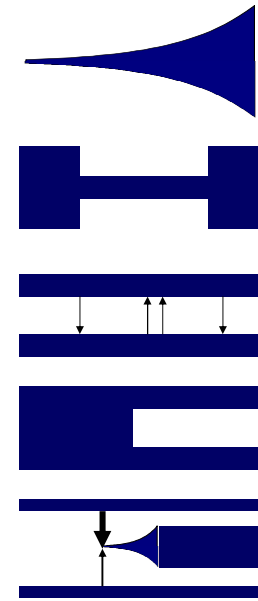


Demographic extensions of the neutral coalescent

- Data from many loci across the genome (from one or more populations) can be used to address demographic questions
 - Population growth, bottlenecks
 - Population structuring
 - Admixture
- Basic questions
 - Have demographic processes influenced genetic variation?
 - How can I summarise geographical structuring?
- Bigger questions
 - What is F_{ST} ?
 - How can we use explicit models to learn about demographic history?

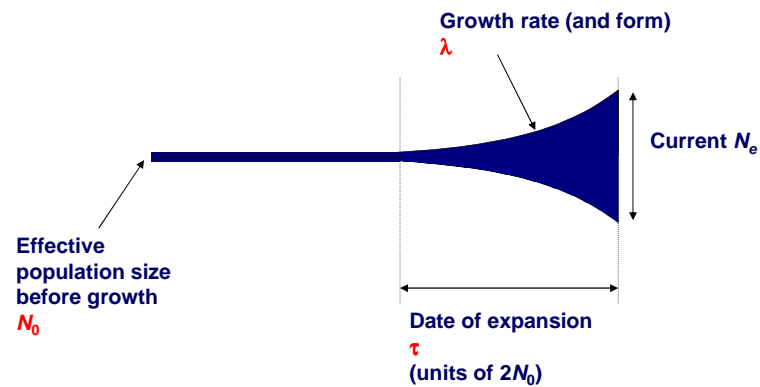
Demographic models

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

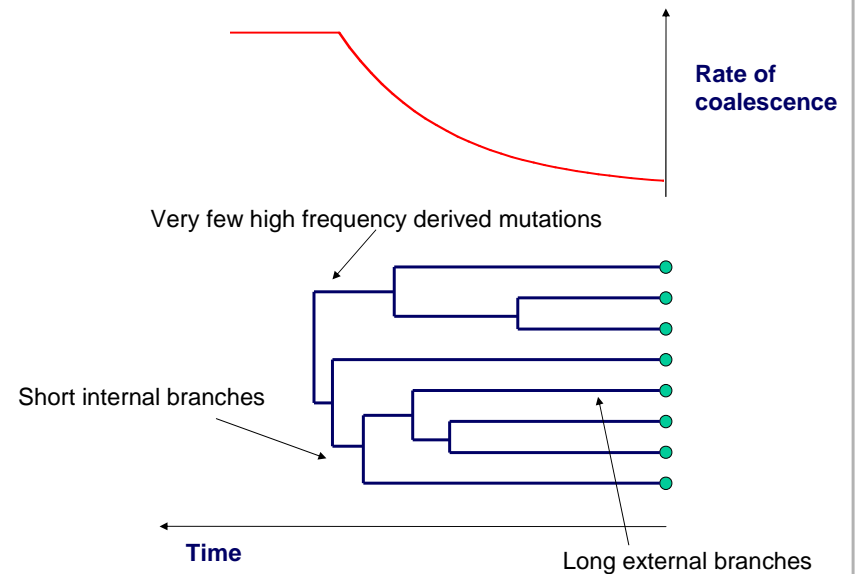


Population growth

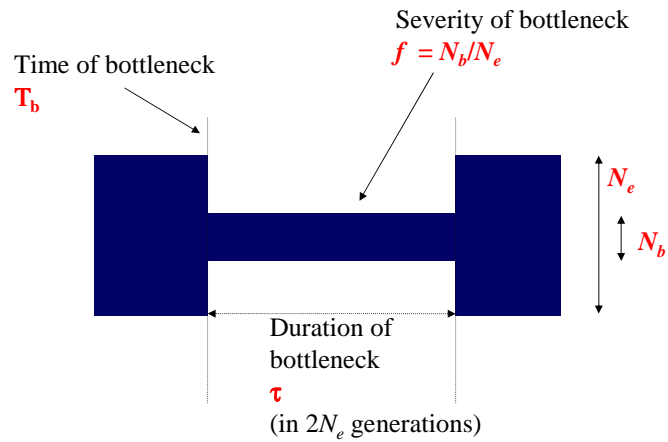
- Exponentially growing populations
 - Humans, *HIV-1* (within patients), *HIV-1* (worldwide)



Gene genealogies in growing populations

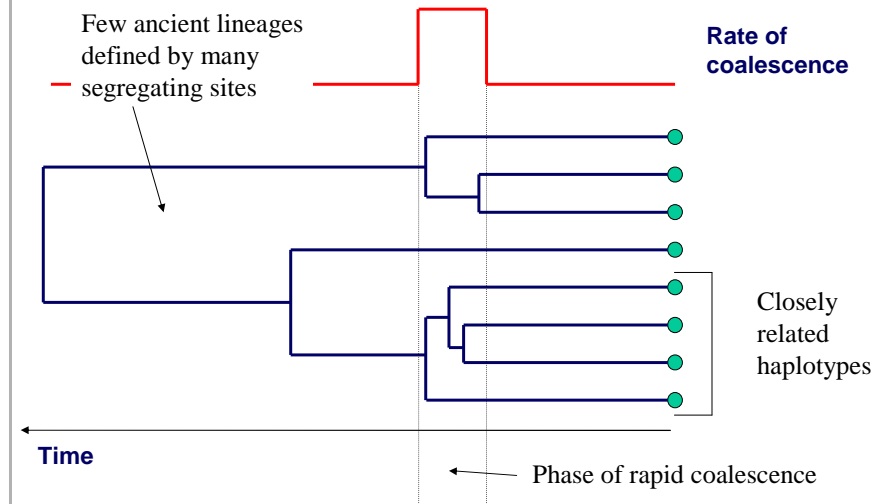


Population bottlenecks



- e.g. out-of-Africa hypothesis
- Strength of bottleneck = τ / f
- If assume no mutations during bottleneck, can treat as single parameter

Gene genealogies during bottlenecks



Detecting growth and bottlenecks

Growth	Bottlenecks
Reduced polymorphism	Polymorphism little affected
Excess of rare mutations	Excess of intermediate-frequency mutations
Negative Tajima D	Positive Tajima D
Negative F_u and $Li D^*$	Positive F_u and $Li D^*$
Fragmented haplotype structure	Strong haplotype structure

Coalescent likelihood methods

GENETREE (Griffiths: www.stats.ox.ac.uk/mathgen)

Batwing (Wilson and Balding: www.maths.abdn.ac.uk/~ijw)

Beaumont (1999) (www.rubic.rdg.ac.uk/~mab/software.html)

Estimators of θ

- Watterson's estimate
 - Counts segregating sites
- Pairwise differences
 - Influenced by intermediate frequency alleles
- The number of external mutations
 - Sensitive to excess of recent mutations
- Fu's (1996) estimator
 - Sensitive to high-frequency derived mutations

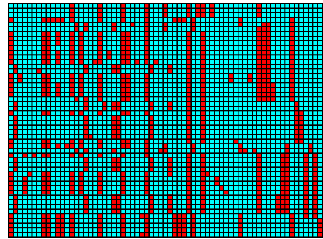
$$\hat{\theta}_W = S \left(\sum_{i=1}^{n-1} 1/i \right)^{-1}$$

$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{ij, i \neq j} k_{ij}$$

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

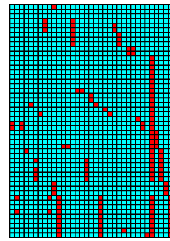
$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

