Population genetic inference

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

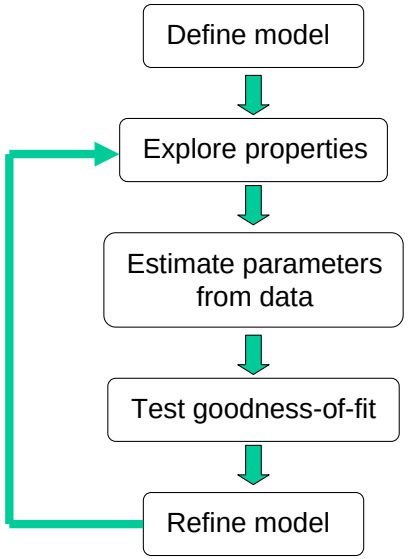
Books

Balding DJ, Bishop M and Cannings C. 2001. Handbook of Statistical Genetics. John Wiley and Sons Ltd.

Casella GC and Burger RL. 1990. Statistical Inference. Duxbury Press, Wadsworth Publishing, Co.

Weir BS. 1990. Genetic Data Analysis. Sinauer

Statistical inference



Issues

rules
parameters
quantities

summary statistics
graphical representation
stochastic simulation

moment methods
likelihood
Bayesian inference

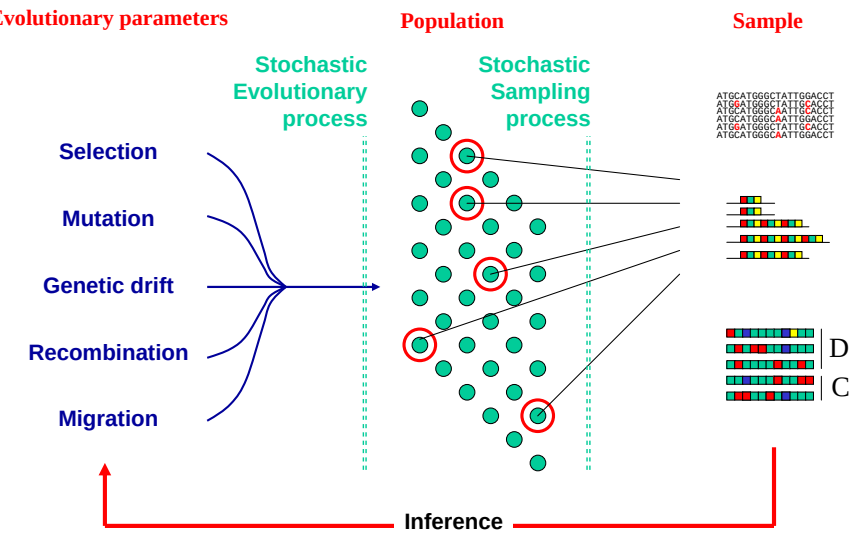
outliers
heterogeneity
comparison of estimators

add parameters

The unusual nature of population genetic data

- Conventional statistical inference
  - Many independent data points
  - Sample space of low dimensions
  - Analytical formulations for inference using all possible information often possible
- Population genetic data
  - Typically a single draw from the evolutionary process
  - Sample space of many dimensions
  - Analytical formulations for inference using ALL possible information usually impossible to derive

