

Population genetic inference

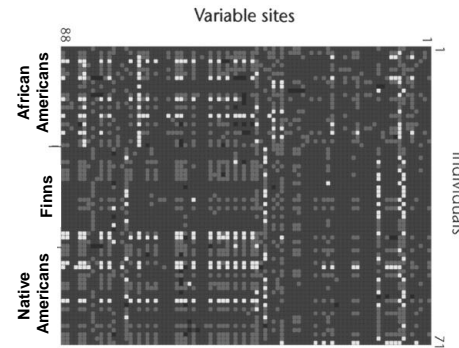
Date	Topic	
28 th Jan	Modelling population genetic data	GM
4 th Feb	Inference	GM
11 th Feb	Model testing	GM
18 th Feb	Extensions of the neutral model	GM
25 th Feb	Recombination	SM
4 th March	Recombination and haplotype structure	GM
11 th March	Fine-scale mapping	JH
18 th March	Genome-wide association studies	JM

Books

Balding DJ, Bishop M and Cannings C. 2001. Handbook of Statistical Genetics. John Wiley and Sons Ltd.
 Casella GC and Berger RL. 1990. Statistical Inference. Wadsworth and Brooks/Cole
 Li W-H. Molecular evolution. Sinauer.
 Weir BS. 1990. Genetic Data Analysis. Sinauer

1

Good questions in population genetics

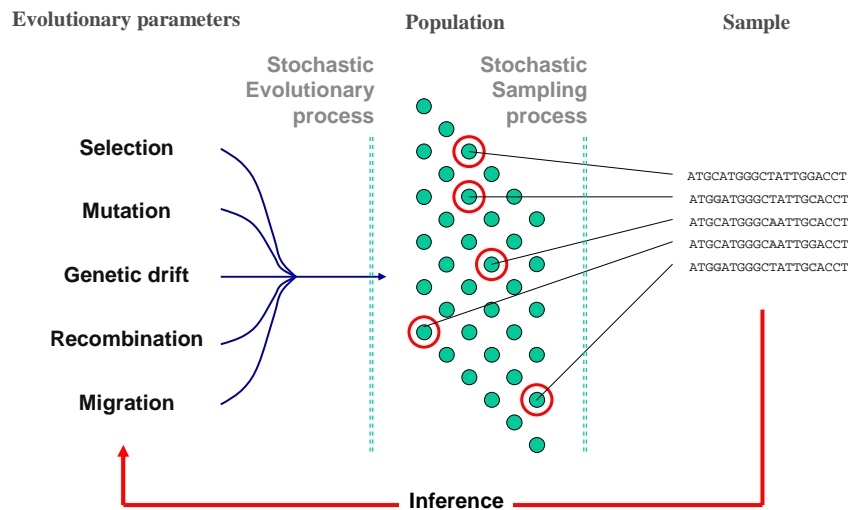


DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene
 Nickerson, *et al.* 1998 *Nature Genetics* 19, 233 - 240

- Is there an association between DNA sequence variation and the disease phenotype?
- What do the sequences tell us about human history?
- How has natural selection shaped diversity in the gene?

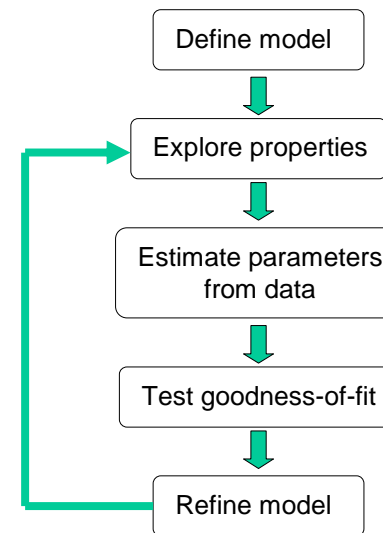
2

Population genetic inference



3

Statistical inference



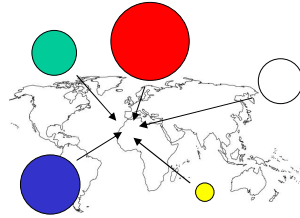
Issues

- rules
- parameters
- quantities
- summary statistics
- graphical representation
- stochastic simulation
- moment methods
- likelihood
- Bayesian inference
- outliers
- heterogeneity
- comparison of estimators
- add parameters

