

## Population genetic inference

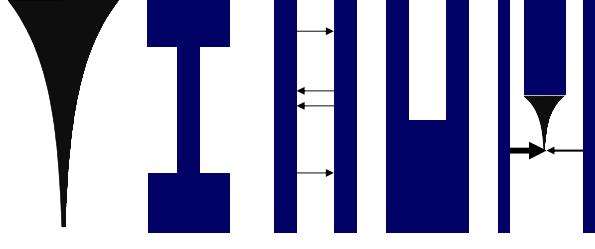
Date	Topic	
22 <sup>nd</sup> Jan	Good questions in population genetics	GM
29 <sup>th</sup> Jan	Principles of population genetic inference	GM
5 <sup>th</sup> Feb	Recombination in the coalescent	JH
12 <sup>th</sup> Feb	Natural selection	GM
19 <sup>th</sup> Feb	Demographic models	GM
26 <sup>th</sup> Feb	Combinatorics of the coalescent	JH
5 <sup>th</sup> March	Population genetics of disease mutations	GM
12 <sup>th</sup> March	Model organisms	GM

## Reading

- Beaumont M.A. 1999. Detecting population expansion and decline using microsatellites. *Genetics* **153**: 2013-2029
- Rousset F. 2001. Inferences from spatial population genetics. pp.239-269 in *Handbook of Statistical Genetics* (Eds. Balding, Bishop and Cannings).

## Demographic models

- Population growth
- Population bottlenecks
- Subdivided populations
- Population splits
- Admixture

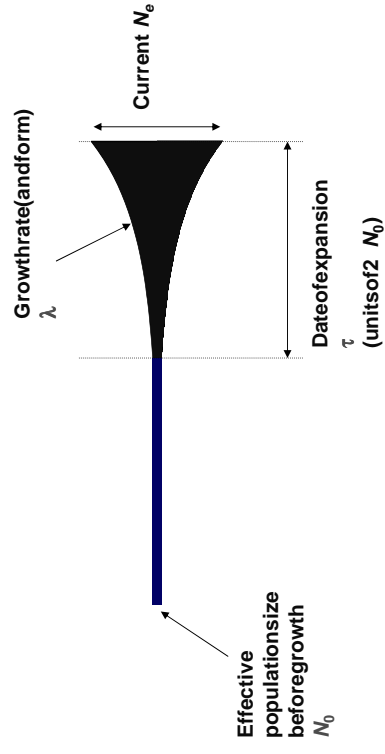


## Themes in demographic inference

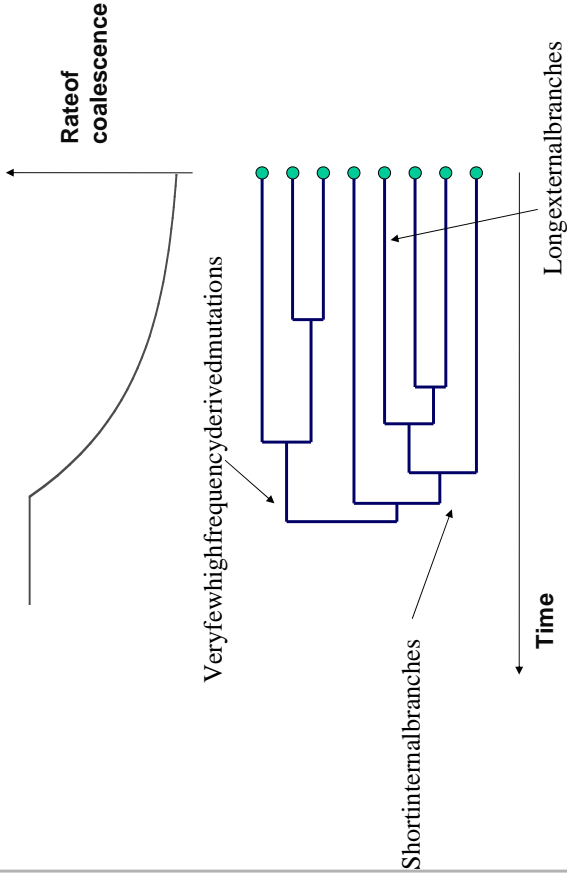
- Models and parameters
  - What is the simplest formulation relevant to my data?
- Genealogies
  - How are genealogies affected by the demographic model?
- Patterns
  - What is the signature of the demographic model?
- Estimation
  - How can I estimate the relevant parameters?