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



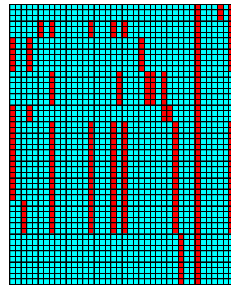
Null model $n=50, \theta=10, \rho=10$

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



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

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

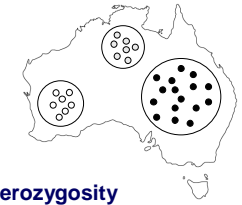


Recent bottleneck: $n=50, \theta=10, \rho=10, 10$ ancestral lineages

$\hat{\theta}_W = 4.2$
 $\hat{\theta}_\pi = 5.8$
 $\hat{\theta}_e = 0.0$
 $\hat{\theta}_H = 6.0$
 $K/S = 0.42$

Describing population subdivision

- Wright's F_{ST} statistic



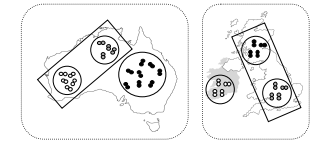
Heterozygosity over all populations

Average heterozygosity within subpopulations

$$F_{ST} = \frac{H_T - \bar{H}_S}{H_T}$$

- 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)

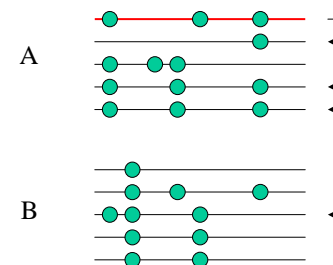
- SNPs

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

^aBased on pairwise diversity

Haplotype structuring: Hudson's S_{nn} statistic

- Measures location of similar haplotypes
- Test by permutation



e.g. Chromosome 1

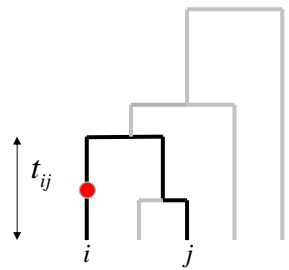
nearest neighbour (nn) haplotype has 2 differences, of these, ¾ are in the same population

Strong differentiation: 100% nn in same population

No differentiation: 50% nn in same population

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

Heterozygosity and pairwise coalescence times for SNPs



Time in tree = T

H = probability mutation occurs on lineage leading to sampled chromosomes

$$E[H_{ij}] = \lim_{\mu \rightarrow 0} \frac{E[t_{ij}e^{-\mu T}]}{E[Te^{-\mu T}]} = \frac{E[t_{ij}]}{E[T]}$$

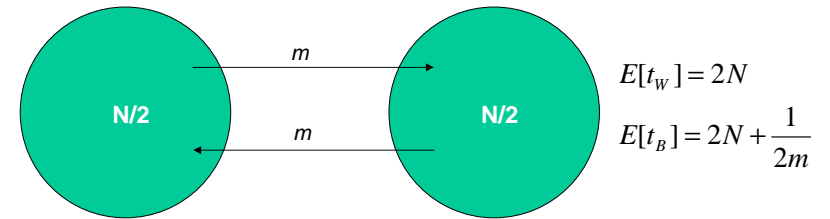
$$F_{ST} = \frac{H_T - \bar{H}_s}{H_T} \approx 1 - \frac{E[t_s]}{E[t_r]}$$

Average coalescence time for pair of chromosomes from same population

Average coalescence time for pair of chromosomes from whole sample

By considering the average pairwise coalescence times under different models, we can understand what F_{ST} is measuring in different situations

