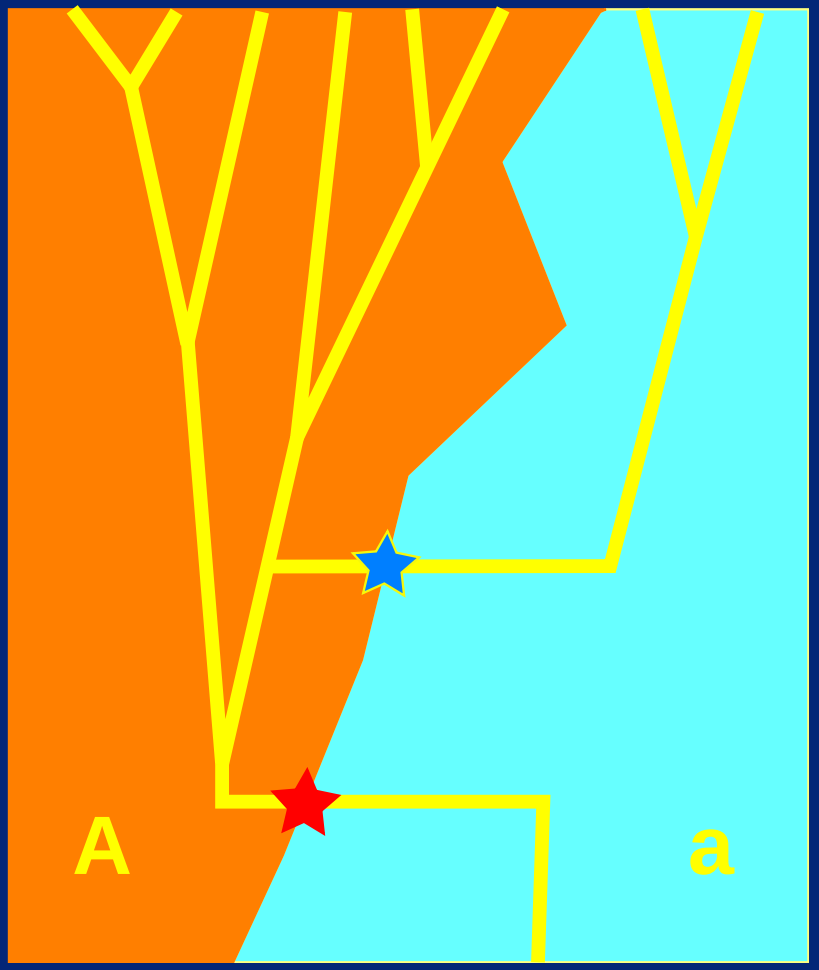
Wakeley and Nielsen (2001) Figure 7.The joint integrated likelihood surface for T and M estimated from the data by Orti et al. (1994). Darker values indicate higher likelihood.