## Population growth

- Exponentially growing populations
  - Humans, HIV-1 (within patients), HIV-1 (worldwide)



## Genealogies in growing populations



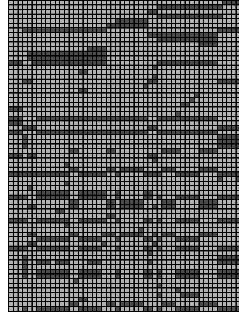
## Estimators of $\theta$

- Watterson's estimate
  - Counts segregating sites
$$\hat{\theta}_W = S \left( \sum_{i=1}^{n-1} \frac{1}{i} \right)^{-1}$$
- Pairwise differences
  - Influenced by intermediate frequency alleles
$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{i,j,i \neq j}^{n-1} k_{ij}$$
- The number of external mutations
  - Sensitive to excess of recent mutations
$$\hat{\theta}_e = \eta_e$$
- Fu's (1996) estimator
  - Sensitive to high -frequency derived mutations
$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

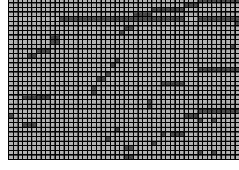
Also: no. haplotypes / no. segregating sites  $K/S$  = measure of haplotype diversity

## Detecting growth

- Low rates of polymorphism relative to stable populations
- Excess of recent mutations
  - Negative Tajima  $D$  and Fu and Li  $D$  statistics
- Fragmented haplotype structure
  - High  $K$  for given  $S$ , low haplotype diversity



$$\begin{aligned} \hat{\theta}_W &= 15.0 \\ \hat{\theta}_\pi &= 16.3 \\ \hat{\theta}_e &= 17.0 \\ \hat{\theta}_H &= 12.7 \\ K/S &= 0.37 \end{aligned}$$



$$\begin{aligned} \hat{\theta}_W &= 7.8 \\ \hat{\theta}_\pi &= 3.9 \\ \hat{\theta}_e &= 13.0 \\ \hat{\theta}_H &= 1.5 \\ K/S &= 0.63 \end{aligned}$$

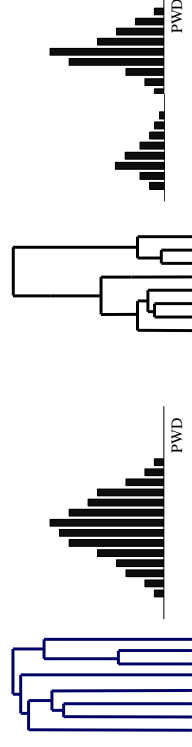
Null model  $n=50, \theta=10, \rho=10$

Growth  $n=50, \theta=10, \rho=10, \lambda=5$

## Estimating growth parameters

- Moment methods
  - Mismatch distribution (Rogers and Harpending 1992)