F_{ST} and island migration

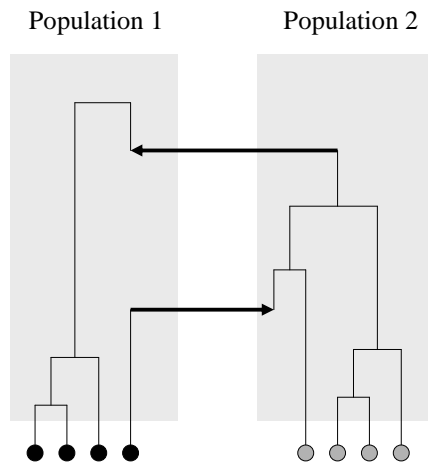


$$E[t_w] = 2N$$

$$E[t_B] = 2N + \frac{1}{2m}$$

$$F_{ST} = \frac{\bar{t} - \bar{t}_w}{\bar{t}} = \frac{1}{1 + 8Nm}$$

Gene genealogies in subdivided populations

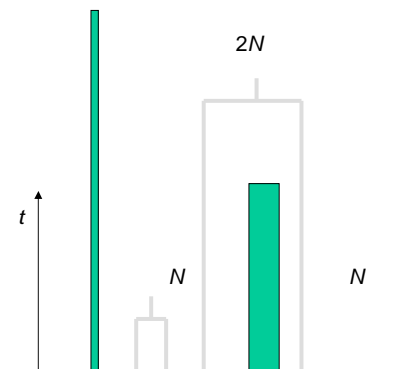


$$\Pr\{\text{coalescence}\} = \frac{n_i(n_i - 1)}{4N_{e(i)}}$$

$$\Pr\{\text{migration}\} = n_i m$$

Key parameter
 $4N_{e(i)}m$

F_{ST} and population splits

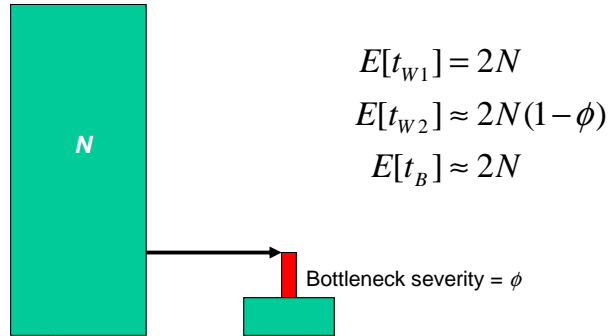


$$E[t_w] = 2N(1 + e^{-t/2N})$$

$$E[t_B] = t + 4N$$

$$F_{ST} = \frac{\bar{t} - \bar{t}_w}{\bar{t}} = \frac{1 + t/2N - e^{-t/2N}}{3 + t/2N + e^{-t/2N}} \approx \frac{t}{4N}$$

F_{ST} and founder events



$$E[t_{W1}] = 2N$$

$$E[t_{W2}] \approx 2N(1-\phi)$$

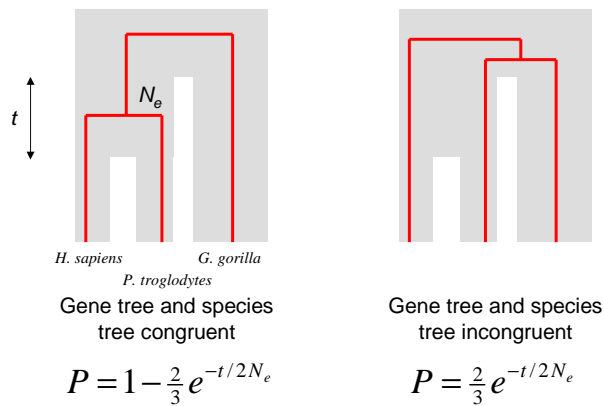
$$E[t_B] \approx 2N$$

$$F_{ST} = \frac{\bar{t} - \overline{t_W}}{\bar{t}} = \frac{\phi}{4 - \phi}$$

What does F_{ST} measure?