The complete model



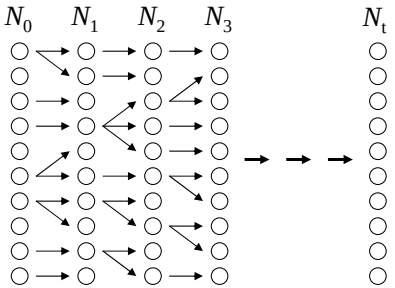
The null model in population genetics

Nothing interesting ever happens in evolution

Modelling the evolutionary process

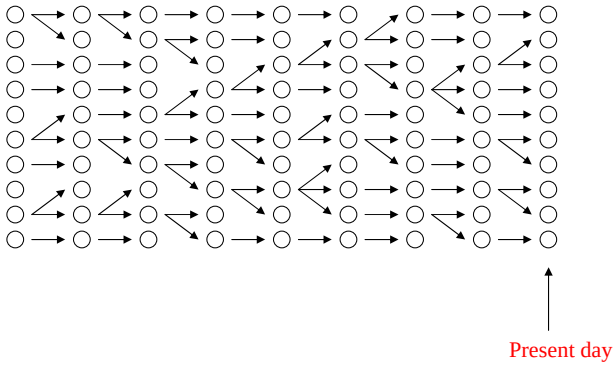
- Start with simplest model:
  - Constant population size
  - No population structure
  - No selection
- Ask:
  - What patterns of genetic variability do we expect?
  - What are the important parameters
  - How can we estimate the parameters?
  - How can we test whether the fitted model is adequate?

The Wright-Fisher population model

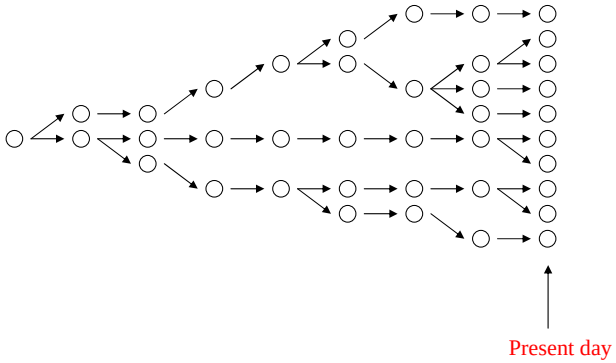


- Diploid Individuals reproduce by sexual reproduction with possibility of selfing
- Mating is random with respect to location and genotype
- Generations are non-overlapping (everyone reproduces simultaneously)
- The population size is constant of size N ($2N$ alleles)
- There is no migration or selection
- Mutations are neutral and occur at a constant rate per generation

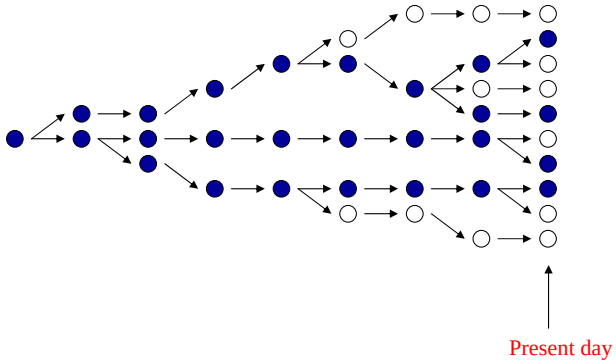
Genes in populations



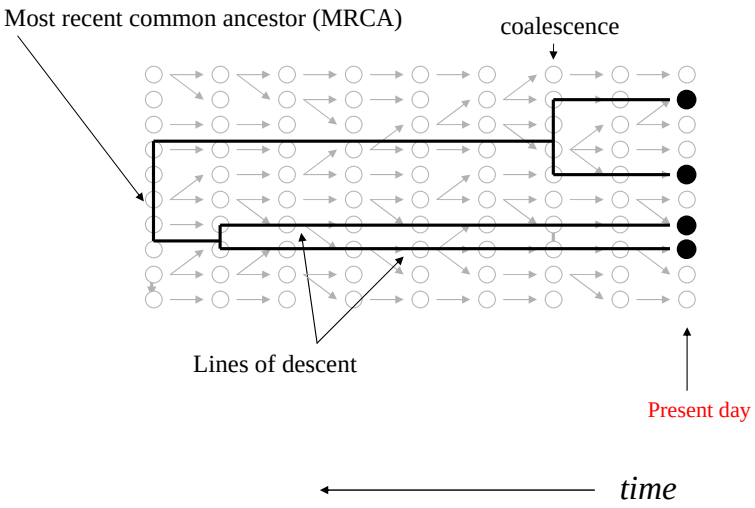
Ancestry of current population



Ancestry of sample



The coalescent: samples in populations

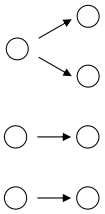


The key ideas behind coalescent theory

- If most mutations are neutral, the fate of an individual within the population is not influenced by the mutations it may have
- We can therefore model the genealogical process independently of the mutation process
- If mutations are neutral, we also need not be concerned about those parts of the population that did not leave descendants in the sample
- We need only model the genealogy of samples of chromosomes from populations
  - *Coalescent theory is a probabilistic description of the genealogical process for samples of chromosomes in large populations*