Variance of pairwise differences reduced in growing populations

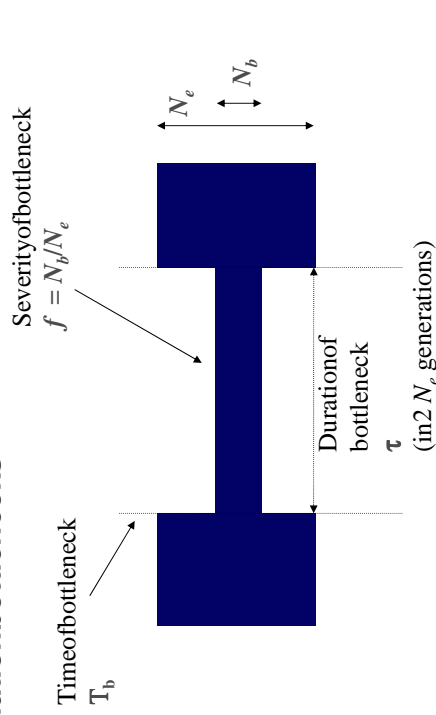


- Coalescent likelihood methods

- Importance sampling
- GENETREE (Griffiths: [www.stats.ox.ac.uk/mathgen](http://www.stats.ox.ac.uk/mathgen))
- Monte Carlo Markov Chain

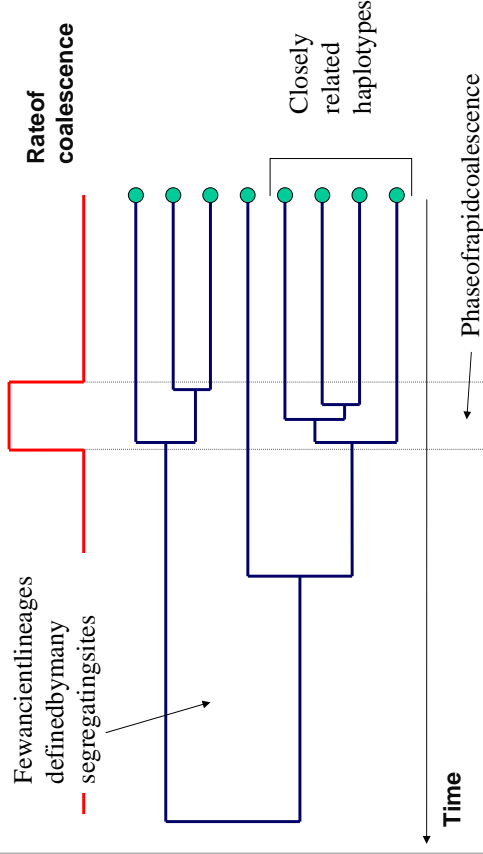
**Batwing** (Wilson and Balding: [www.maths.abdn.ac.uk/~tjw/](http://www.maths.abdn.ac.uk/~tjw/))  
 Beaumont (1999) ([www.rubic.rdg.ac.uk/~mab/software.html](http://www.rubic.rdg.ac.uk/~mab/software.html))

## Population bottlenecks



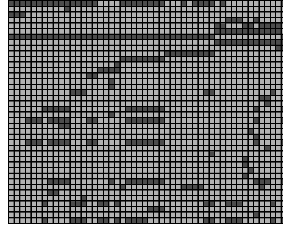
- e.g. out-of-Africa hypothesis
- Strength of bottleneck =  $\tau / f$
- If assumptions during bottleneck, cannot assign parameter

## Genealogies during bottlenecks



## Detecting bottlenecks

- Rate of polymorphism reduced
  - Subset of ancestral polymorphism persists
- Excess of mutations at intermediate frequency
  - Positive Tajima  $D$  and Fu and Li  $D$  statistics
- Strong haplotype structure