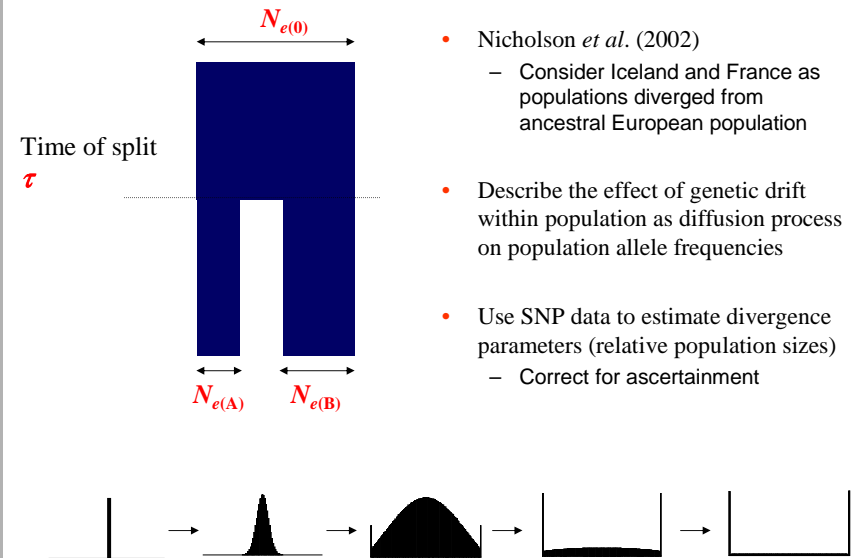
- F_{ST} can be used as a statistic to summarise patterns of differentiation between populations
- HOWEVER – the interpretation of F_{ST} depends critically on which model applies to the populations of interest
 - Migration rates
 - Time since separation
 - Founder events
- Explicit modelling of population histories allows us to distinguish between different demographic scenarios
 - Population splits in early human history
 - Founder events in Icelandic history
 - Recent human population structure
 - Admixture in island populations