4

Example: Structure in human populations (Rosenberg *et al.* 2002)

- Questions
 - Is there significant natural structuring to genetic variation in humans?
 - Does this structuring coincide with geographical boundaries?
- Data
 - 377 autosomal microsatellite loci in 1056 individuals from 52 populations
- Model
 - K 'Hidden' populations in linkage and Hardy-Weinberg equilibrium
- Estimation
 - Estimate population allele frequencies
 - Most likely value of K
 - Posterior probability for each individual
- Model testing



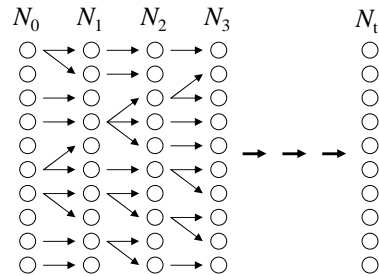
5

The null model in population genetics

Nothing interesting ever happens in biology

6

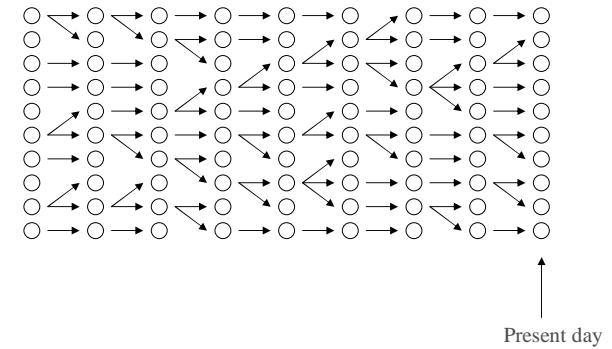
The Wright-Fisher population model



- 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

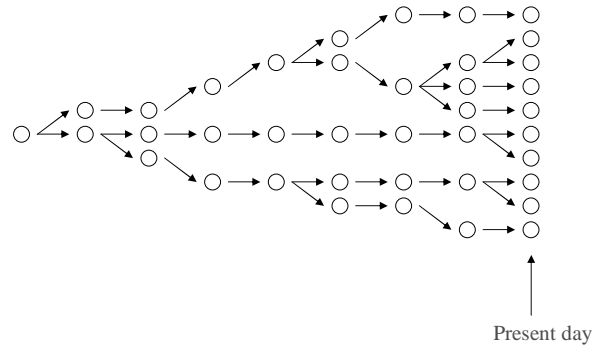
7

Genes in populations



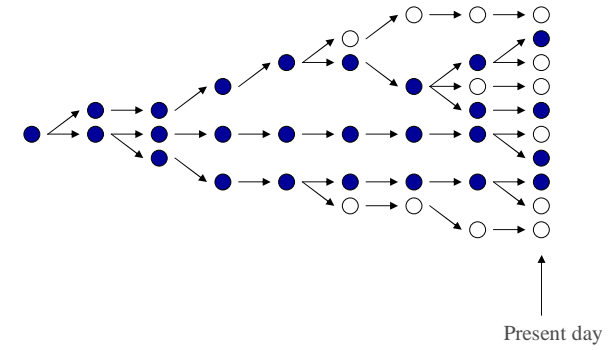
8

Ancestry of current population



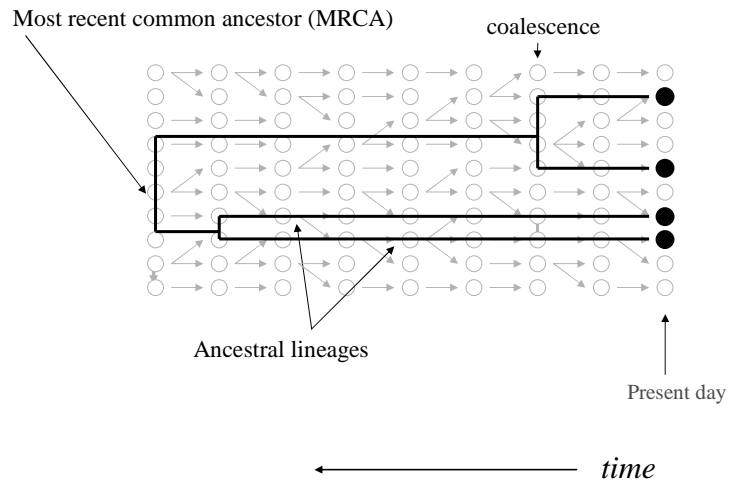
9

Ancestry of sample



10

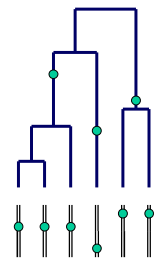
The coalescent: samples in populations



11

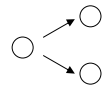
The neutral coalescent

- Premised on neutral assumption