$$\hat{\theta}_W = 8.9$$

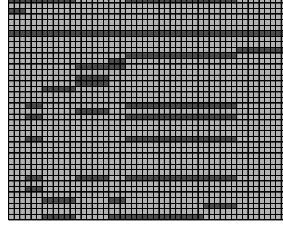
$$\hat{\theta}_\pi = 7.1$$

$$\hat{\theta}_e = 16.0$$

$$\hat{\theta}_H = 3.8$$

$$K / S = 0.70$$

No bottleneck:  $n=50, \theta=10, \rho=10$



$$\hat{\theta}_W = 4.2$$

$$\hat{\theta}_\pi = 5.8$$

$$\hat{\theta}_e = 0.0$$

$$\hat{\theta}_H = 6.0$$

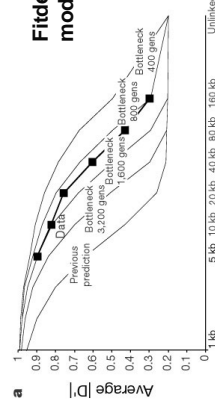
$$K / S = 0.42$$

Recent bottleneck:  $n=50, \theta=10, \rho=10$ , 10 ancestral lineages

## Estimating bottleneck parameters

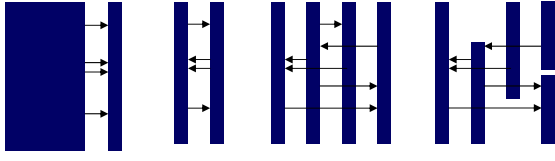
- Bottlenecks increase variance in pairwise differences
  - microsatellite measures (e.g. Reich and Goldstein 1998)
- Coalescent methods to detect population declines
  - e.g. Beaumont (1999)
- Bottlenecks increase linkage disequilibrium
  - Reich *et al.* (2001)

Figure 10.10: Average  $D$  vs. time for different bottleneck scenarios. The plot shows Average  $D$  on the y-axis (ranging from 0.1 to 1.0) and time in generations on the x-axis (ranging from 0 to 140). The scenarios are: Previous prediction, Bottleneck 500 gens, Bottleneck 1000 gens, Bottleneck 2000 gens, Bottleneck 4000 gens, and Data. The data points are shown as black squares, and the predictions are shown as lines.



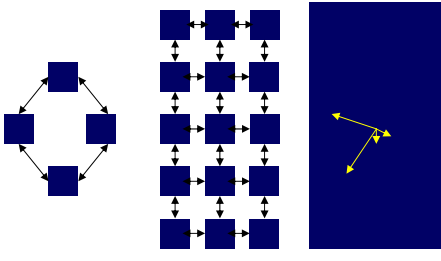
## Subdivided populations: demes

- The island model
- The 2-island model
- The  $n$ -island model
- Metapopulations

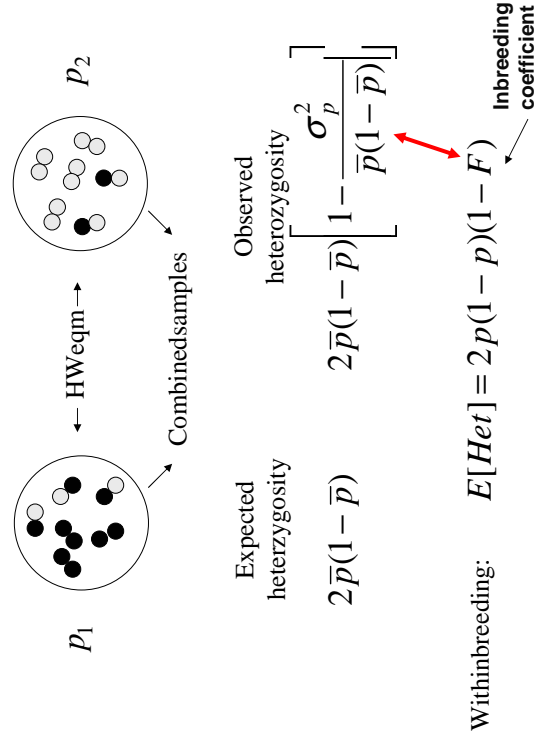


## Subdivided populations: isolation by distance

- Stepping-stone models
- Circular stepping-stones
- Lattice models
- Continuous-space models