Population splits in early human history

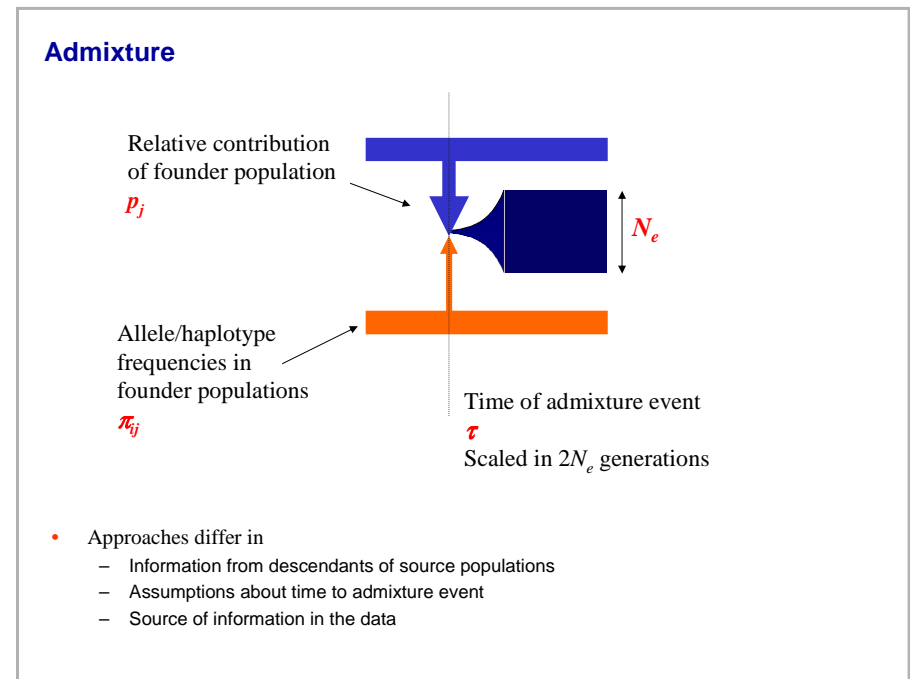
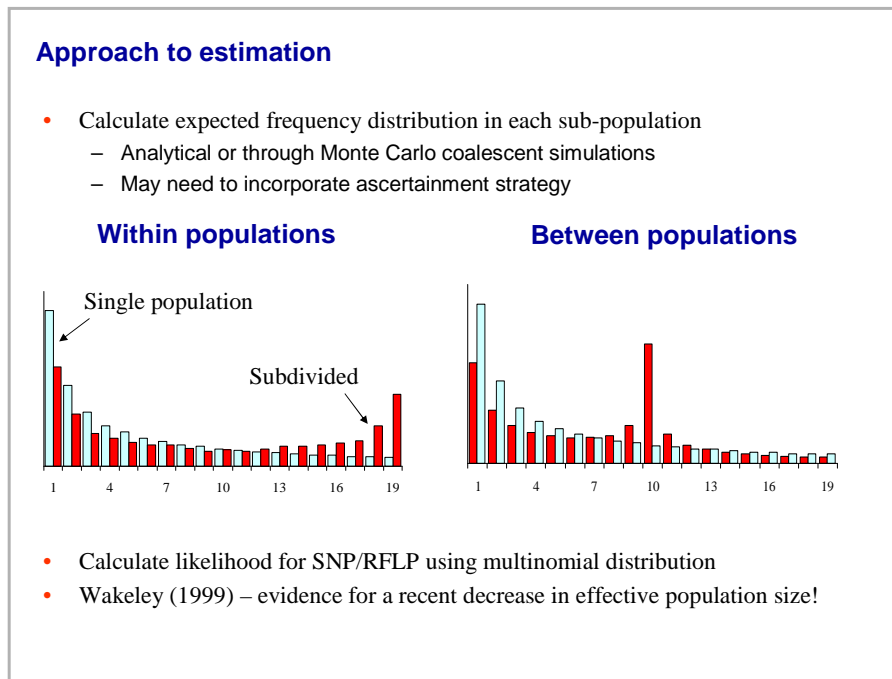