- Implication is that all information about underlying evolutionary processes is in the underlying gene genealogy
- Coalescent theory formulated in early 1980s by Kingman, and developed by Hudson, Griffiths, Tavaré, Donnelly, etc. describes genealogical history of samples from populations
- Major challenge for contemporary theoretical community is developing statistical inference from population genetic data



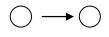
12

The genealogical process for two chromosomes



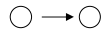
Probability from same parent
(coalescence)

$$= \frac{1}{2N}$$



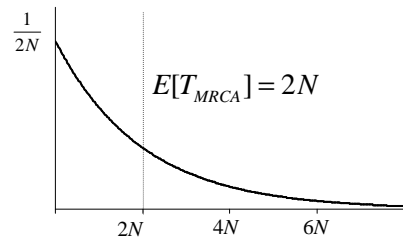
Probability from different parents

$$= 1 - \frac{1}{2N}$$



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 $T_{MRCA} < 2N$

13

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

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

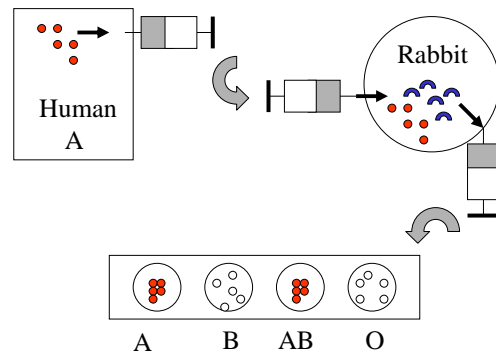
$$= 4Nu$$

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

$$\theta = 4Nu$$

14

Serological techniques for detecting variation



15

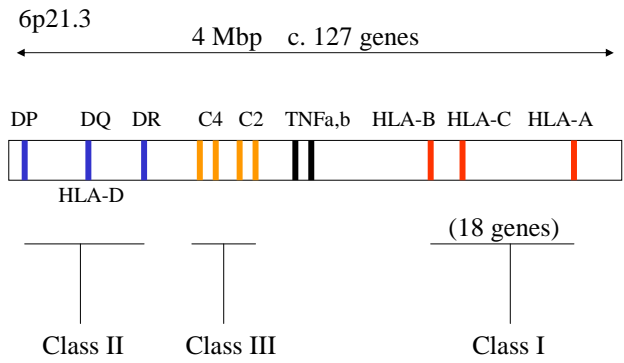
Polymorphic blood groups in the white English population (no. types)