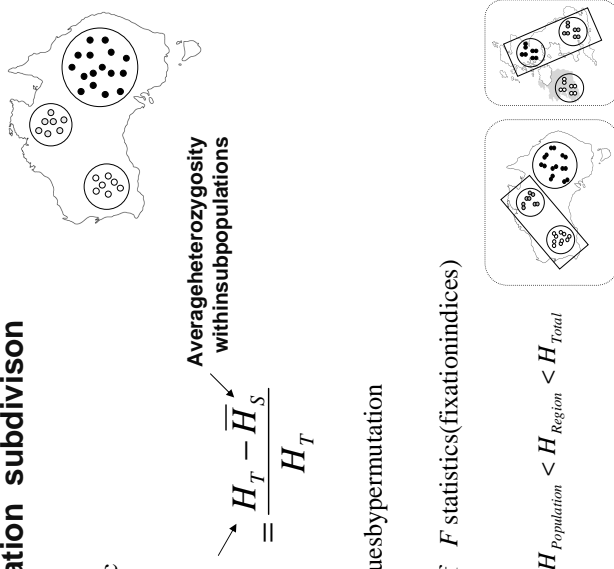


## The inbreeding effect on population structure



## Describing population subdivision

- Wright's  $F_{ST}$  statistic
- Heterozygosity over all populations
- Detect significant values by permutation
- Hierarchical nature of  $F$  statistics (fixation indices)



$$H_{Individual} < H_{Subpopulation} < H_{Population} < H_{Region} < H_{Total}$$

## $F_{ST}$ in natural populations

Allozymes	Organism	$H_T$	$\bar{H}_s$	$F_{ST}$
	Human (major races)	0.130	0.121	0.069
	Human (Yanomama)	0.039	0.036	0.077
	House mouse	0.097	0.086	0.113
	Jumping rodent	0.037	0.012	0.676

Nei (1975)

## • SNPs

Organism	$H_T$	$\bar{H}_s$	$F_{ST}$
Human (major races)	0.195	0.201	0.067
<i>Drosophila melanogaster</i> <sup>a</sup>	0.0154	0.0151	0.023

<sup>a</sup>Based on pairwise diversity

## What are $F$ statistics?

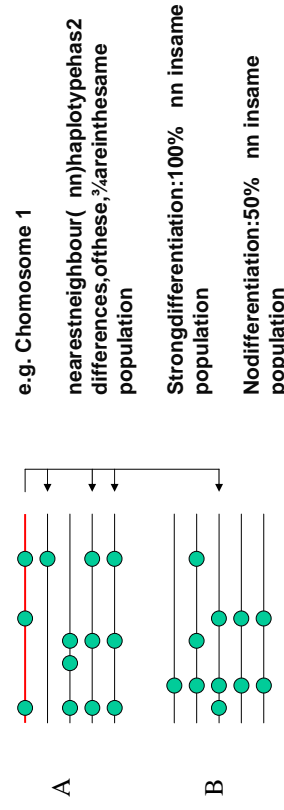
- A summary statistic measure of population differentiation?
  - Account for stochastic sampling process in estimation procedure
- A function of the underlying evolutionary model?
  - A *meta-parameter*
  - Account for stochastic sampling process AND stochastic evolutionary process in estimation procedure
- For certain explicit models of population differentiation the expectation of  $F_{ST}$  can be equated with model parameters

$$F_{ST} = \frac{1}{1 + 4N_e m} \quad \text{For Wright's island model ONLY}$$

- General framework of analysis of variation = AMOVA

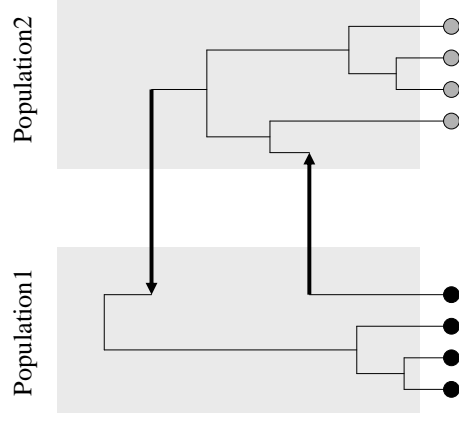
## Haplotype structuring: Hudson's $S_{nn}$ statistic

- Measure location of similar haplotypes
- Test by permutation



$$S_{nn} = \frac{1}{n} \sum_{\text{chromosomes}} \text{Proportion nearest neighbours in same population}$$