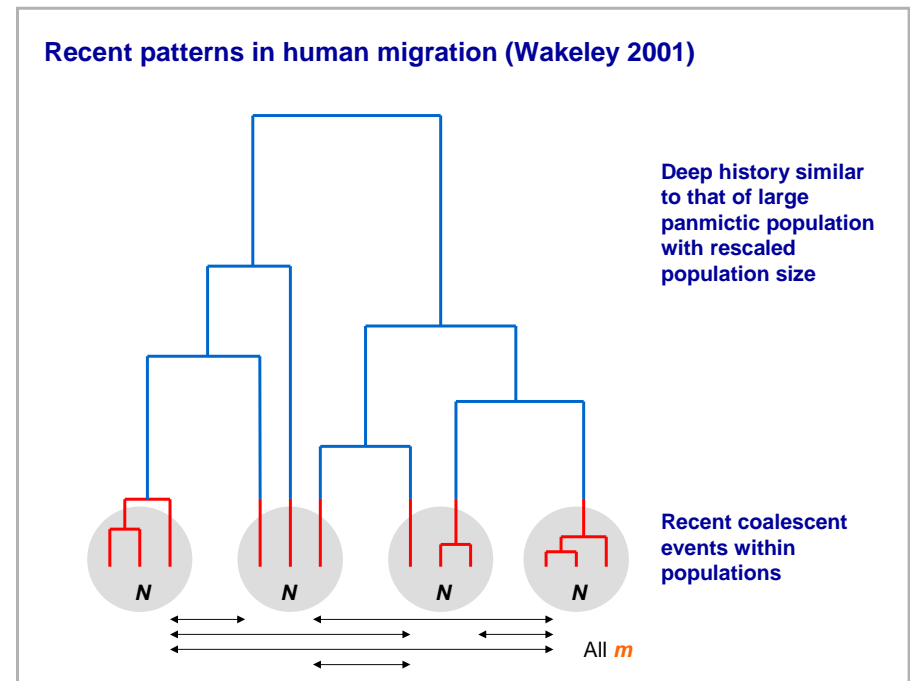
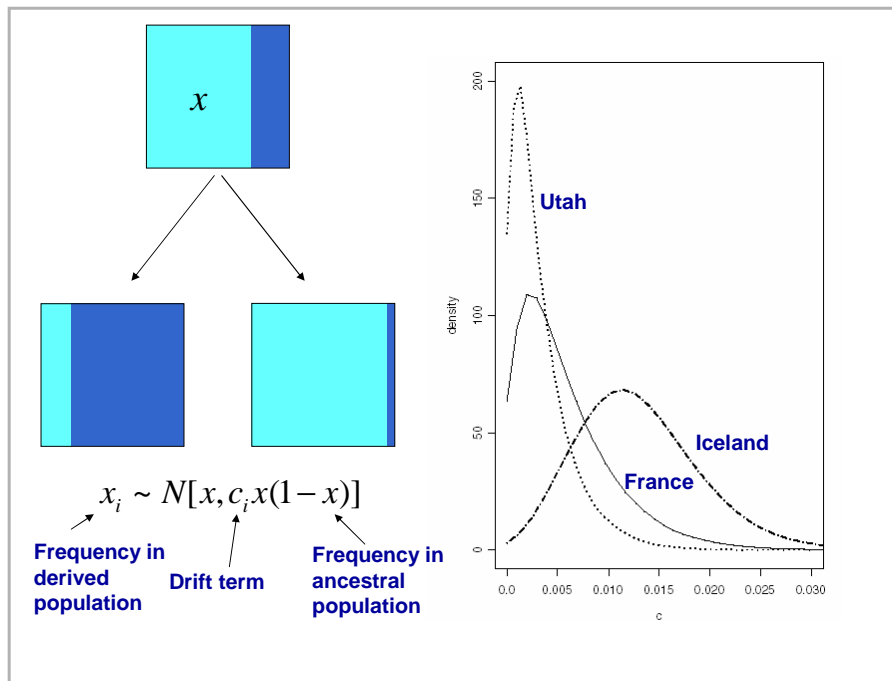