ABO	(4)	Kidd	(3)
Rh	(7)	Dombrock	(2)
MNS	(6)	Auberger	(2)
P	(3)	Xg	(2)
Secretor	(2)	Sd	(2)
Duffy	(3)	Lewis	(2)

$$\Pr\{2 \text{ people same blood type}\} \approx 3 \text{ in } 10,000$$

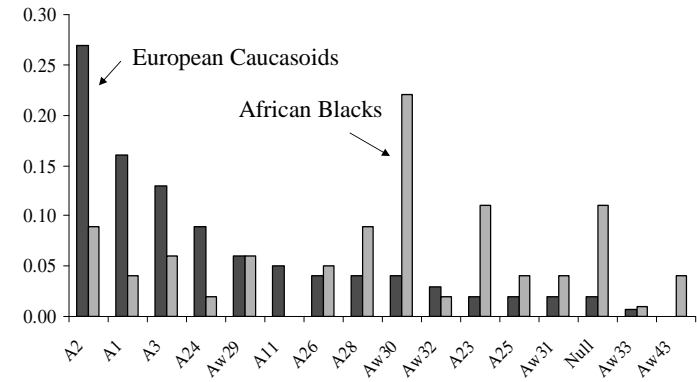
16

HLA diversity at the MHC locus



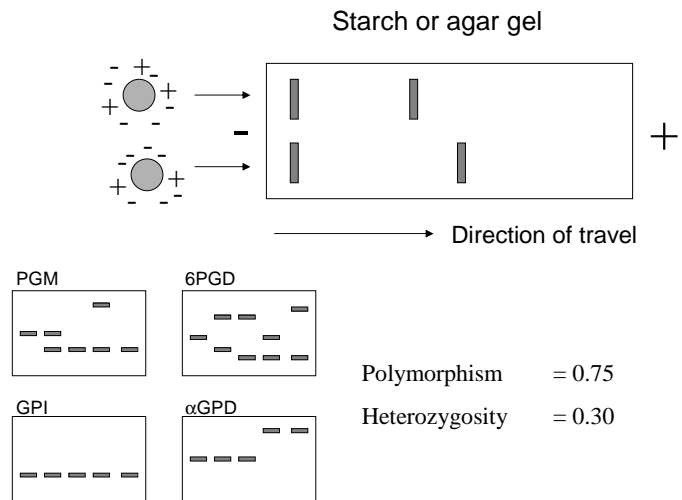
17

HLA-A



18

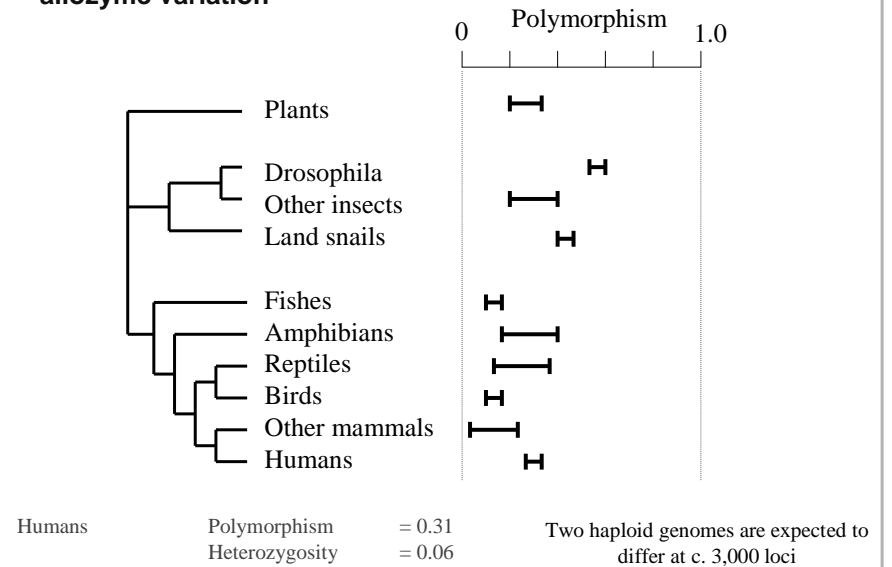
Protein electrophoresis



Lewontin and Hubby (1966)
Harris (1966)