Kingman (1982)

The genealogical process for two chromosomes



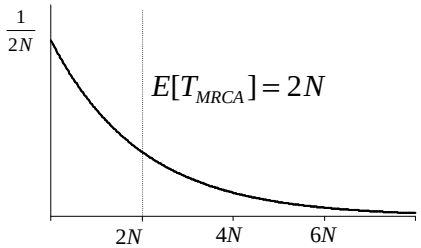
Probability from same parent (coalescence) $= \frac{1}{2N}$

Probability from different parents $= 1 - \frac{1}{2N}$

Probability of coalescence t generations ago $= \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$

Not coalesced for first t-1 generations

Coalesce in next generation

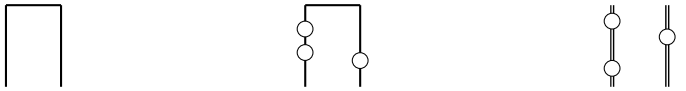


63% of outcomes have $T_{MRCA} < 2N$

Adding mutations

- Mutations occur randomly at a rate proportional to the product of the time to coalescence and the mutation rate

Genealogy Mutations DNA sequences



- Expected number of differences between a pair of sequences

$$E[\pi] = 2 \times u \times E[T_{MRCA}]$$
$$= 4Nu$$

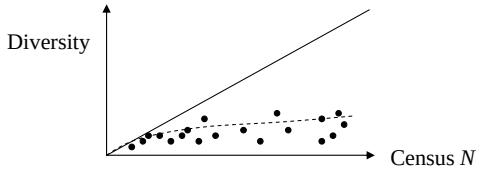
- The product $4Nu$ is so important in population genetics, it is usually written as a single parameter

$$\theta = 4Nu$$

Our first parameter!

Census population size and effective population size

- Levels of polymorphism vary less between species than the census population size



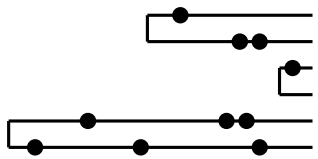
- The rate of genetic drift varies due to
  - Inbreeding, skewed sex ratios, fluctuating population size, variation in family size

- Many biologically realistic complications can be modelled by a coalescent process with a smaller EFFECTIVE population size

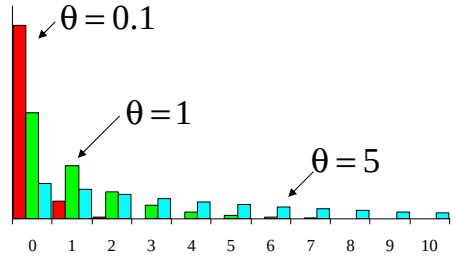
$N \rightarrow N_e$ $\quad E[\pi] = 4N_e u$ $\quad \theta = 4N_e u$

Variation in pairwise diversity

- Variance due to
  - Distribution of coalescence times (geometric)
  - Randomness of mutation process (Poisson)



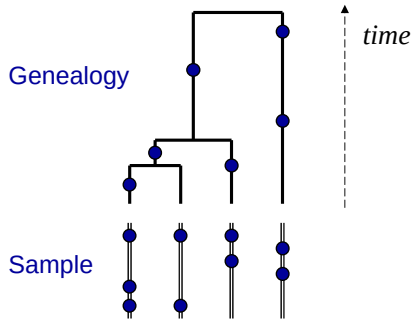
$$\Pr\{k \text{ mutations}\} = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^k$$
$$\theta = 4N_e u$$



Number of differences

The n-coalescent

- *Assume*
  - Lineages coalesce independently
  - No more than one coalescent event can occur in a single generation: in effect $N_e \to \infty$

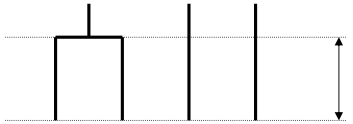


Coalescence times with n sequences

$$\Pr\{\text{coalescence given } n \text{ lineages}\} = \frac{n(n-1)}{2}\frac{1}{2N_e}$$