## Gene genealogies in subdivided populations

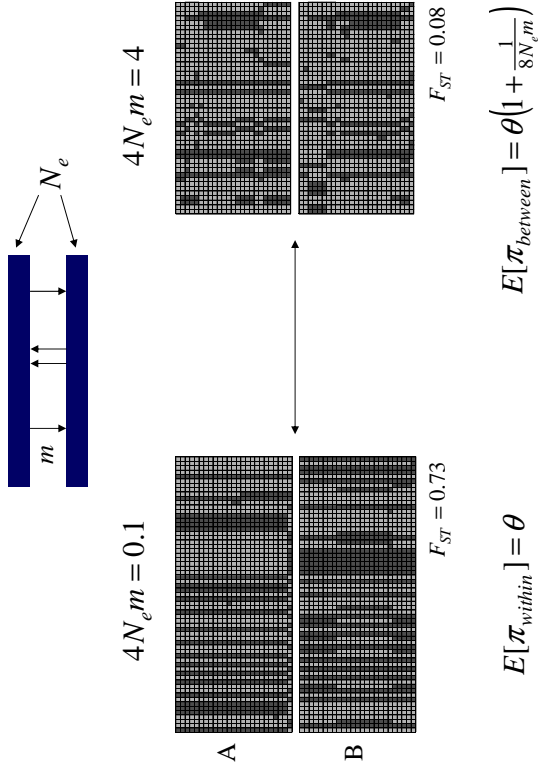


$$\Pr\{\text{coalescence}\} = \frac{n_i(n_i - 1)}{4N_e(t)}$$

$$\Pr\{\text{migration}\} = \frac{n_j m}{4N_e(t)}$$

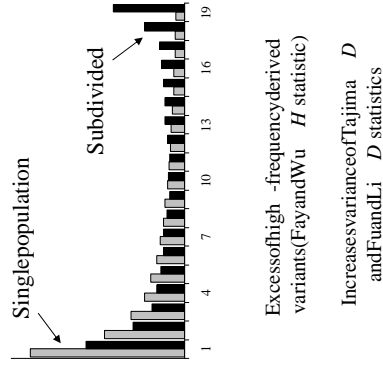
Key parameter  
 $4N_e(t)m$

## Subdivision in the symmetric 2-deme model

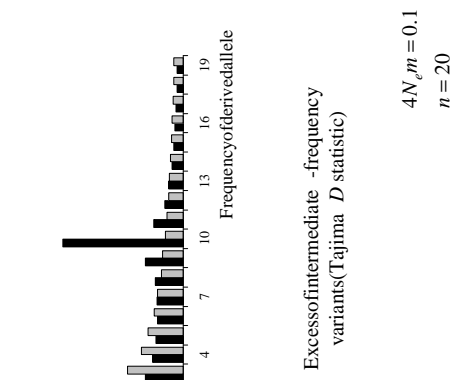


## Population subdivision and frequency spectrum neutrality tests

### Within populations

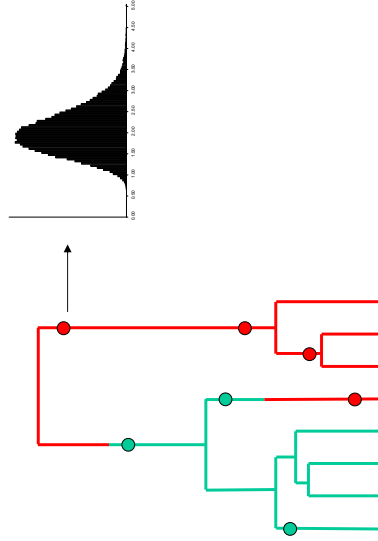


### Between populations



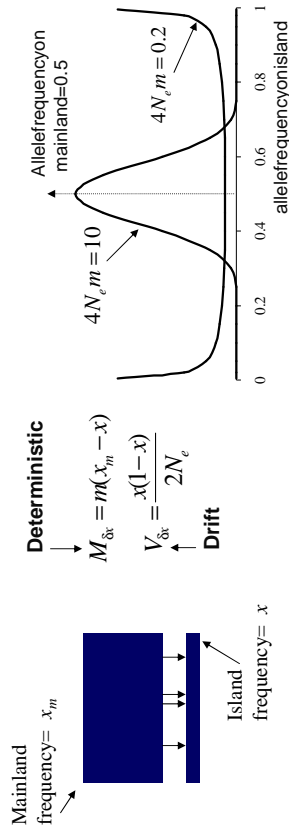
## Coalescent inference in the 2-population model

- Importance sampling routine based on coalescent transition probabilities
  - Bahlo and Griffiths (199?)
- Coloured coalescent graphs



## Wright's island model

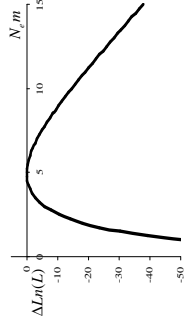
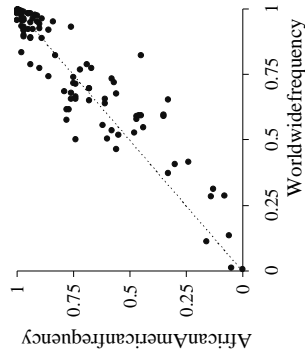
- Wright's distribution of allele frequencies in the island model