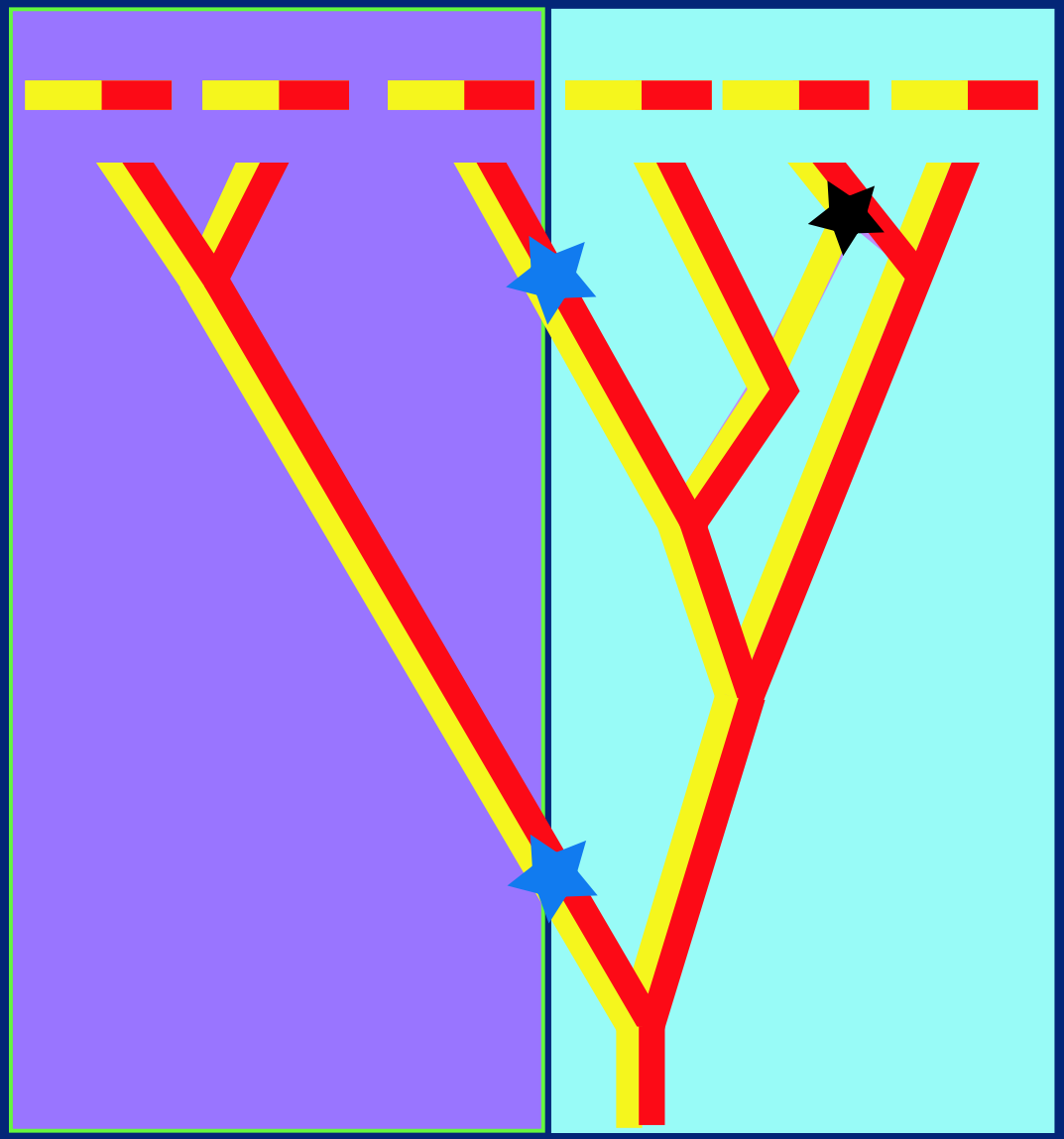
Selection coefficient estimation

Krone and Neuhauser (1999), Felsenstein (unpubl)

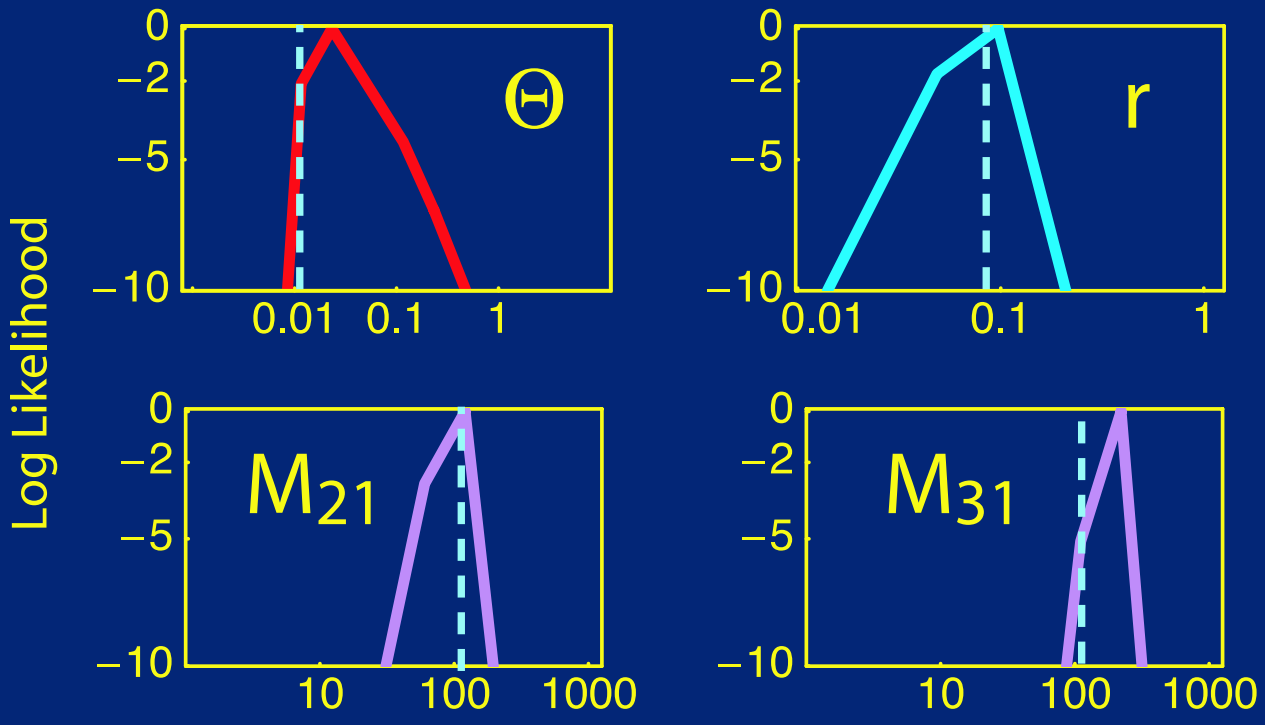




Joint estimation of recombination rate and gene flow



Joint estimation of recombination rate and migration



Any questions?

Pointers to software through
`http://evolution.gs.washington.edu/lamarc/popgensoftware.html`