Number of pairs of lineages

Probability of a given pair coalescing



$$E[T_{co}] = \frac{4N_e}{n(n-1)}$$

$$E[\text{no. mutations}] = E[T_{co}] \times n \times u$$

Number lineages

Total mutation rate

$$= \frac{\theta}{n-1}$$

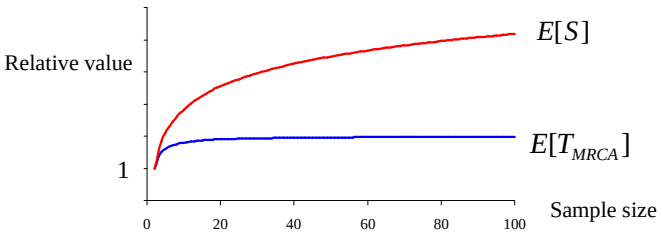
Properties of the n-coalescent

- The total number of segregating sites is the sum over each coalescent interval

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$ Watterson (1975)

- The time until the MRCA for n sequences is

$$E[T_{MRCA}] = 2\left(1 - \frac{1}{n}\right)$$



The variance in the number of segregating sites

- The number of segregating sites is a compound distribution

$$\text{Var}(S) = uE[T_{tot}] + u^2 \text{Var}[T_{tot}]$$

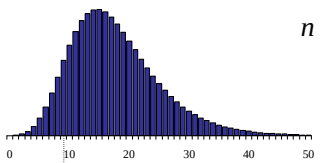
- Due to the independence of successive coalescent events, the variances in coalescence times are additive

$$\text{Var}[T_{tot}] = \sum_{i=2}^{n} i^2 \text{Var}[T_{co}(i)]$$

$$\text{Var}[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

- The full distribution can be calculated by a simple recursion (Tavaré, 1984)

$\theta = 5$, $n = 20$ → $E[S] = 17.7$, $Var[S] = 57.6$

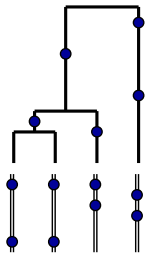


Mutations, alleles and haplotypes

- Infinite-allele model
  - Each mutation creates a new allele
  - Equivalent to a new haplotype if NO recombination

$$E[K] = 1 + \frac{\theta}{1+\theta} + \frac{\theta}{2+\theta} + ... + \frac{\theta}{n-1+\theta}$$

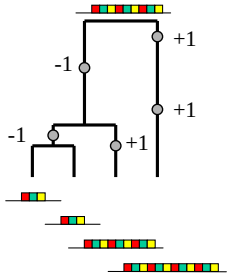
Ewens (1972)



- Microsatellites
  - Step-wise mutation model

$$E[Var(L)] = N_e\mu$$

Moran (1975)
Slatkin (1995)



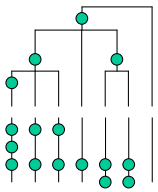
Summary statistics

- Good properties of summary statistics
  - Include most (all) information in the data
  - Different statistics should use different information
  - Expectations and variances should have simple relationship to model parameters

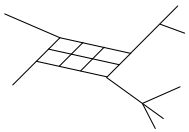
Statistic	Symbol	Expectation
Average pairwise diversity	$\pi = \frac{2}{n(n-1)}\sum_{ij}\pi_{ij}$	θ
Number of segregating sites	S	$\theta\sum_{i=1}^{n-1} 1/i$
Number of haplotypes	K	$\theta\sum_{i=0}^{n-1} 1/(i+\theta)$ (no recombination)

Graphical representation

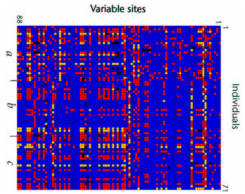
- Gene trees
  - For case of no recombination
  - Maximum possible resolution given data
  - Times proportional to coalescent estimates



- Networks
  - Split tree decomposition
  - Represent ambiguity in 'genealogy'
  - Good for back mutations and rare recombination

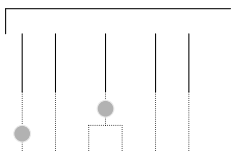


- Haplotype plots
  - ignore non-polymorphic sites
  - Can be used with recombination



Stochastic simulation

- Generate random samples from Fisher-Wright population
  - Joint simulation of genealogies and mutations (Griffiths & Tavaré, 19??)