- Combine with multinomial sampling properties
  - allele counts at single locus follows Dirichlet-multinomial distribution
  - full-likelihood inference possible for unlinked SNPs
  - also applies to infinite-island model

## Example: SNP frequencies in African Americans

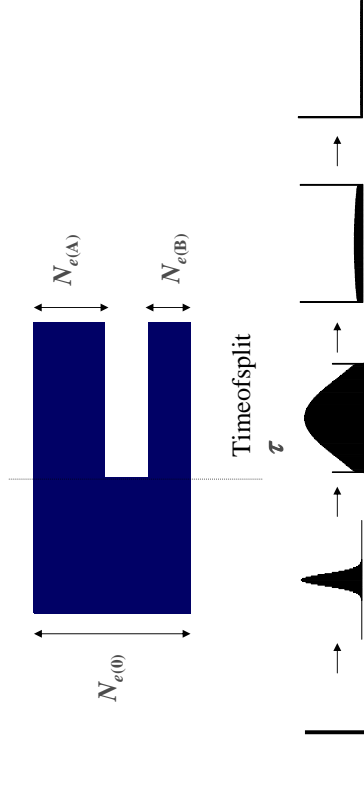
- Goddard *et al.* (2000)
  - 114 SNPs in 33 genes
  - 190 African American samples
  - Assume worldwide frequency is equivalent to that of immigrant alleles



$$\widehat{N_e m} = 5.0$$

## Other formulations: split-time models

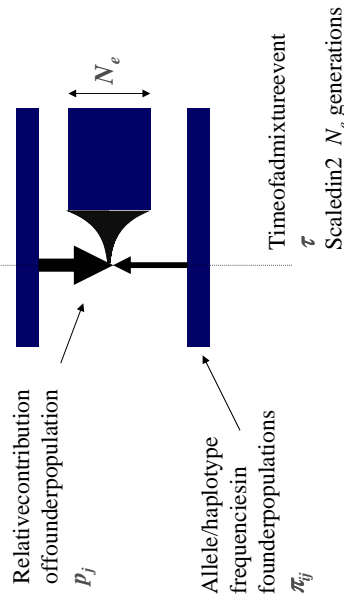
- Assuming populations diverged from ancestral population
  - Estimated divergence time scaled in rate of drift
  - Allele frequency change described by diffusion approximation
  - Bayesian approach integrates over possible ancestral allele frequencies (Nicholson and Donnelly 2002)