- Can estimate divergence times and ancestral population sizes from multiple data sets
- Chen and Li (2001) – 53 autosomal regions, 22 give incongruent trees
- Yang (2002) – Using estimate of mutation rate, $t = 1.1$ MY, $N_e = 12,000$ -21,000

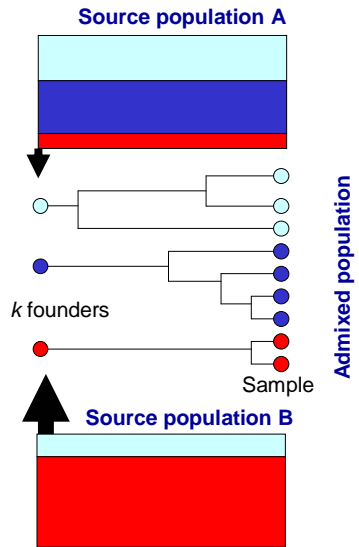
Recent population splits in human history



- Nicholson *et al.* (2002)
 - Consider Iceland and France as populations diverged from ancestral European population
- Describe the effect of genetic drift within population as diffusion process on population allele frequencies
- Use SNP data to estimate divergence parameters (relative population sizes)
 - Correct for ascertainment



Urn-model approach to genetic drift



$$L(\text{data}) = \sum \Pr(k | \tau)$$

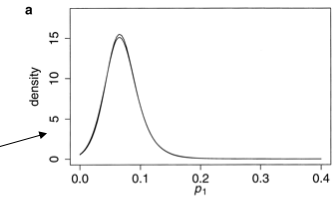
$$\times \Pr(\text{founders} | \pi_A, \pi_B, p, k)$$

$$\times \Pr(\text{Data} | \text{founders})$$

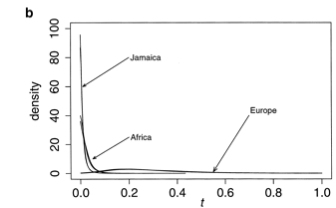
Admixture modelling I

- 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

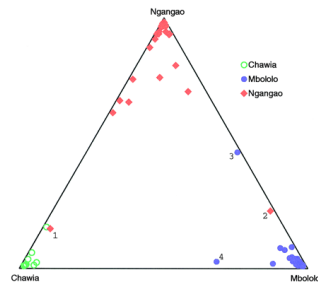


Time to admixture event (scaled by population sizes) Suggests model inadequacies



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

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

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

- Over time random mating returns population to equilibrium

$$D_0 = \frac{1}{4} \delta_A \delta_B \quad D_t = D_0 (1-r)^t$$

- Disequilibrium between unlinked loci can persist for several generations, while Hardy-Weinberg equilibrium is achieved instantly