Event	Rate
Coalescence	$\frac{n(n-1)}{2}$
Mutation	$n\frac{\theta}{2}$

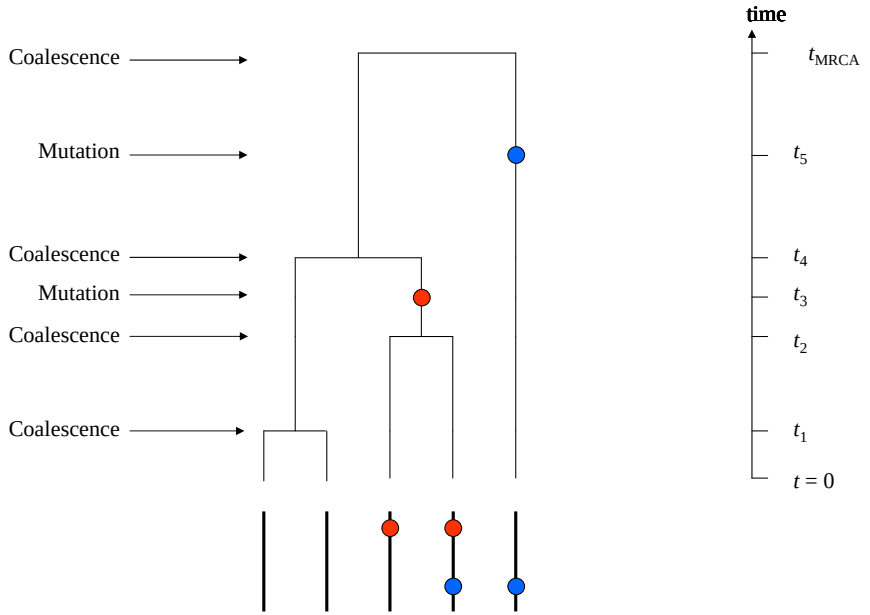
- Time until next event = exponentially distributed with rate equal to the sum of all possibilities

$$\phi(\tau) = \lambda e^{-\lambda\tau}$$

$$\lambda = \frac{n(n-1)}{2} + \frac{n\theta}{2}$$

- Probability next event is a mutation = probability of mutation divided by total probability of events

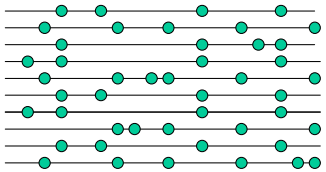
$$\Pr\{\text{mutation}\} = \frac{\theta}{\theta + n - 1}$$



Estimating parameters

- Moment methods
  - Equate observed statistic with theoretical expectation
  - Derive variances (analytical or by simulation)
- Likelihood
  - Calculate relative probability of observing data given different values of parameters
- Bayesian inference
  - Use prior information about parameters to influence estimate

Data set
$n = 10$
$S = 14$



Moment estimators of θ

- Pairwise diversity

$$b_1 = \frac{n+1}{3(n-1)}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

$$E[\pi] = \theta$$

$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{ij} \pi_{ij}$$

$$Var(\hat{\theta}_\pi) = b_1\theta + b_2\theta^2$$

Tajima (1983)

- Number of segregating sites

$$a_1 = \sum_{i=1}^{n-1} 1/i$$

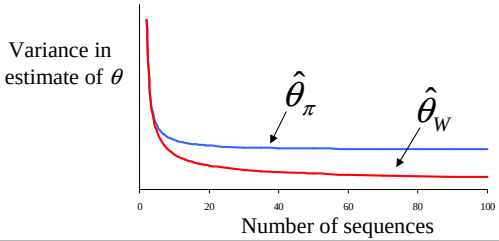
$$a_2 = \sum_{i=1}^{n-1} 1/i^2$$

$$E[S] = \theta a_1$$

$$\hat{\theta}_W = S / a_1$$

$$Var(\hat{\theta}_W) = \theta / a_1 + \theta^2 a_2 / a_1^2$$

Watterson (1975)