## Distinguishing gene-flow from divergence

- Any differentiated populations can be interpreted as
  - Totally isolated: diverged from ancestral population (non-equilibrium)
    - Estimate population sizes and  $4N_e\mu$
  - Partially isolated: recurrent migration (equilibrium)
    - Estimate population sizes and time since split
- Which?
- MCMC approach to distinguishing models
  - Nielsen and Wakeley (2001)
    - [www.biom.cornell.edu/Homepages/Rasmus\\_Nielsen/files.html](http://www.biom.cornell.edu/Homepages/Rasmus_Nielsen/files.html)
  - e.g. variance in pairwise differences under recurrent migration Wakeley (1996)

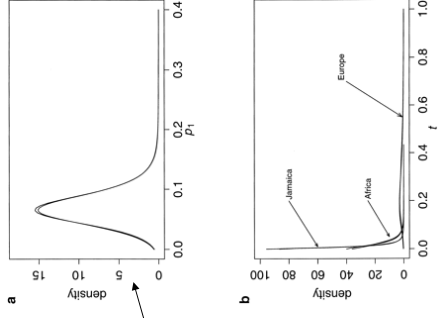
## Admixture



- Approaches differ in
  - Information from descendants of source populations
  - Assumption about time to admixture event
  - Source of information in the data

## Admixture modelling I

- Information from source populations, time to admixture estimated
- Fully-Bayesian method to estimate parameters
  - Chikhi *et al.* (2001)

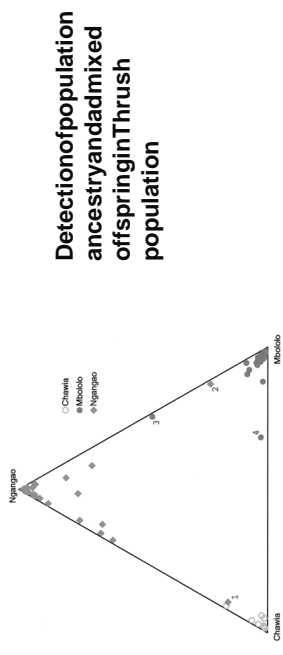


Proportion of European contribution to Jamaican population

Time to admixture event (scaled by population sizes) Suggests model in adequacies

## Admixture modelling II

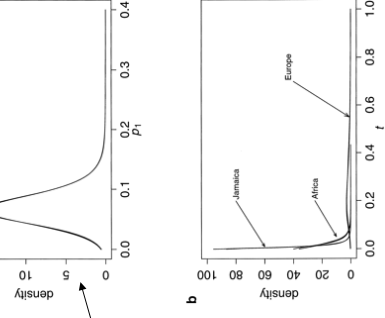
- Information from single population, assume recent admixture
  - Aim to detect unacknowledged admixture in population samples
- Fully Bayesian method: posterior probabilities of ancestry for each individual
  - Pritchard *et al.* (2000)



Detection of population ancestry and admixed offspring in Thrush population

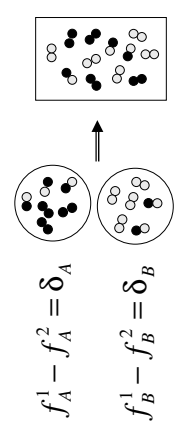
## Admixture and linkage disequilibrium

- Combination of two previously differentiated populations generates associations between alleles
- Over time random mating returns population to equilibrium
- Disequilibrium between unlinked loci can persist for several generations, while Hardy-Weinberg equilibrium is achieved instantly



## Admixture and linkage disequilibrium

- Combination of two previously differentiated populations generates associations between alleles
- Over time random mating returns population to equilibrium
- Disequilibrium between unlinked loci can persist for several generations, while Hardy-Weinberg equilibrium is achieved instantly



$$f_A^1 - f_A^2 = \delta_A$$

$$f_B^1 - f_B^2 = \delta_B$$

$$D_0 = \frac{1}{4} \delta_A \delta_B$$

$$D_t = D_0 (1-r)^t$$