19

The phylogenetic distribution of allozyme variation



20

The rise of the neutral theory

Kimura (1968); King and Jukes (1969)

- Observations
 - Constancy of rate of molecular evolution (the molecular clock)
 - More important regions of proteins evolve at a slower rate than less important domains
 - High levels of protein polymorphism
 - High rates of molecular evolution (about 1.5×10^{-9} changes per amino acid per year)
- Theoretical considerations
 - Segregation load of balanced polymorphisms
 - Haldane's cost of natural selection

21

Kimura's neutral theory

- The majority of changes in proteins and at the level DNA which are fixed between species, or segregate within species, are of no selective importance
- The rate of substitution is equal to the rate of neutral mutation

$$k = f_{neutral} \mu$$

- The level of polymorphism in a population is a function of the effective population size and the neutral mutation rate

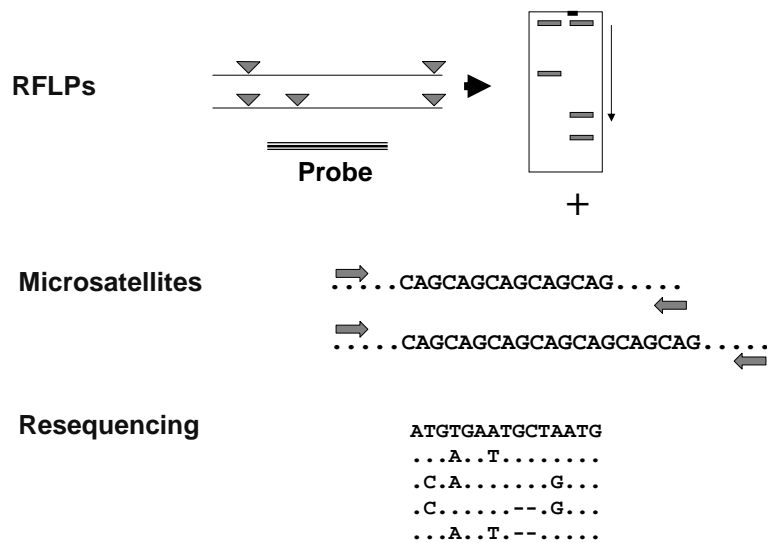
$$\pi = 4N_e \mu$$

- Polymorphisms are transient rather than balanced



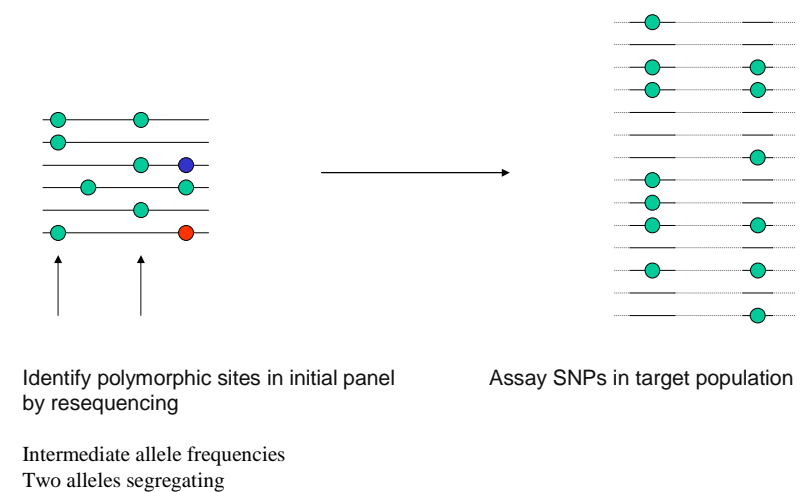
22

Detecting DNA sequence variation



23

SNP analyses (Single Nucleotide Polymorphisms)