$\hat{\theta}_\pi = 5.9$ $\hat{\theta}_W = 5.0$

$Var(\hat{\theta}_\pi) = 12.1$ $Var(\hat{\theta}_W) = 6.5$

Likelihood estimation of θ

- Using the recursion of Tavaré (1984)
  - Number of segregating sites only

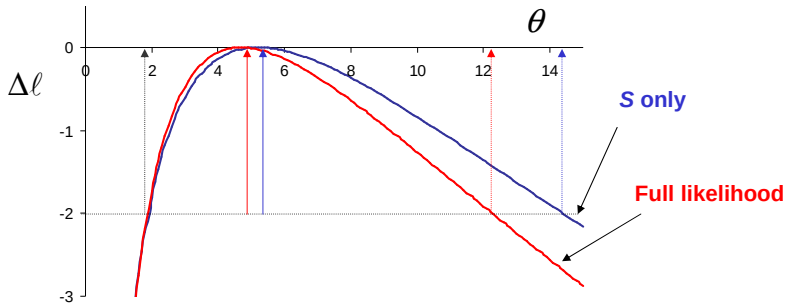
$\hat{\theta} = 5.2$

$2U = 1.9 - 14.4$

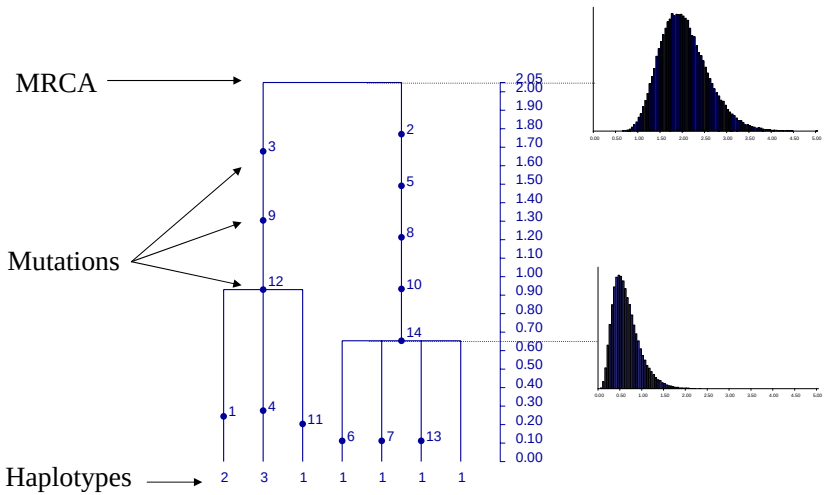
- Using the full-likelihood method of Griffiths (??)
  - Implemented in GENETREE software

$\hat{\theta} = 4.7$

$2U = 1.9 - 12.2$



Estimating ages of events



estimated by GENETREE

Bayesian estimation

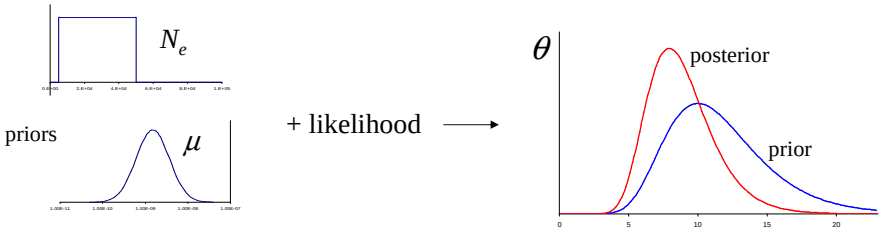
- May have prior information on the effective population size, or rate of mutation that we wish to incorporate
- Calculate the probability of θ given the data

Prior probability for θ **Likelihood**

Posterior probability for θ

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)}$$

Probability of data



Strengths and weaknesses of coalescent theory

- Very flexible, simulations are easy to implement irrespective of population and mutational models
- Deals explicitly with basic unit of empirical population genetics research
- Full likelihood analysis within the coalescent framework uses all possible information

- Some types of natural selection are difficult to incorporate
  - Coalescence depends on allelic state and rest of population
- Full likelihood inference is computationally intensive