24

Patterns of variation at the DNA level

- Synonymous & nonsynonymous mutations

Arg Gln Val	Arg Gln Val
AGA CAA GTA	AGA CAA GTA
↓	↓
CAG CGA GTA	AGA CAG GTA
Arg Arg Val	Arg Gln Val

e.g. *D. simulans*

$\pi_{\text{total}} = 0.010$ per site
 $\pi_{\text{silent}} = 0.038$
 $\pi_{\text{noncoding}} = 0.023$

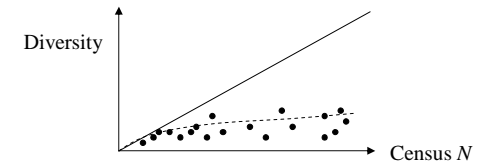
- Nucleotide variation v. protein variation?

	Humans	<i>D. melanogaster</i>
Allozyme	6%	14%
Nucleotide	0.1%	1%

25

Census population size and effective population size

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



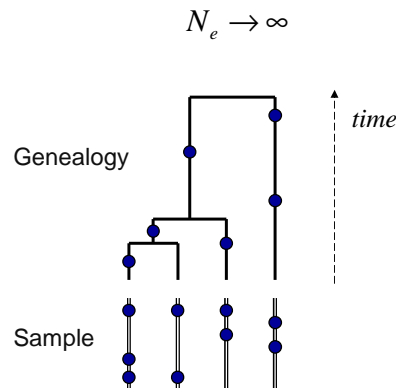
- The rate of genetic drift varies due to
 - Inbreeding, skewed sex ratios, fluctuating population size, variation in family size
- Many biologically realistic complications can be modelled by a coalescent process with a smaller EFFECTIVE population size

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

26

The n-coalescent

- Assume
 - Lineages coalesce independently
 - No more than one coalescent event can occur in a single generation: in effect



27

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

A diagram showing a horizontal bar with four vertical lines extending upwards from it, representing coalescence events. A double-headed vertical arrow indicates the time interval from the bar to the top of the lines.

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

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

Number lineages Total mutation rate

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

28

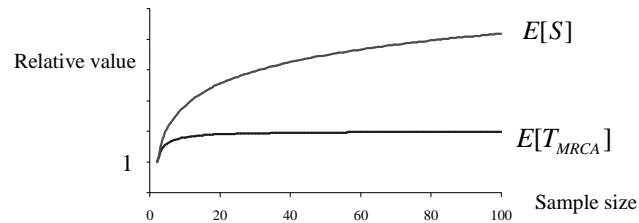
Properties of the n-coalescent

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

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

- The expected time until the MRCA for n sequences is

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



29

The variance in the number of segregating sites

- The number of segregating sites is a compound distribution

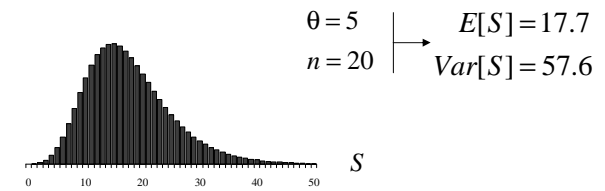
$$\text{Var}(S) = uE[T_{tot}] + u^2\text{Var}[T_{tot}]$$

- 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)



30

Mutations, alleles and haplotypes

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

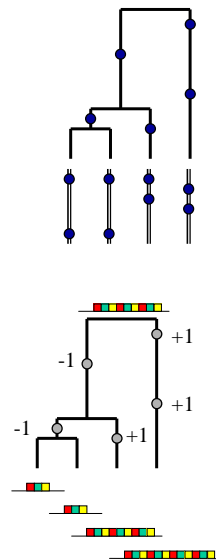
$$E[K] = 1 + \frac{\theta}{1+\theta} + \frac{\theta}{2+\theta} + \dots + \frac{\theta}{n-1+\theta}$$

Ewens (1972)

- Microsatellites
 - Step-wise mutation model

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

Moran (1975)
Slatkin (1995)



31

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

32